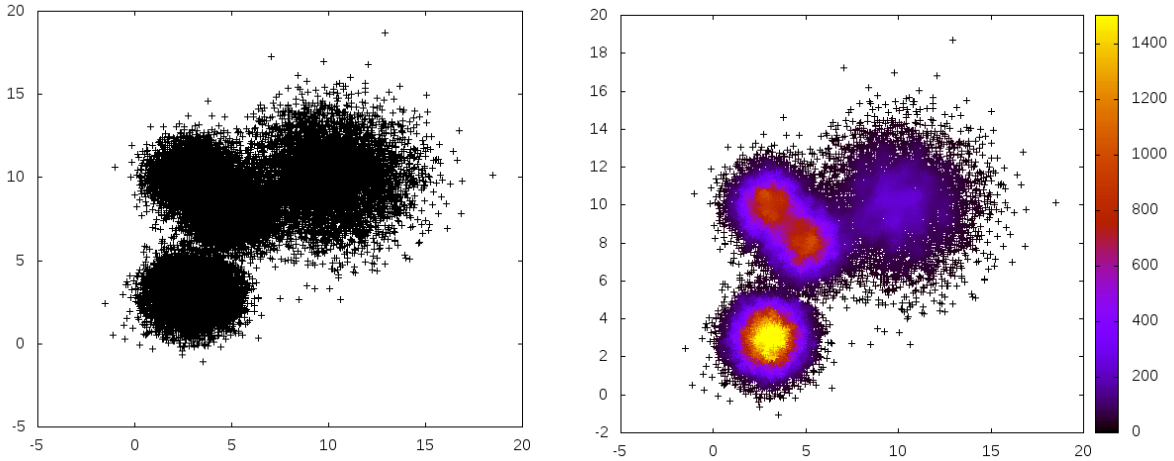
RECONSTRUCTING HIGH DIMENSIONAL PROBABILITY LANDSCAPES

Reconstructing complex landscapes in phase space

Phase spaces of complex physical systems
are **high-dimensional**

How to chart a probability density in such space ?

In D=2, one can easily produce a **density map**

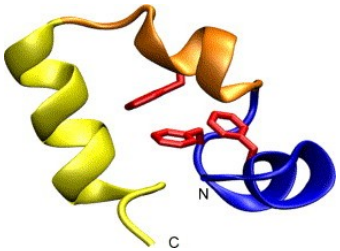


In large D, one can try to project data in dimension d=2 $\Pi^d : \mathbf{x}_i \in \mathbb{R}^D \mapsto \mathbf{y}_i \in \mathbb{R}^d$

the “data loss” measured by preservation of distance relations: $\mathcal{L}(\Pi^d) = \sum_i ||\mathbf{x}_i - \mathbf{y}_i||^2$

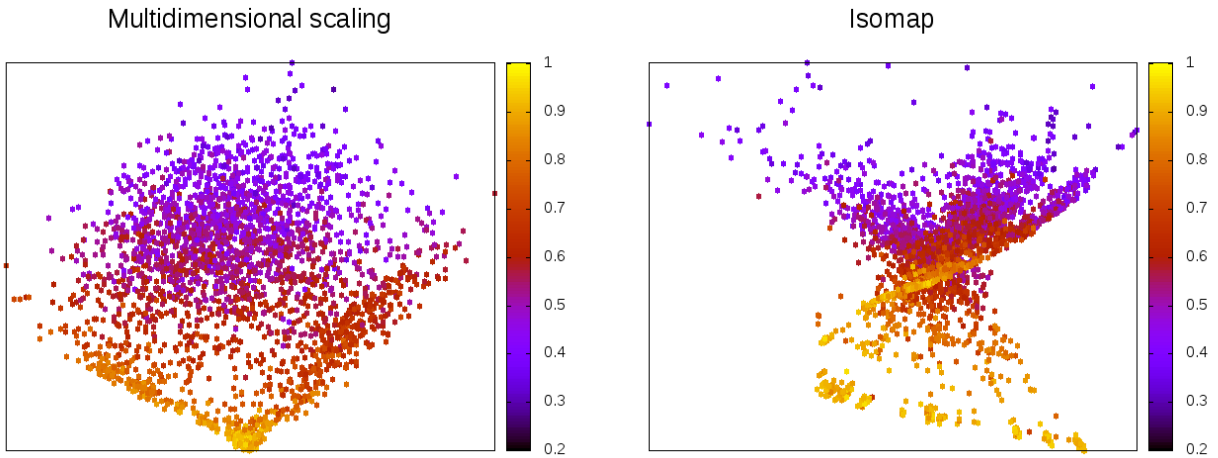
when the space is high-dimensional, usually high loss!

Example: phase space landscape of folding protein



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

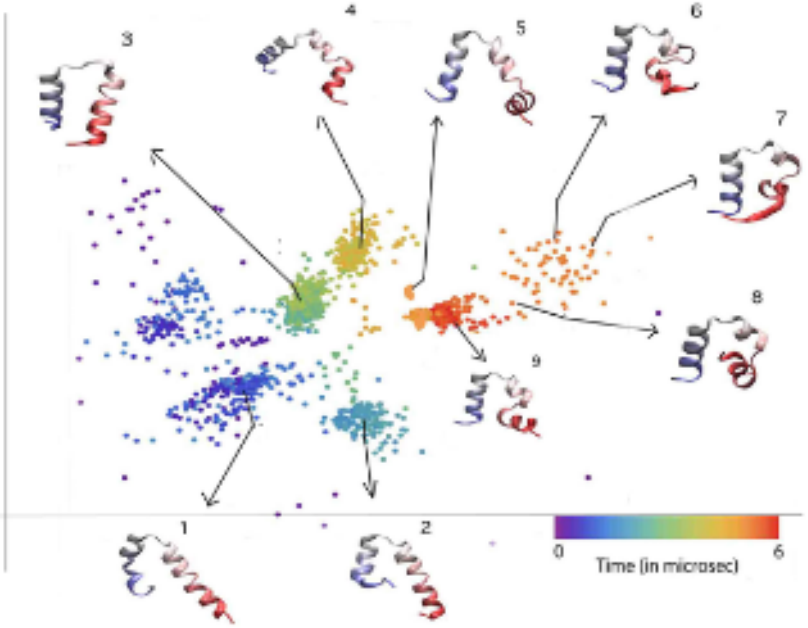
We can try to project the data in 2D with several methods



The resulting maps are quantitatively (and qualitatively) inaccurate

Charting complex data landscapes

From molecular dynamics...



characterize complex free energy landscapes

...to general data

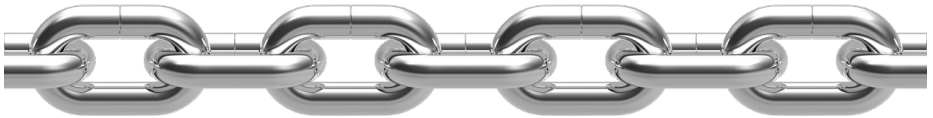
characterize structures in complex data spaces

An intrusion into data analysis

From physics

to data analysis

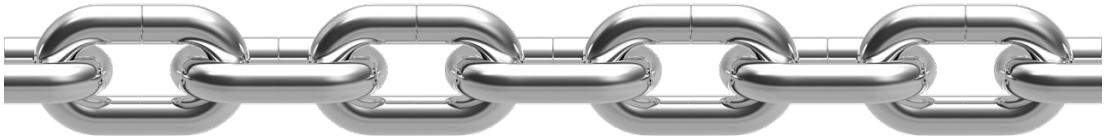
A chain of methodological developments



DPC PaK TWO-NN Hidalgo

A toolkit of methods

Density peak clustering

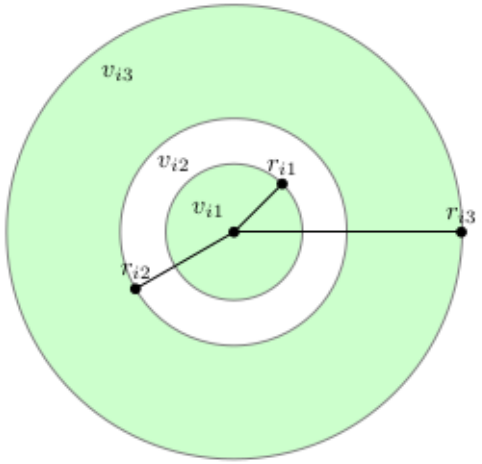


DPC

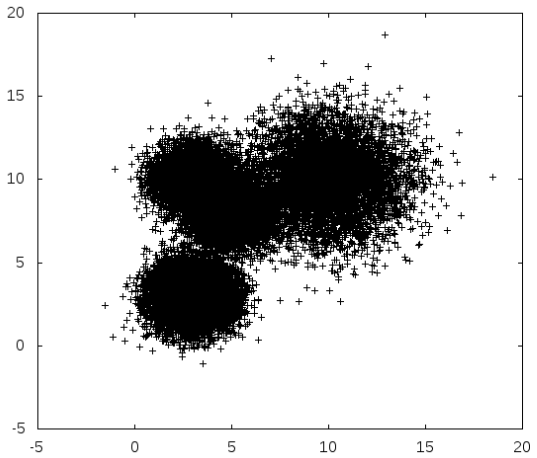
A novel clustering approach

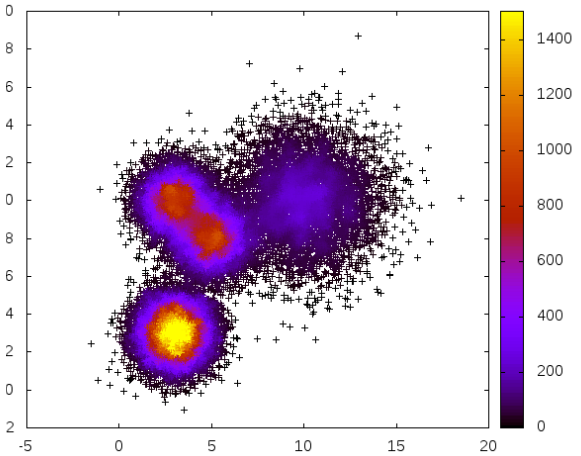
Density peak clustering

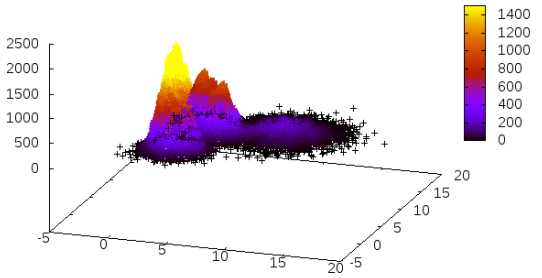
- Data can be thought of as samples of a density distribution
- Reconstruct the probability density of the data with proper *density estimator*
- K-nearest-neighbor: Assume ρ ≈ const in small region around each point
- For each point *i*, consider its *k* nearest neighbors at distances $r_{i1}, r_{i2},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



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

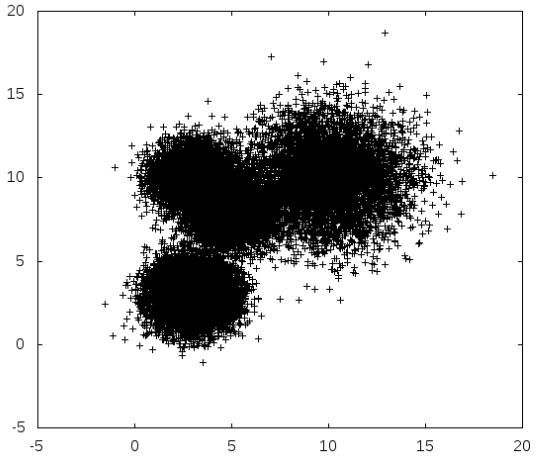


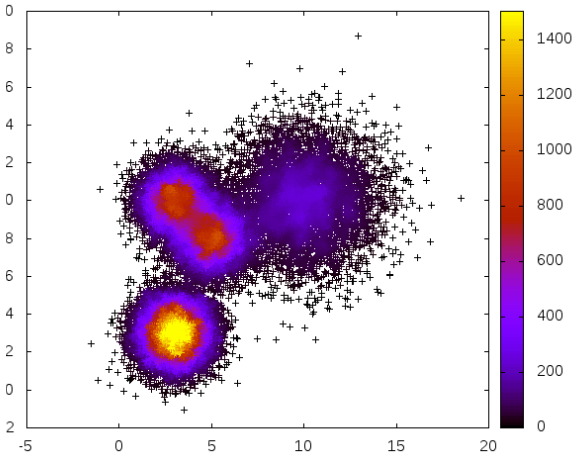


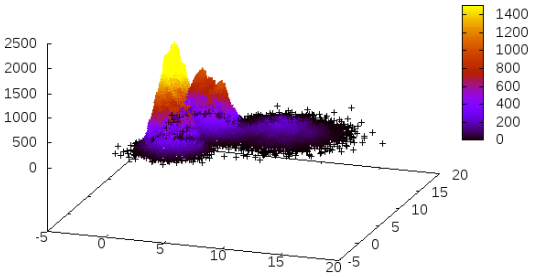


Density-based clustering

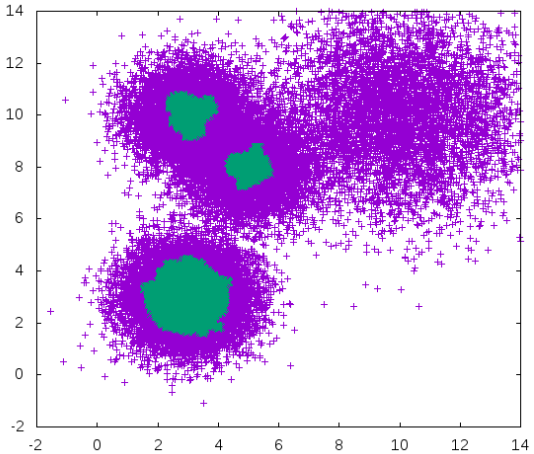
- Reconstruct the probability density of the data

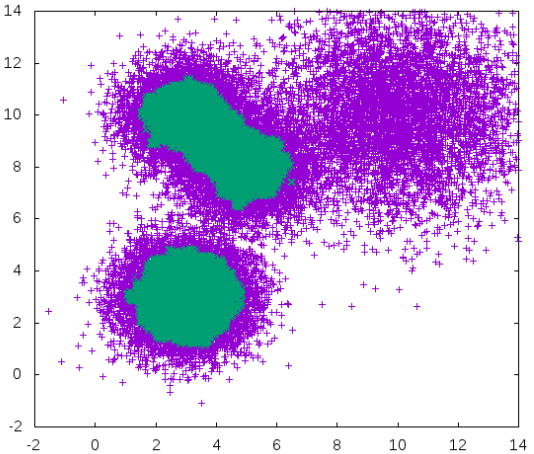


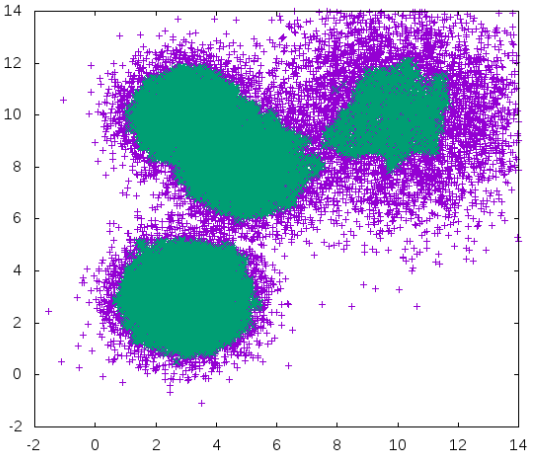




- Then look for disconnected regions of high density







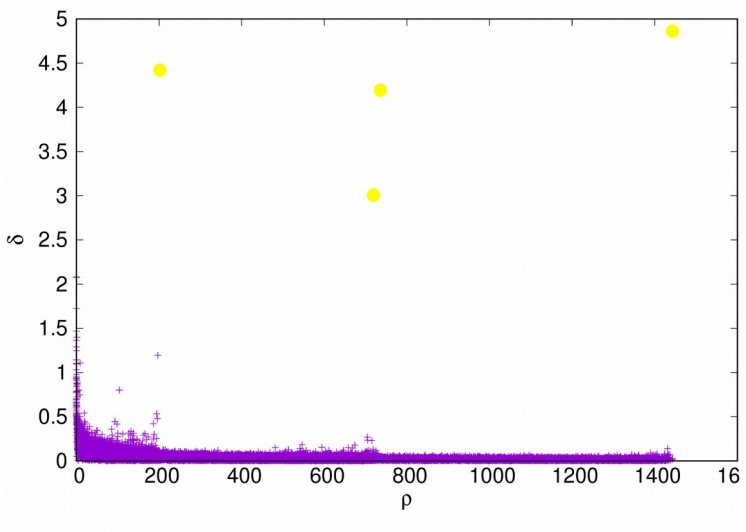
- What is high? Results depend on the chosen density threshold
- Cannot resolve features at different density scalesl

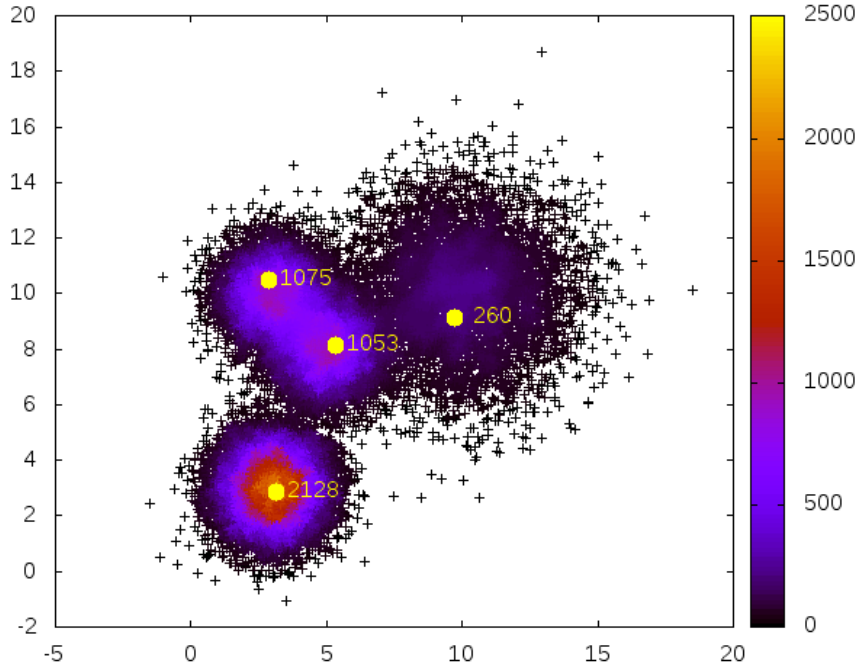
Density Peak Clustering: the topography of data

A Rodriguez, A Laio, Science 344, 1492 (2014)

Characterize a density distribution by finding its maxima and saddle points

Look for density peaks, i.e. local maxima in the density





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

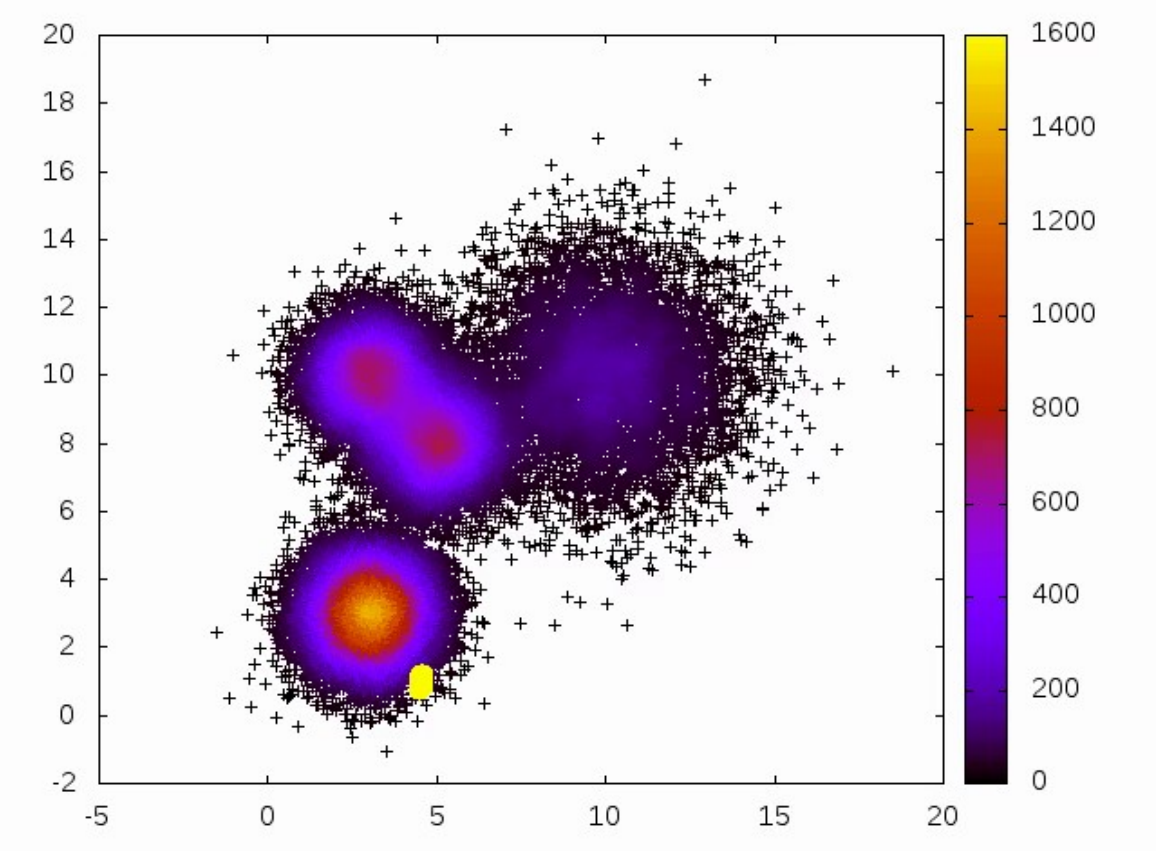
Compute for all points min distance from point at higher density $\delta_i = min_{j:\rho_j > \rho_i} d_{ij}$

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



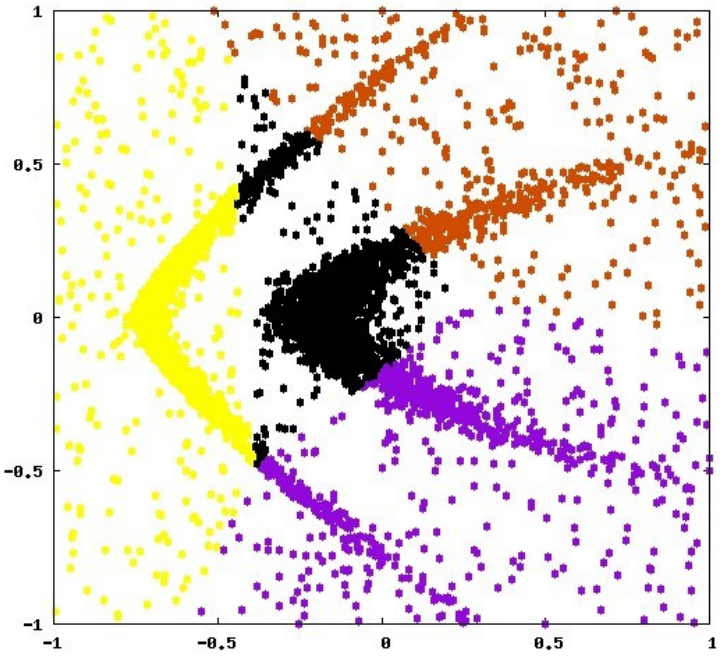
Density-peak clustering

A Rodriguez, A Laio, Science 344, 1492 (2014)

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

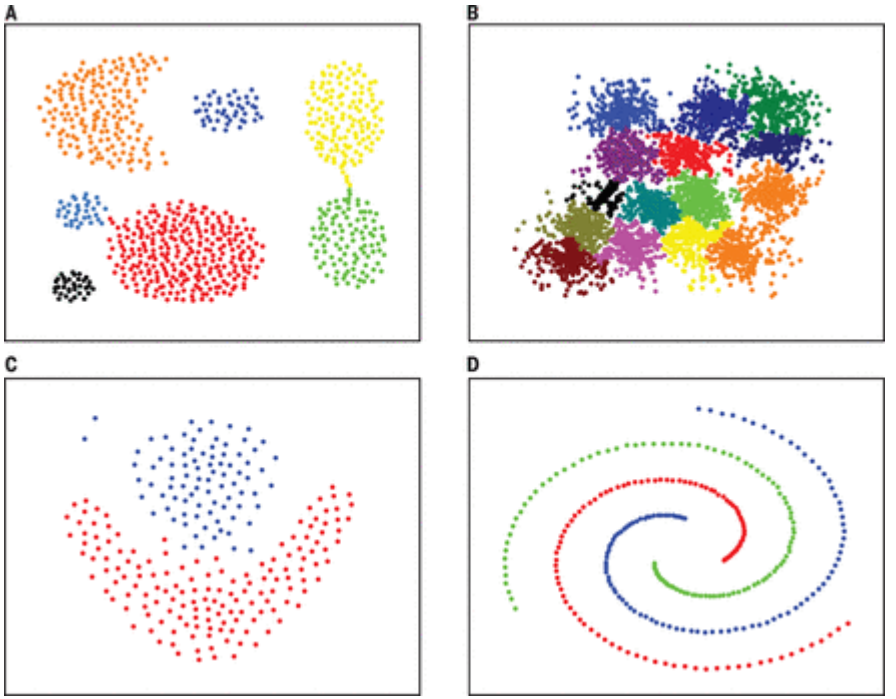
This assignation rule allows to retrieve clusters of arbitrary shape

K-means



DPC

r_i

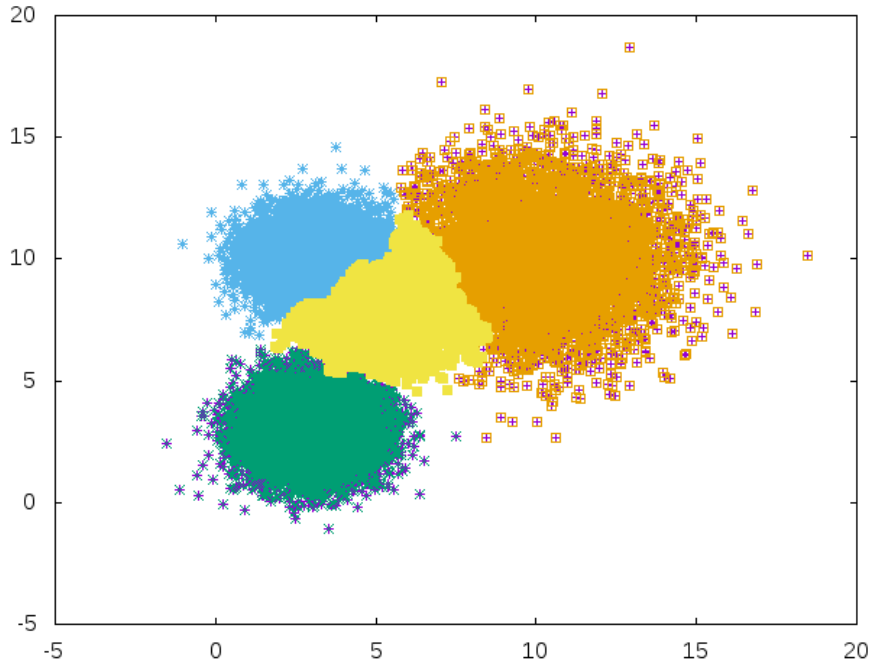


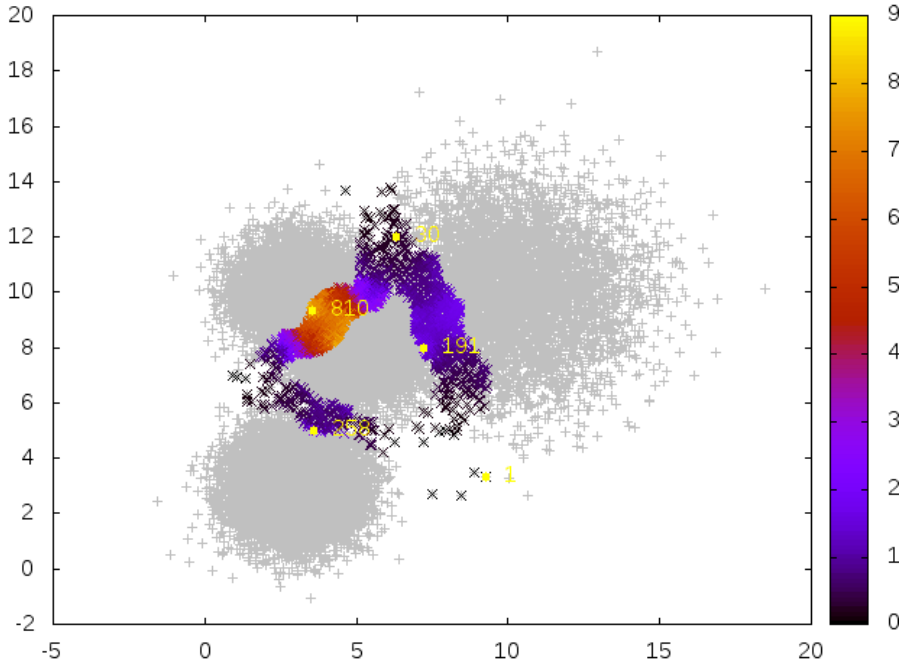
Density-peak clustering

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

After assigning points to peaks, find *border points*: their neighborhood
Contains points assigned to different clusters

Saddle points are density maxima maxima on the borders between peaks





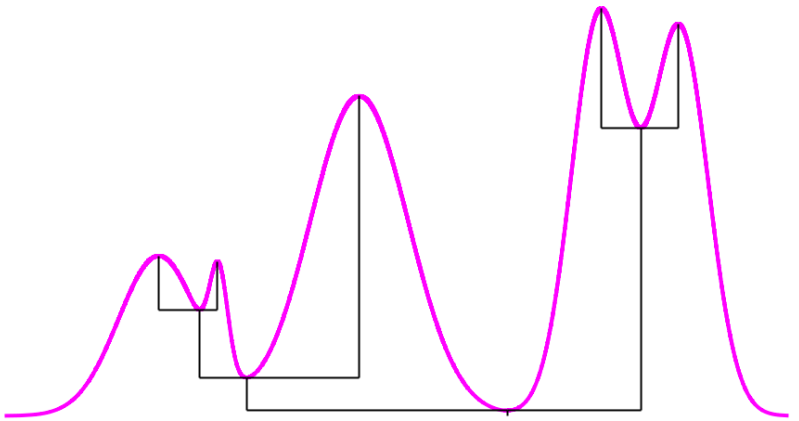
Density-peak clustering

Find hierarchical structure

Peaks with high saddle point between them are “subpeaks” of larger peak

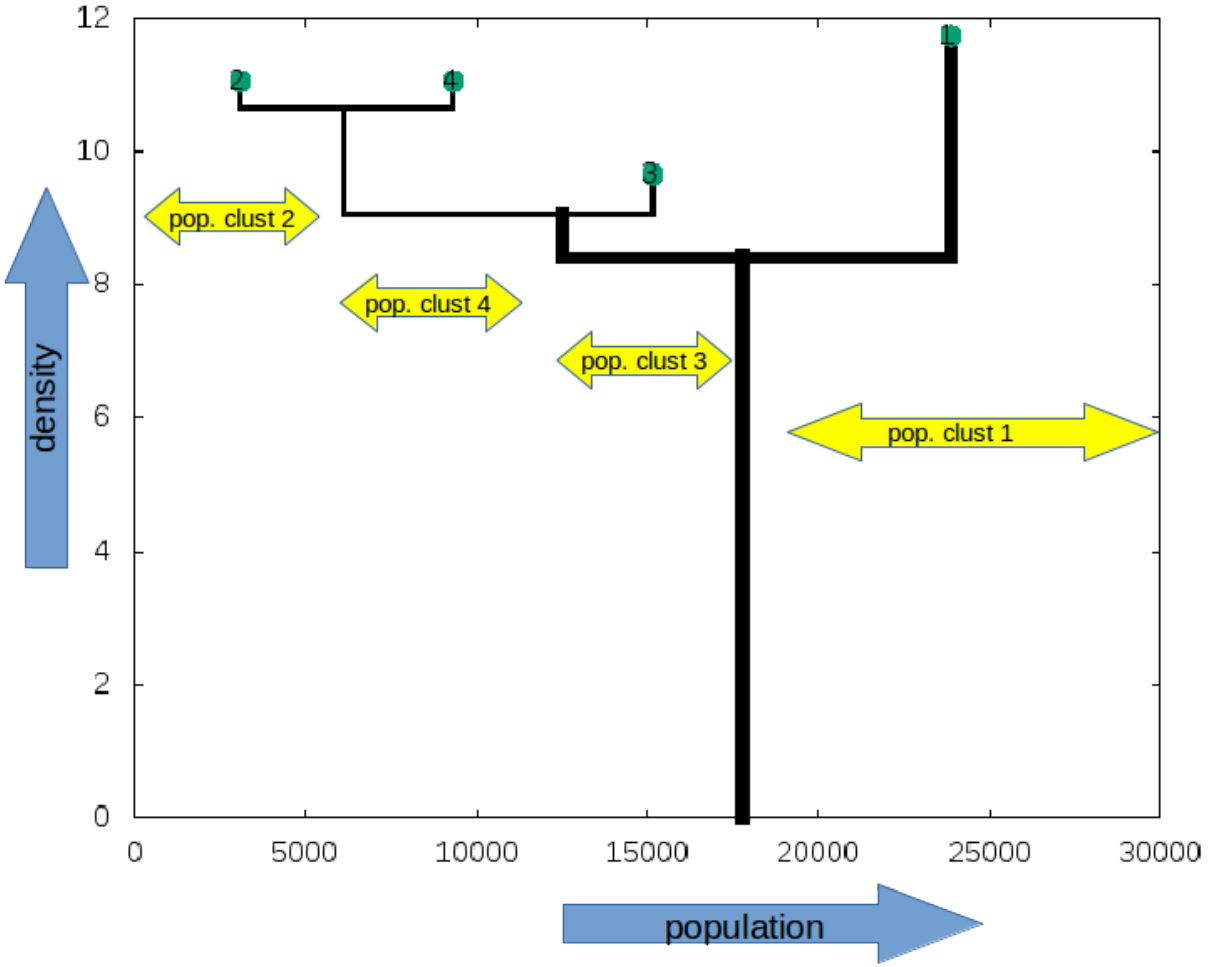
Retrieve such hierarchical structure with single linkage algorithm:

- rank saddle points by their density values
- loop over saddle points and merge two peaks at each step



Density-peak clustering

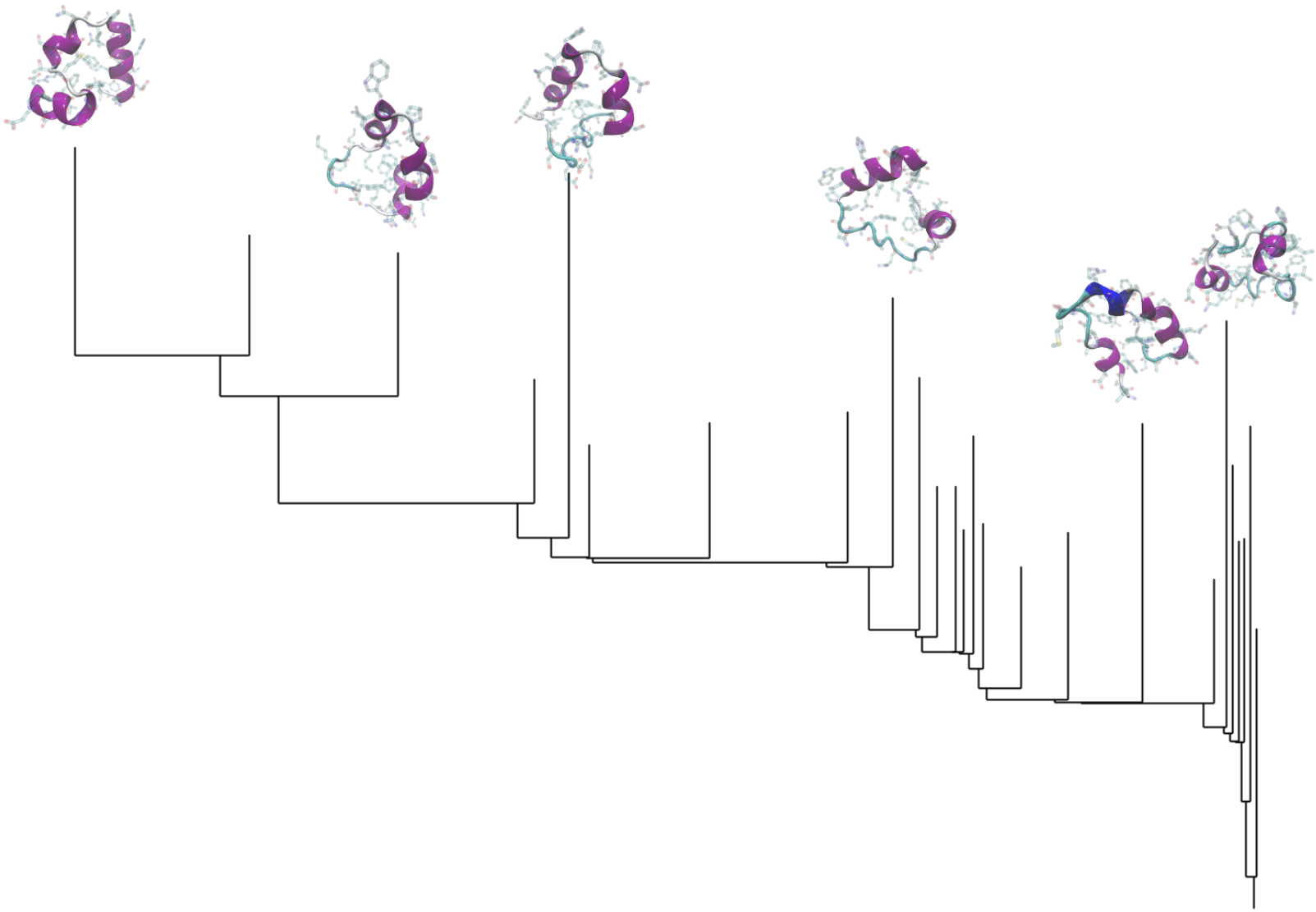
A compact representation of the probability density: **density dendrogram**



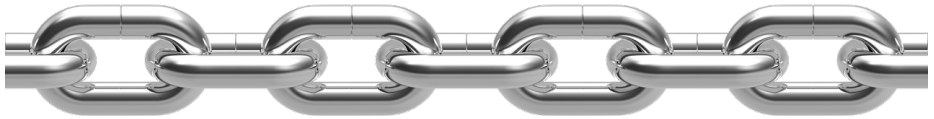
Information about the size and density of the peaks, and their (hierarchical) relations

Density-peak clustering

A compact representation of the probability density: **density dendrogram**



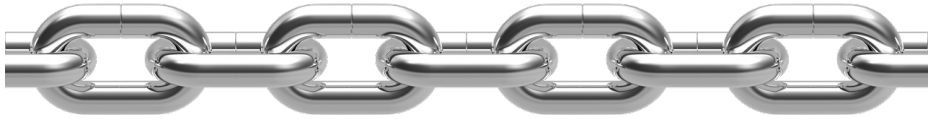
The problem of density Estimation



DPC

- Original DPC had two **one key problem:** dependency on free parameter *k*
- Fixed *k* leads to inaccurate ρ and δρ , hence wrong assessment of the statistical significance of the clusters

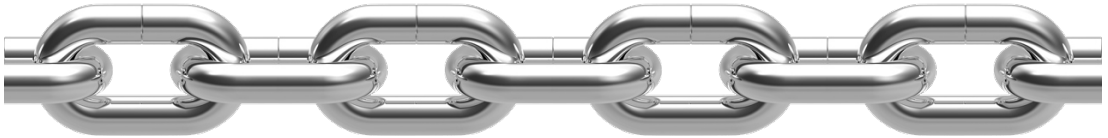
These problems can be solved by means of improved density estimation technique



DPC **PAk**

Density Estimation: PAk

A Rodriguez, M D'Errico, E. Facco and A Laio, JCTC 14 (3), 1206 (2017)



DPC **PAk**

A novel parameter free density estimation approach

What is the right *k?*

Bias – variance tradeoff: $\rho = \frac{k}{V_{ik}}$ $\delta\rho = \frac{\sqrt{k}}{V_{ik}}$

k too small: large error in the estimate

k too large: density is not constant over V

consider point *i* and neighboring point *j*

- If the density at *i* and *j* is different (M1):

$$\mathcal{L}_{M1}(\{v_{ik}\},\{v_{jk}\}|\rho,\rho') = \rho^k e^{-\rho V_{ik}} \rho'^k e^{-\rho' V_{jk}}$$

- If the density at *i* and *j* is the same (M2):

$$\mathcal{L}_{M2}(\{v_{ik}\},\{v_{jk}\}|\rho) = \rho^{2k} e^{-\rho(V_{ik}+V_{ik})}$$

Pointwise Adaptive k-NN

Compare M1 (same ρ) and M2 (different ρ,ρ') for point *i* and its neighbors *j*

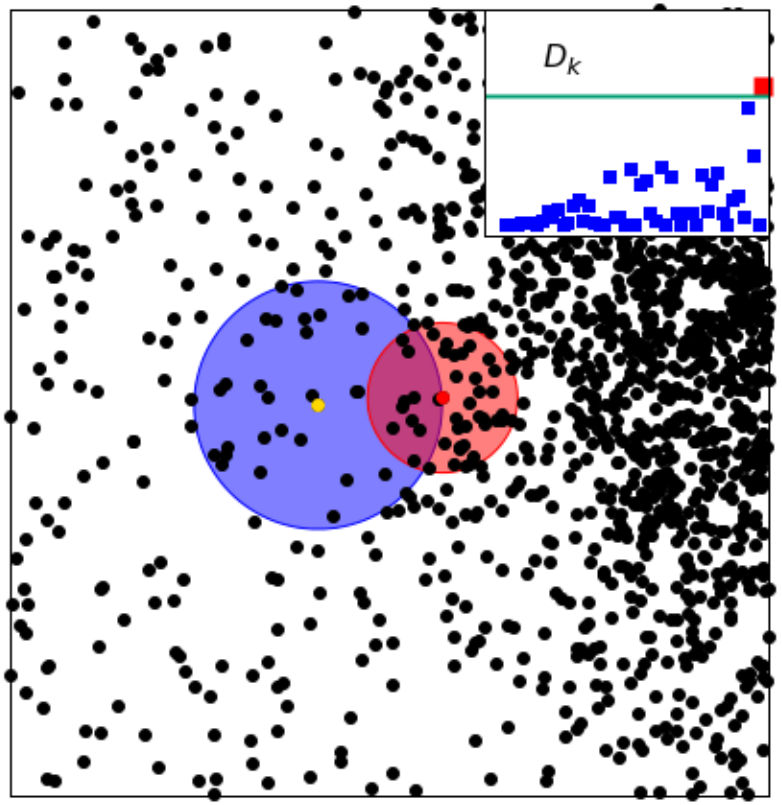
start from first neighbor, then second ….

for each *j* , maximise $\mathcal{L}_{M1}, \mathcal{L}_{M2}$

perform likelihood ratio test to compare M1,M2

$$D = -2\log(\mathcal{L}_{M1}/\mathcal{L}_{M2})$$

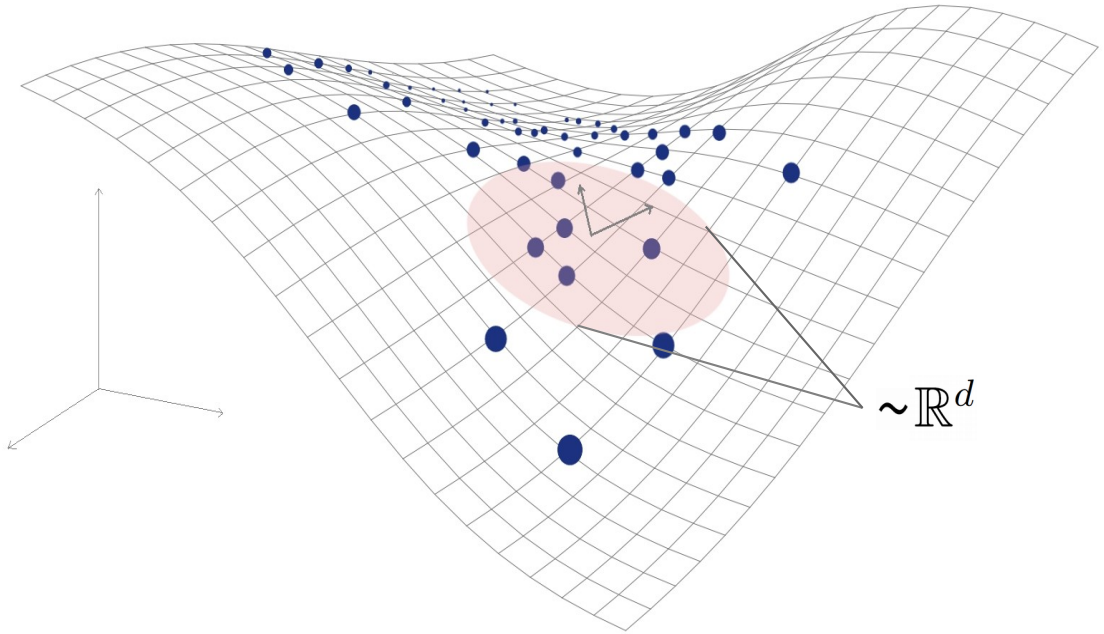
stop when D is large, i.e. M2 is significantly more likely ($p < 10^{-7}$)



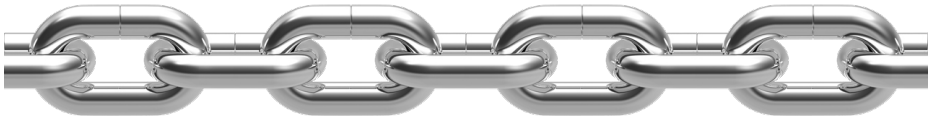
What is the right *d?*

The data actually lie on hypersurface of lower dimension than D

the density should be evaluated on this hypersurface

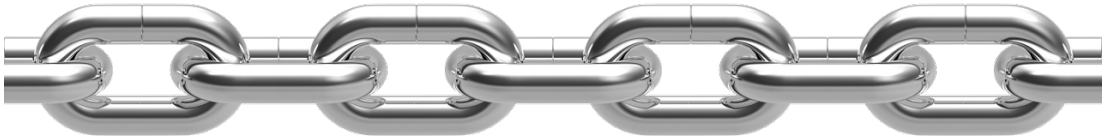


find good ID estimator!



DPC **Pak** **TWO-NN**

Intrinsic dimension Estimation: TWO-NN



DPC Pak **TWO-NN**

A novel intrinsic dimension estimator

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)

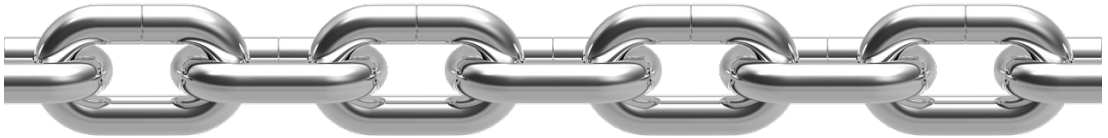
- In principle, one should evaluate simultaneously both d and $\rho(\mathbf{X})$!
- TWO-NN idea: **decouple the estimation problem by finding suitable function of the distances that depends only on *d***
- Assumption: $\rho(\mathbf{X})$ is constant on the scale of the first two neighbors
- Then if d_{i1}, 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

Reconstruction of a probability landscape



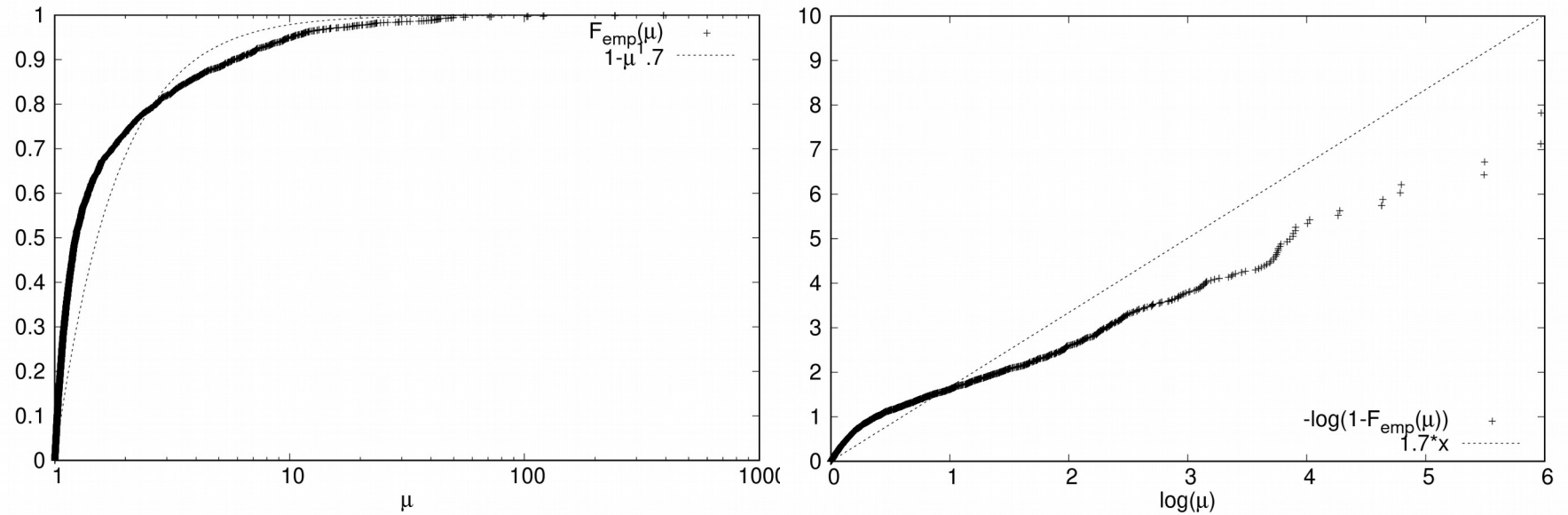
DPC Pak TWO-NN

We reconstruct a probability density, its intrinsic dimension, and its peaks in a high-dimensional space.

The reconstruction of the density effectively takes place in a low-dimensional space, without the need of collective variables

The problem of multiple IDs

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

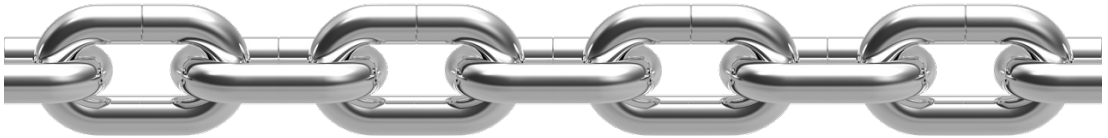


- 1) the density is strongly varying even on the scale of the first two neighbors
- 2) the dimension is not uniform in the dataset

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: Hidalgo



DPC Pak TWO-NN **Hidalgo**

A method to discrimimate regions of different ID in a data set

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 $P(\mu_i) = \sum_{k=1}^{K} p_k d_k \mu_i^{-d_k - 1}$

- 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 fixed by trying increasing values in [1,K_{max}] and performing a model selection test e.g. likelihood ratio 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 $\mathcal{L}(\boldsymbol{\mu}|\mathbf{d}, \mathbf{p}, \mathbf{Z}) = \prod_{i=1}^{N} p_{Z_i} d_{Z_i} \mu_i^{-dz_i - 1}$
- Estimate jointly $\mathbf{d}, \mathbf{p}, \mathbf{Z}$
- **Heterogeneous ID algorithm (hidalgo)**

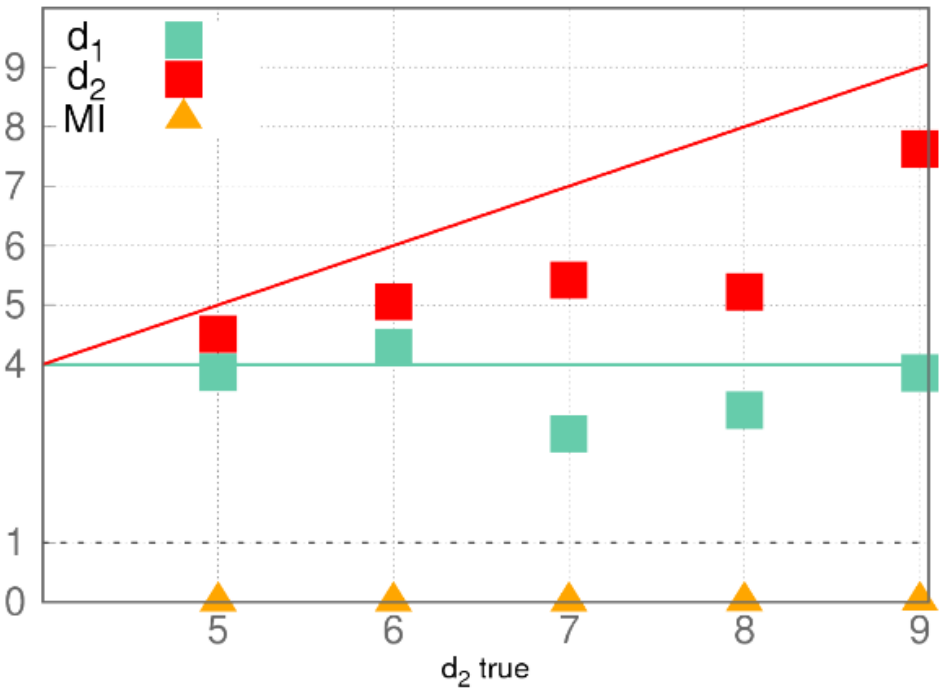
Hidalgo

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

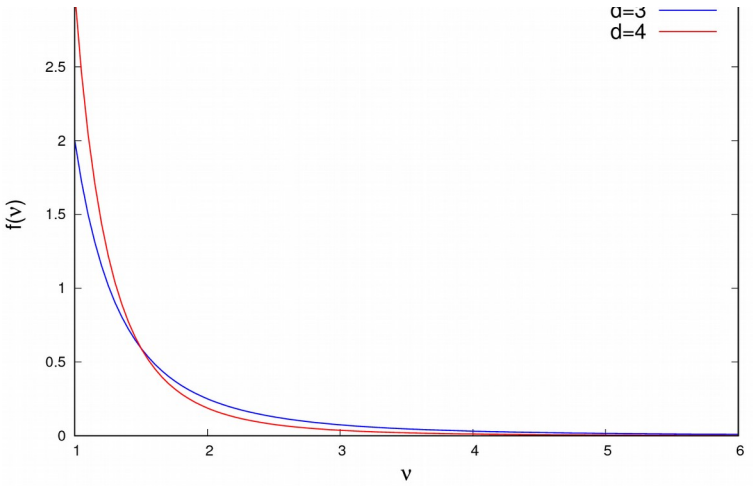


Why?

Pareto distributions with different *d*
are highly overlapping

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

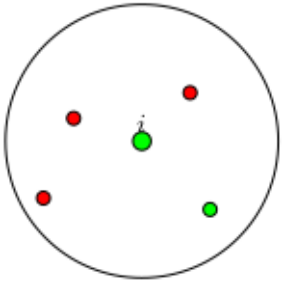
Difficult to assing Z if μ value is not predictive



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

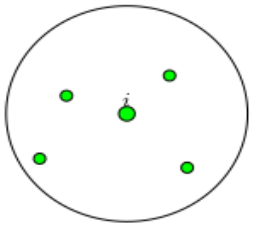
$$\mathcal{L}(n_i^{in}|\mathbf{Z}) = \frac{\zeta^{n_i^{in}}(1-\zeta)^{n_i^{out}}}{\mathcal{Z}}$$

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

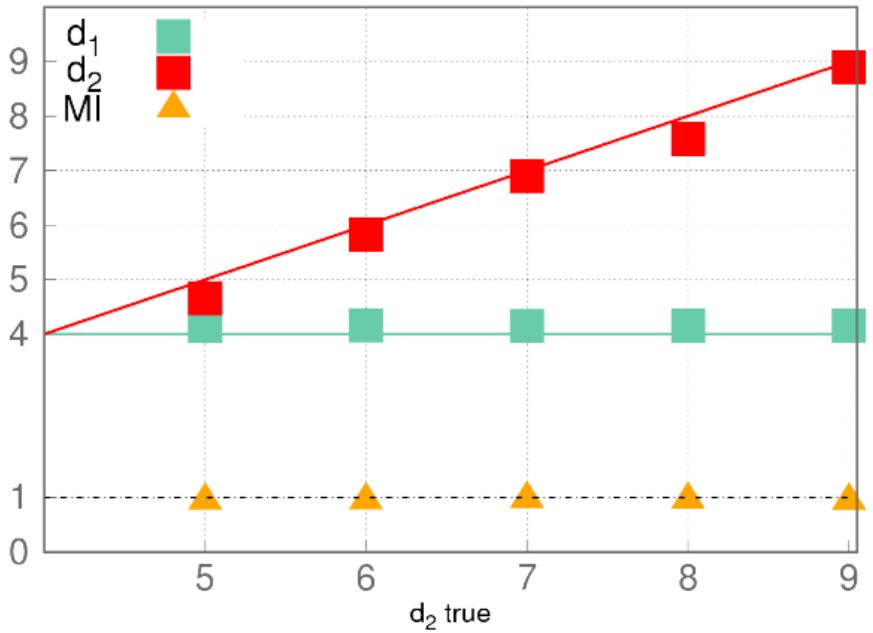
Hidalgo

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***

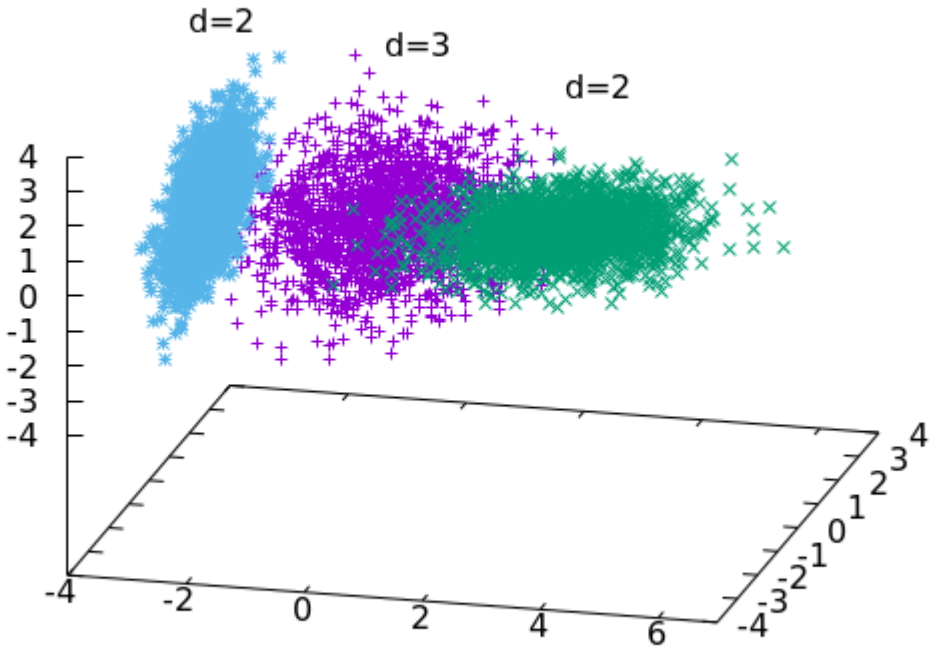


Hidalgo

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

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

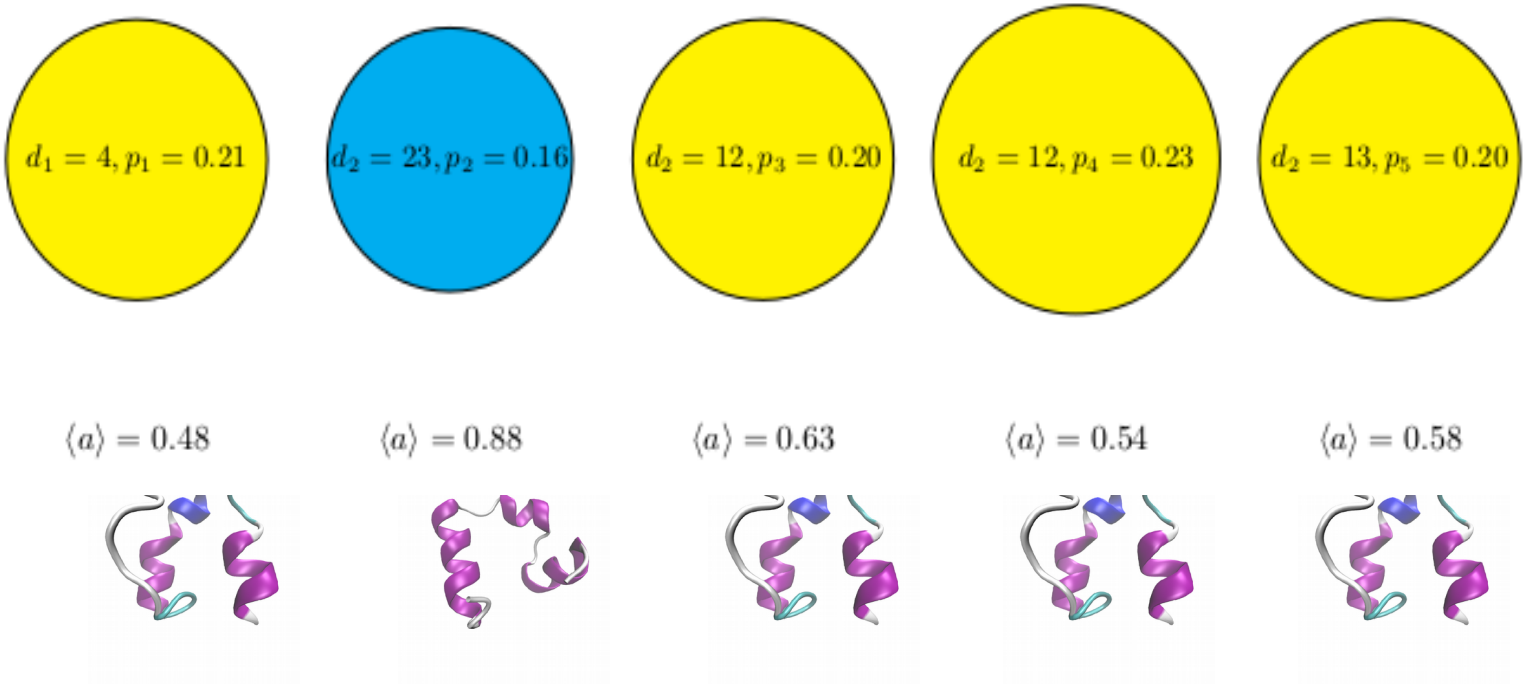
Divide space into regions of uniform intrinsic dimension



Using the information on the **Z,** the different **d** could be used for a more Precise density computation

Hidalgo

Villin example



The folded state is recognized from its higher ID

We can identify it only with topological information

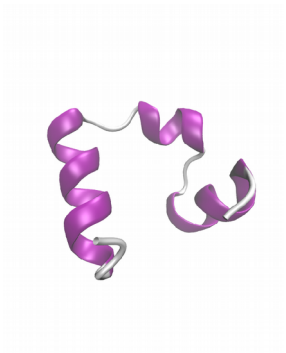
Charting high dimensional data spaces

A chain of methodological developments

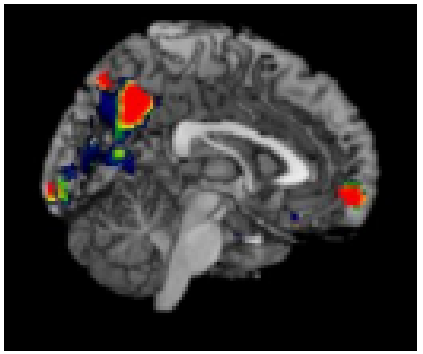
DPC PAk TWO-NN Hidalgo

A toolkit of methods

A lot of applications...



Molecular dynamics



Brain Imaging

Protein classification

Acknowledgments

Alessandro Laio

Maria D’Errico

Elena Facco

Alex Rodriguez

Thank you for your attention!!