6) Intrinsic Dimension Estimation :-

→ Trivial ID estimation using PCA

↓

Gap in eigenvalue spectrum

↓

ID = δ

↳ But this is an upper bound

→ Clear exception :- circle as a data manifold

↓

δ = 2, but ID = 1

→ ID estimation is very important for estimating density. If we say ID = 2, but it is = 1, density estimation will be wrong.

→ The dimension of fractal is somehow the same thing as ID.

→ Fractals became so imp. ∵ they are strongly entangled with chaotic dynamical systems.

↳ makes deterministic systems appear stochastic

→ In dynamical systems, we have a map,

s.th. at

$$x^{t+1} = f(x^t)$$

assumption of Markov process

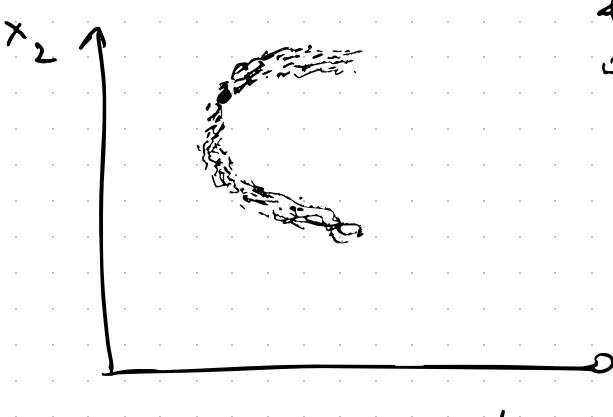
↓

can be deterministic / stochastic

↓

→ $x^{t+1} = f(x^t, \eta)$

→ In a Poincare Map, we fix a 2D plane & put a point on it, when our traj. crosses this plane. This is VERY DIFF from a 2D projection. →



→ Poincare maps are typically 100 dim. objects

→ our pts. are often confined in a compact region.

→ first strong hint of the manifold hypothesis.

→ Poincaré maps gave us the first hint of the fact that data pts. exists on a manifold that is much lower dim. than the embedding space.

→ ID can be non-integers.

Dimensionality of manifold can also be non-integers / fractal

→ So all techniques used to estimate fractal dimension can be used for ID estimation.

↳ Some of them are:

(1) Correlation Dim. Estimator
(2) Box-Counting Estimator

→ We can estimate ID without any explicit dim. reduc^n.

⊙ Correlation Dim Estimator :-

→ count how many data points are at a distance within each other smaller than the threshold → {R}

$$\forall i,j \rightarrow \| x^i - x^j \| \cdot$$

(Cartesian dist)

How many $\| x^i - x^j \| \leq R$

↓

N(R)

↳ monotonically dec func of R

∴ if R ↑, N(R) ↓

$$N(R) \sim A R^{ID} \rightarrow A \text{ (prefactor)}$$

↳ if ID = 1, N(R) grows linearly

ID = 2, N(R) grows quadratically

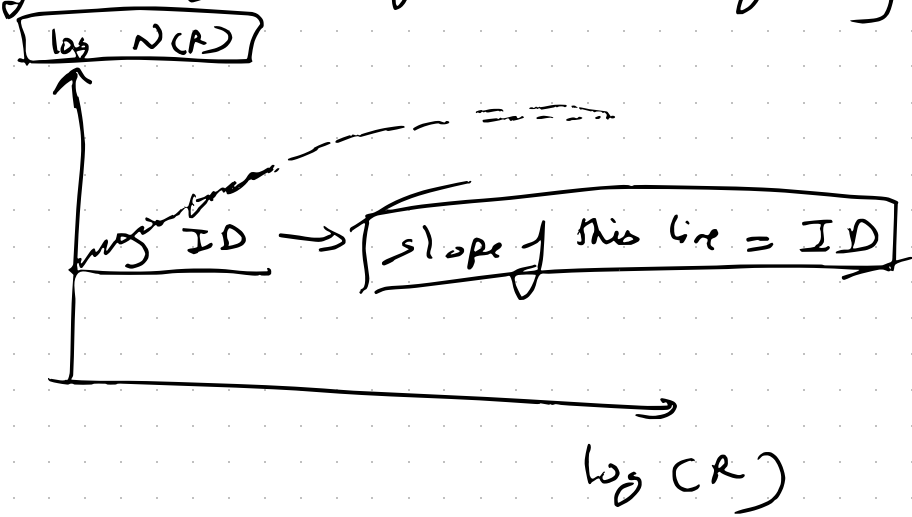
etc.

→ $N(R) \sim A R^{ID}$

↳ defines the correlation dim.

→ estimate $N(R)$ from data, & then do a (log - log) plot

∴ $\log N(R) \simeq \log A + ID \log(R)$



→ Imagine we do the following expt. :-

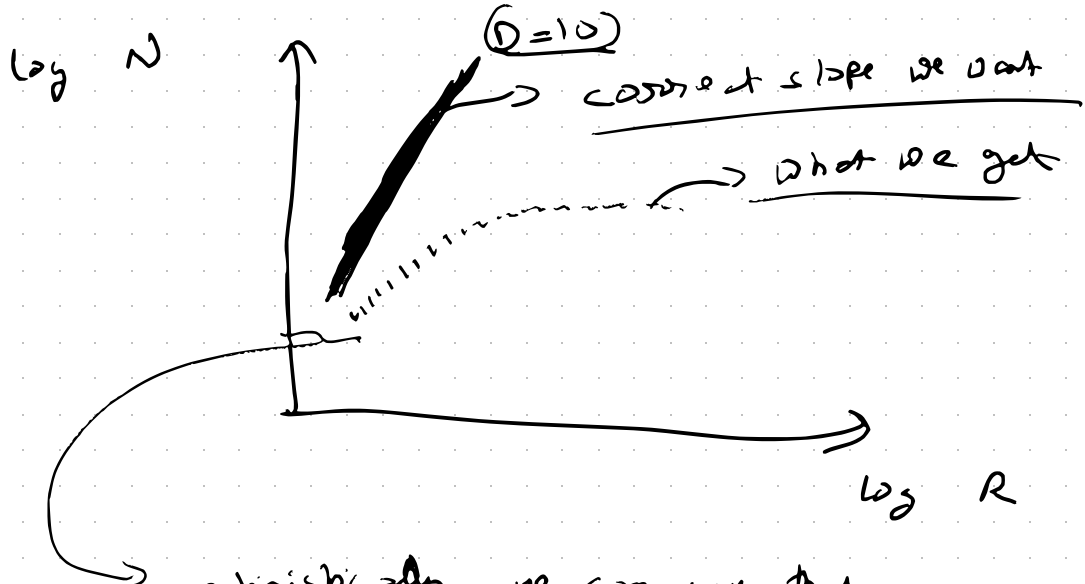
① We harvest data pts in a 10-D space from a Gaussian distrib^n of variance = 1

$X \in \mathbb{R}^{10}$

$x_i \sim N(0, 1)$

(Normal distrib^n)

# of data pts. = 1000 (say)



optimistically, we can say that as

$R \to 0$, slope $\to 10$, but we never really reach it.

⇒ What is the origin of this discrepancy?

prefactor = A → density

$N(R) = \rho \, R^{ID} \, \Omega_{ID}$

(density $\int dx_1 \, dx_2 \ldots$)

(Vol^n of unit sphere in ID dimensions)

↳ in dim = 2 $\quad \Omega = \pi$

dim = 3 $\quad \Omega = \frac{4}{3}\pi$

($\pi R^2 \to R=1 \to \pi$)

$\ln N(R) = \ln \rho + ID \ln R + \ln \Omega_{ID}$

Assumption :- density is const. on this the dist. R

↳ if density varies the expression is much more complex. This leads to the discrepancy.

→ It can be shown that the scaling is correct for $\lim_{R \to 0}$, ∵ density would be const. in an infinitesimally small neighborhood.

$$\Rightarrow \boxed{ID = \lim_{R \to 0} \frac{d \ln(N)}{d \ln R}}$$

(i.e., $\log N(R) = \log A + ID \log R$ is valid only for very small R)

↳ So in neighborhood of very small R, this scaling is exact, even if density is pos^n dependent.

→ Unfortunately taking $R \to 0$ in practice is very hard ∵ ID needs to be very small & we have MANY data points.

→ This problem is not very relevant for fractals ∵ they were usually studied in low dim space.

→ This led to the origin of other techniques to measure fractal dimension / ID.

→ To recap, origin of problem:

$$\log N \simeq \log f + ID \log R$$

$\log N$ scales with density, so if density varies we're screwed ∵ we're trying to measure ID & f & ID are entangled.

⇒ How to disentangle γ & ID?

2) Two-NN estimator :-

→ for every point (i), we find the first & second nearest neighbours.

→ we asume the distance, r, of the these 2 nearest neighbours to (i).

$$r_i^{(1)} = \min_{j \neq i} \| x^i - x^j \|$$

$$r_i^{(2)} = \min_{\substack{j \neq i \\ j \neq 1^{st} \text{ nearest neighbour}}} \| x^i - x^j \|$$

→ For each (i), estimate

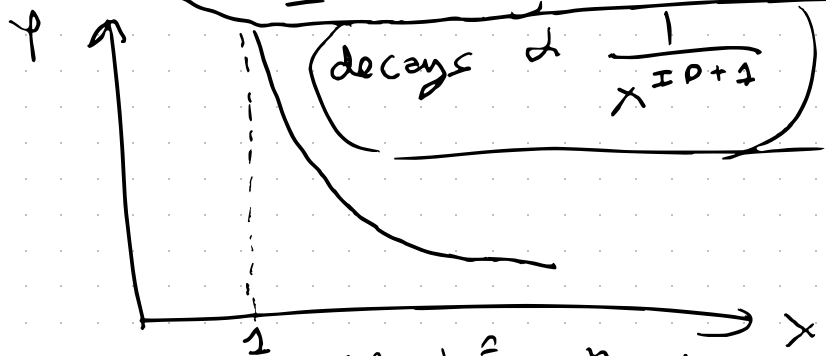
$$\mu_i = \frac{r_i^{(2)}}{r_i^{(1)}} > 1 \quad (\text{always, by def}^n)$$

→ it can be proved that regardless of ($f \rightarrow$ density), μ_i can be thought of as harvested from a probability density that is a Pareto distrib^n parameterized by the ID.

$$\mu_i \sim \text{Pareto}(ID)$$

→ Pareto Distrib^n :- prob. density.

$$f_{ID}(x) = \frac{ID}{x^{ID+1}}, \quad x \geq 1$$
$$= 0, \quad x < 1$$



→ power-law distrib^n with decay as $(ID+1)$.

→ $\int_{1}^{+\infty} p_{ID}(x)\, dx = 1$

normalized over the density,
Pareto distrib^{n} should be $= 1$

→ regardless of f, μ_i can be thought to be harvested from Pareto distrib^{n}.

$$\mu_i \sim \text{Pareto}(ID), \forall f$$

☺ Qualitative reason why f (density) prefactor disappears from the prob. density :- we considers μ^i, which is a ratio btw^{n} $r^{(2)}$ & $r^{(1)}$.

→ if density is high, & data pts are close by, both $r^{(1)}$ & $r^{(2)}$ are typically small.

⟹

→ if we are in the region of very low density, $r^{(2)}$ & $r^{(1)}$ are both typically large.

→ ∴ though dist. scales with density ratio of dist, r such that density factor disappears

∴ Ratio of distances (μ_i) scales in a manner which is density independent,

→ ∴ as each $\mu_i \sim \text{Pareto}(ID)$ we can infer ID as described now.

→ $$P(\mu_i \mid ID) = \frac{ID}{\mu_i^{ID+1}}$$

(prob. of observing μ_i, given ID)

→ for N data pts we have several observations of (μ_i)

$$P(\mu_1 \ldots \mu_N \mid ID) = \prod_i \frac{ID}{\mu_i^{ID+1}}$$

assuming each observation of μ_i is indep. Strictly speaking this is NOT true, ∵ many μ_i's are of the nearest neighbours themselves but for large (N) & for simplicity's sake we can assume this

→ simple eg. of parameter inference given a model,

→ Here, the only parameter to be inferred is ID.

→ Let's use Bayesian Inference :-

$$\therefore P(\mu_1 \dots \mu_N \mid ID)$$

$$= \prod_i \frac{ID}{\mu_i^{ID+1}}$$

(prior of ID ≈ const. term)

$$= \frac{P(ID \mid \bar{\mu})\, P(ID)}{}$$

(maximize (P) w.r.t. (ID))

↓

typical procedure for maximum likelihood estimation (MLE)

↓

Find (ID) which maximizes $P(ID \mid \bar{\mu})$

↓

($\bar{\mu} \equiv$ given our observations)

→ we do a typical procedure of maximizing $\log P(ID|\bar{\mu})$, instead of $P(ID|\bar{\mu})$

$\therefore L(ID) = \log P(ID|\bar{\mu})$ $\quad [\bar{\mu} = \mu_1, \ldots, \mu_N]$

$$= \sum_{i=1}^{N} \log \frac{ID}{\mu_i^{ID+1}}$$

$$= N \log ID - (ID+1) \sum_{i=1}^{N} \log \mu_i$$

To maximize this :

$$\frac{\partial L}{\partial ID} = \frac{N}{ID} - \sum_i \log \mu_i = 0$$

$$\therefore \boxed{ID = \frac{N}{\sum_i \log \mu_i} = \frac{1}{\langle \log \mu \rangle}}$$

TL;DR :

(1) compute 1^{st} & 2^{nd} N.N. $(r_i^{(1)})$ $(r_i^{(2)})$ $\forall i$

(2) compute $\mu_i = \frac{r_i^{(2)}}{r_i^{(1)}}$ $\forall i$

(3) $$ID = \frac{1}{\langle \log \mu \rangle}$$

TWO-NN estimator algorithm

(4) much more numerically robust than correlation-dim estimator, ∵:

prob. distrib. $f(\mu)$ is indep. of f

& " " (N) " dep. on f.

TWO-NN Algorithm →

(1) Find first & second NN for every data pt. (i)

$r_i^{(1)}$ $r_i^{(2)}$

(2) Compute $\mu_i = \dfrac{r_i^{(2)}}{r_i^{(1)}}$ $\forall i$

(3) $\mu_i \sim$ Pareto (ID); the local density cancels out.

↓

allows us to disentangle ID & density estimates

$$ID = \frac{1}{\langle \log \mu \rangle}$$

⇒ Derivation of 2-NN :-

(·) Let's say we harvest data points in a region where the density is constant and is equal to (ρ).

(·) The vol^{m} of this region is (v).

$$P(n \mid v\ \rho) \equiv \text{Poisson distrib}^{n}$$

(prob. of harvesting (n) points, given v & ρ)

$$\therefore P(n \mid v\ \rho) = \frac{(\rho v)^{n}}{n!} e^{-\rho v}$$

$\rightarrow n = 0, 1, \ldots, \infty$

we can harvest any # of data pts

$$\sum_{n=0}^{\infty} P(n \mid v\rho) = 1$$

(normalization cond^{n})

→ if we fix the size of a region, this tells us how likely it is we observe a given # of pts. in this region.

→ All ID / density estimates depend on this prob. distrib^n.

→ The reason of & ID are entangled is because they appear together in this prob distrib^n (C^∞: $V = \Omega_{ID}$ R^{ID})

So, ID & ρ are entangled in the prob distrib^n.

V is always some form of a hypersphere in our derivations.

(1) Poisson distrib^n → prob. of having k pts from a region given the density.

→ Postulate of our derivation :- # of data that fall within a region of vol^{n} (V) with density (γ) are harvested acc. to a Poisson distrib^{n}.

→ What is the prob. of observing 0 data pts. in a region?

$$P(n=0 \mid \gamma V) = e^{-\gamma V} = P(\gamma V \mid 0)$$

→ Now we can use Bayes' theorem to estimate that density of vol^{n} V given that we have 0 data points in a region. This is the key ingredient of density of ID estimation.

⇒

i^{th} data pt

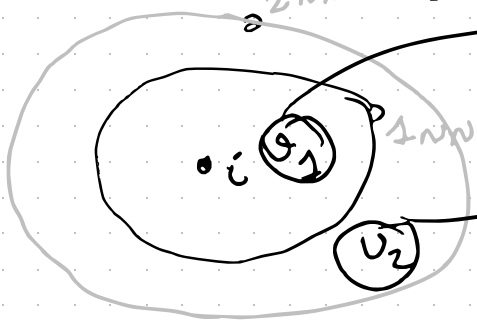
↳ draw a sphere with radius = dist to NN

centered at i^{th} data pt.

→ vol of sphere bto 2 i^{th} data pt

↳ its NN reduces ⇒ density ↑.

→ Consider 2 diff regions of empty space:



vol of empty space first NN (V_1)

empty space / vol in 2nd shell (V_2)

$$P(\rho v = \eta) = e^{-\eta} \quad [\eta = \rho v \text{ is written}]$$

$$\rightarrow \int_0^\infty e^{-\eta} d\eta = 1$$

(normalization)

$\rightarrow$ estimate $P(v|\rho)$

prob density of v given ρ

$\rightarrow P(x) \rightarrow y = a \cdot x \Rightarrow$ we want $P(y)$

$\therefore dx\, P(x) = dy\, P(y)$

$$P(y) = P(x) \frac{dx}{dy}$$

$$P(y) = P\left(\frac{y}{a}\right) \cdot \frac{1}{a}$$

$\Rightarrow x = \eta \rightarrow$ & $y = v$ & do the same process we get $P(v|\rho)$

$$\therefore \boxed{P(\vartheta|\rho) = \rho e^{-\rho\vartheta}}$$

$\rightarrow \vartheta_1 \sim \exp(\rho)$

$\vartheta_2 \sim \exp(\rho)$

$\rightarrow$ We want $P\left(\lambda = \frac{\vartheta_2}{\vartheta_1}\right)$

$$\therefore \boxed{P\left(\lambda = \frac{\vartheta_2}{\vartheta_1}\right) = \int_0^\infty d\vartheta_1 \int_0^\infty d\vartheta_2 \; \rho e^{-\rho\vartheta_1} \rho e^{-\rho\vartheta_2} \; \delta\left(\frac{\vartheta_2}{\vartheta_1} - \lambda\right)}$$

How? — delta-func?

---

① $x, y \rightarrow$ 2 random vars $\rightarrow$ harvested from $P(x)$ prob. density

② let's say we have another var, $s = f(x, y)$

⊙ how do we estimate $P(s)$?

$$P(s) = \int dx \int dy \; P(x) P(y) \, \delta(f(x,y) - s)$$

$$\Rightarrow \therefore P\left(\lambda = \frac{v_2}{v_1}\right) = \int dv_1 \int dv_2 \; p e^{-pv_1} p e^{-pv_2} \, \delta\left(\frac{v_2}{v_1} - \lambda\right)$$

$\Rightarrow$ (marginal prob distrib^n)

say $\frac{v_2}{v_1} = x \leftrightarrow \delta(x - \lambda)$

$$\therefore P(\lambda) = \int_0^\infty dv_1 \int_0^\infty dx \; v_1 \, e^{-pv_1} e^{-px v_1} \cdot p^2 \, \delta(x - \lambda)$$

$$= \frac{1}{1 + \lambda^2}$$

(first integrate over x $\therefore$ δ func^n disappears)

$\therefore$ $P(\lambda) = \dfrac{1}{1+\lambda^2}$

ρ has disappeared

while $P(v_1)$ $P(v_2)$ depend on ρ

$P\left(\dfrac{v_2}{v_1}\right)$ is indep. of ρ.

This allows us to disentangle ID & density estimation.

$\therefore v_1 = \mathcal{L}_{ID}(x^{(1)})$ ID (vol$^{\supset}$ hypersphere in dim ID)

$$\therefore v_2 = S_{ID}\left(\left(r^{(2)}\right)^{ID} - \left(r^{(1)}\right)^{ID}\right)$$

$$\therefore \lambda = \frac{v_2}{v_1} = \left(\frac{r^{(2)}}{r^{(1)}}\right)^{ID} - 1$$

$$\frac{r^{(1)}}{r^{(1)}} = \mu$$

$$\therefore \lambda = \mu^{ID} - 1$$

monotonic func^n that can be easily inverted

→ We know $P(\lambda) = \frac{1}{1+\lambda^2}$

∴ what is $P(\mu)$?

$$P(\mu) = P(\lambda(\mu)) \frac{d\lambda}{d\mu}$$

∴ one can go from $P(\lambda) \to P(\mu)$

$$\therefore P(\mu) = \frac{ID}{\mu^{ID} + 1}$$

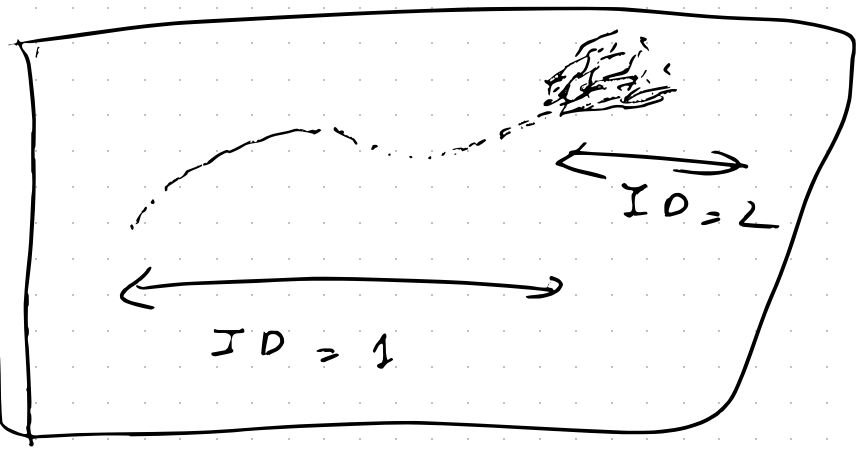
← Pareto distrib^n

∴ This is basically the derivation of the TWO-NN estimator.

→ Estimating ID is still a key open research problem.

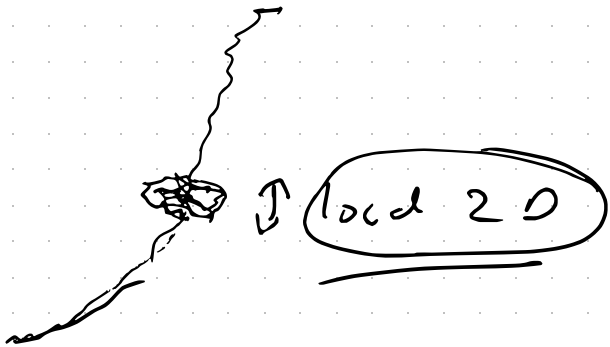
↓

∵ ID in diff. regions of our dataset can be diff. So pos^n dep. ID estimation is strongly desired.



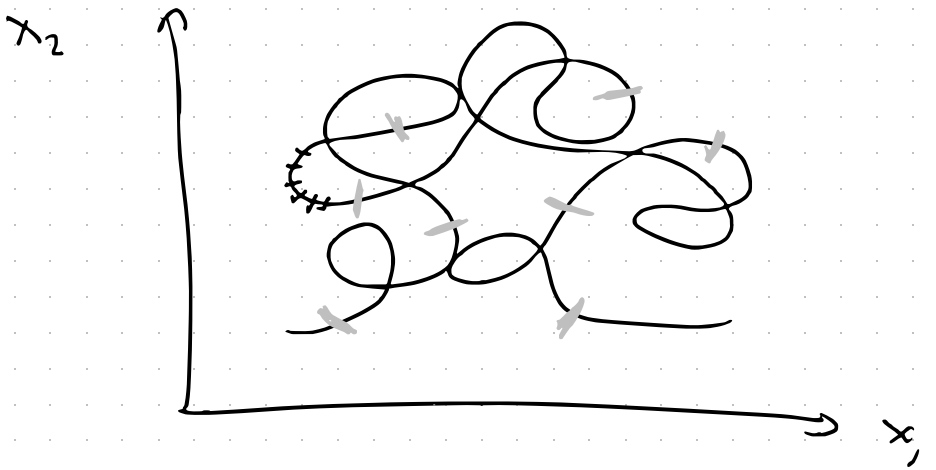
→ Another issue with ID estimation :-

data pts where manifold is 2D on a really local scale whereas the rest is 1D.



→ There are many more ID estimators each with its own advantages & disadvantages.

⇒ ID in time-series data :



→ in ID estimation often we assume Poisson distrib^n, we assume uncorrelated data. For time-series, data may be correlated;

→ let's look at $(x_2 - x_1)$ phase space traj. from an nD sim^n above.

→ If we harvest points very frequently, ID = 1 (if traj. in Hamiltonian dynamics, & ID = 1.x, if there's some stochasticity)

(∵ points are correlated)

→ if we harvest data pts. infrequently (grey pts.), in such a way that they are approximately independent, then the true ID of our manifold emerges