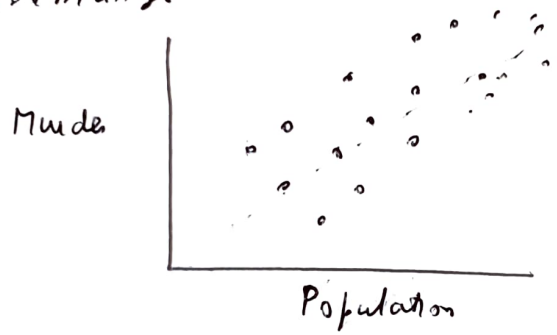
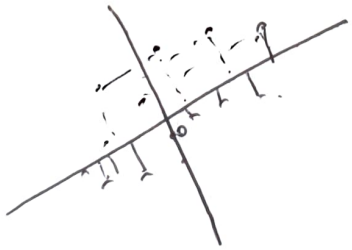
SOUMYA BANERJEE

PCA NOTES

	Murder	Population → (Features)
California	20	10,000
Nevada	30	20,000
Wisconsin	50	60,000
⋮	⋮	⋮

↓ Normalize

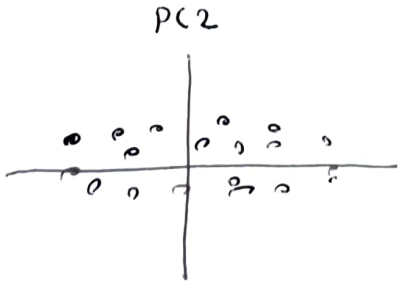




CONCEPT

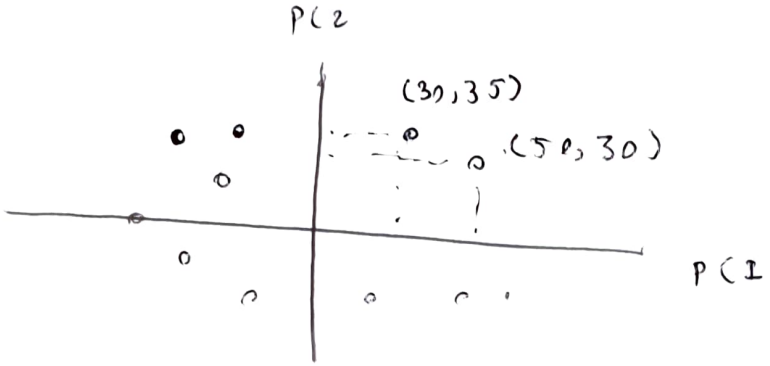
as close as possible to the data.

line along which data varies the most



new axes

how to interpret
PC1 - Violence / safety
PC2 - urbanization



scores

$$Z_1 = \Phi_{11} X_1 + \Phi_{21} X_2 + \dots +$$

$$Z_2 = \Phi_{12} X_1 + \Phi_{22} X_2 + \dots$$

loadings

$$Z_1 = (1.1) X_1 + (0.5) X_2 + \dots$$

$$Z_2 = 0.1 X_1 + 1.6 X_2 + \dots$$

CONCEPT – the absolute values do not matter. What does is relative to each other.

	PC1	PC2
Murder	(0.5)	−0.4
Assault	(0.5)	−0.2
Urban Pop	0.27	(0.8)
Rape	(0.55)	0.1

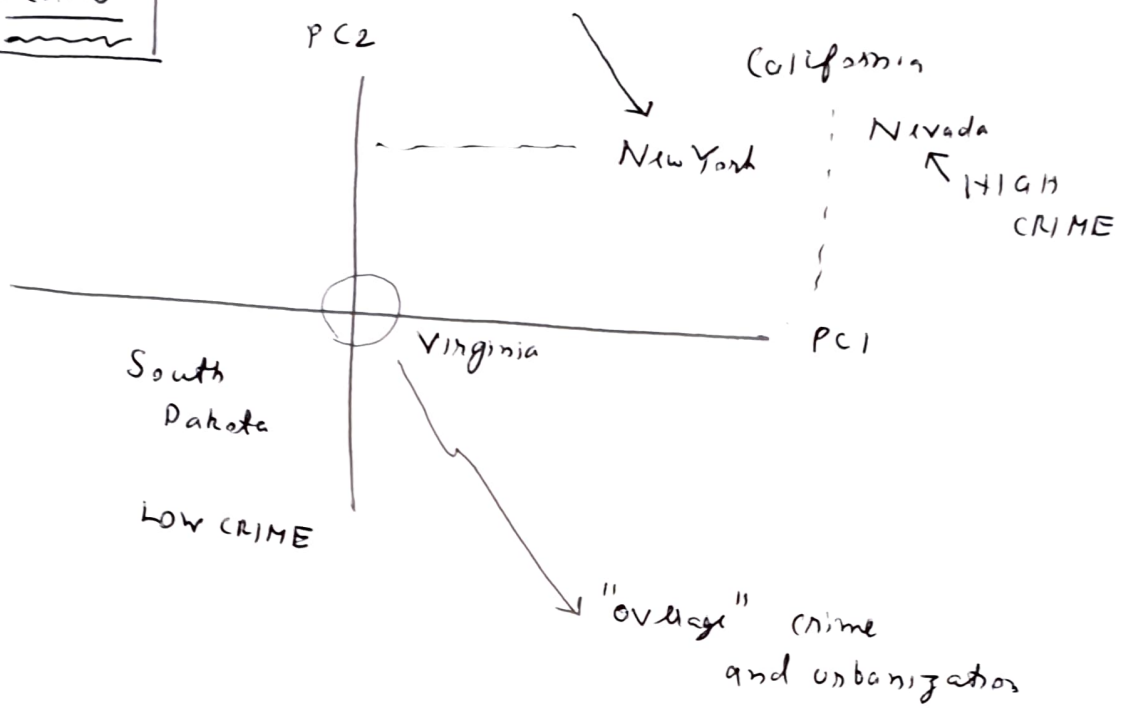
Loadings

crime $PC_1 = 0.5 \times \text{Murder} + 0.5 \times \text{Assault} + (0.2 \times \text{Urban}) \dots + 0.5 \times \text{Rape}$

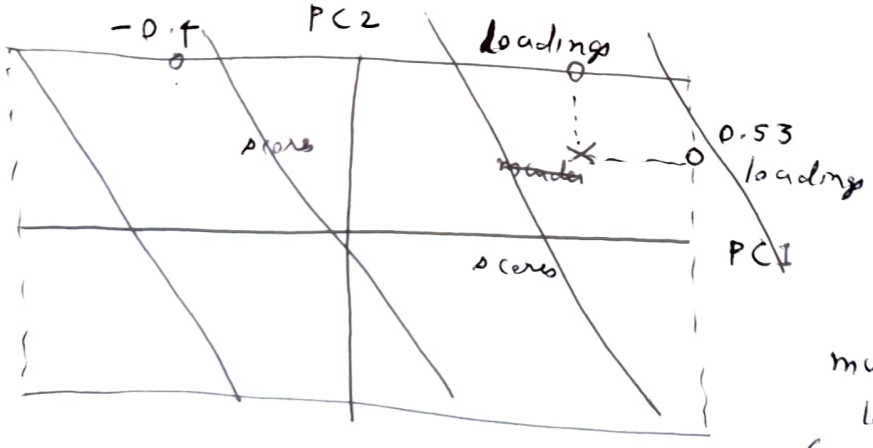
population $PC_2 = -0.4 \times \text{Murder} + (-0.2) \times \text{Assault} + (0.8 \times \text{Urban}) + 0.1 \times \text{Rape}$

"read" PCA plot

scores



loadings

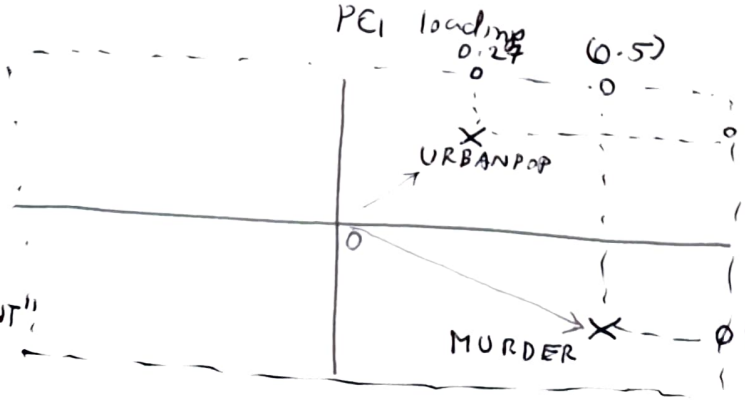


murder loading (0.5, -0.4)

CONCEPT

URBANPOP very far from MURDER

"DIFFERENT" CONCEPT



Urban Pop (0.27, 0.87)

CONCEPT - The principal components scores and loadings are "encoding" different concepts.

We did not "tell" it these concepts but it "found" it.

UNSUPERVISED

ENCODING

→ words

→ pictures

https://projector.tensorflow.org

- how to convert pictures to numbers?
- how to convert words to numbers?

→ how these are used in modern LLMs.

→ $Z_1 = 0.5 x_1 + 0.7 x_2 + \ldots$

LINEAR MODEL

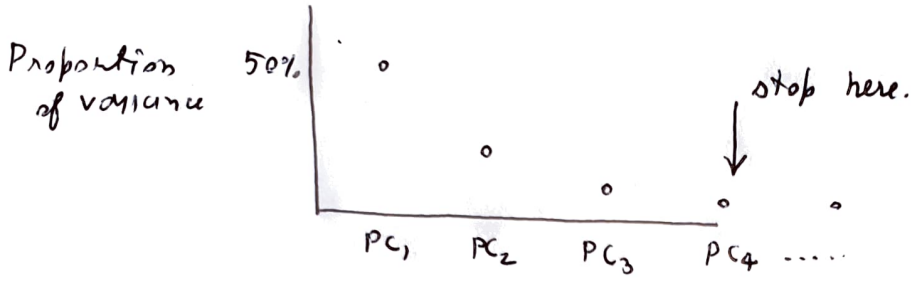
CONCEPT

PCA is linear

how to make non-linear?

50% variance PC_1

45% variance PC_2



PRACTICALS - images

genomic data