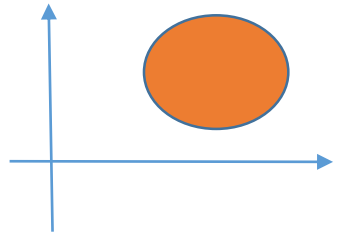
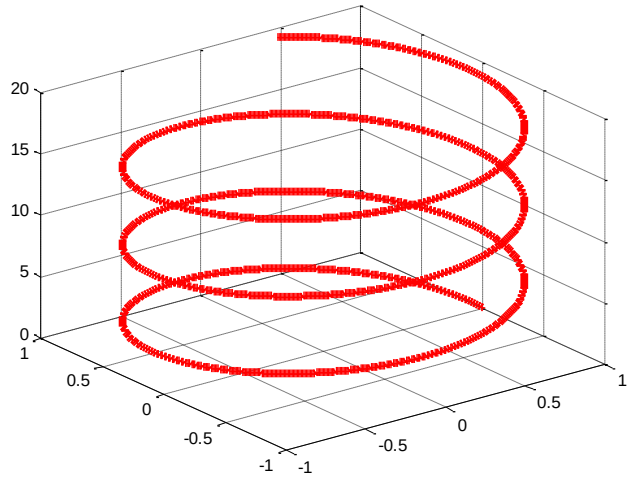
Lecture 5:

Intrinsic dimension estimation

The dimension of a subspace is the number of linearly independent vectors → any vector in this subspace can be described by a set of numbers whose number is equal to the dimension

Naive definition: the dimension of a set (subset) equals to the dimension of 'minimum-dimensional' linear subspace with which contains the set

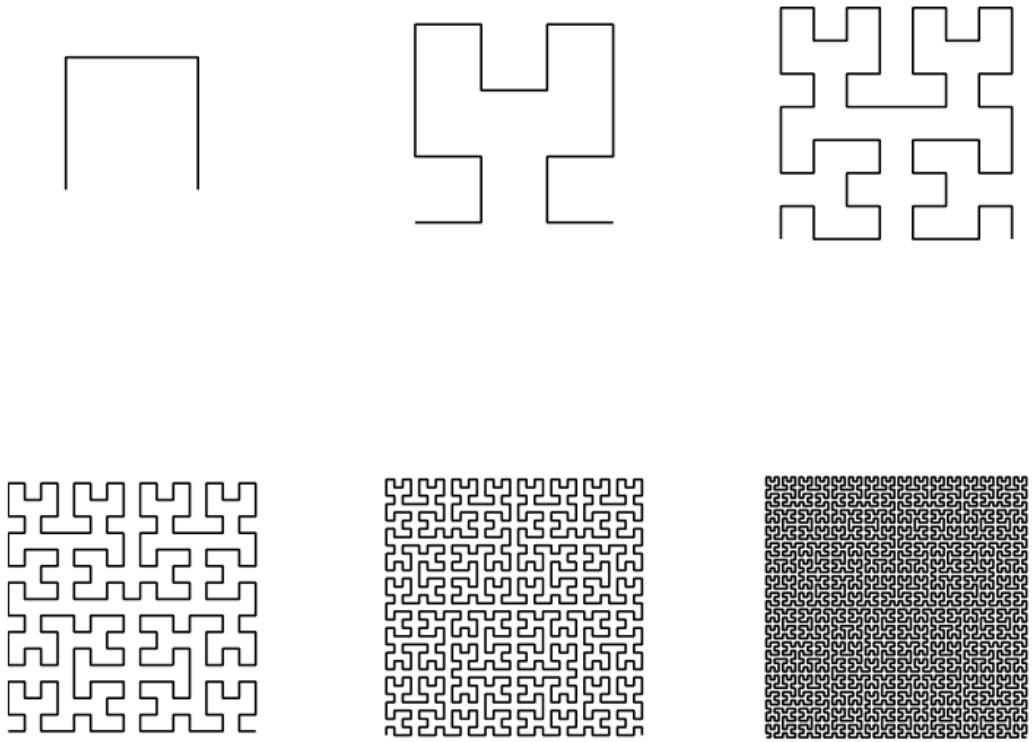




This curve in R^3 does not lie in any two-dimensional linear subspace but the curve's points can be described by one variable (t)

$$x(t) = \begin{pmatrix} \sin t \\ \cos t \\ t \end{pmatrix}$$

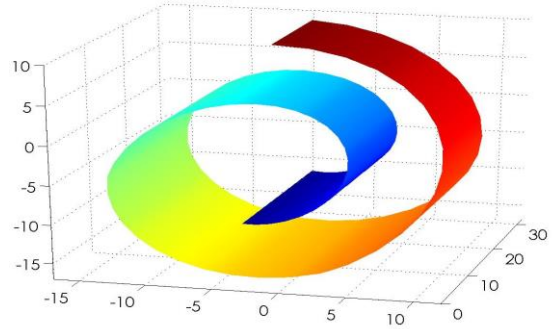
Intuitively: this curve in R^3 has **'intrinsic' dimension** equal to 1 - the curve can be 'straightened'



A family of Hilbert's space-filling curves:

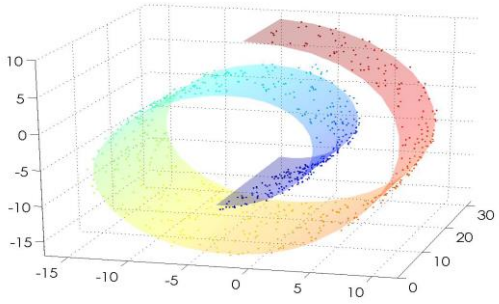
the one-dimensional curves from the family evolve iteratively and progressively fill a two-dimensional square - first six iteration steps

Let $C \subset R^p$. How define the concept of **'true'** intrinsic dimension of the set?



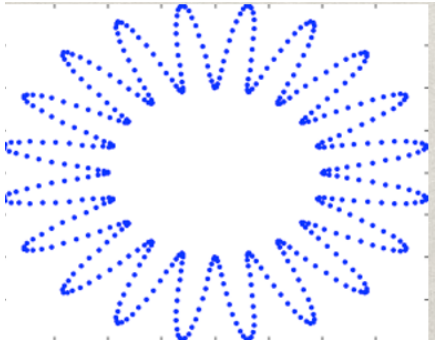
How estimate intrinsic dimension of a set when we

know only finite number of points from the set C?

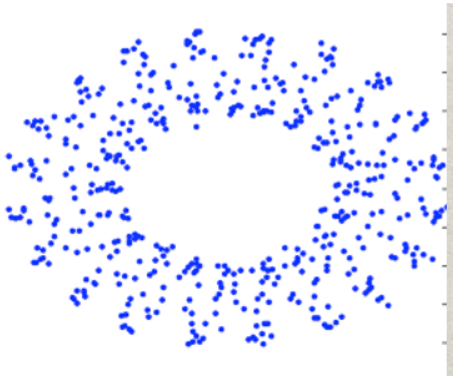


finite number of noisy points?

How define the concept of intrinsic dimension of a finite dataset and estimate this?



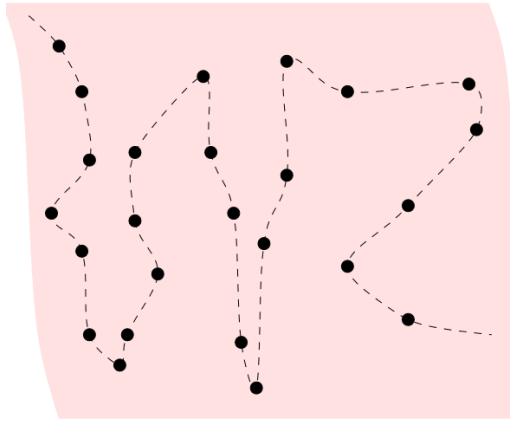
noisy dataset?



Intrinsic dimension of the '**solid**' set is defined in math using various geometrical notions: topological dimension, Hausdorff dimension, Hausdorff-Besicovitch dimension, Minkowski dimension, etc.

It is very difficult to estimate such defined intrinsic dimension if only a finite set of points is available: a priori a discrete sample has dimension zero.

We can make a 'solid set' of any dimension which 'pass' through the sample points: so, the sample may have dimension one according to one criterion, dimension two according to a different one, and so on



Curve or surface?

Intrinsic dimension of the '**solid**' based on arbitrary math definition does not exceed the dimension p of the ambient space

Intrinsic dimension (informal definitions)

General definitions:

- intrinsic dimension of a phenomenon: **the number of independent variables** that explain satisfactorily that phenomenon

- intrinsic dimension: the minimum number of parameters required to account of the observed properties

- intrinsic dimension: the optimal number of variable needed to capture the salient features of a generic dataset

- intrinsic dimension: the minimum number of parameters needed to represent the data sampled from the data structure without information loss

Definitions based on ‘data model’:

- latent variable model: the sampled p-dimensional points are actually be governed by a few latent variables. Intrinsic dimension: the minimal number of latent variables needed to describe the points

- data points are sampled from an unknown **geometric shape** (solid set). In latent variable model, this set consists of points corresponding to all values of latent variables (feature space) → **statistical problem of estimating** the unknown intrinsic dimension of this solid set from the sample

- the sampled points are **i.i.d. observations** over random vector whose **support** can be parameterized using a relatively small number of variables. Intrinsic dimension: the minimal number of such parameters

Math: Hausdorff intrinsic dimension

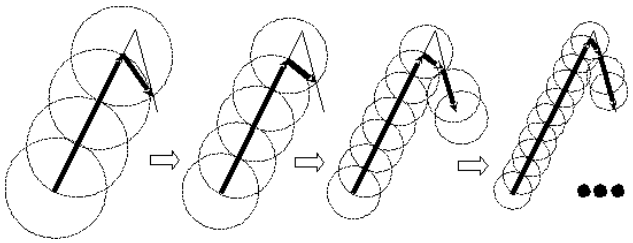
Let $C \subset R^p$ - bounded set and $S(X, r)$ – the r-ball centered at X

$\mathbf{S} = \{S(X_i, r_i)\}$ - set of balls such that

- $r_i \le r$
- $U_i S(X_i, r_i) \supset \mathbf{C}$

$\Gamma_H^d(r) = \inf_{\mathbf{S}} \sum_i (r_i)^d$ - infimum over all 'covering' sets **S** of balls – **minimal cover**

$$\Gamma_H^d = \lim_{r \to 0} \Gamma_H^d(r)$$



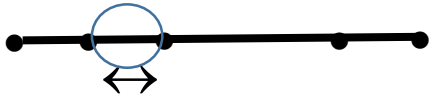
There exist the value D_H such that

- $\Gamma_H^d = \infty$ if $d < D_H$
- $\Gamma_H^d = 0$ if $d > D_H$
- nonzero constant if $d = D_H$

The value D_H – **Hausdorff intrinsic dimension** of the set **C**

Example 1: $C = [0, 1]$, $r = 1/(2n)$

Minimal cover consists of n balls



$\{S(i/n,1/(2n)), i = 0, 1, \ldots, (n-1)\}$

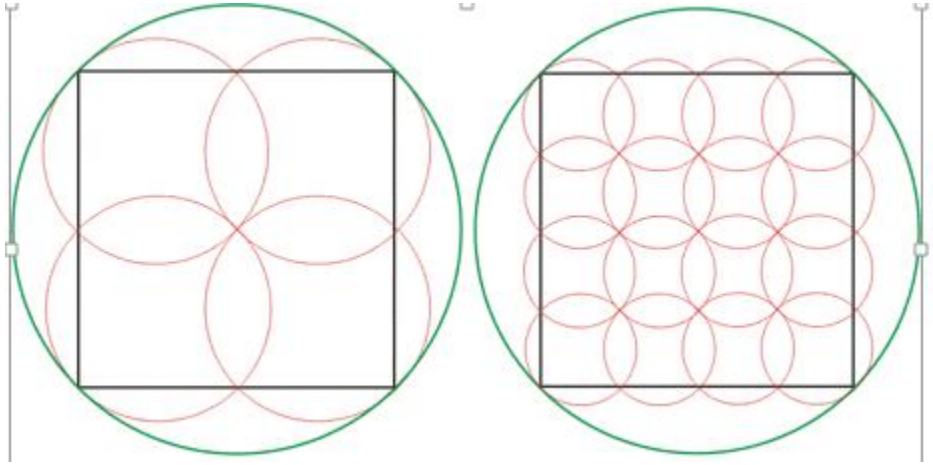
$\rightarrow \infty$ if $d < 1$

$\Gamma_H^d(1/(2n)) = \sum_{i=1}^{n} \left(\frac{1}{2n}\right)^d = \left(\frac{1}{2}\right)^d \times n^{(1-d)}$ $\rightarrow 0$ if $d > 1$

$D_H = 1$

$\rightarrow 1/2$ if $d = 1$

Example 2: $C = [0, 1]^2$



$D_H = 2$

Kolmogorov capacity intrinsic dimension

Capacity intrinsic dimension

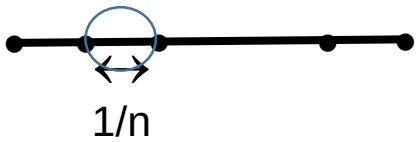
Box-counting intrinsic dimension

S = $\{S(X_i, r)\}$ - set of r-balls such that $\bigcup_i S(X_i, r) \supset \mathbf{C}$

$N(C, r)$ – minimum number of r-balls which cover the set C

Kolmogorov capacity intrinsic dimension of the set **C**: $D_{Cap} = \lim_{r \to 0} \frac{\ln N(C, r)}{\ln\left(\frac{1}{r}\right)}$

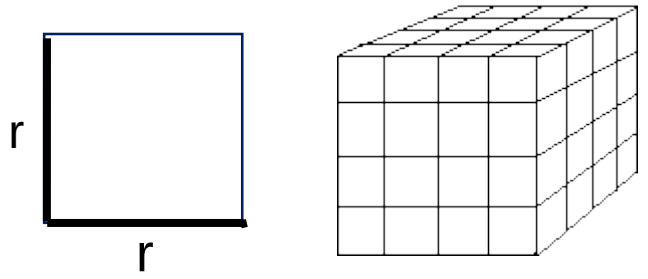
C = [0, 1], r = 1/(2n)



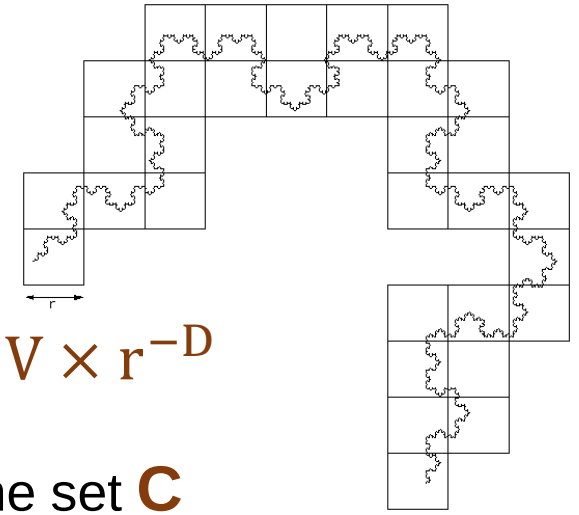
$$D_{Cap} = \lim_{r \to 0} \frac{\ln N(C, r)}{\ln\left(\frac{1}{r}\right)} = \lim_{n \to \infty} \frac{\ln n}{\ln(2n)} = \lim_{n \to \infty} \frac{\ln n}{\ln n + \ln 2} = 1$$

Hausdorff-Besicovitch intrinsic dimension

$E_r = \{[k\times r, k\times r + r)^p, k = 0, \pm 1, \pm 2, \dots\}$ – set of cubes with edge r



$N(C, r) = \#\{Q \in E_r: C \cap Q \neq \varnothing\}$



Let there exist the numbers V and D such that $\lim\limits_{r\to 0} \frac{N(C, r)}{V\times r^{-D}} = 1$ $N(C, r) \sim V \times r^{-D}$

The value $D(C) = D_{HB}(C)$ – **Hausdorff-Besicovitch intrinsic dimension** of the set **C**

The value $V(C) = V_{HB}(C)$ – **Hausdorff-Besicovitch volume** of the set **C**

Math: If the set **C** is measurable with respect to Jordan, then $D_{HB}(C)$ and $V_{HB}(C)$ equal to topological dimension $D_{top}(C)$ and volume $V(C)$, respectively

Correlation intrinsic dimension

$\{X_1, X_2, \dots, X_n\}$ – sample, Let n - sample size, $m(n) = n(n-1)/2$ – the number of pairs in the sample

$C_n(r) = \frac{1}{m(n)} \sum_{1 \le i < j \le n} I\{\|X_i - X_i\| \le r\}$ - 'correlation integral' $I(A)$ - indicator of event A

Correlation intrinsic dimension $D_{corr}(C) = \lim_{r \to 0} \lim_{n \to \infty} \frac{\ln C_n(r)}{\ln r}$

Direct applying of various intrinsic dimension definition for estimating purposes is impossible:

if C finite set consisting of n points then $N(C, r) \leq n$ and $D_H(C) = D_{HB}(C) = D_{cap}(C) = 0$

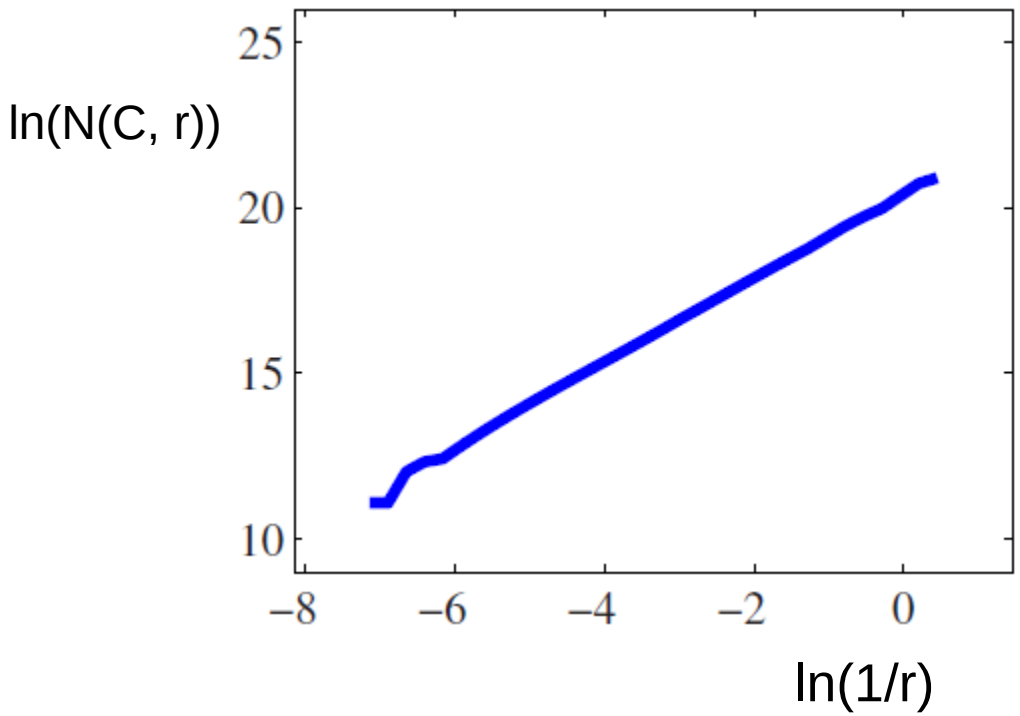
Intrinsic dimension estimation techniques founded on different notions of intrinsic dimension are based on their various approximations

$$D_{Cap} = \lim_{r \to 0} \frac{\ln N(C, r)}{\ln\left(\frac{1}{r}\right)}$$

$\ln(1/r) \to \infty$ and $\ln(N(C, r)) \to \infty$ as $r \to 0$

$$\lim_{r \to 0} \frac{\ln N(C,r)}{\ln\left(\frac{1}{r}\right)} = \lim_{r \to 0} \frac{\frac{\partial \ln N(C,r)}{\partial r}}{\frac{\partial \ln\left(\frac{1}{r}\right)}{\partial r}} = \lim_{r \to 0} \frac{\partial \ln N(C,r)}{\partial \ln\left(\frac{1}{r}\right)} \approx \frac{\ln N(C,r_2) - \ln N(C,r_1)}{\ln\left(\frac{1}{r_2}\right) - \ln\left(\frac{1}{r_1}\right)}$$

$$D_{Cap} = \lim_{r \to 0} \frac{\ln N(C, r)}{\ln\left(\frac{1}{r}\right)}$$



$D_{cap}(C) \approx$ Slope of log-log plots

The slope of log-log plots can be estimated using Least Squares technique from calculated values $\ln N(C, r)$ and $\ln r$ for a few small values of r

Intrinsic dimension estimation techniques

Intrinsic dimension estimation techniques are founded on different approaches and based on different notions of intrinsic dimension

Most of existing reviews about Intrinsic dimension estimation methods use various classification principles to observe such methods:

- global methods
- local methods
- pointwise methods

or

- projection methods
- geometric methods
- probabilistic methods

etc.

After that, various concrete procedures are described

Why we estimate an Intrinsic dimension (informally)

In Data analysis:

we want to find low-dimensional descriptions (low-dimensional features) of high-dimensional data

- ✓ only for available points (training dataset) already sampled from Data space (consists of all the points that can be sampled someday)
- ✓ for all points from Data space, including Out-of-Sample points

- Dimensionality reduction problem

To solve the Dimensionality reduction problem, we should know how many such features is needed - this number is informally called **Intrinsic dimension**

In Dimensionality reduction algorithms, this number d (Intrinsic dimension) is assumed to be known (given)

Dimensionality reduction algorithms find ‘the best’ d features under given d

Intrinsic dimension estimation from the Dimensionality reduction point of view (1):

'Trial and error' approach

Dimensionality Reduction: given dimension $q < p$ and p-dimensional dataset $\mathbf{X}_{(n)} = \{X_1, X_2, \dots, X_n\}$, find q-dimensional dataset (reduced feature dataset) $\mathbf{Y}_{(n)} = \{Y_1, Y_2, \dots, Y_n\} \subset R^q$ such that ***faithfully represents*** the sample $\mathbf{X}_n$

The term *faithfully represents* is determined by chosen certain **minimized** cost function $L_q(\mathbf{Y}_{(n)} | \mathbf{X}_{(n)})$ which reflects the desired properties of the sample that should be preserved (local data geometry, proximity relations, geodesic distances, angles, etc.)

Dimensionality Reduction algorithm: given sample $\mathbf{X}_n$, dimension q, cost function $L_q(\mathbf{Y}_{(n)} | \mathbf{X}_{(n)})$

finds q-dimensional dataset $\mathbf{Y}^*_{(n)} = \text{argmin } L_q(\mathbf{Y}_{(n)} | \mathbf{X}_{(n)})$

$L(q) = L(q|\mathbf{X}_{(n)}) = \min L_q(\mathbf{Y}_{(n)} | \mathbf{X}_{(n)}) = L_q(\mathbf{Y}^*_{(n)} | \mathbf{X}_{(n)})$ - 'achieved' minimum

Intrinsic dimension estimation: considering quantities $\{L(q)\}$ for various q, to choice minimal ‘appropriate’ value q^* that is taken as an estimator D for unknown ‘true’ intrinsic dimension of Data space

The ‘appropriate’ value q^* is not $\operatorname{argmin} L(q)$:

$$\min\{L_{q+1}(\mathbf{Y}_{(n)}|\mathbf{X}_{(n)}):\ \mathbf{Y}_{(n)} \subset R^{q+1}\} \leq \min\{L_{q+1}(\breve{Y}_{(n)}|\mathbf{X}_{(n)}):\ \breve{Y}_{(n)} = \{\begin{pmatrix} y_1 \\ \cdots \\ y_q \\ 0 \end{pmatrix}\} \subset R^{q+1}\}$$

$$= \min\{L_q(\mathbf{Y}_{(n)}|\mathbf{X}_{(n)}):\ \mathbf{Y}_{(n)} \subset R^{q}\}$$

$$\operatorname{argmin}_q L(q) = p$$

Intrinsic dimension estimation from the Dimensionality reduction point of view (2):

‘Low-dimensional features preserve information contained in high-dimensional data’

‘Low-dimensional features allows to recover initial high-dimensional data’

there exists a recovering mapping $g: y \in R^q \to g(y) \in R^p$ (y - reduced feature of vector X) such that

$$g(y) \approx X$$

Such recovering mapping exists if q is ‘true intrinsic dimension’ ~ the data can be described by q variables

Let X and X' - near p-dimensional points ($X \approx X'$) and y and y' - their q-dimensional features

$$X' \approx g(y') = g(y) + J_g(y)\times(y' - y) + o(y' - y) \approx X + J_g(y)\times(y' - y) + o(y' - y)$$

$$\approx X + J_g(y)\times(y' - y)$$

Taylor expansion:

$J_g(y)$ - $p\times q$ Jacobian matrix of the mapping $g: R^q \to R^p$

$X' \approx X + J_g(y) \times (y' - y)$ - for all near points x'

$L(X) = \{X + J_g(y) \times t: t \in R^q\}$ - **q-dimensional** affine subspace in R^p passing through point X

All sample points from small neighborhood of the selected point X lie approximately on q-dimensional affine subspace

Intrinsic dimension estimation:

- having breaking the sample points into clusters (regions) consisting of near points, to find affine subspace with ‘minimal dimensionality’ for each cluster which fits its points in the best way
- to combine the found dimensions of clusters into a single number that is taken as an estimator D for ‘true’ intrinsic dimension of Data space

More general breaking the Data space into small regions (Voronoi tesselation of Data space) - the solutions to certain *Computational geometry* tasks: to approximate unknown Data space by certain geometrical structures such as

- simplicial complex
- tangential Delaunay complex
- finitely many affine subspaces called ‘flats’
- k-means and k-flats, etc.

‘Low-dimensional features allows to recover initial high-dimensional data’

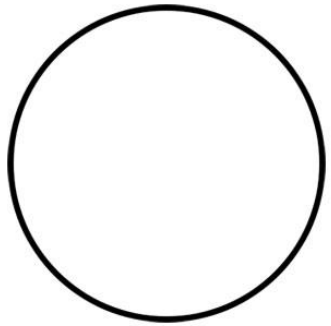
Data space **X** consisting of all points that ‘can be sampled someday’ **can be parameterized by**

q-dimensional parameter $y \in \mathbf{Y}$ where **Y** - Feature space ‘consisting of features of all points from the Data space’:

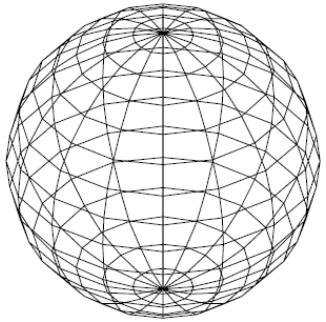
$$\mathbf{X} = \{X = g(y) \in R^q : y \in Y \subset R^q\}$$ with smooth mapping $g: R^q \to R^p$

Data space - q-dimensional smooth surface in ambient space R^p

Very strong assumption:



1-D sphere in 2-D



2-D sphere in 3-D

can not be described
using single mapping g

More realistic assumption: Data space can be described as q-dimensional smooth surface in ambient space R^p **‘locally’ -** strict assumption is based on **manifold concept**

Intrinsic dimension estimation: uses Data surface/Data manifold structure of the Data space

Intrinsic dimension estimation from the Dimensionality reduction point of view (3):

‘sampled points are i.i.d. observations over random vector whose support is unknown Data space’

Statistical structure of the sample depends on intrinsic dimension of the distribution support

Intrinsic dimension estimation:

- estimation of certain statistical characteristics of this random vector which depend on ‘true’ intrinsic dimension
- finding the intrinsic dimension from the constructed estimators

Intrinsic dimension estimation from direct use of Math intrinsic dimension definitions

Intrinsic dimension definitions use certain covers by small balls or cubes.

Main drawback: It is impossible to use the limiting transition along a tending to zero ball radius/cube edge r if only a finite set of points from solid set is available

An approach:

- calculating the until-limit values of used quantities for a few values of r
- constructing approximations to the limit value based on these until-limit values

Intrinsic dimension estimation from the Dimensionality reduction point of view (1):

'Trial and error' approach

Example 1: Intrinsic dimension estimation via PCA Dimensionality reduction procedure

'global method', 'projection method'

- $\{X_1, X_2, \dots, X_n\} \subset R^p$ – sample
- $\Sigma = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}) \times (X_i - \overline{X})^T$ – p×p sample covariance matrix
- eigenvalues $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p$ of the matrix Σ

Cost function: $\frac{1}{n}\sum_{i=1}^{n}\|X_i - \mathrm{Pr}_L(X_i)\|^2 \rightarrow \min$ under condition Dim L = q

$$L(q) = \min\left(\frac{1}{n}\sum_{i=1}^{n}\|X_i - \mathrm{Pr}_L(X_i)\|^2\right) = \sum_{k=q+1}^{p}\lambda_k$$

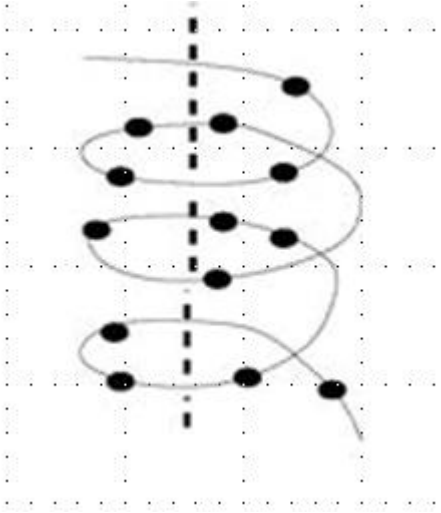
Commonly used choice of the most appropriate dimension:

$$D_{PCA} = q^* = \min\left\{q: \frac{\sum_{k=1}^{q} \lambda_k}{\sum_{k=q+1}^{p} \lambda_k} > \text{chosen threshold}\right\}$$

Another form of the same rule: $$D_{PCA} = q^* = \min\left\{q: \frac{\sum_{k=1}^{q} \lambda_k}{\sum_{k=1}^{p} \lambda_k} > \text{another chosen threshold}\right\}$$

Main drawback: PCA can overestimate the true intrinsic dimension for 'curved' sets:

$D_{PCA} = 3$ for 1-dimensional 'nonlinear' curve in R^3



Linear PCA technique is usually used as preliminary procedure:

- it finds affine linear subspace with minimal dimension that contains data
- after that, another 'nonlinear' technique should be used for estimating the intrinsic dimension of the 'PCA-projected data'

Intrinsic dimension estimation from the Dimensionality reduction point of view (2):

all sample points from small neighborhood of the selected point X lie approximately on q-dimensional affine subspace

Example 2: Intrinsic dimension estimation via 'local' PCA Dimensionality reduction procedure

'local method', 'projection method'

$X \in \mathbf{X}$ - selected point, $U(X)$ - **small** neighborhood of the point X, D - unknown intrinsic dimension

ε-neighborhood:

$$U(X) = \{X' \in \mathbf{X}_n : \|X' - X\| \leq \varepsilon\}$$

k Nearest Neighbors

$$X_{(1)}, X_{(2)}, \dots, X_{(n)} \in \mathbf{X}_n:$$

$$\|X_{(1)} - X\| \leq \|X_{(2)} - X\| \leq \dots \leq \|X_{(n)} - X\|$$

$$U(X) = \{X_{(1)}, X_{(2)}, \dots, X_{(k)}\}$$

The points $X' \in U(X)$ lie approximately on D-dimensional affine subspace

The points $X' \in U(X)$ lie approximately on **D**-dimensional affine subspace

$U(X)$ - neighborhood of the point X:

(1) should be sufficiently small to provide proximity the neighborhood' points to linear affine subspace

(2) $Card(U(X))$ should be sufficiently big to estimate the subspace

ε-neighborhood:

we manage the accuracy of approximation,

but do not control the number of points

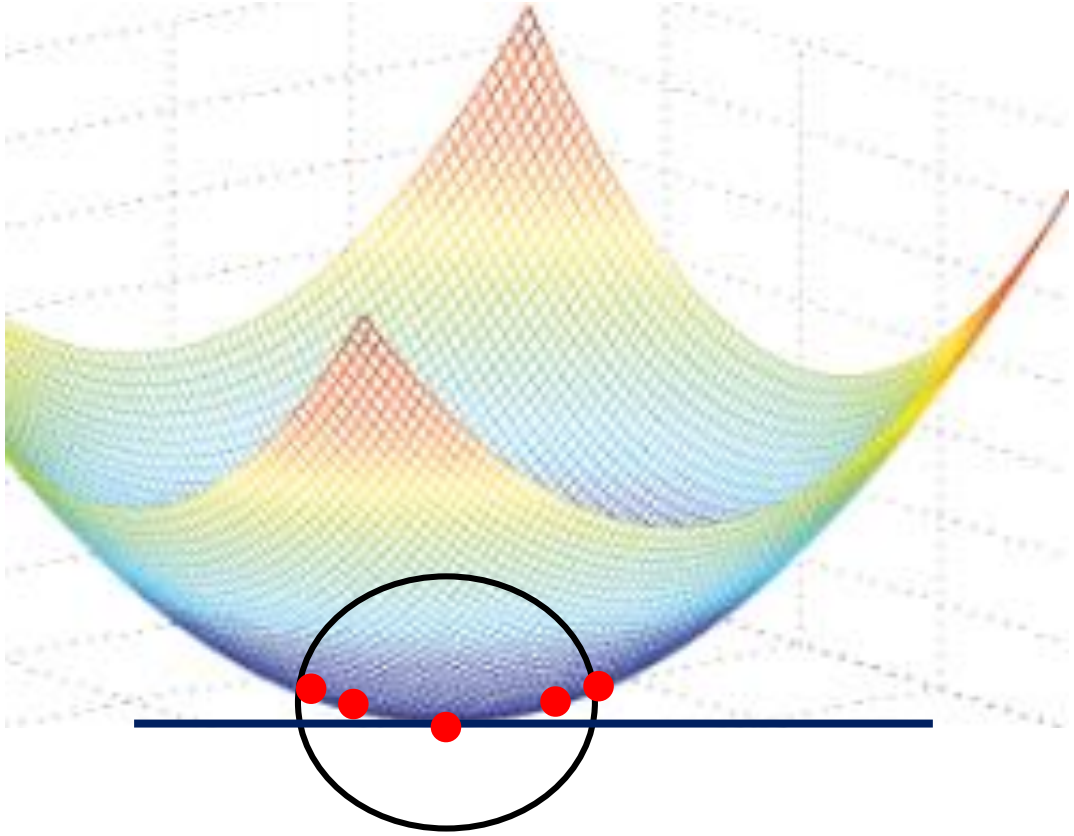
k Nearest Neighbors

we choose in advance the number of points (k) but

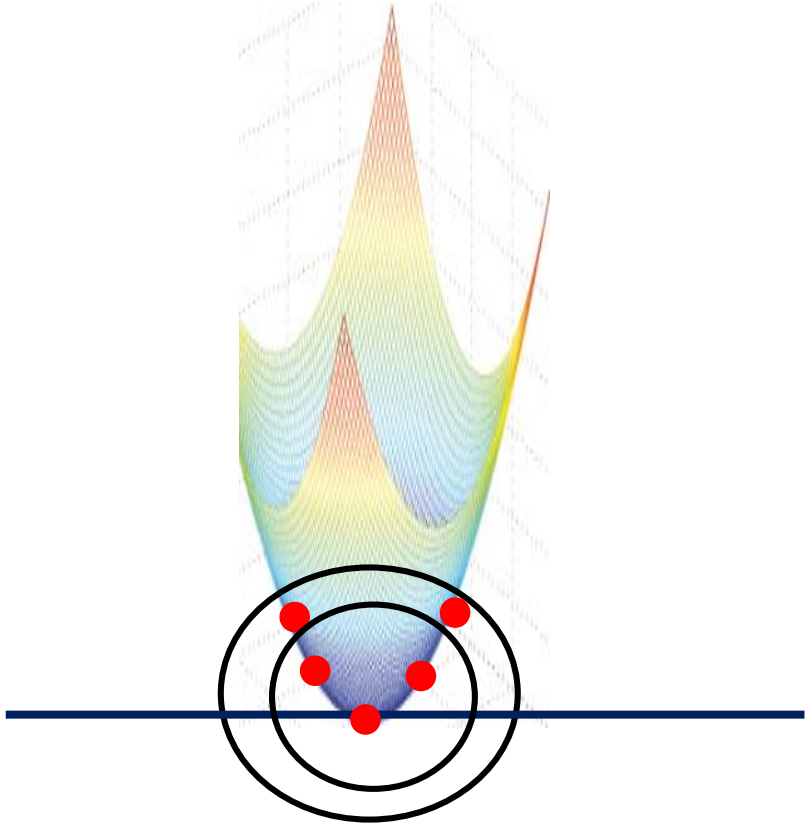
do not control the accuracy of approximation

Property (1) depends on sample size n

Property (2) depends on 'curvature' of the Data surface at selected point



Small curvature



Big curvature

Trade-off when we choose neighborhood' parameters (ε or k)

$\{X_1, X_2, \dots, X_n\} \subset R^p$ – sample

$X_i \to U(X_i)$ - neighborhood of the point X_i, $n_i = Card(U(X_i))$, $i = 1, 2, \dots, n$

Apply the PCA to each dataset $U(X_i)$:

$\Sigma_i = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (X_j - X_i) \times (X_j - X_i)^T$ – p×p matrix

Eigenvalues: $\lambda_{i1} \geq \lambda_{i2} \geq \dots \geq \lambda_{ip}$ of the matrix Σ

Denote: $\lambda_t = \frac{1}{n} \sum_{i=1}^{n} \lambda_{it}$ - averaged value of t^{th} eigenvalue, $t = 1, 2, \dots, p$

Based on ‘Trial and error’ approach, Intrinsic dimension is estimated by

$$D_{PCA} = q^* = \min\left\{q: \frac{\sum_{k=1}^{q} \lambda_k}{\sum_{k=q+1}^{p} \lambda_k} > \text{chosen threshold}\right\}$$

or:

$$D_{PCA} = q^* = \min\left\{q: \frac{\sum_{k=1}^{q} \lambda_k}{\sum_{k=1}^{p} \lambda_k} > \text{another chosen threshold}\right\}$$

Intrinsic dimension estimation from the Dimensionality reduction point of view (2):

all sample points from small neighborhood of the selected point X lie approximately in q-dimensional affine subspace

Example 3: Intrinsic dimension estimation via geometrical properties of the linear subspace

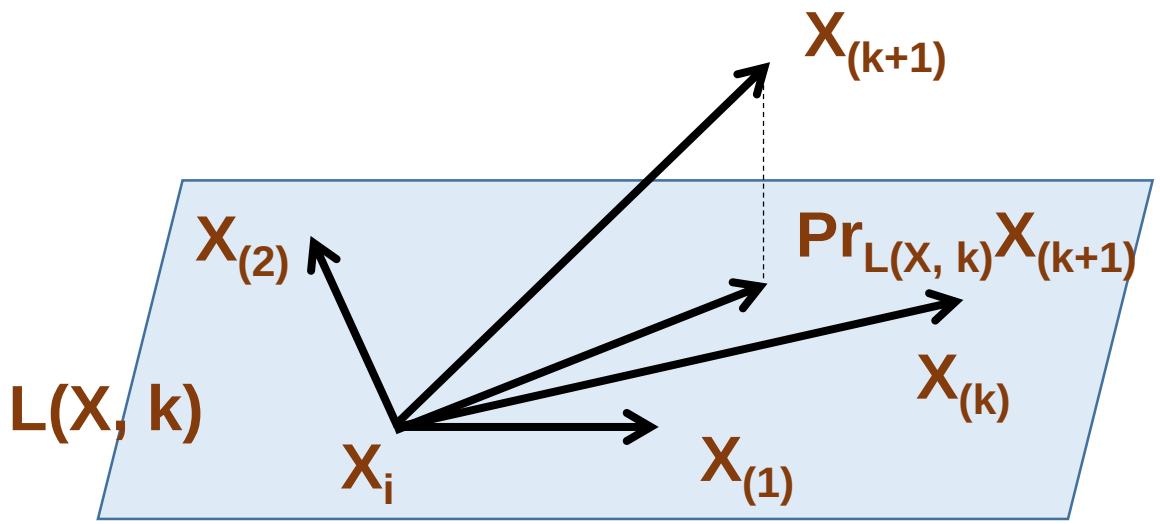
$X = X_i \in \mathbf{X}$ - selected sample point, D - true unknown intrinsic dimension and $X_{(1)}, X_{(2)}, \dots, X_{(k)} \in \mathbf{X}_n$ - its k Nearest Neighbors, excluding the point $X = X_i$ itself, that lie in the subspace

$$L(X, k) = \mathrm{Span}(X_{(1)} - X, X_{(2)} - X, \dots, X_{(k)} - X)$$

Let $k \leq D$ and assume that the points $X_{(1)}, X_{(2)}, \dots, X_{(k)}$ lie in 'general position' - it means that vectors $X_{(1)} - X, X_{(2)} - X, \dots, X_{(k)} - X$ are linear independent ones

If dataset points are sampled from Data space randomly and independently of each other, then a probability of the opposite 'patological' case equals to zero

Then $Dim(L(X, k)) = k$ and the vectors $X_{(1)} - X$, $X_{(2)} - X$, … , $X_{(k)} - X$ form basis in $L(X, k)$,



Then $(k+1)$-th Nearest Neighbor $X_{(k+1)}$ lies approximately in the subspace $L(X, k)$

this means that an angle $\alpha_i(k)$ between vector $X_{(k+1)} - X_i$ and subspace $L(X_i, k)$ is small

$$\alpha_i(k) = \arccos \frac{\left(X_{(k+1)} - X_i, Pr_{L(X_i,k)}\left(X_{(k+1)}\right) - X_i\right)}{\left|X_{(k+1)} - X_i\right| \times \left|Pr_{L(X_i,k)}\left(X_{(k+1)}\right) - X_i\right|}$$

Let

$$\alpha(k) = \frac{1}{n}\sum_{i=1}^{n}|\alpha_i(k)|$$

be averaged angle

Then Intrinsic dimension is estimated by

$$D = \min\{k: \alpha(k) < \text{chosen threshold}\}$$

Intrinsic dimension estimation from the Dimensionality reduction point of view (3):

'sampled points are i.i.d. observations over random vector whose support is unknown Data space'

Example 4: regression approach

Let Data space $\mathbf{X}$ has intrinsic dimension d and $\{X_1, X_2, \dots, X_n\} \subset R^p$ be the sample consisting of n vectors randomly extracted from Data space $\mathbf{X}$ independently of each other according to the probability measure with density $f(x)$ defined on $\mathbf{X}$

Consider a ball $B(X, \varepsilon)$ centered at $X \in \mathbf{X}$ with small radius ε and $X' \in \mathbf{X}_n$ - sample point randomly extracted from Data space $\mathbf{X}$. Denote $P(X, \varepsilon) = P\{X' \in B(X, \varepsilon)\}$ a probability that random point X' falls into the ball $B(X, \varepsilon)$

$$P(X, \varepsilon) = P\{X' \in B(X, \varepsilon) \cap \mathbf{X}\} = \int_{B(X,\square)\square\mathbf{X}} f(x)dx$$

Since Data space $\mathbf{X}$ has intrinsic dimension d, the points $X' \in B(X, \varepsilon) \cap \mathbf{X}$ lie near d-dimensional affine subspace $L(X)$ passing through point X. Then

$$P(X, \varepsilon) = P\{X' \in B(X, \varepsilon) \cap \mathbf{X}\} \approx \int_{B(X, \varepsilon) \cap L(X)} f(x)dx$$

The density $f(x)$ is smooth function on $x \in \mathbf{X}$, thus $f(x) \approx f(X)$ under $x \in B(X, \varepsilon)$. Thus:

$$P(X, \varepsilon) \approx \int_{B(X, \varepsilon) \cap L(X)} f(x)dx \approx f(X) \times Vol(B(X, \varepsilon) \cap L(X))$$

$B(X, \varepsilon) \cap L(X)$ - d-dimensional ball $B_d(\varepsilon)$ in R^d with radius ε - has a volume

$Vol(B_d(\varepsilon)) = \varepsilon^d \times Vol(B_d(1)) = \varepsilon^d \times V(d)$ $V(d) = Vol(B_d(1)) = \frac{\pi^{d/2}}{\Gamma\left(\frac{d}{2}+1\right)}$, Γ - gamma function

Thus, $P(X, \varepsilon) \approx f(X) \times \varepsilon^d \times V(d)$

$X = X_i \in \mathbf{X}$ - selected sample point, and $X_{(1)}, X_{(2)}, \ldots \in \mathbf{X}_n$ - its Nearest Neighbors, excluding the point $X = X_i$ itself

Consider a ball $B(X_i, r_k)$ centered at X_i with radius $r_k = \|X_{(k)} - X_i\|$

Then: exactly k points from n fall into the ball $B(X_i, r_k)$

Math: Let A be a random event with probability is P. Suppose that in N independent trials the event A occurs M times. Then $\frac{M}{N} \approx P$

Then: $\frac{k}{n} \approx P\{X' \in B(X_i, r_k)\} = P(X_i, r_k) \approx f(X_i) \times (r_k)^d \times V(d)$

$$\frac{k}{n} \approx f(Xi) \times (r_k)^d \times V(d)$$

$$\ln n - \ln k \approx \ln f(X_i) + d \times \ln r_k + \ln V(d)$$

$$\ln k \approx A_i + d \times T_k(i) \qquad A_i = \ln n - \ln V(d) - \ln f(X_i), \quad T_k(i) = -\ln \left\| X_{(k)} - X_i \right\|$$

$$\ln k \approx A_i + d \times T_k(i) \qquad k = 1, 2, \ldots \qquad i = 1, 2, \ldots, n$$ - regression equations

Then Intrinsic dimension d is estimated using standard Least Squares technique

Intrinsic dimension estimation from the Dimensionality reduction point of view (3):

'sampled points are i.i.d. observations over random vector whose support is unknown Data space'

Example 5: Maximum likelihood approach

$X \in \mathbf{X}$ - selected sample point, $B(X, t)$ - a ball centered at X with radius t, $0 \le t \le R$, R - small radius,

$$N(t, X) = \sum_{i=1}^{n} I\{\|X_i - X\| \le t\} = \sum_{i=1}^{n} I\{X_i \in B(X, t)\} = \#\{X_i \in B(X, t)\}$$

- **random process** on time t

$N(t, X)$: increasing random process

- piecewise-constant process with jumps 1 at points $X_{(1)}, X_{(2)}, \ldots \in \mathbf{X}_n$ - Nearest Neighbors of point X

- binomial random process: for each moment t:

$N(X, t)$ - binomial random variable = number of successes in n tests with probability of success

$$P(X, t) \approx f(X) \times t^d \times V(d)$$

Math (Poisson approximation): Let A be a random event with probability is P. If N is big, P is small and $\lambda = N \times P$ then P{in N independent trials the event A occurs M times} $\approx \frac{\lambda^M}{M!} e^{-\lambda}$

$N(X, t)$ is approximated by Poisson process with rate $\lambda(t) = f(X) \times V(d) \times d \times t^{d-1}$

$N(X, t)$ is observed on segment $0 \le t \le R = X_{(k)}$: likelihood function is

$$F(f(X), d) = \exp\left\{\int_0^R \ln \lambda(t) dN(X, t) - \int_0^R \lambda(t)\, dt\right\}$$

and should be maximized over unknown parameters $f(X)$ and d

Maximum likelihood estimator for dimension d at point X is

$$D_R(X) = \left[\frac{1}{N(X,R)} \sum_{j=1}^{N(X,R)} \ln \frac{R}{r_j(X)} \right]^{-1} \qquad r_j(X) = \left\| X_{(j)} - X \right\|, \; j = 1, 2, \ldots$$

Let we use k Nearest Neighbors $X_{(1)}, X_{(2)}, \ldots, X_{(k)}$ of point X and $R = X_{(k)}$:

$$D_k(X) = \left[\frac{1}{k-1} \sum_{j=1}^{k-1} \ln \frac{r_k(X)}{r_j(X)} \right]^{-1}$$

Intrinsic dimension is estimated by

$$D_k = \frac{1}{n} \sum_{i=1}^{n} D_k(X_i)$$

Maximum likelihood estimator for unknown density $f(x)$:

$$\hat{f}_R(x) = N(x, R) \times [V(D_R(x))]^{-1} \times R^{-D_R(x)}$$

Let we use k Nearest Neighbors $X_{(1)}, X_{(2)}, \dots, X_{(k)}$ of point X and $R = X_{(k)}$:

$$\hat{f}_k(x) = (k - 1) \times [V(D_R(x))]^{-1} \times (r_k(X))^{-D_R(x)}$$

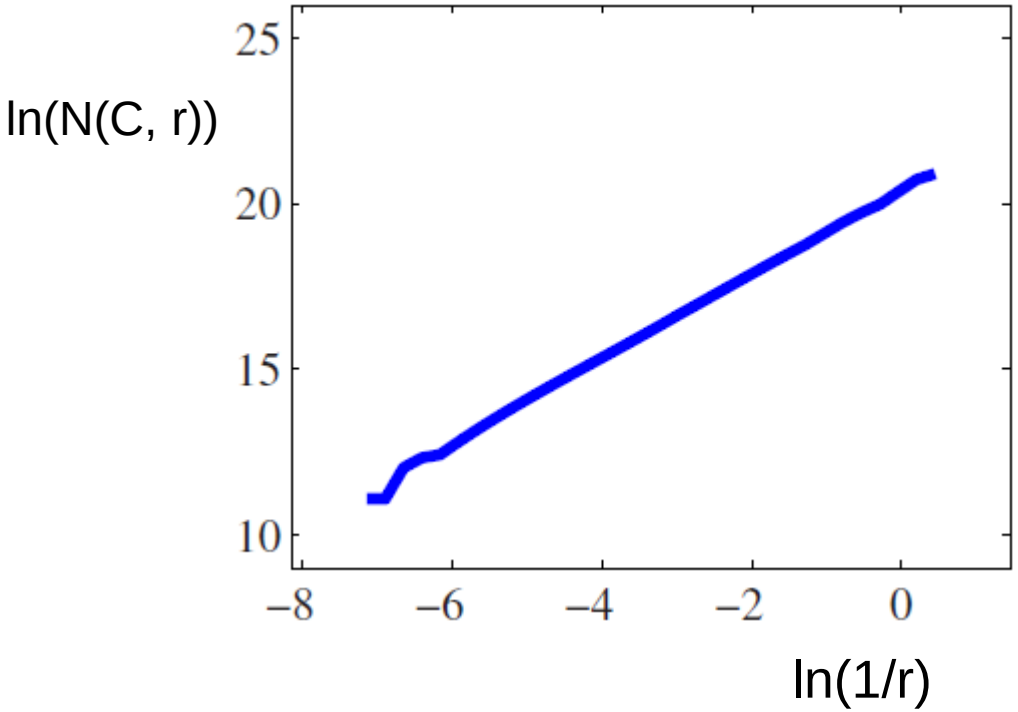
Entropy $J(f) = \int f(x) \log f(x) dx$ is estimated as

$$\widehat{J(f)} = \frac{1}{n} \sum_{i=1}^{n} \log(\hat{f}_R(X_i)) \quad \text{or} \quad \widehat{J(f)} = \frac{1}{n} \sum_{i=1}^{n} \log(\hat{f}_k(X_i))$$

Intrinsic dimension estimation from direct use of Math intrinsic dimension definitions

Example: Kolmogorov capacity intrinsic dimension

$$D_{Cap} = \lim_{r \to 0} \frac{\ln N(C, r)}{\ln\left(\frac{1}{r}\right)} \qquad D_{Cap} = \lim_{r \to 0} \frac{\ln N(C,r)}{\ln\left(\frac{1}{r}\right)} == \lim_{r \to 0} \frac{\partial \ln N(C,r)}{\partial \ln\left(\frac{1}{r}\right)} \approx \frac{\ln N(C,r_2) - \ln N(C,r_1)}{\ln\left(\frac{1}{r_2}\right) - \ln\left(\frac{1}{r_1}\right)}$$



$D_{cap}(C) \approx$ Slope of log-log plots

The slope of log-log plots can be estimated using Least Squares technique from calculated values $\ln N(C, r)$ and $\ln r$ for a few small values of r