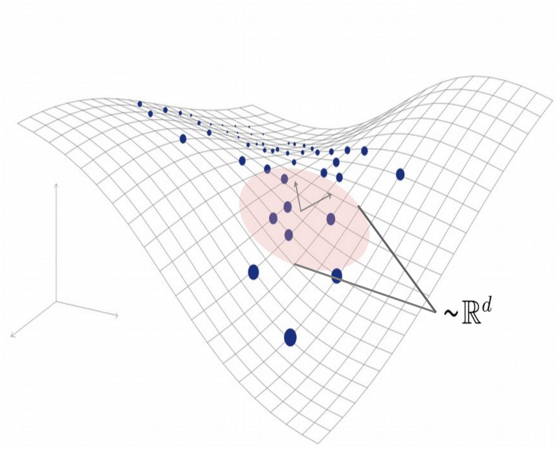
Michele Allegra

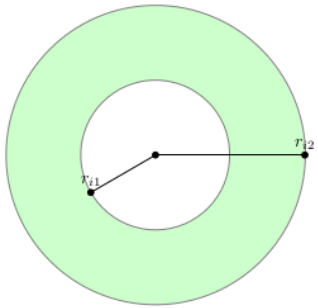
Clustering by the local intrinsic dimension

Overview

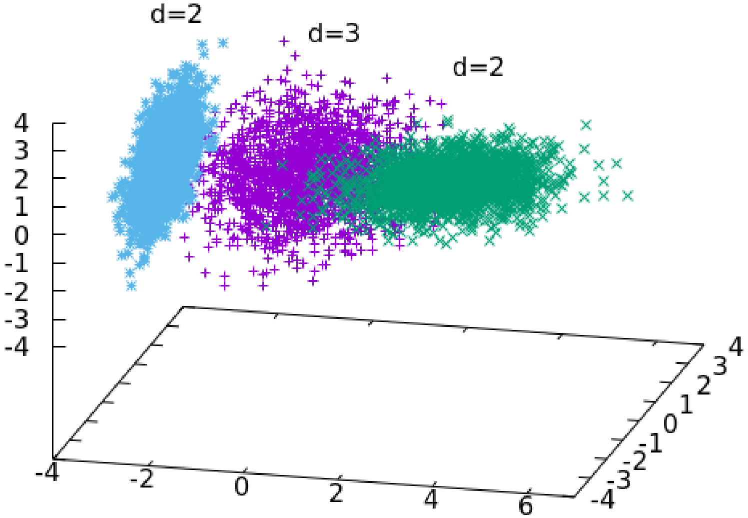
The intrinsic dimension of a dataset



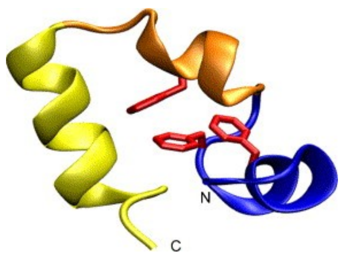
The TWO-NN approach for ID estimation



The case of variable ID

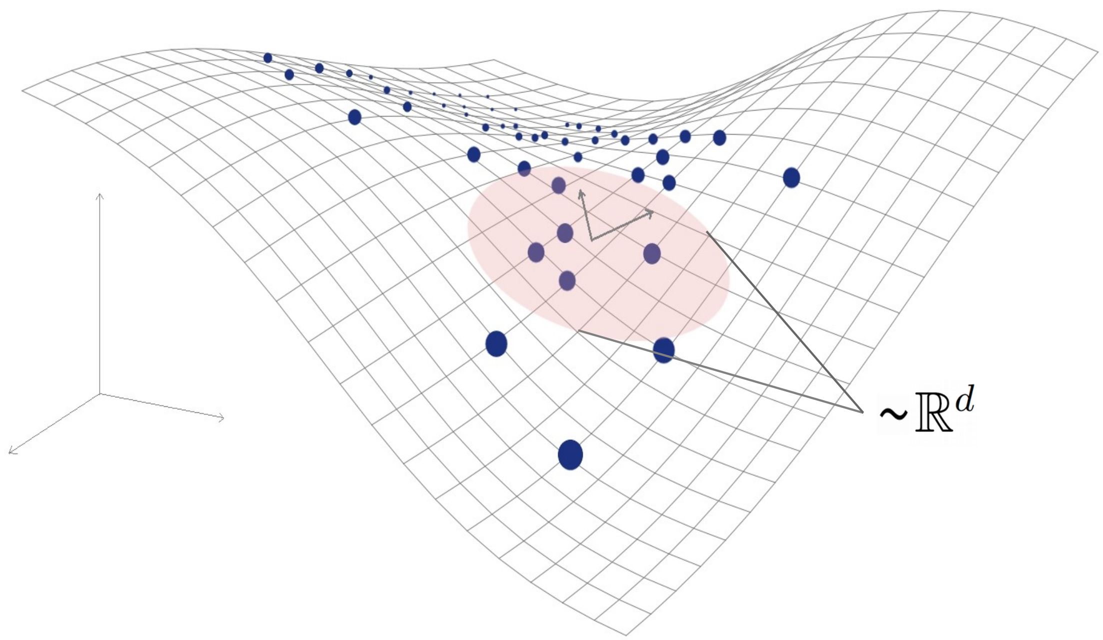


Application to a molecular dynamics trajectory



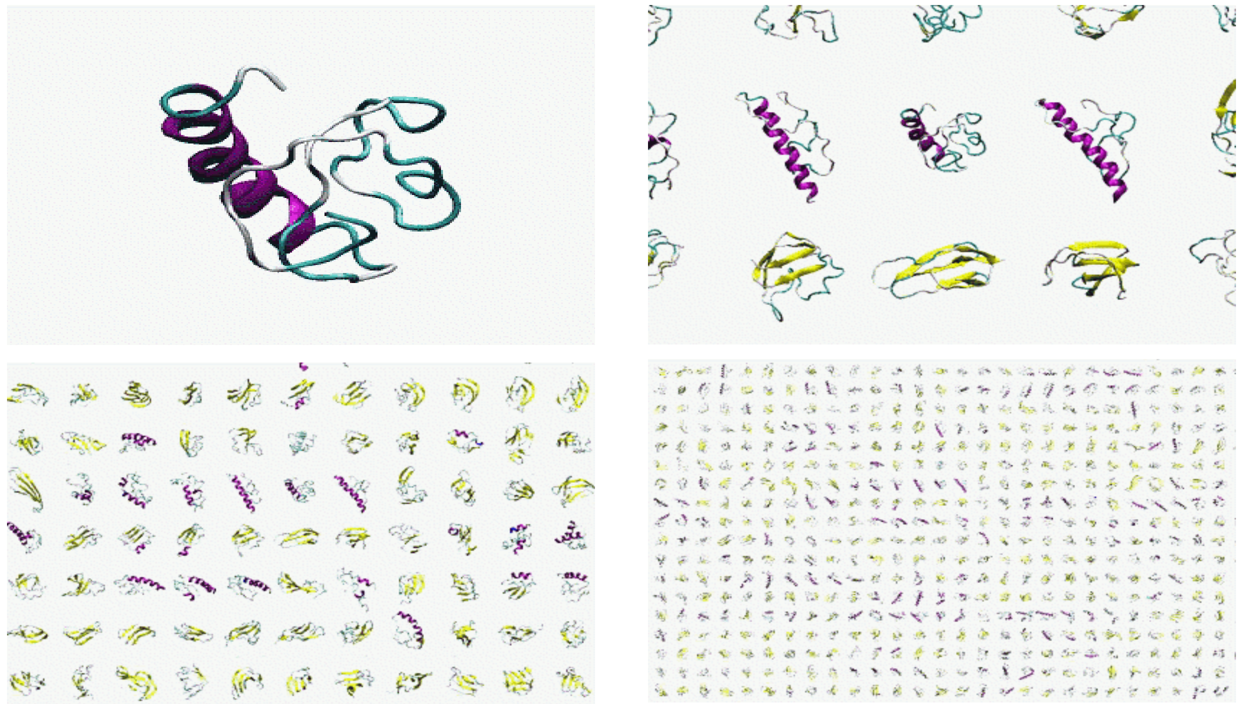
Intrinsic dimension

- Data are defined in a space with D variables
- However, the data lie on hypersurface of lower dimension d < D
- This dimension is called ***intrinsic dimension***



Intrinsic dimension

The state of a molecule is described by 6N variables



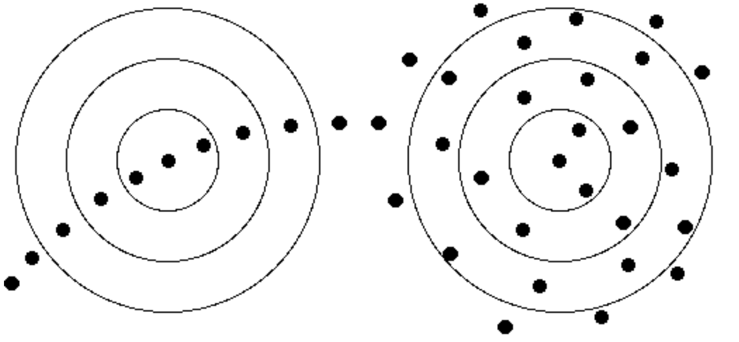
$$\underbrace{\boxed{\mathbb{R}}\boxed{\mathbb{R}}\boxed{\phantom{\mathbb{R}}}\boxed{\phantom{\mathbb{R}}}\boxed{\phantom{\mathbb{R}}}\boxed{\phantom{\mathbb{R}}}\boxed{\phantom{\mathbb{R}}}\boxed{\phantom{\mathbb{R}}}\boxed{\phantom{\mathbb{R}}}\boxed{\phantom{\mathbb{R}}}\boxed{\phantom{\mathbb{R}}}\boxed{\mathbb{R}}}_{6\mathrm{x}N}$$

Due to soft and hard constraints, the independent phase space directions are d << 6N

ID estimation

- Data are sampled from a distribution with density *ρ(X)*
- **If *ρ(X)* is onstant, distances between points in the dataset follow scaling laws that depend only on *d***
- Example: correlation dimension
    - If ρ(X) is constant, # of points at distance < ε from point *i* scales as

$$N_i(\varepsilon) \sim \varepsilon^d$$



    - *d* can be estimated with simple linear fit
    - when *ρ(X)* is variable, the scaling is violated, estimation fails dramatically

ID estimation: TWO-NN

E Facco, M D’Errico, A Rodriguez, A Laio, Scientific Reports 7, 12140. (2017)

- TWO-NN: estimating the ID in case of (strongly) variable density

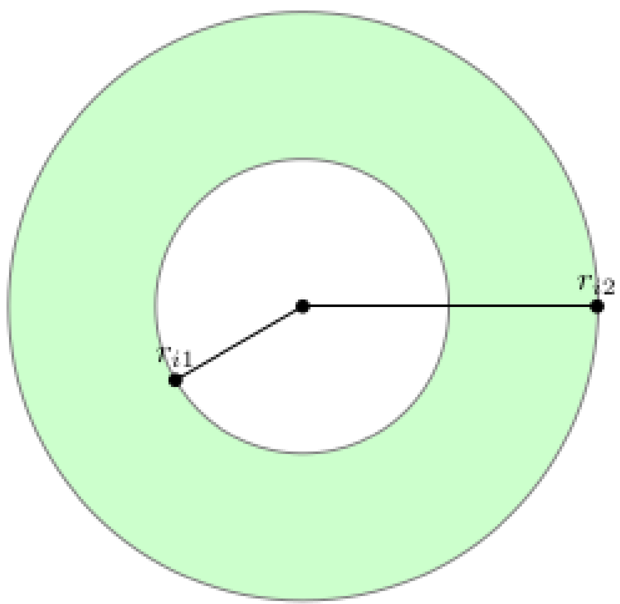
  Make two **broad assumptions**:

  - **H1)** the data points x_i are **independent samples** from a density ρ(x).
  - **H2) local uniformity:** ρ(x) ~ const. in the region containing the first 2 neighbors of x_i

- r_{i1}, r_{i2} distances of 1st and 2nd neighbor of point i

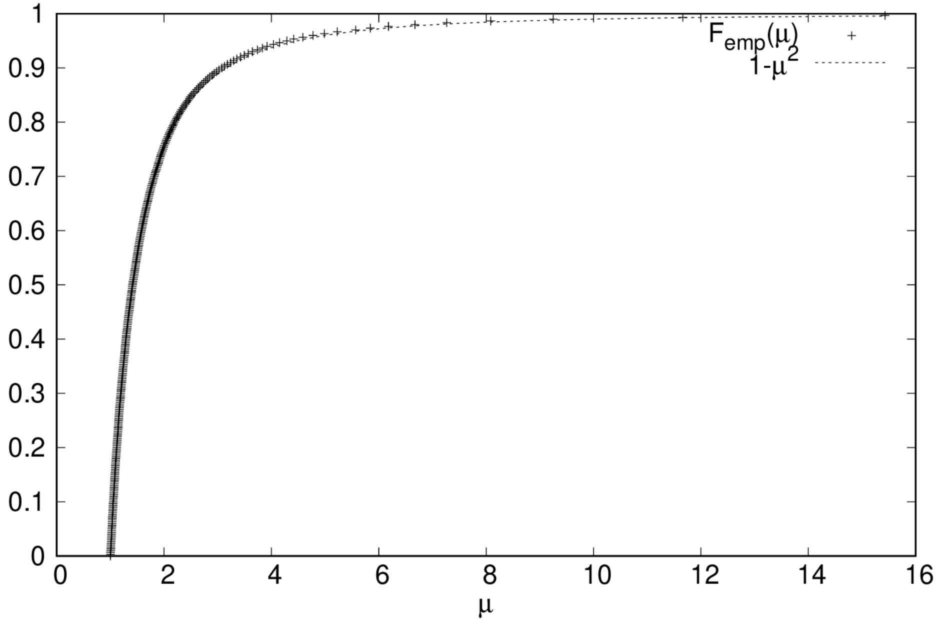
- $\mu_i = d_{i2}/d_{i1}$ follows a **Pareto distribution: $P(\mu) = d\mu^{-d}$**

- **The distribution of μ depends only on *d***

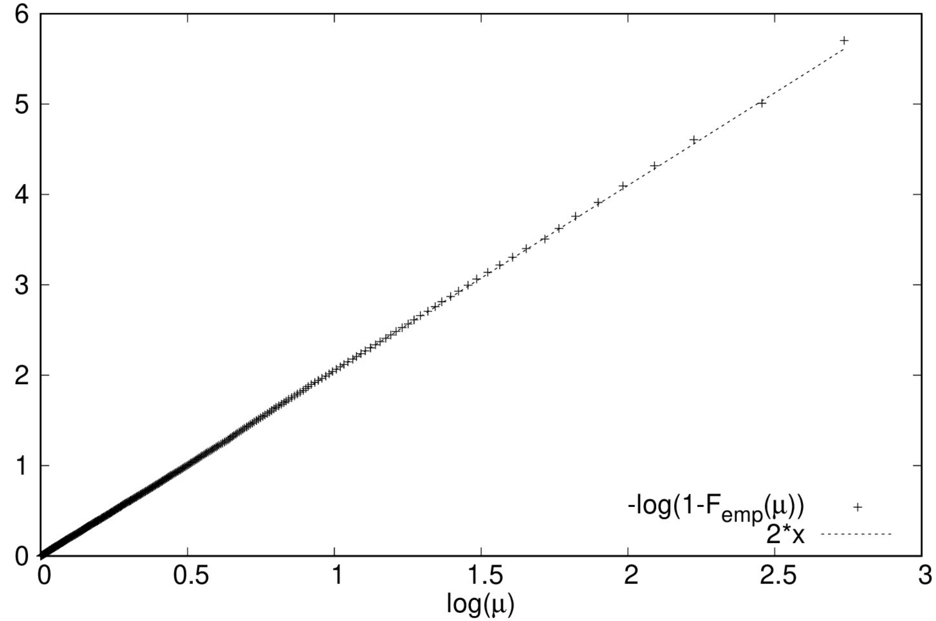


ID estimation: TWO-NN

- r_{i1}, r_{i2} distances of 1st and 2nd neighbor of point i
- $\mu_i = d_{i2}/d_{i1}$ follows a **Pareto distribution: $P(\mu)=d\mu^{-d}$** → **$F(\mu)=1-\mu^{-d}$**
- Fit the empirical cumulative distribution of the μ_i and estimate *d*
- Equivalently, linear fit on $\log(1-F(\mu))=-d \log\mu$



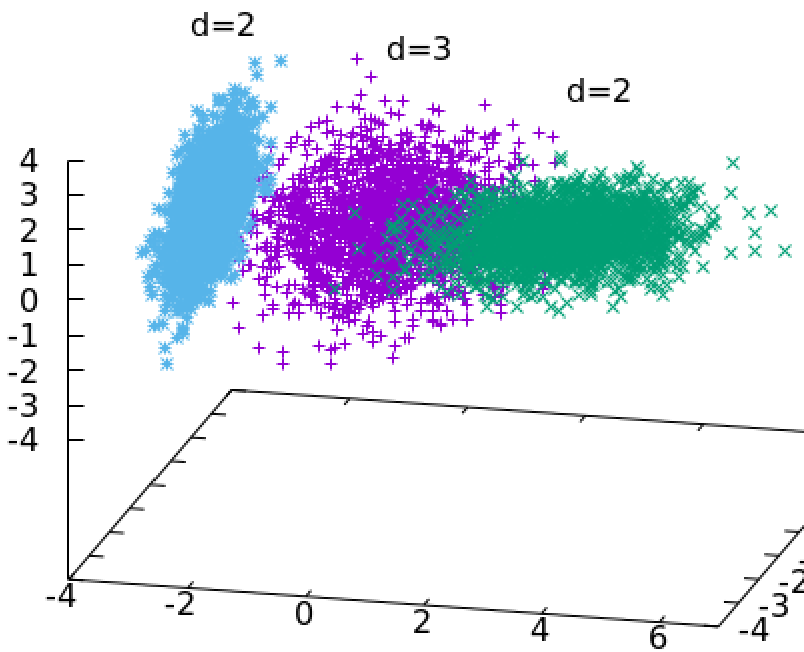
fit $F_{emp}(\mu)$ with $1-\mu^{-d}$



fit $\log(1-F_{emp}(\mu))$ with $-d \log\mu$

The problem of multiple IDs

the data may lie on several manifolds, each with different ID



Simple example: just merge two datasets with different ID

Is this an artificial oddity or a common situation?

Extending TWO-NN to multiple IDs

- TWO-NN assumptions:
  - **H1)** the data points x_i are **independent samples** from a density ρ(x).
  - **H2) local uniformity:** ρ(x) ~ const. in the region containing the first 2 neighbors of x_i

Additional assumption:

- H3) the distribution ρ(x) has support on K manifolds with different IDs $\mathbf{d}=d_1,\ldots,d_K$

- Under H1), H2), H3) the distribution of μ is simply a **mixture of Pareto distributions**

$$\mathcal{L}(\boldsymbol{\mu}|\mathbf{d},\mathbf{p}) = \prod_{i=1}^{N}\sum_{k=1}^{K} p_k d_k \mu_i^{-d_k-1}$$

Extending TWO-NN to multiple IDs

Estimate parameters **p,d** with Bayesian approach

- Fix $P_{prior}(\mathbf{d},\mathbf{p})$
- Compute posterior distribution $P_{post}(\mathbf{d},\mathbf{p}) \propto \mathcal{L}(\boldsymbol{\mu}|\mathbf{d},\mathbf{p}) P_{prior}(\mathbf{d},\mathbf{p})$
- Average $\mathbf{d}^e, \mathbf{p}^e = \langle \mathbf{d}, \mathbf{p} \rangle_{post}$

- to sample the posterior, we must introduce latent variables **Z**=$Z_1,\ldots,Z_K$
manifold membership of each point

$$\mathcal{L}(\boldsymbol{\mu}|\mathbf{d},\mathbf{p},\mathbf{Z}) = \prod_{i=1}^{N} p_{Z_i} d_{Z_i} \mu_i^{-d_{Z_i}-1}$$

- Estimate jointly **d,p,Z** by Gibbs Sampling of the posterior distribution

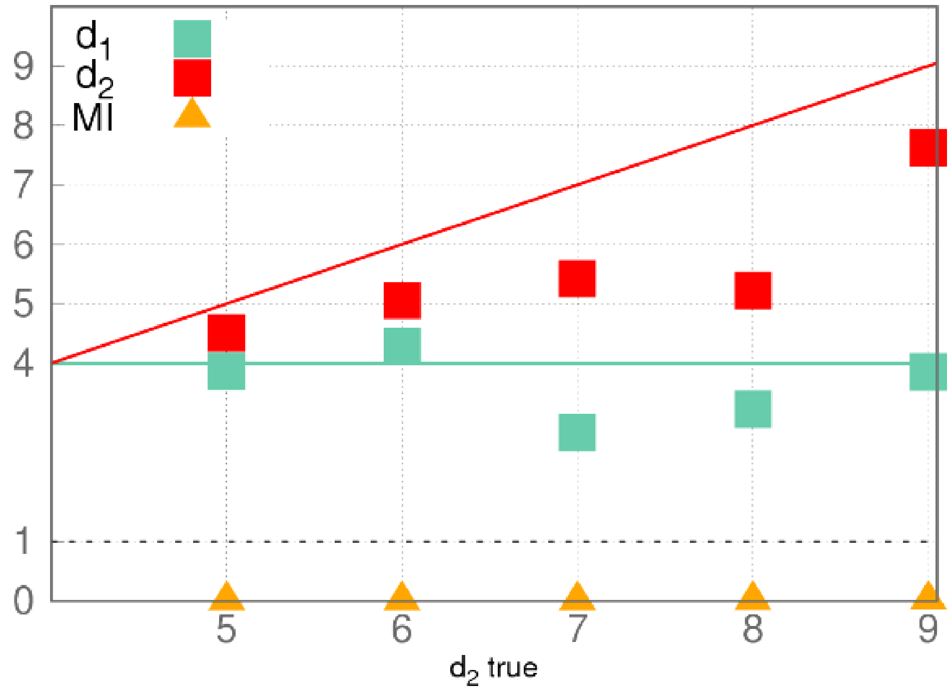
Extending TWO-NN to multiple IDs

Little problem: this approach does not work!

Two manifolds of dimension d_1=4 and d_2=5,..,9 (Gaussian ρ)

estimation of d_1 and d_2 is inaccurate

estimation of Z is completely wrong
(mutual information MI between true and estimated membership Z is 0)



Extending TWO-NN to multiple IDs

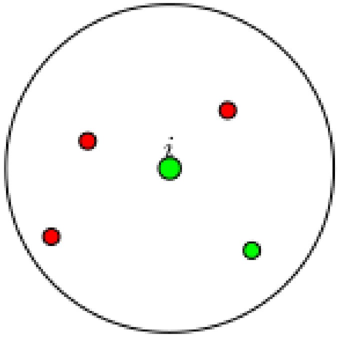
Let the neighborhood of point *i* be defined by its first *q* neighbors

n_i^{in} # neighbors with same Z as *i*

n_i^{out} # neighbors with diffferent Z

We get non-uniform neighborhoods: $n_i^{out} > n_i^{in}$

Problem in correctly estimating Z!



One more assumption:

H4) the manifolds have a small intersection:

neighborhoods must be approximately uniform

We enforce this through **additional term in the likelihood**

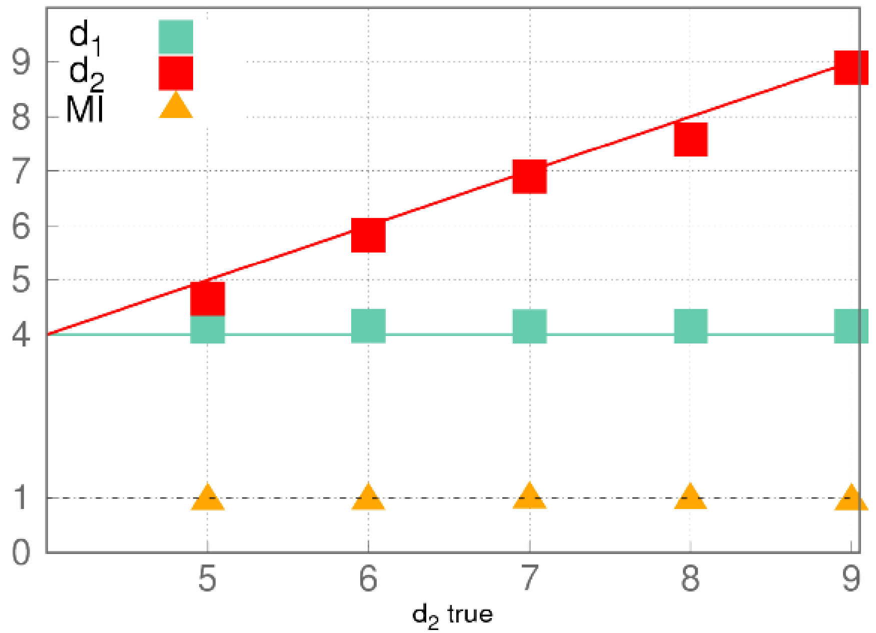
Extending TWO-NN to multiple IDs

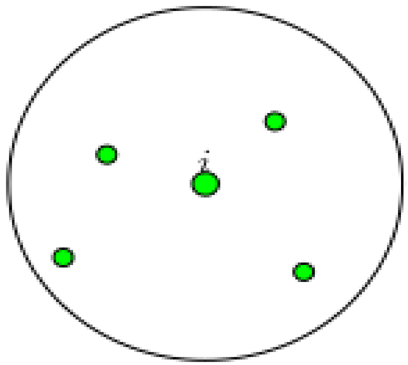
We enforce uniform neighborhoods through **additional term in the likelihood**

$$\mathcal{L}(n^{in}|\mathbf{Z}) = \prod_i \frac{\zeta^{n_i^{in}}(1-\zeta)^{n_i^{out}}}{\mathcal{Z}}$$

$\zeta > \frac{1}{2}$ Probability that two neighbors are in the same manifold

Now we get uniform neighborhoods and correct estimates!





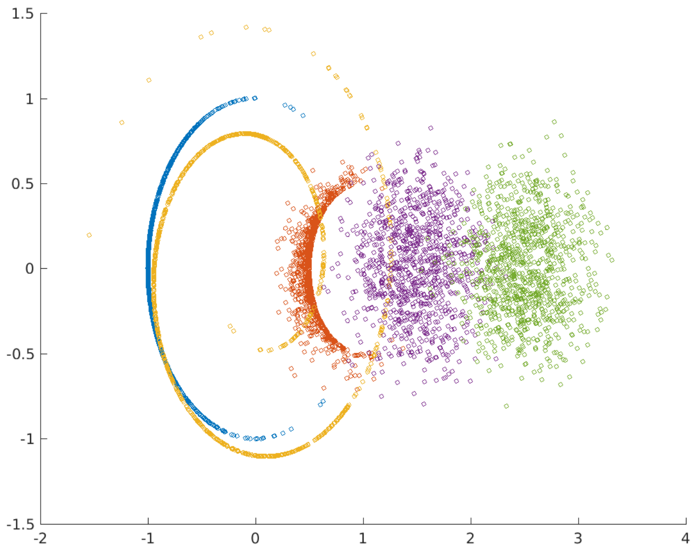
Heterogeneous ID algorithm (Hidalgo)

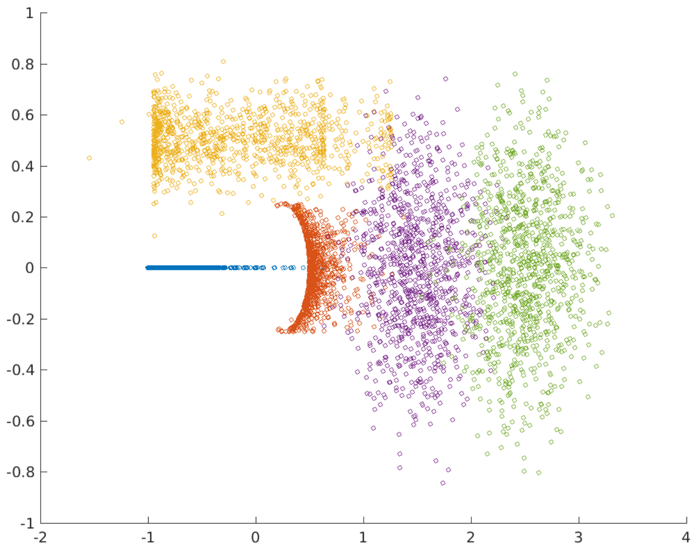
M Allegra, E Facco, A Laio and A Mira, arXiv:1902.10459 (2019)

Find regions (manifolds) of different ID in the data

Works also for nonlinear and topologically complex manifolds

E.g. circle in d=1, swiss roll in d=4, torus d=2, sphere d=5, sphere d=9





Heterogeneous ID algorithm (Hidalgo)

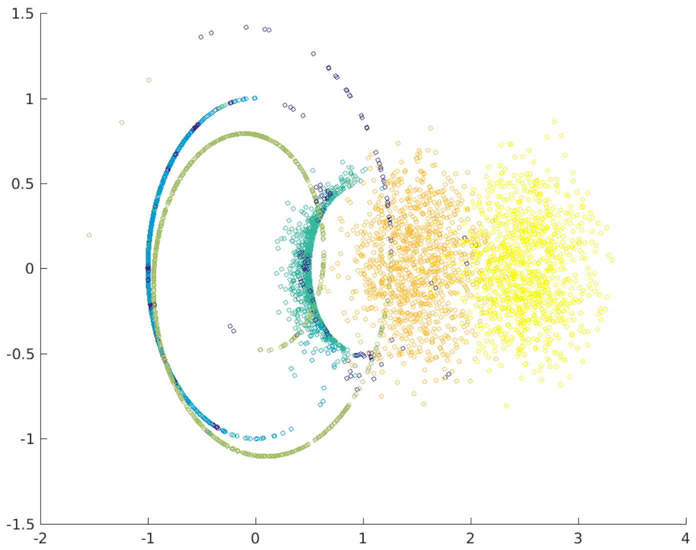
M Allegra, E Facco, A Laio and A Mira, arXiv:1902.10459 (2019)

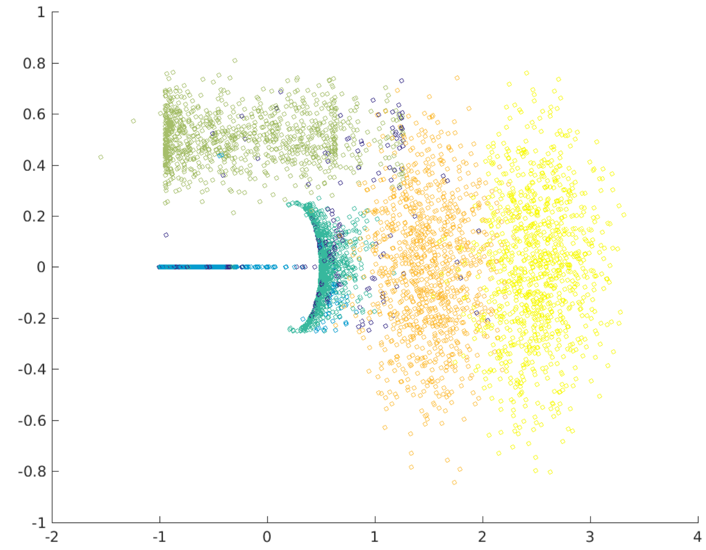
Find regions (manifolds) of different ID in the data

Works also for nonlinear and topologically complex manifolds

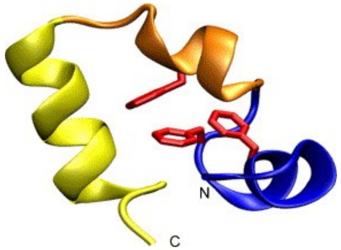
Circle d=1, swiss roll in d=4, torus d=2, sphere d=5, sphere d=9

Estimated dimensions 0.9,2.0,4.1,5.2,8.5





Real example: phase space of folding protein



- consider a simulation of unfolding/refolding villing headpiece
- for each of the N ~ 32,000 configurations, D=32 dihedral angles.

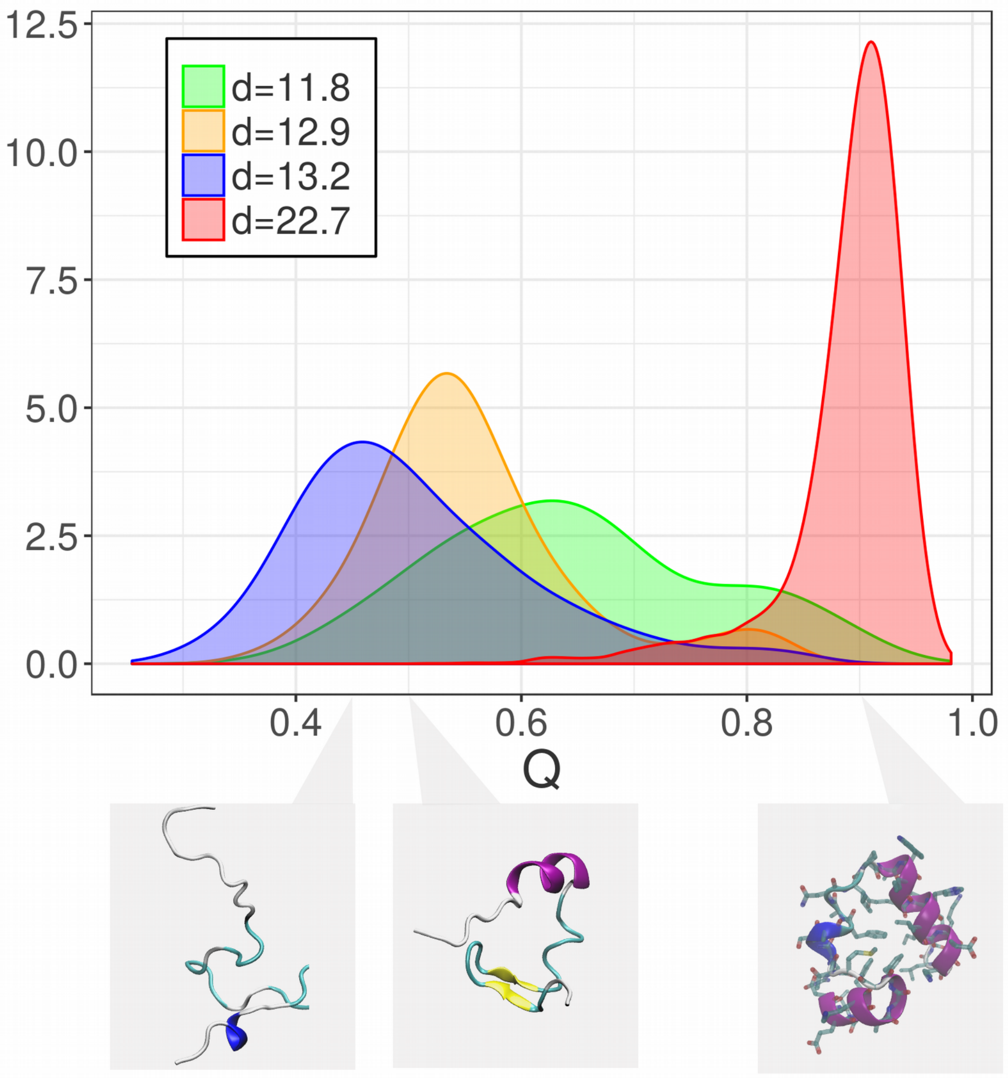
We find four manifolds,

- three with low dimensions d=11.8,d=12.9, d=13.2
- one with high dimension d=22.9

Which configurations are assigned to the different manifolds?

- Consider **q=fraction of native contacts (=degree of folding)**

Example: phase space of folding protein

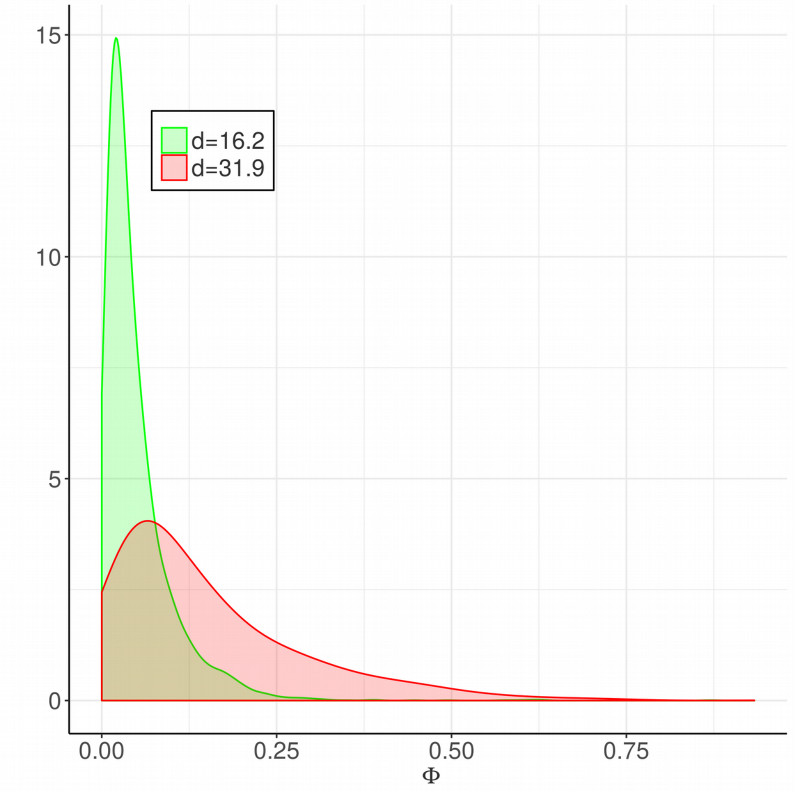


- **Folded configurations are in the high-dimensional manifolds**
- The local ID is able to discriminate between folded and unfolded configurations

The effective # of phase space directions the system can explore varies in the two states

Example: brain imaging time series

- Consider 202 fMRI images of the brain during visuospatial task
- N=40,000 brain voxels; for each voxel, BOLD time series with D=202 points
- **We find two manifolds with dimensions** d=31.9, d=16.1
- Consider Φ, “clustering frequency”, measuring how many times a voxels participates to transient coherent patterns



Companies with high Φ involvement are preferentially assigned to the high dimensional manifolds

Φ is related to task involvement

Conclusions

- We extended a recently developed ID estimator, TWO-NN, to the case where the ID is variable in a single dataset
- The method rests on quite weak assumptions (local uniformity of density and dimension)
- We find regions of different local ID in the data
- In real data, we find large variations of the ID, highlighting relevant structure in the data
- ID estimation is not just a preliminary step, but can highlight structure in the data

Acknowledgments

Alessandro Laio

Antonietta Mira

Elena Facco

Thanks for the invitation!

Aldo Glielmo

Aalto University

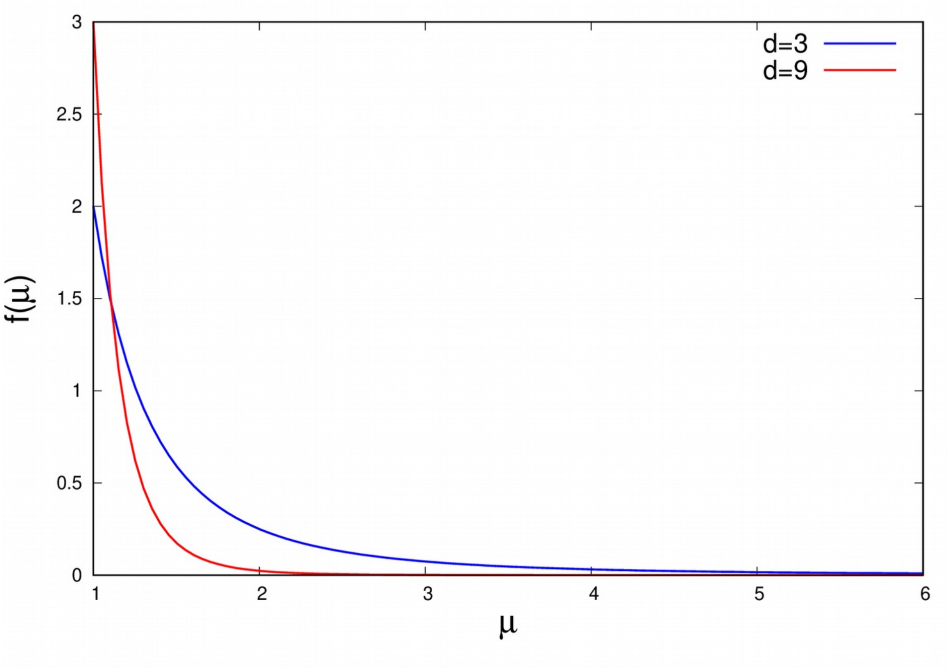
Thank you for your attention!!

Extending TWO-NN to multiple IDs

What is the problem?

Pareto distributions with different *d* are highly overlapping!

The Z assigment in the Gibbs sampling, based on μ, is not reliable



ID estimation: projective approach

- Project *D*-dimensional data into lower dimension *d* : $\Pi^d : \mathbf{x}_i \in \mathbb{R}^D \mapsto \mathbf{y}_i \in \mathbb{R}^d$
- Try different *d* and evaluate for each a “loss function” $\mathcal{L}(\Pi^d)$
- $\mathcal{L}(\Pi^d)$ measures the “data loss” occurring in the projection.

$\mathcal{L}(\Pi^d) = \sum_i ||\mathbf{x}_i - \mathbf{y}_i||^2$ preservation of original distance relations

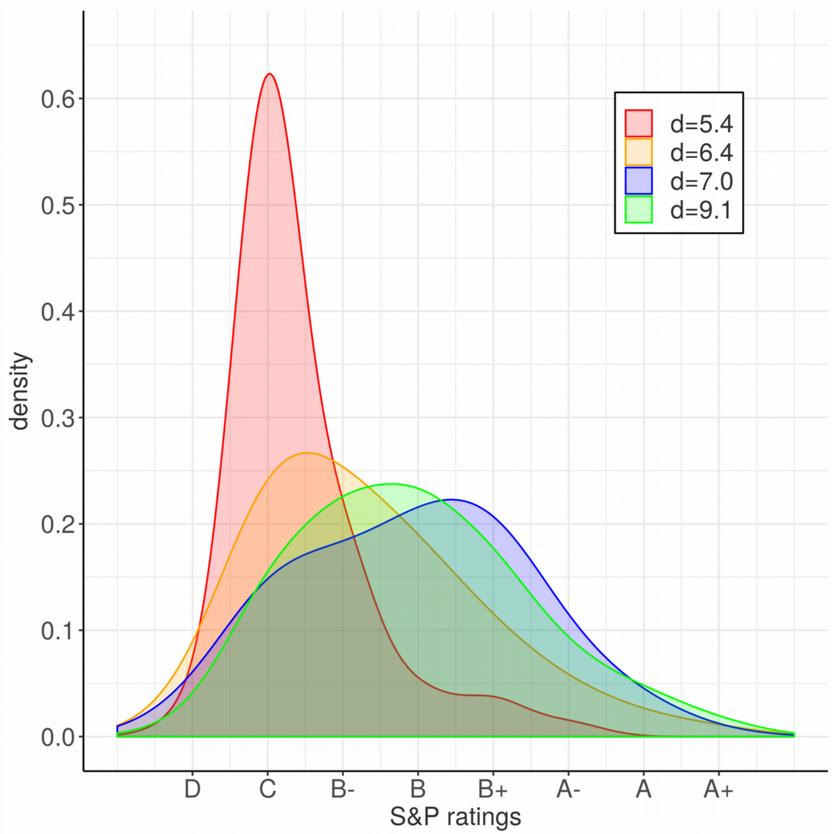
$\mathcal{L}(\Pi^d) = \sum_i \mathbf{x}_i \mathbf{x}_i^T - \mathbf{y}_i \mathbf{y}_i^T$ preservation of original covariance matrix

- d is “estimated” from tradeoff between dimension reduction and data loss
- Problem (1): Computationally burdensome (search for optimal projection for each *d*)
- Problem (2): robust ID estimates only if $\mathcal{L}(\Pi^d)$ has large gap as a function of *d*

  if no gap, the estimation can be rather arbitrary

Example: companies balance sheets

- consider D=38 balance sheet variables for N=8000 companies
- **We find four manifolds with dimensions** d=5.4,d=6.4,d=7.0, d=9.1
- Consider the financial risk of the companies assigned to different manifolds

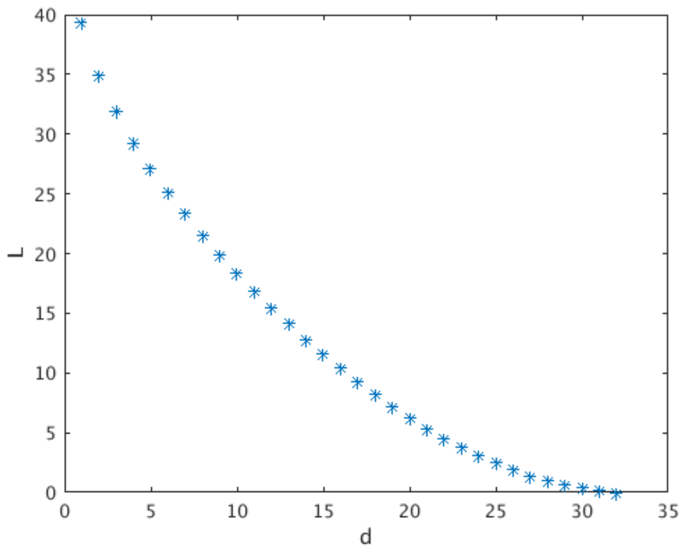


Companies with higher risk are preferentially assigned to low dimensional manifolds!

ID estimation: projective approach

- Example: Principal Component Analysis (PCA)
- Prjoects data onto linear subspace spanned by first *d* eigenvalues of

  covariance matrix X^TX. Loss: $\mathcal{L}(\Pi^d) = ||\sum_i \mathbf{x}_i\mathbf{x}_i^T - \mathbf{y}_i\mathbf{y}_i^T||$
- Typical data:

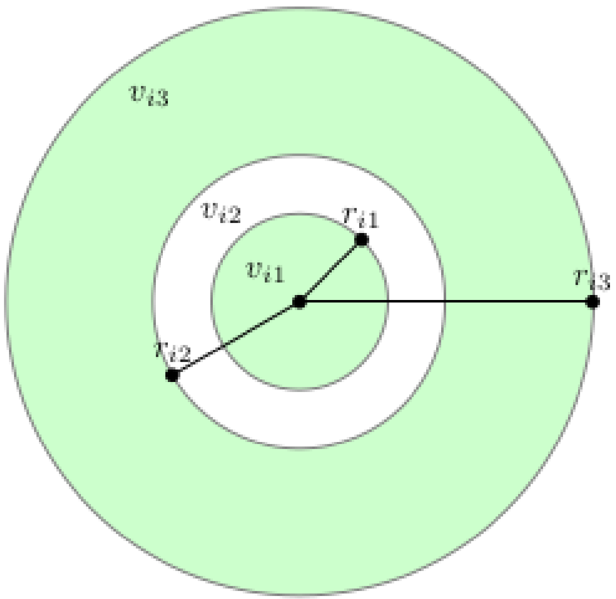


- How can one select an appropriate *d*?

ID estimation: TWO-NN

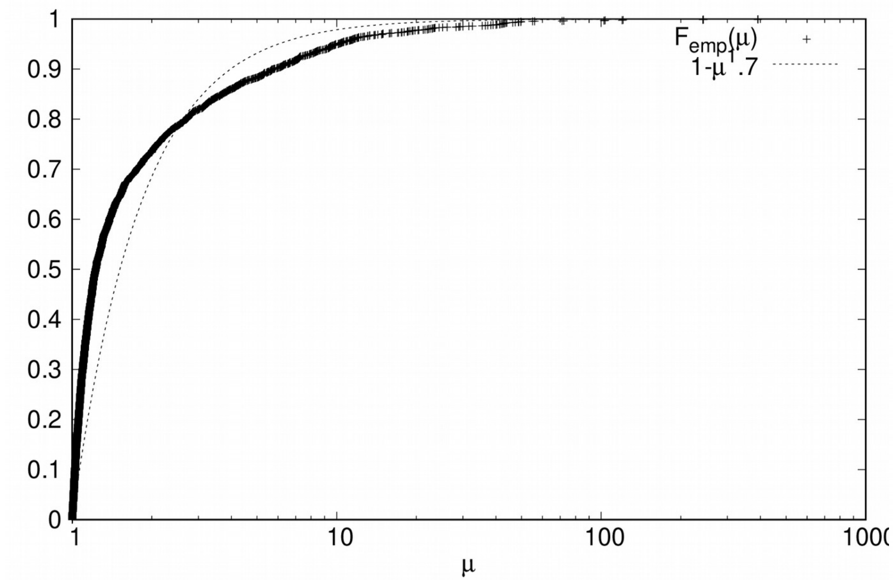
E Facco, M D’Errico, A Rodriguez, A Laio, Scientific Reports 7, 12140. (2017)

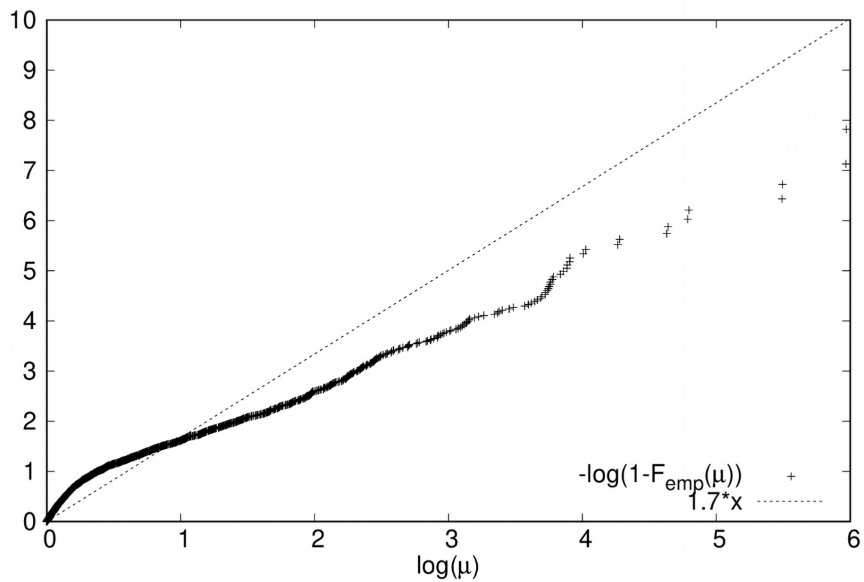
- points are sampled independently
- ρ constant over region A
- n=# of points in a region A
- n follows Poisson law $P(n)=(\rho V)^n/n!\ \exp(-\rho V)$
- Consider hypershepical shells defined by first and second neighbor of a point
- $f(v_{i1},v_{i2})= \exp(-\rho v_{i2})\ dv_{i1}\ dv_{i2}$
- derive $f(r_{i1},r_{i2})$
- derive $f(r_{i2}/r_{i1})$



Is the ID uniform?

Sometimes the model fails...





- 1) the density is strongly varying even on the scale of the first two neighbors
- 2) **the dimension is not uniform in the dataset**