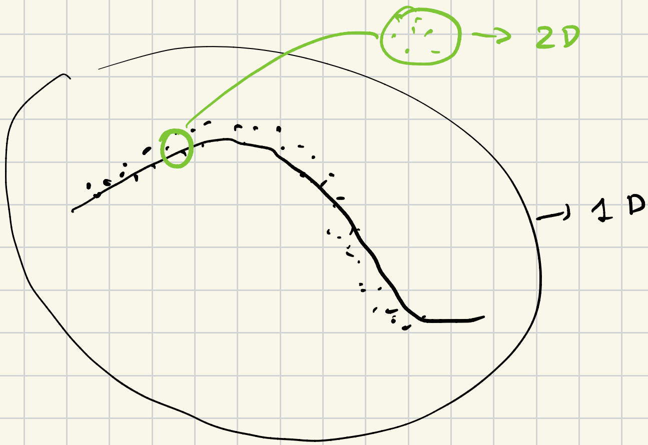
Intrinsic Dim.

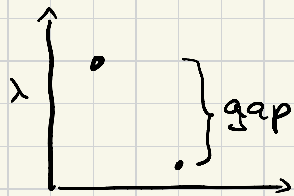
a) Minimum Number of variables needed for represent the data with minimum information loss

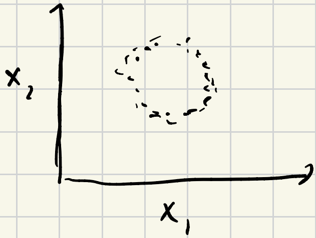
b) The dimension of the manifold in which our data lies.

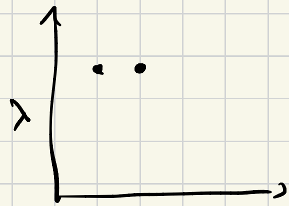


PCA is a method that performs Id = d estimation

$d = 1$



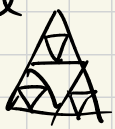


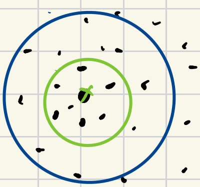


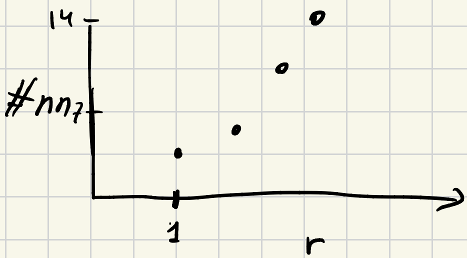
I_d (PCA) = 2

From the data ; learn directly the I_d

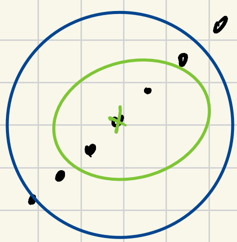
Fractal

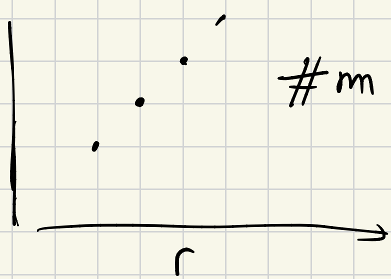






$\# nn \propto r^2$





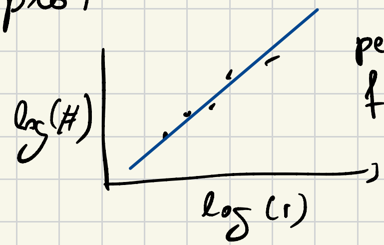
$\# nn \propto r^1$

$$\boxed{\#\, nn \propto r^d}$$

a) fix a set of different values of r

b) for each data point : count the number of neighbors within r (take the average)

c) log vs log plot



perform a linear fit

d) Slope of the fit $\rightarrow$ Id

Foreach data point

a) Compute the distances from its K-NN

b) Plot the log of the rank as a function of the log of the distances

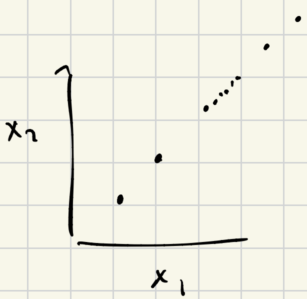
c) linear fit

d) slope will be d

$$\underbrace{\log(\#)}_{\text{count}} = \underbrace{\log \rho}_{} + \underbrace{d}_{Id} \log(\underbrace{r}_{\text{distances}})$$

$$\# = \rho \, r^d$$

We need to disentangle Id & ρ estimations



Two-NN

$$\mu^i = [1, \infty)$$

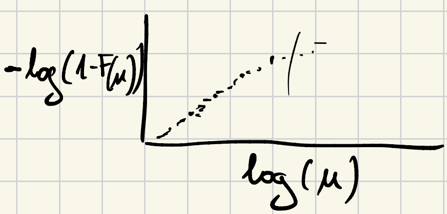
$$\mu^i = \frac{r_2^i}{r_1^i} \qquad \text{i.i.d} \quad p(\mu) = \underbrace{\mu^{-d-1} \cdot d}_{\text{pareto distribution}}$$

ρ within r_2^i can be considered constant

$$\mu^i = \frac{\cancel{\rho}\, r_2^i}{\cancel{\rho}\, r_1^i}$$

(a) $$F(\mu) = \int_1^{\mu} \mu^{*\,-d-1} \cdot d \; d\mu^* = 1 - \mu^{-d}$$

$$\frac{-\log(1 - F(\mu))}{\log(\mu)} = d$$



(D) ML

$$p(\mu) = \mu^{-d-1} \cdot d$$

$$\log \mathcal{L} = \sum \log\left(\mu_i^{-d-1} \cdot d\right) = N \log(d) - (d+1) \cdot \sum_i \log(\mu_i)$$

$$\frac{\partial \log \mathcal{L}}{\partial d} = \frac{N}{d} - \sum_i \log(\mu_i) = \emptyset$$

$$d = \frac{N}{\sum_i \log(\mu_i)}$$