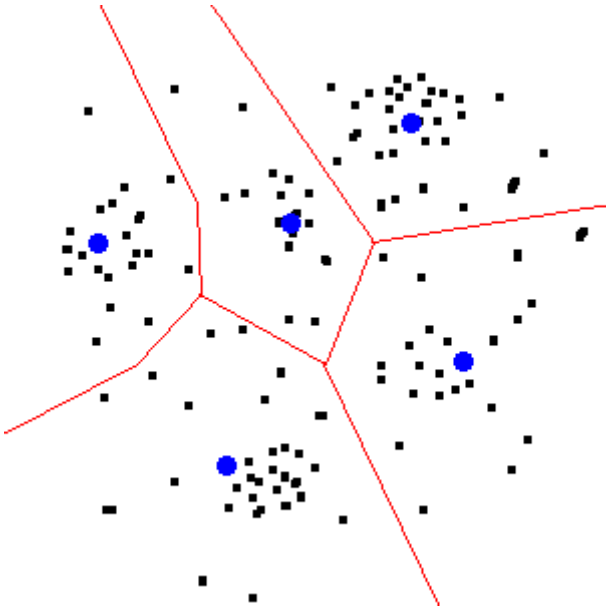
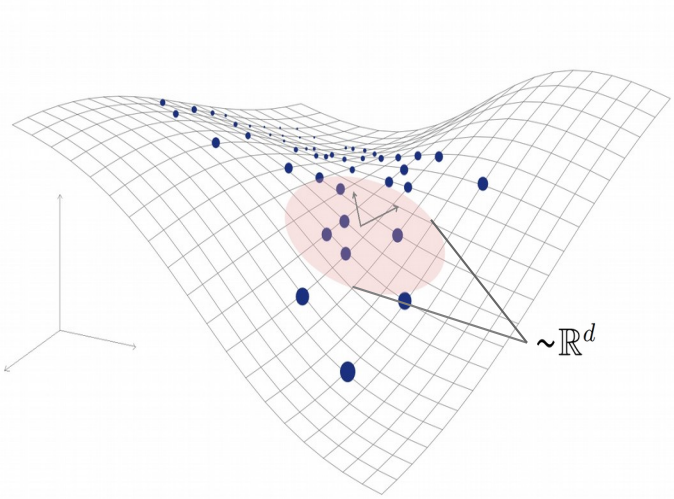
DATA CLASSIFICATION BASED ON THE LOCAL INTRINSIC DIMENSION

classification and intrinsic dimension



- **Classification**
- hard in high dimension (many variables): computational problems, sampling issues...

- # of independent directions of variation can be lower: *intrinsic dimension (ID)*
- Accounting for ID can improve classification schemes

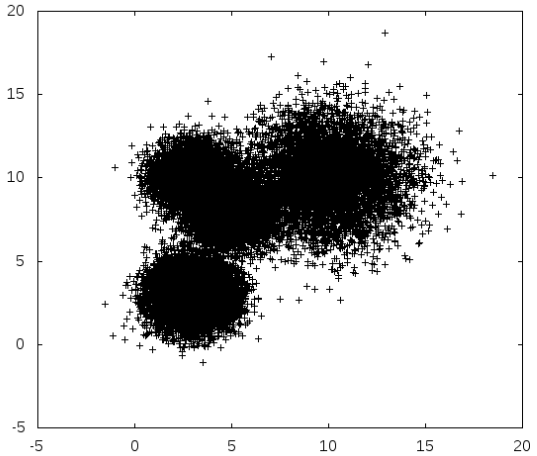


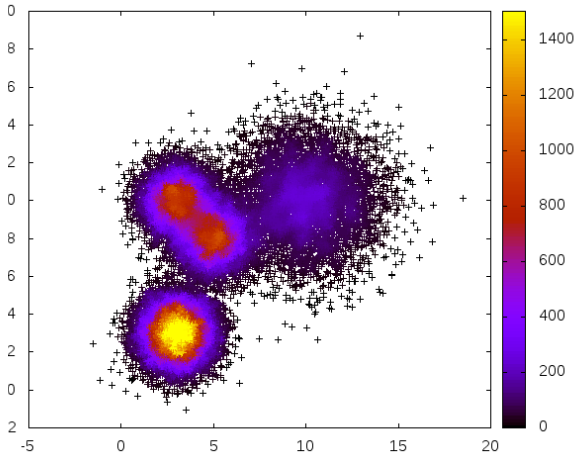
Our journey: from classification to intrinsic dimension and back

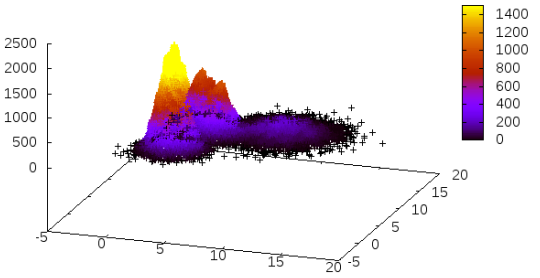
- We started with a classification problem (clustering algorithm)
- This required accurate ID estimation
- We developed a method to estimate the ID
- We realized than often the ID is not constant within a dataset
- This in turn allows for rough, topologically-based data classification

Density-based clustering

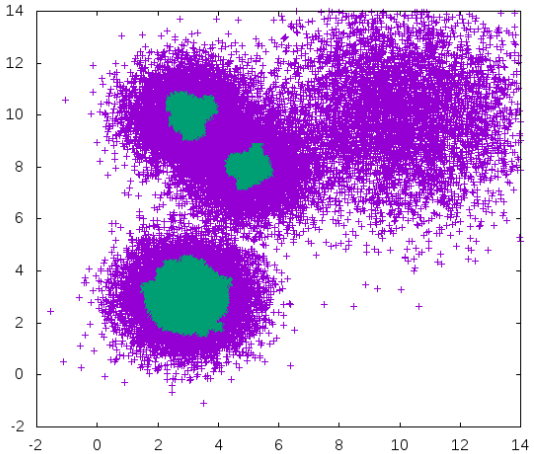
- Reconstruct the probability density of the data

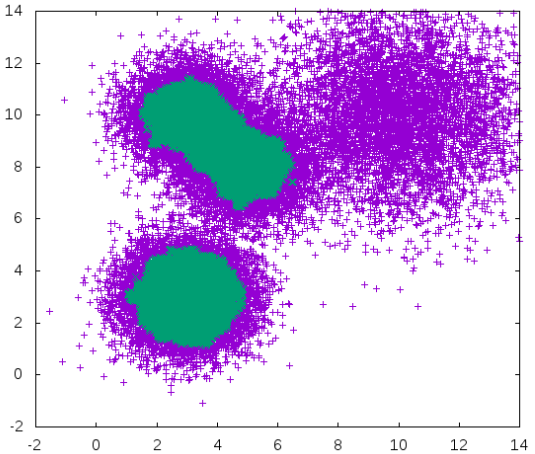


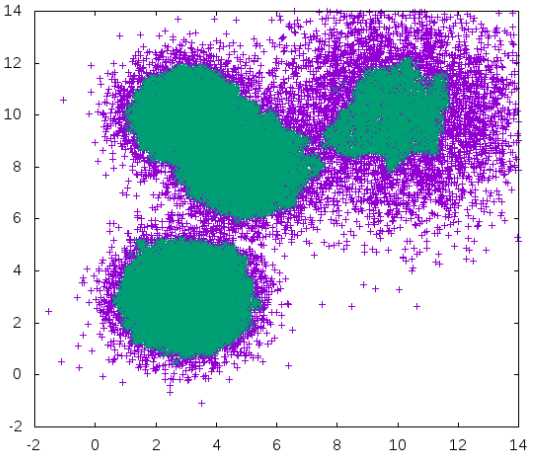




- Then look for disconnected regions of high density





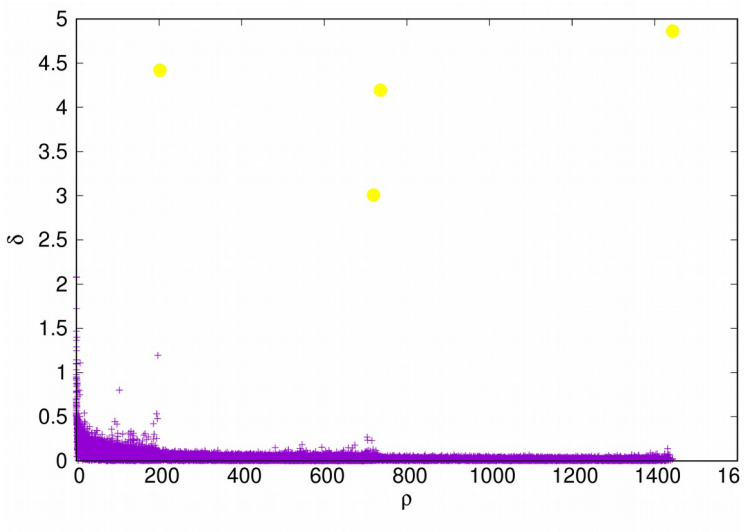


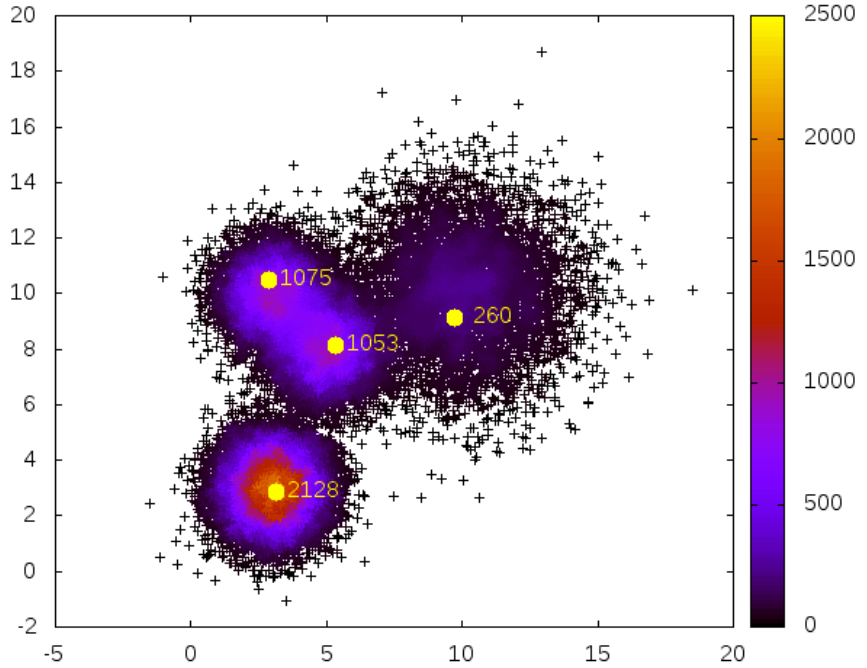
- What is high? Results depend on the chosen density threshold
- Cannot resolve features at different density scalesl

Density Peak Clustering

A Rodriguez, A Laio, Science 344, 1492 (2014)
M d'Errico, E Facco, A Laio, A Rodriguez, arXiv:1802.10549 (2018)

Cluster around density peaks, i.e. local maxima in the density





Original algorithm: density peaks are far from any point with higher density

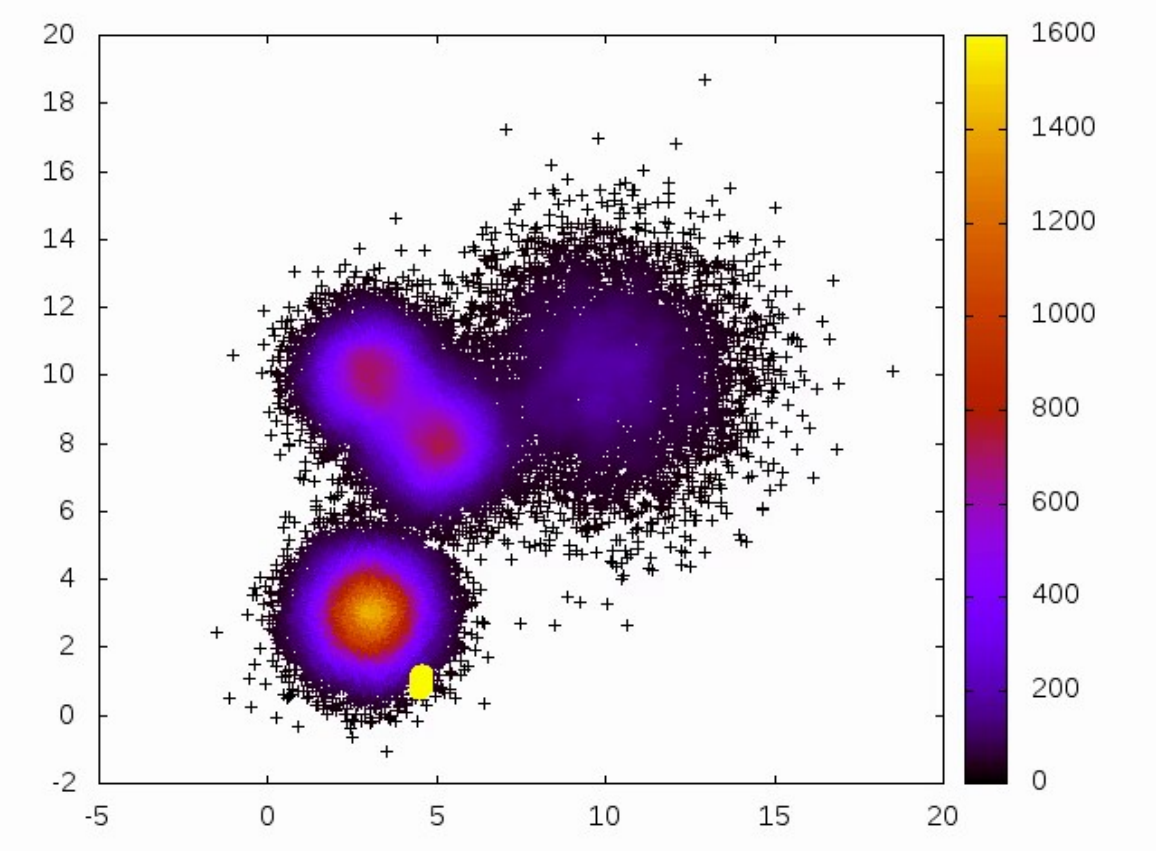
Compute for all points min distance from point at higher density $\delta_i = min_{j:\rho_j > \rho_i} d_{ij}$

Peak are outliers in decision graph ρ_i vs δ_i :

Density-peak clustering

Points are assigned to peaks by following a path of increasing density leading to one of the peaks.

One jumps from a point to a point with higher density



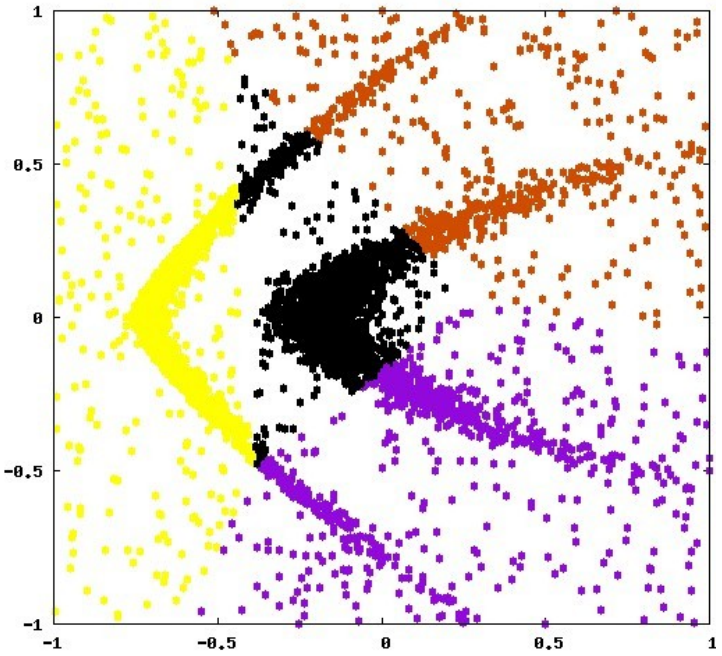
Density-peak clustering

A Rodriguez, A Laio, Science 344, 1492 (2014)
M d'Errico, E Facco, A Laio, A Rodriguez, arXiv:1802.10549 (2018)

Points are assigned to clusters by following a path of increasing density leading to one of the peaks.

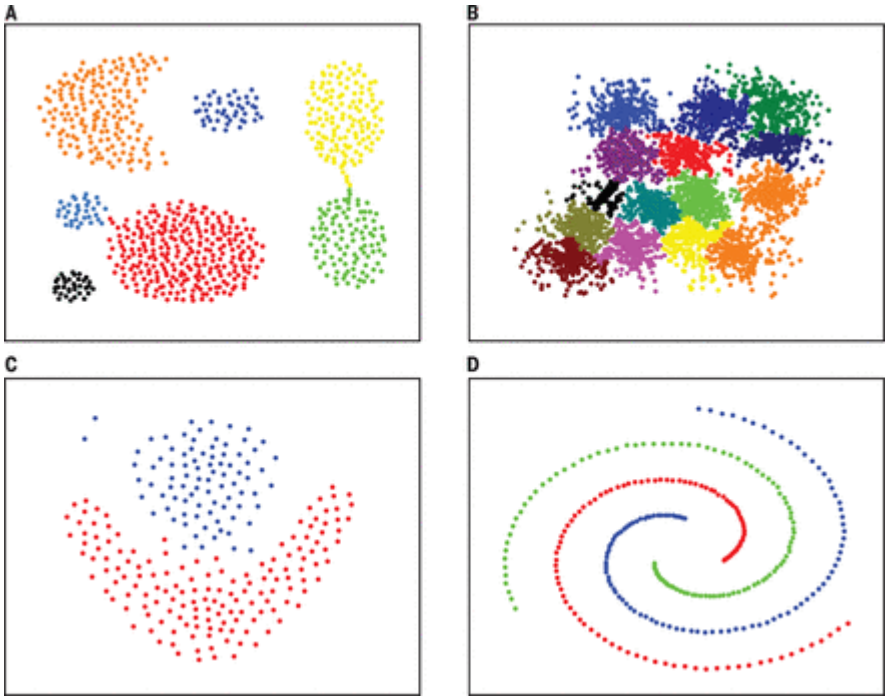
This assignation rule allows to retrieve clusters of arbitrary shape

K-means



r_i

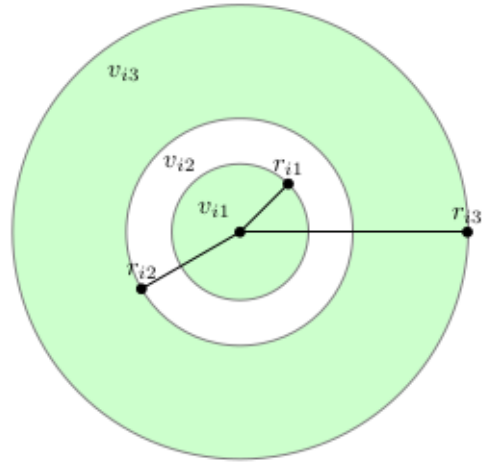
DPC

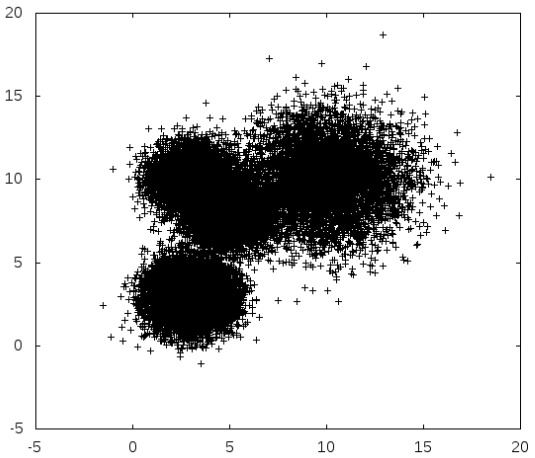


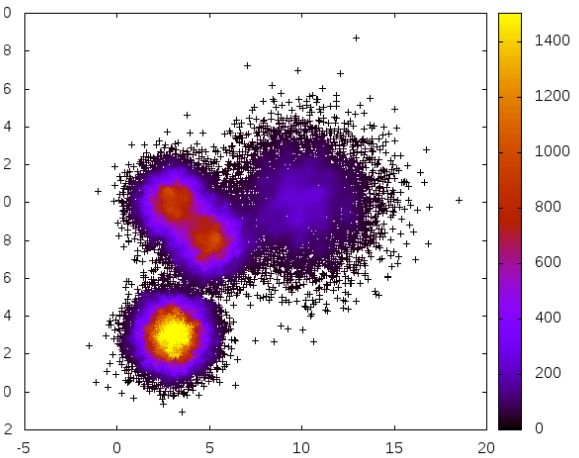
Density estimation

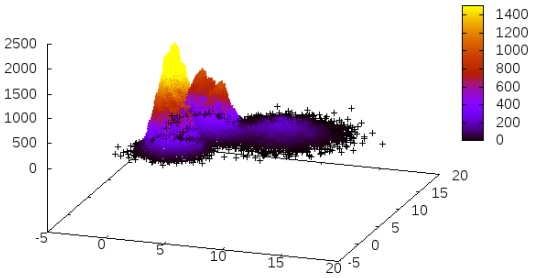
- Data can be thought of as samples of a density distribution
- Reconstruct the probability density of the data with proper *density estimator*
- K-nearest-neighbor: Assume ρ ≈ const in small region around each point
- For each point *i*, consider its *k* nearest neighbors at distances $r_{i1}, r_{i2}, r_{i3}, \ldots$
- density= *k*/volume of sphere containing the *k* points

$$\rho = \frac{k}{V_{ik}} \qquad \delta\rho = \frac{\sqrt{k}}{V_{ik}} \qquad V_{ik} = \omega_d r_{ik}^d$$





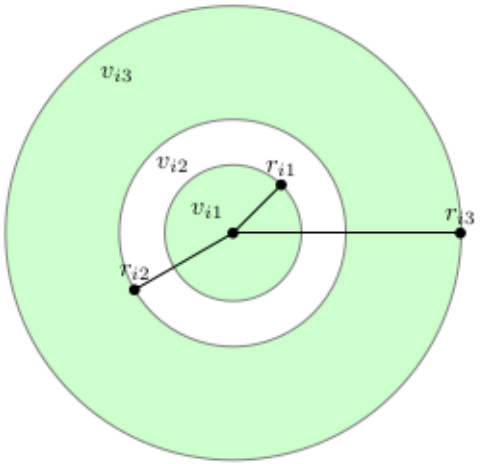




Density Estimation: PAk

A Rodriguez, M D'Errico, E Facco, A Laio, JCTC (2018)

- K-nearest-neighbor: Assume ρ ≈ const in small region around each point
- For each point *i*, consider its *k* nearest neighbors at distances $r_{i1}, r_{i2}, r_{i3}, \ldots$
- density= *k*/volume of sphere containing the *k* points



$$\rho = \frac{k}{V_{ik}} \qquad \delta\rho = \frac{\sqrt{k}}{V_{ik}} \qquad V_{ik} = \omega_d r_{ik}^d$$

- Two problems:
- 1) **what is right *k*?**
- 2) **what is right *d*?**

Density Estimation: PAk

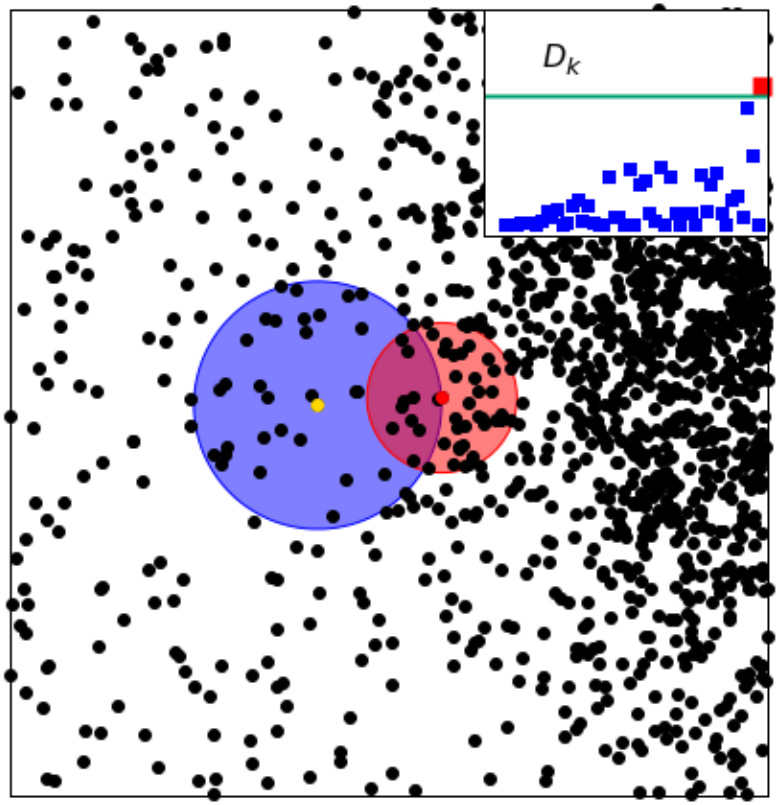
what is right *k*?

k too small: large error in the estimate

k too large: density is not constant over V

Solution:

- Adapt *k to* each point so that the constant density assumption always holds
- **Pointwise Adaptive k-NN (PAk) estimator**

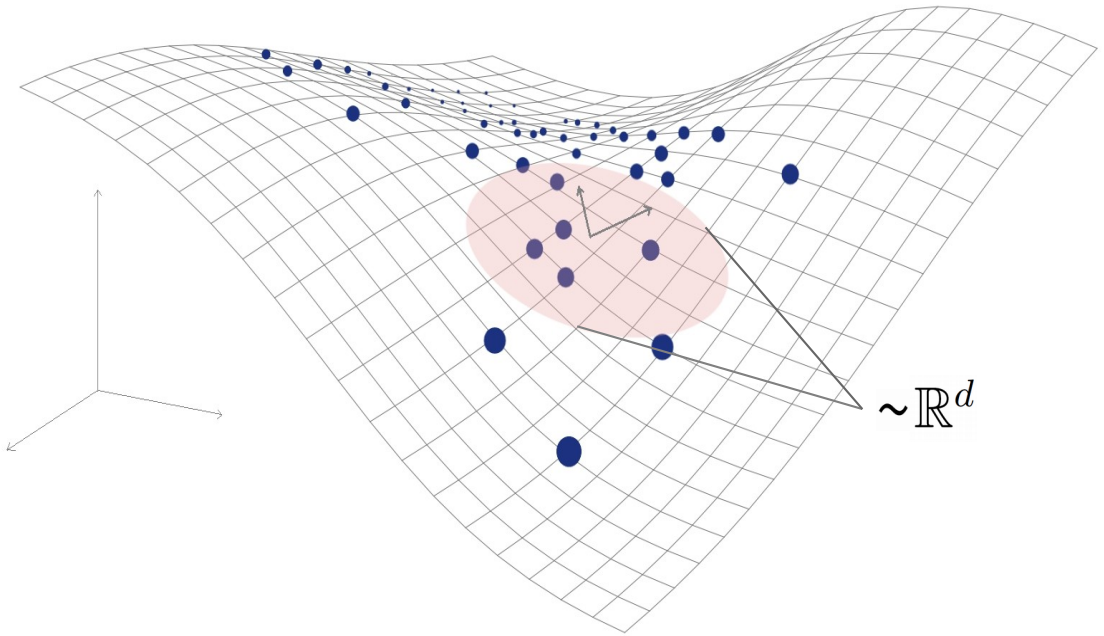


Intrinsic dimension

Problem 2): what is right *d*?

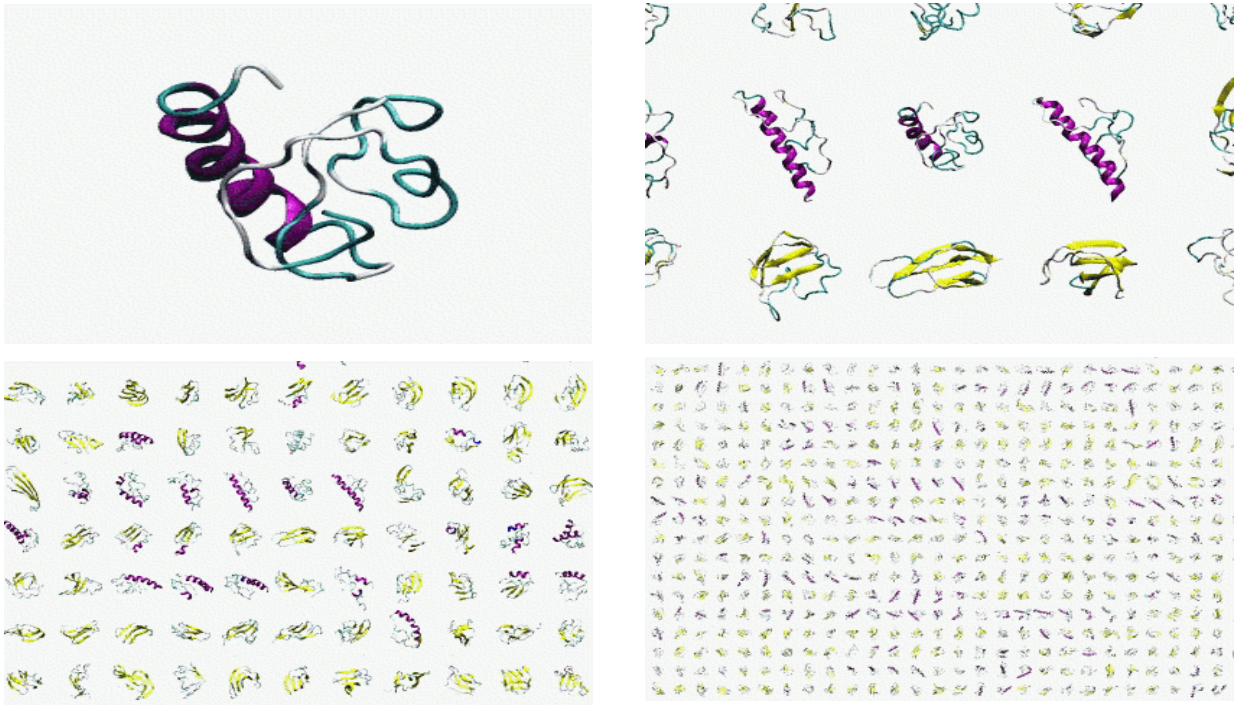
The data actually lie on hypersurface of lower dimension than D

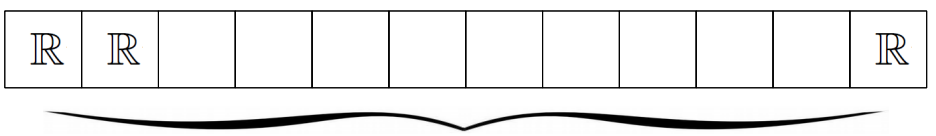
The density should be evaluated on this hypersurface



Intrinsic dimension

Example: molecular dynamics

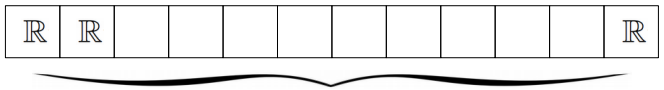




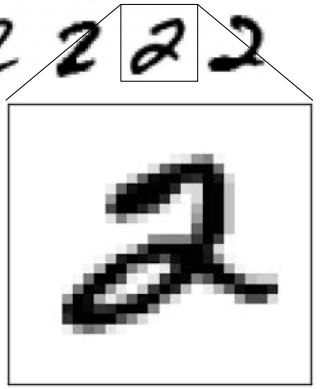
3xN

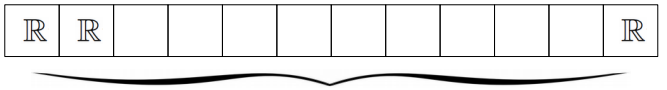
Intrinsic dimension

Example: images



20x20





64x64

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. Examples:

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

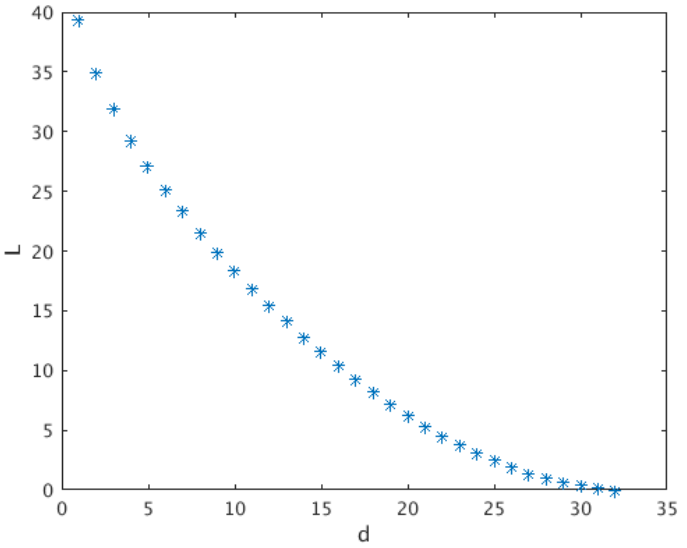
$$\mathcal{L}(\Pi^d) = \sum_i \mathbf{x}_i \mathbf{x}_i^T - \mathbf{y}_i \mathbf{y}_i^T$$ preservation of original covariance matrix

- 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

ID estimation: projective approach

- Example: Principal Component Analysis (PCA)
- Prjoects data onto linear subspace spanned by first *d* eigenvalues of covariance matrix. $X^T X$ Loss: $\mathcal{L}(\Pi^d) = || \sum_i \mathbf{x}_i \mathbf{x}_i^T - \mathbf{y}_i \mathbf{y}_i^T ||$
- on the villin headpiece simulation:



- How can one select an appropriate *d*

ID estimation: statistical approach

- assumes that the data are sampled from a distribution with density $\rho(\mathbf{X})$
- distances between points in the dataset follow a scaling law that depends
- on $\rho(\mathbf{X})$ and *d*
- If the dependence on $\rho(\mathbf{X})$ can be removed, then *d* can be estimated from the scaling
- Example: correlation dimension
    - The number of points at distance $< \epsilon$ from point *i* scales as

$$N_i(\epsilon) = \sum_j \theta(d_{ij} < \epsilon) \approx \epsilon^d / \rho(\mathbf{X}_i)$$

    - If $\rho(\mathbf{X})$ is constant, $N(\epsilon) = \sum_{ij} \theta(d_{ij} < \epsilon) \sim \epsilon^d / \rho$
    - *d* can be estimated with simple linear fit
- However, when $\rho(\mathbf{X})$ is variable the estimation fails dramatically

ID estimation: TWO-NN

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

- In principle, one should evaluate simultaneously both d and $\rho(\mathbf{X})$!
- TWO-NN idea: **decouple the estimation problem by finding suitable function of the distances that depends only on *d***
- Assumption: $\rho(\mathbf{X})$ is constant on the scale of the first two neighbors
- Then if d_{i1}, d_{i2} are distances from 1st and 2nd neighbor of point i,
- their ratio $\mu_i = \frac{d_{i2}}{d_{i1}}$ follows a Pareto distribution: $f(\mu_i) = d\mu_i^{-(d+1)}$
- depends only on *d*, not on $\rho(\mathbf{X})$!
- **Collect the μ for each point. Fit their emprical distribution and estimate *d***
- The ID is inferred from the μ collectively

ID estimation: TWO-NN

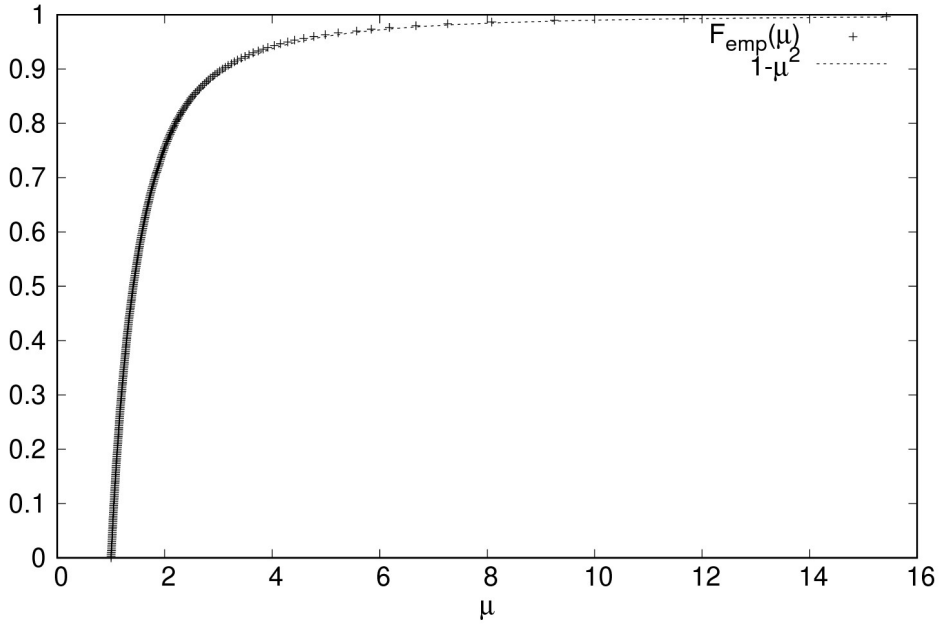
E Facco, M D'Errico, A Rodriguez, A Laio, Scientific Reports 7, 12140. (2017)

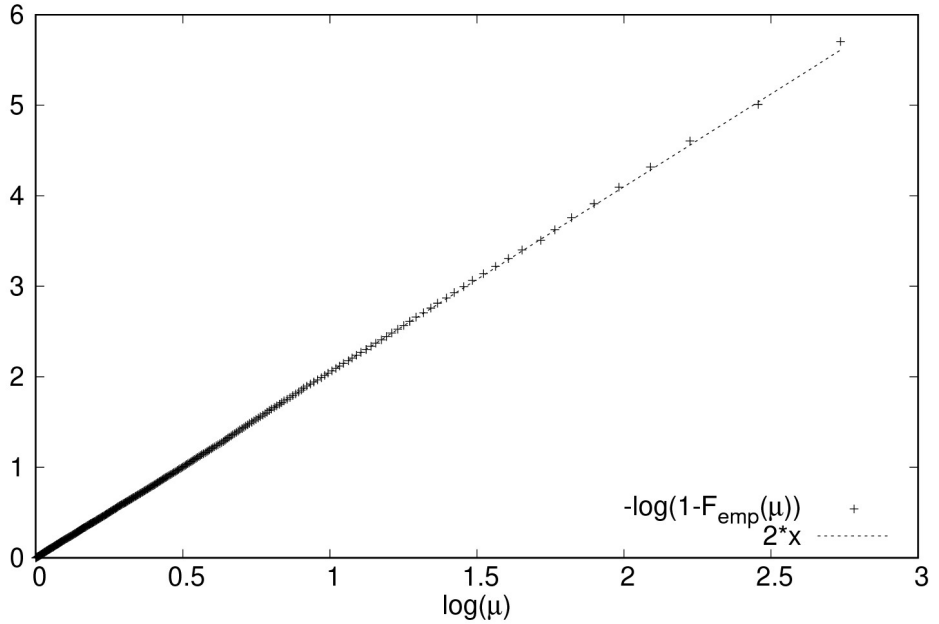
- In principle, one should evaluate simultaneously both d and $\rho(\mathbf{X})$!
- TWO-NN idea: **decouple the estimation problem by finding suitable function of the distances that depends only on *d***
- Assumption: $\rho(\mathbf{X})$ is constant on the scale of the first two neighbors
- Then if d_{i1}, d_{i2} are distances from 1st and 2nd neighbor of point i,
- their ratio $\mu_i = \frac{d_{i2}}{d_{i1}}$ follows a Pareto distribution: $f(\mu_i) = d\mu_i^{-(d+1)}$
- depends only on *d*, not on $\rho(\mathbf{X})$!
- **Collect the μ for each point. Fit their emprical distribution and estimate *d***
- The ID is inferred from the μ collectively

ID estimation: TWO-NN

There are several ways of fitting:

- One can fit the empirical cumulative distribution of μ with $F(\mu) = 1 - \mu^{-d}$
- Equivalently, linear fit on $\log(1 - F(\mu)) = -d \log \mu$

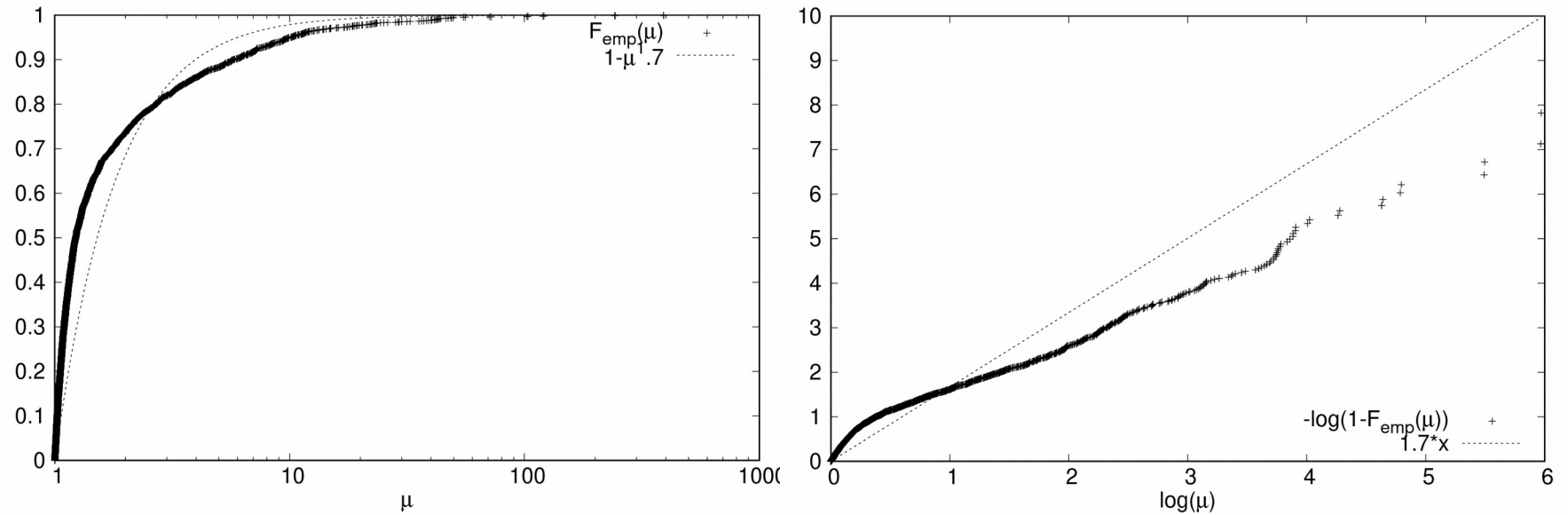




- If the model is satisfied, then the distribution of the μ is well fitted (check χ^2)

The problem of multiple IDs

If the fit is not good, it means the model fails



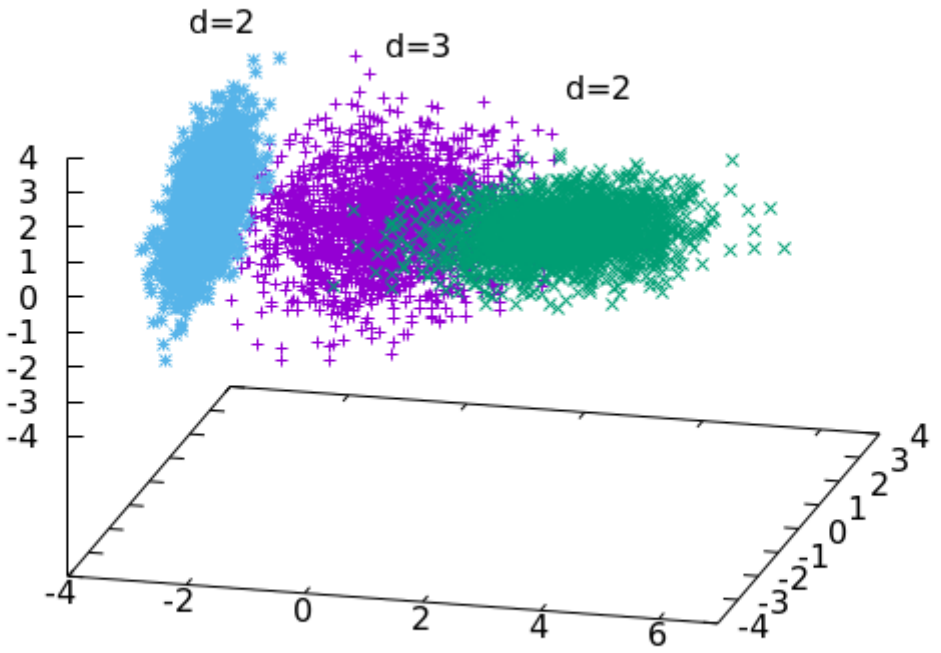
- 1) the density is strongly varying even on the scale of the first two neighbors
- 2) the dimension is not uniform in the dataset

The data may lie on several manifolds $\mathcal{M}_1, \ldots, \mathcal{M}_K$ **, each with different ID**

How to deal with this heterogeneous ID case?

Heterogeneous ID

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



Hidalgo

- H1) data sampled from manifolds of different ID
- H2) ρ is uniform on scale of the first neighbors

- **Under H1), H2) one can still predict the expected distribution of the** μ

- Assume point sampled from $\mathcal{M}_1, \ldots, \mathcal{M}_K$ with different probablities $\mathbf{p} = p_1 \ldots p_K$

- mixture of Pareto distributions $P(\mu_i) = \sum^{K} p_k d_k \mu_i^{-d_k - 1}$

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

- Then we can again estimate $\mathbf{d} = d_1 \ldots d_K, \quad \mathbf{p} = p_1 \ldots p_K$

- K is not estimated as a parameter

- Fix K by trying increasing values in [1,K_{max}] and performing a model selection test

Hidalgo

- To estimate parameters, fix inferential approach
- A) frequentist: $\mathbf{d}^e, \mathbf{p}^e = argmax(\mathcal{L}(\boldsymbol{\mu}|\mathbf{d}, \mathbf{p}))$
-
- B) Bayesian
  - Fix $P_{prior}(\mathbf{d}, \mathbf{p})$
  - Compute mean $\mathbf{d}^e, \mathbf{p}^e = \langle \mathbf{d}, \mathbf{p} \rangle_{post}$ $P_{post}(\mathbf{d}, \mathbf{p}) \propto \mathcal{L}(\boldsymbol{\mu}|\mathbf{d}, \mathbf{p}) P_{prior}(\mathbf{d}, \mathbf{p})$

- Because of the sum over k, hard to work with $\mathcal{L}(\boldsymbol{\mu}|\mathbf{d}, \mathbf{p}) = \prod_{i=1}^{N} \sum_{k=1}^{K} p_k d_k \mu_i^{-d_k - 1}$
- Introduce latent variables $\mathbf{Z} = Z_1, \ldots, Z_N$: manifold membership of each point
- Likelihood is seen as marginal over $\mathcal{L}(\boldsymbol{\mu}|\mathbf{d}, \mathbf{p}, \mathbf{Z}) = \prod_{i=1}^{N} p_{Z_i} d_{Z_i} \mu_i^{-d_{Z_i} - 1}$
- Estimate jointly $\mathbf{d}, \mathbf{p}, \mathbf{Z}$
- **Heterogeneous ID algorithm (hidalgo)**

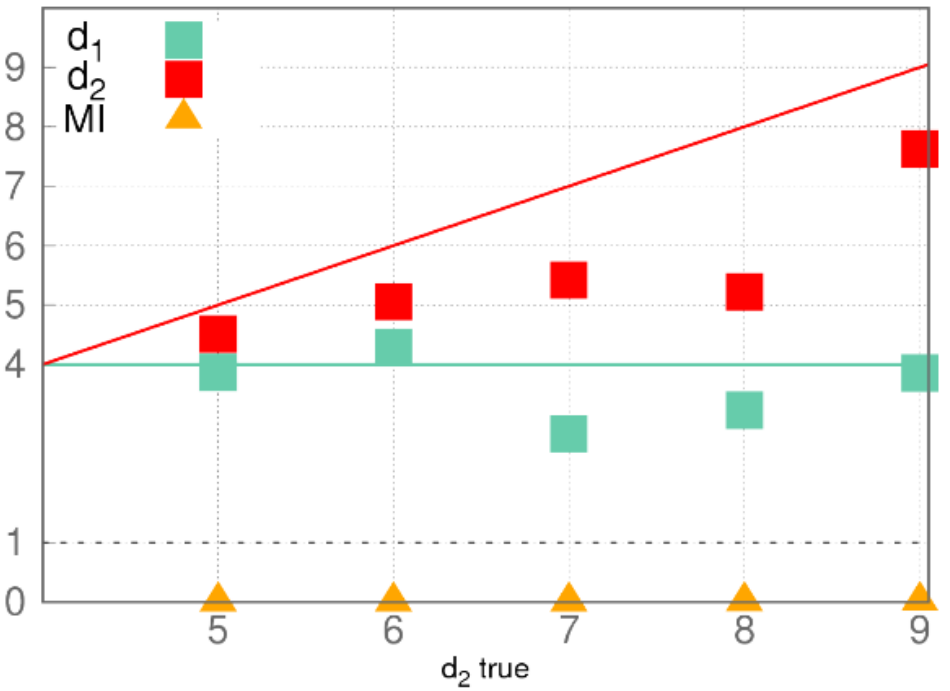
Hidalgo

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

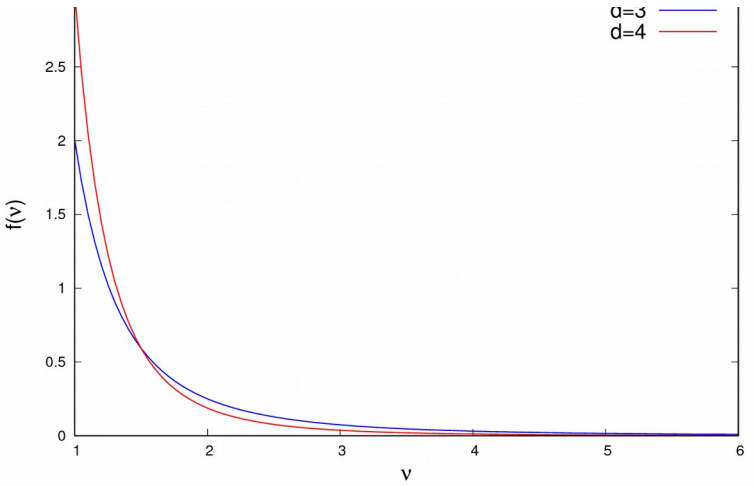


Why?

Pareto distributions with different *d*
are highly overlapping

The Z assignation is based only on the μ
of each point

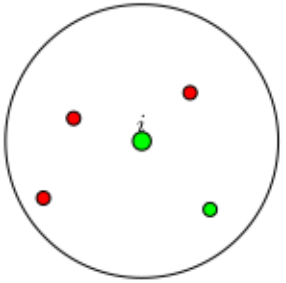
Difficult to assing Z if μ value is not predictive



Hidalgo

We get non-uniform neighborhoods

Neighboring points have different Z



We must assume that the manifolds are separated, with at most a (small) intersection

This implies that the neighborhoods must be approximately uniform

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

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

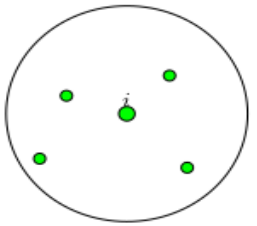
$$\mathcal{L}(n_i^{in}|\mathbf{Z}) = \frac{\zeta^{n_i^{in}}(1-\zeta)^{n_i^{out}}}{\mathcal{Z}}$$

$\zeta > \frac{1}{2}$ Parameter that controls degree of uniformity

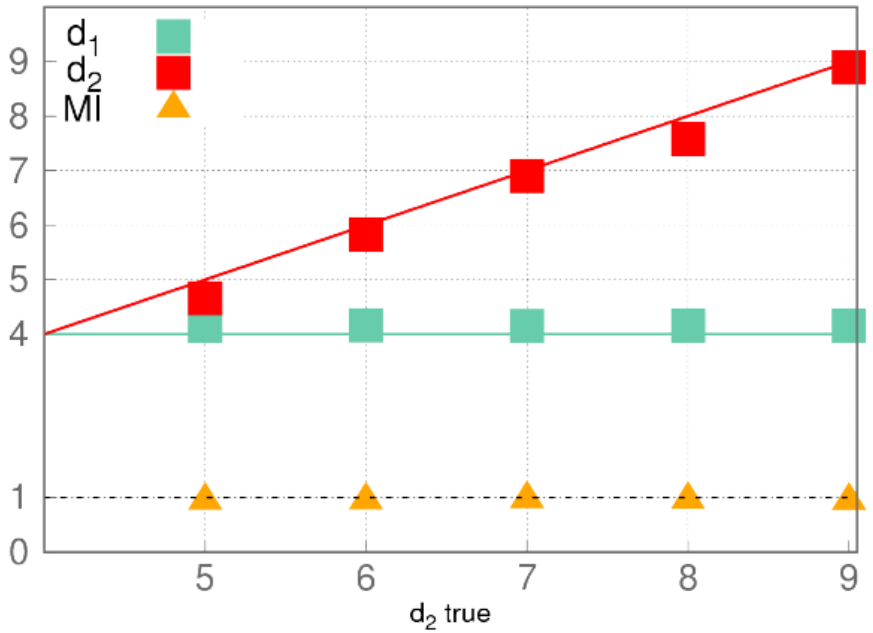
Hidalgo

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

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



Now we get correct estimates of both ***d,p*** and ***Z***

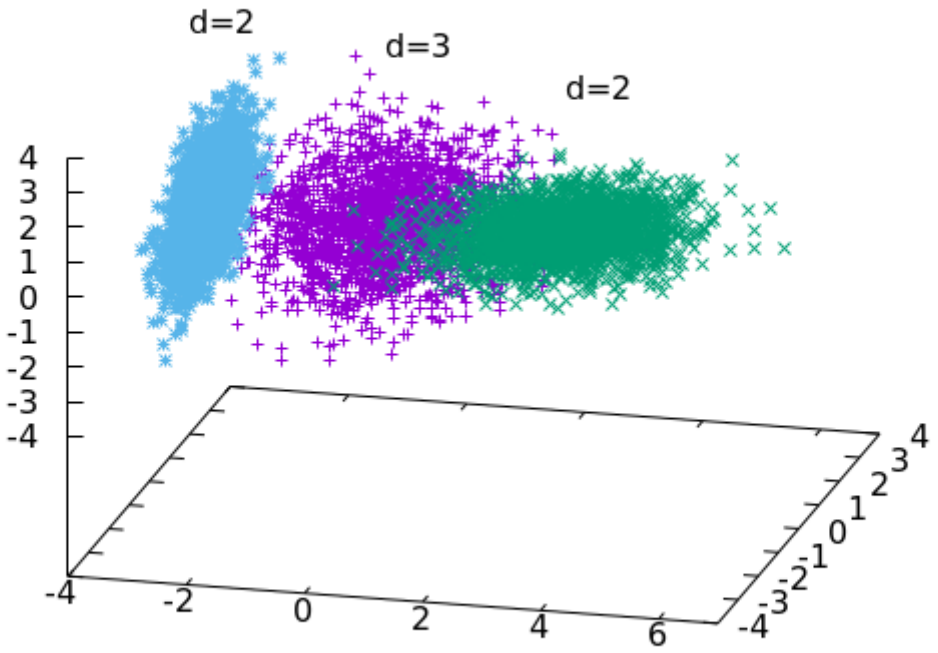


Hidalgo

M Allegra, E Facco, A Laio and A Mira, in prep. (2018)

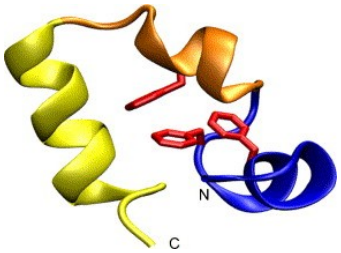
We achieve a **global topological description of the data space**

Divide space into regions of uniform intrinsic dimension



Hidalgo

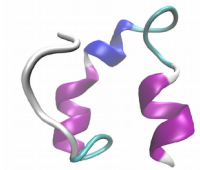
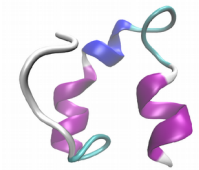
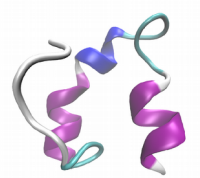
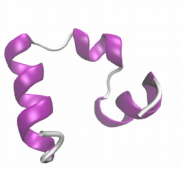
Example: molecular dynamics



- consider a MD of unfolding/refolding villing headpiece
- for each of the N ~ 32000 configurations, D=32 dihedral angles.

We find four manifolds

d=12	d=13	d=13	d=23	
Q=0.53	Q=0.58	Q=0.64	Q=0.89	Fraction of native contacts

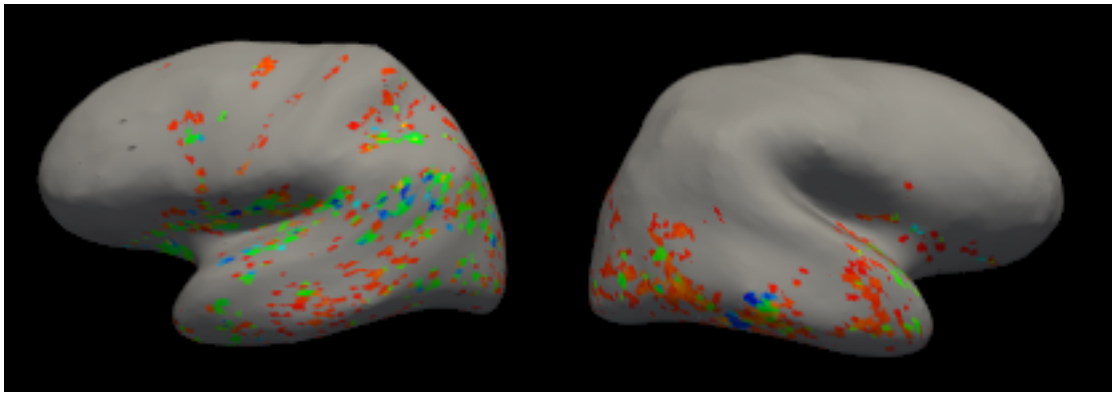
The folded state is recognized from its higher ID!

Hidalgo

Example: fMRI time series

- consider ~30000 time series corresponding to BOLD signal of each voxel in an fMRI experiment
- for each of the N ~ 30000 time sieries, D=202 values

We find two manifolds: d=16, d=32



Red: high-ID voxels

Blue: “task-relevant” voxels

Green: intersection

Task-relevant voxels are in the manifold with higher ID

The low-dimensional manifold nostly includes “noise” voxels

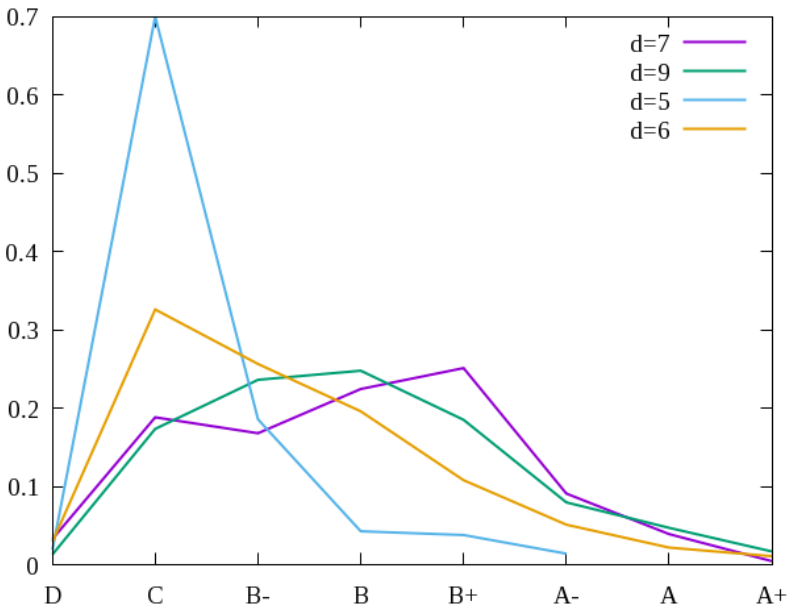
Hidalgo

Example: firms from Compustat

- consider ~8000 firms in the Compustat Database
- for each of the firms, D=31 balance sheet variables

We find four manifolds: d=5, d=6, d=7, d=9

We compute S&P ratings for the different manifolds



Lower dimension tends to have lower ratings!

Data classification based on the intrinsic dimension

- The problem of clustering led us to the problem of density estimation;
  the problem of density estimation led us to the problem of ID estimation
- We developed a reliable ID estimator, TWO-NN, that limits the issue of density variations
- We realized that often the ID is not constant in the dataset:
  we extended the statistical framework of TWO-NN to comply with this case
- We developed Hidalgo, a method that finds groups of points (manifolds) of different ID in the manifold
- Applications of Hidalgo to real datasets reveal that the topological information given by the ID discriminates points differing in important features

Acknowledgments

Alessandro Laio

Alex Rodriguez

Elena Facco

Antonietta Mira

Thank you for your attention!!