

Intro to
Dimensionality
Reduction

Data Matrix

Recall

The columns commonly represent attributes/properties of the data

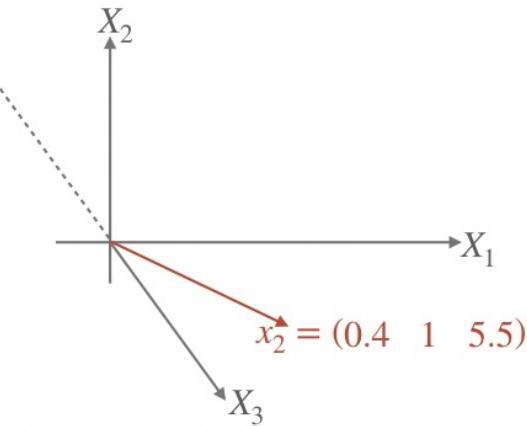
The rows commonly represent entities and their observed values for each attribute

$$D = \begin{array}{c|ccccc} & X_1 & X_2 & X_3 & X_4 \\ \hline x_1 & 0.2 & 23 & A & 5.7 \\ x_2 & 0.4 & 1 & B & 5.4 \\ x_3 & 1.8 & 0 & C & 5.2 \\ x_4 & 5.6 & 50 & A & 5.1 \\ x_5 & -0.5 & 34 & A & 5.3 \\ x_6 & 0.4 & 19 & B & 5.4 \\ x_7 & 1.1 & 11 & A & 5.5 \end{array}$$

Geometric View of Data

Each row is a vector in
m-dim space

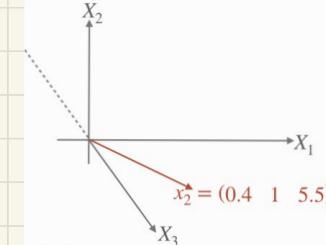
Here, we think of x_2 as
a 3-dimensional vector


$$D = \begin{array}{cccc} & X_1 & X_2 & X_3 \\ x_1 & 0.2 & 23 & 5.7 \\ x_2 & 0.4 & 1 & 5.4 \\ x_3 & 1.8 & 0 & 5.2 \\ x_4 & 5.6 & 50 & 5.1 \\ x_5 & -0.5 & 34 & 5.3 \\ x_6 & 0.4 & 19 & 5.4 \\ x_7 & 1.1 & 11 & 5.5 \end{array}$$

Geometric View of Data

What problems can occur in
high dim space

Here, we think of x_2 as
a 3-dimensional vector

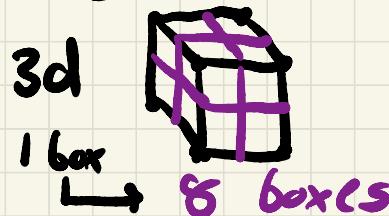
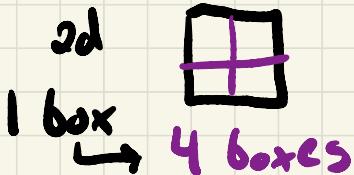


| | X_1 | X_2 | X_3 |
|-------|-------|-------|-------|
| x_1 | 0.2 | 23 | 5.7 |
| x_2 | 0.4 | 1 | 5.4 |
| x_3 | 1.8 | 0 | 5.2 |
| x_4 | 5.6 | 50 | 5.1 |
| x_5 | -0.5 | 34 | 5.3 |
| x_6 | 0.4 | 19 | 5.4 |
| x_7 | 1.1 | 11 | 5.5 |

Geometric View in high dim

Some problems

- intuition from 2 & 3-dim doesn't always work
- volumes scale in non-intuitive ways
- near and far become harder to differentiate
- computing distances becomes more expensive
- algorithms become slower



Example: Volume of Ball w/ radius r

- 2D:
- 3D:
- d-dim:

Volume of d-dim ball w/ radius r

$$V_B(r, d) = \left(\frac{\pi^{\frac{d}{2}}}{\Gamma(\frac{d}{2} + 1)} \right) r^d$$

$$\Gamma\left(\frac{d}{2} + 1\right) = \begin{cases} \left(\frac{d}{2}\right)! & \text{if } d \text{ is even} \\ \sqrt{\pi} \left(\frac{d!}{2^{(d+1)/2}}\right) & \text{if } d \text{ is odd} \end{cases}$$

Let $r=1$, compute volume of a d-ball for:

- $d=4$

- $d=12$

Volume of d-dim ball w/ radius r

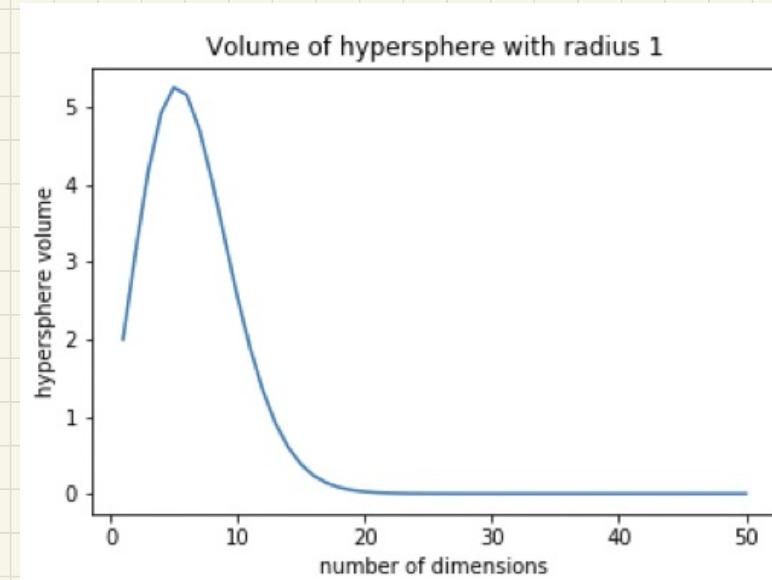
Let $r=1$, compute volume of a d-ball for:

- $d=2 V_B(1, 2)$

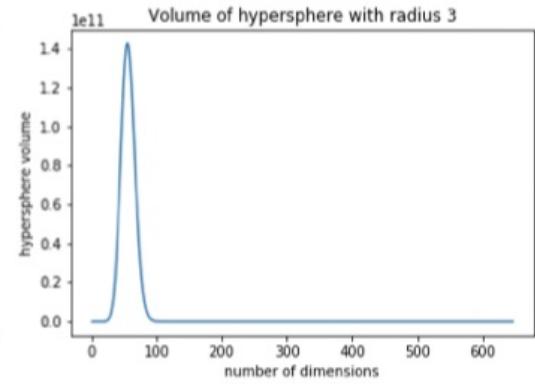
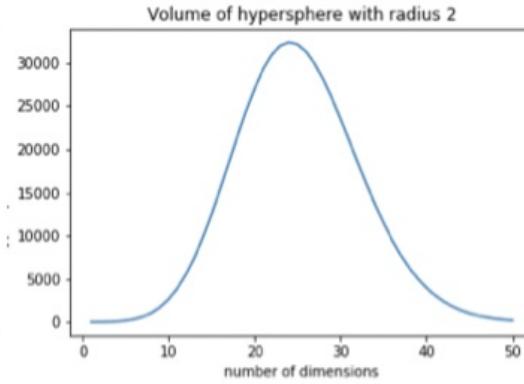
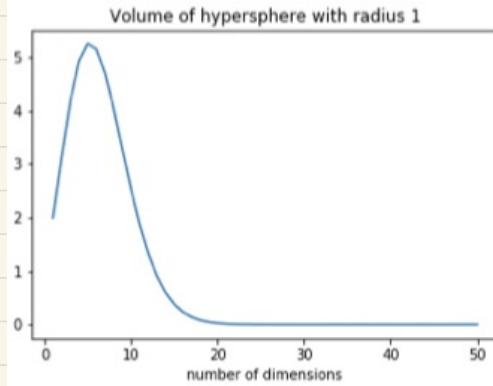
- $d=3 V_B(1, 3)$

- $d=4 V_B(1, 4) \approx 4.93$

- $d=12 V_B(1, 12) \approx 1.34$

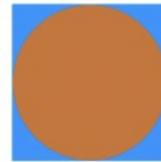


Volume of d-dim ball w/ radius r



Other non-intuitive props:

- As dimensionality increases, the volume of a sphere with radius 1 increases up to a point, and vanishes as $d \rightarrow \infty$
- Consider a hypersphere inscribed inside a hypercube. As dimensionality increases, all the volume of the hyperspace is in the corners of the cube!
- The fraction of points with density at least α times the peak density of a multivariate normal distribution decreases rapidly (i.e., majority of the density is in the tail regions of the distribution!)
- The distance between the nearest neighbor of a point and the farthest neighbor of a point get closer together as dimensionality increases, making it difficult to find true “nearest neighbors”



How to solve?

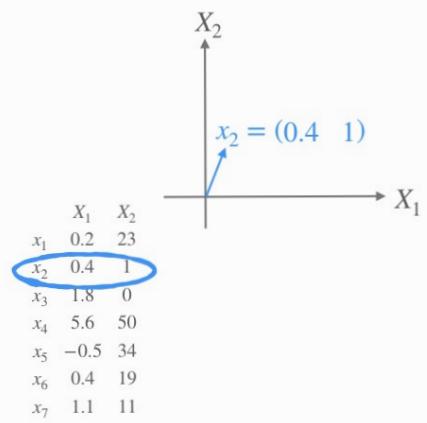
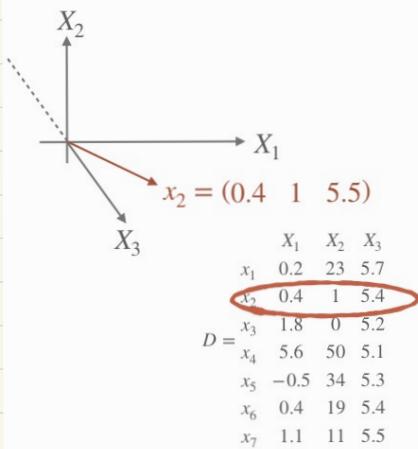
Dim Reduction Brain Storm

How can we reduce dim?

(+/- of that idea)

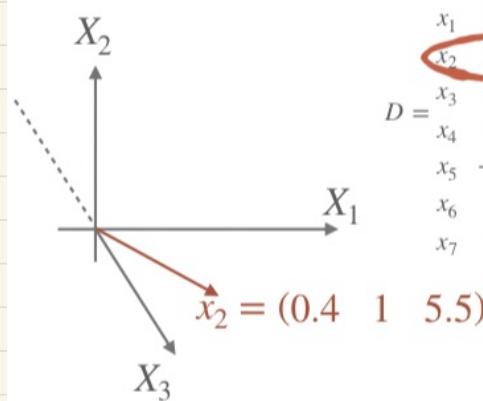
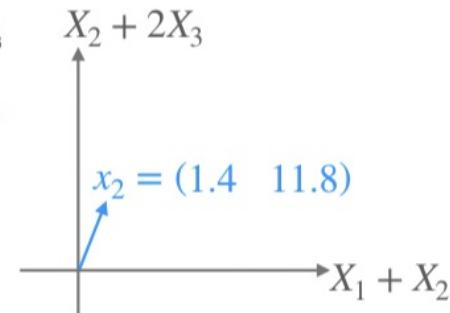
Feature Selection

select subset of attrs (feature selection)

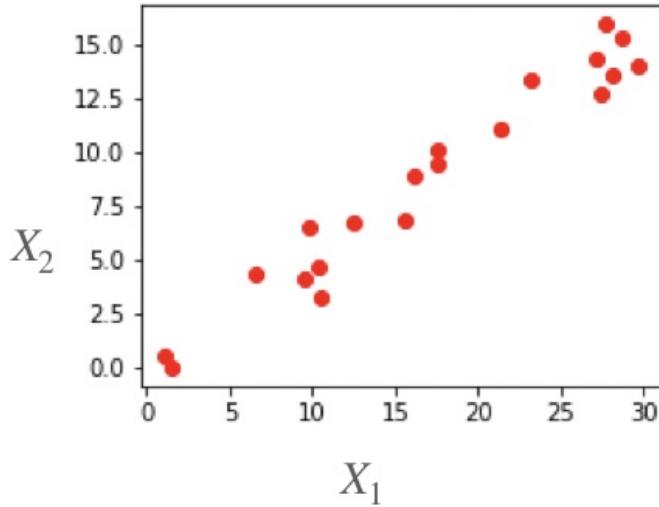


Different Representation

Use all attrs to create new attrs

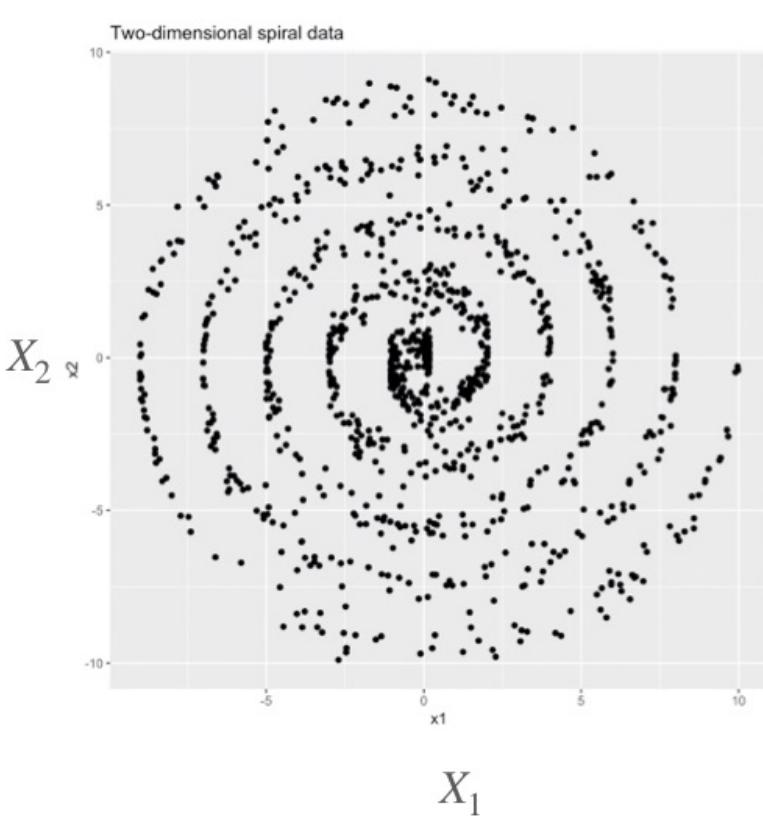

$$D = \begin{array}{l} \begin{array}{c|ccc} & X_1 & X_2 & X_3 \\ \hline x_1 & 0.2 & 23 & 5.7 \\ x_2 & 0.4 & 1 & 5.4 \\ x_3 & 1.8 & 0 & 5.2 \\ x_4 & 5.6 & 50 & 5.1 \\ x_5 & -0.5 & 34 & 5.3 \\ x_6 & 0.4 & 19 & 5.4 \\ x_7 & 1.1 & 11 & 5.5 \end{array} \\ \xrightarrow{\quad} \begin{array}{c|cc} & X_1 + X_2 & X_2 + 2X_3 \\ \hline x_1 & 23.2 & 34.4 \\ x_2 & 1.4 & 11.8 \\ x_3 & 1.8 & 10.4 \\ x_4 & 55.6 & 60.2 \\ x_5 & 33.5 & 44.6 \\ x_6 & 19.4 & 29.8 \\ x_7 & 12.1 & 22 \end{array} \end{array}$$


Low dim Rep - Often, data lies close to a low dim space



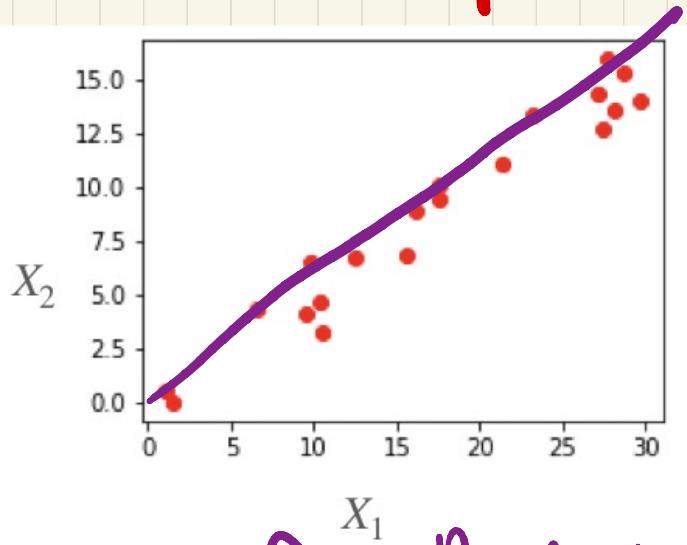
Low dim Rep

- Often, data lies close to a low dim space



Low dim Rep

- Often, data lies close to a low dim space



Idea find line
that represents
points pretty well

Linear Dim Reduction:

- Linear Discriminant Analysis (LDA)

- Principal Component Analysis (PCA)

Next
time