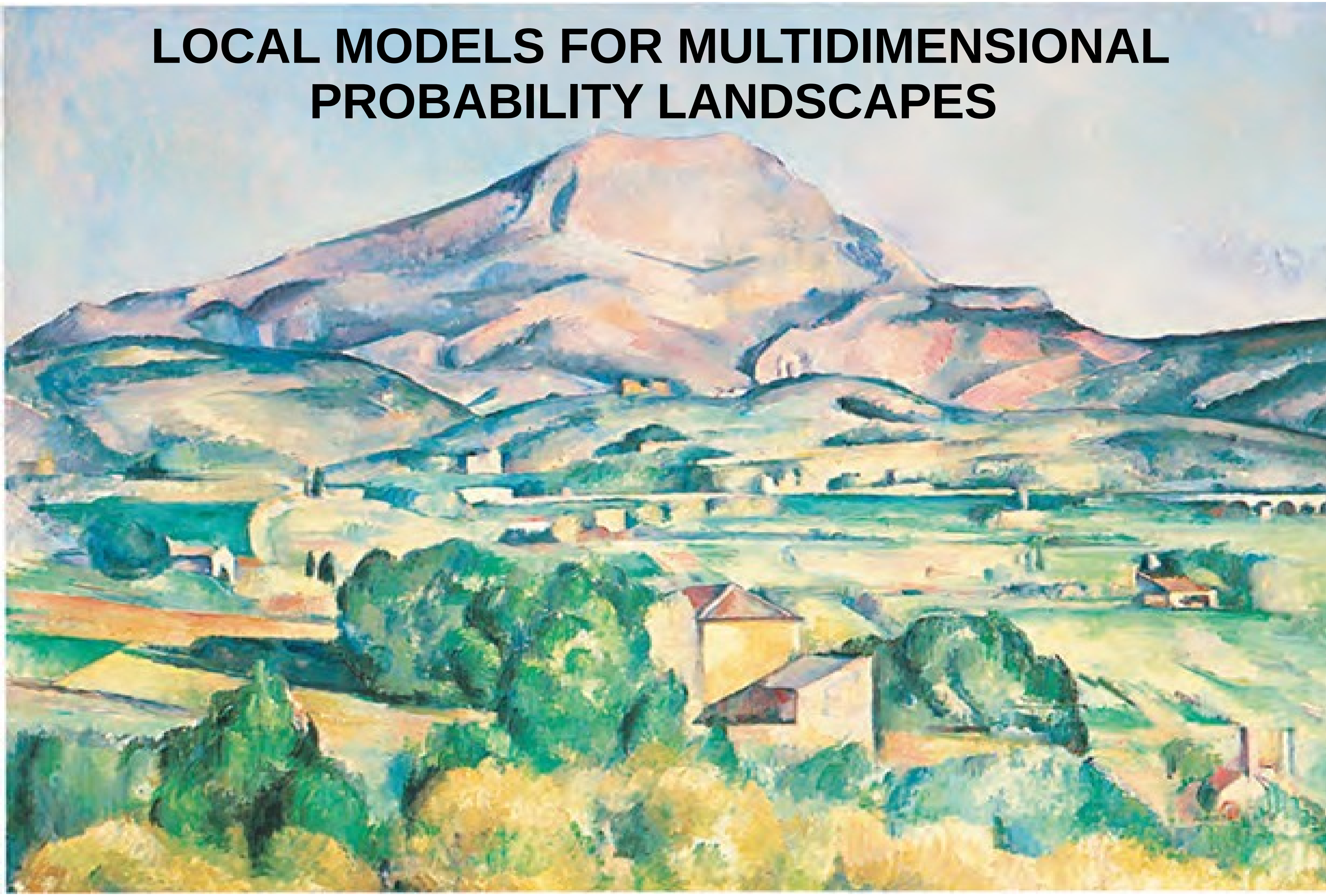
LOCAL MODELS FOR MULTIDIMENSIONAL PROBABILITY LANDSCAPES



Outline

- **Motivation: a rigorous basis for a clustering method**
- **A local model of nearest-neighbor distances under the assumption of locally constant density**
- **Density estimation**
- **Intrinsic dimension estimation**
- **An extension of the model for heterogeneous intrinsic dimension**

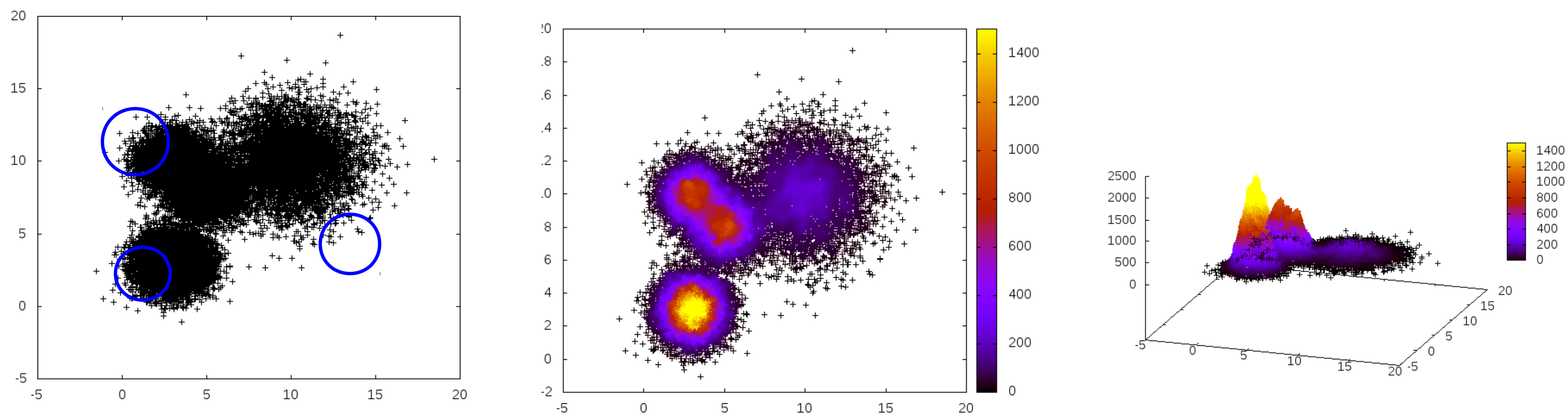
Density peak clustering

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

Find modes (peaks) of a density distribution

Reconstruct density around each point with ε-ball counting:

$$\rho(x_i) = \sum_j \chi(j \in B_\varepsilon(x_i))$$



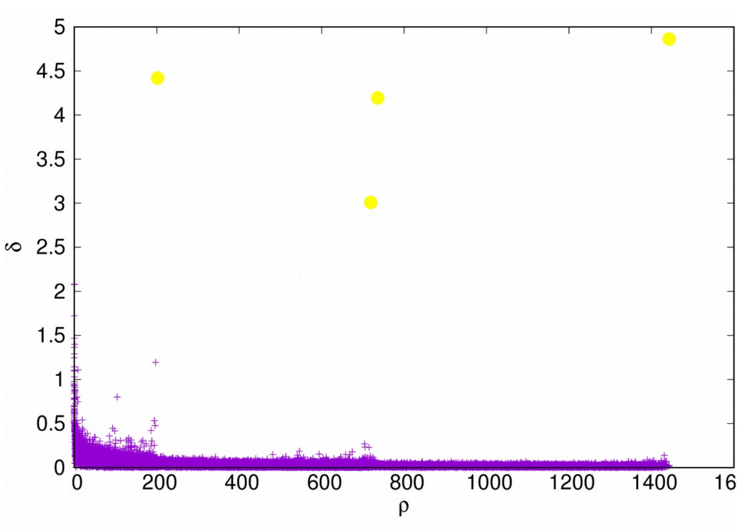
Density Peak Clustering

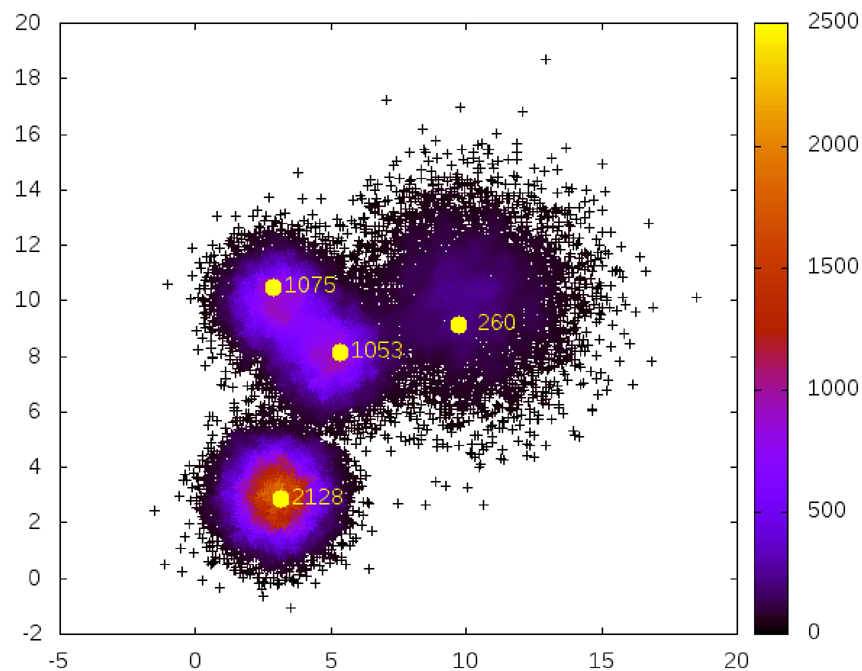
ρ maxima are far from points with lower ρ

Compute minimum distance from point at higher ρ

$$\delta_i = min_{j:\rho_j > \rho_i} d_{ij}$$

Peak are outliers in *decision graph* ρ_i vs δ_i

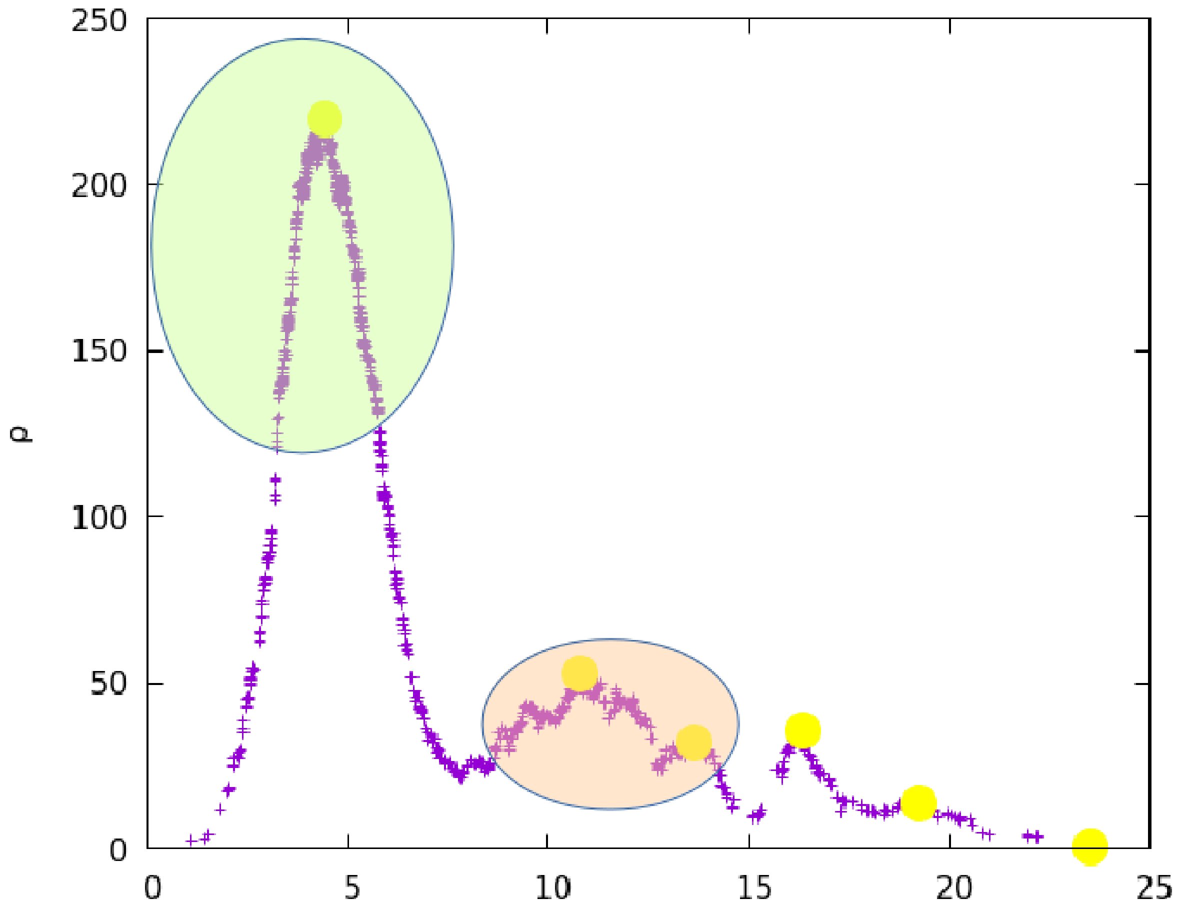




Validating DPC

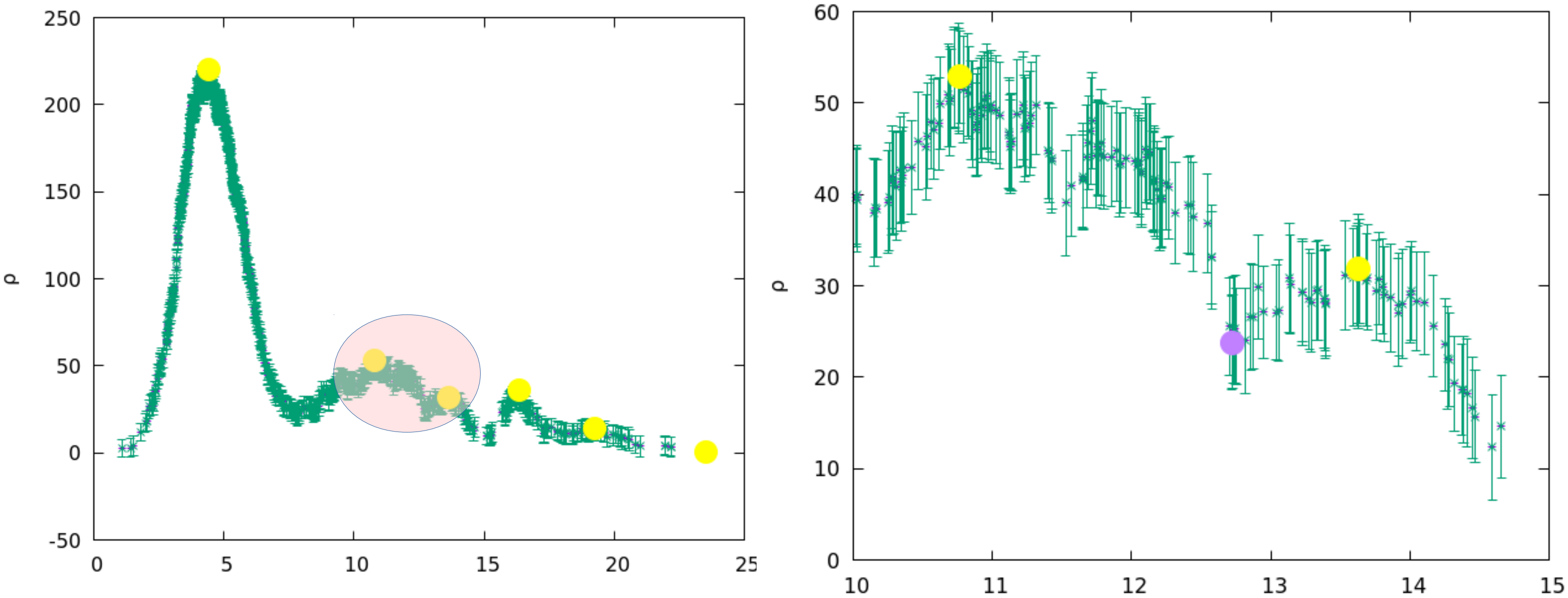
“Significance” of the peaks?

Peaks or density fluctuations?



Validating DPC

- Find **peaks and saddle points**
- Compute **error Δρ**
- Significance: $$\rho_{peak} - \rho_{saddle} \geq z(\sqrt{(\Delta\rho_{peak})^2 + (\Delta\rho_{saddle})^2})$$



A local model for the data

No global model for the data

Only two **broad assumptions**:

H1) the data points x_i are **independent samples** from a density ρ(x).

H2) local uniformity: for all x_i, there exists (small) k such that ρ(x) ~ const. in the region containing the first *k* neighbors of x_i

→ **local model** for the distribution of neighbor distances around each point

\# points within a small region around each point follows **Poisson process** with **parametric dependence on ρ**

$$prob(\mathcal{N}(\mathcal{A}) = n) = \frac{\rho(x_i)^n vol(\mathcal{A})^n}{n!} e^{-\rho(x_i) vol(\mathcal{A})}$$

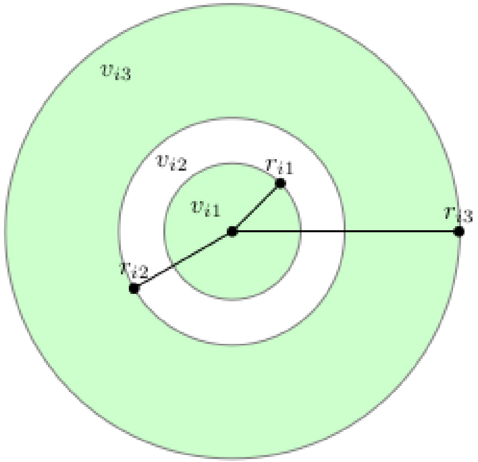
Estimating the density

***k* nearest neighbors** of *i* at distances $r_{i1}, r_{i2}, r_{i3}, \ldots$

hyperspherical shells S_j enclosed between neighbors

volumes $v_j = \omega_d r_{ij}^d - \omega_d r_{i(j-1)}^d$

distribution of shell volumes V_j follows from Poisson process



$$prob(\mathcal{N}(s_j) = 0) = e^{-\rho(x_i)v_j} = prob(V_j > v_j)$$

$$\mathcal{L}(V_j = v_j) = \rho(x_i)e^{-\rho(x_i)v_j}$$

- Considering all volumes $\mathcal{L}(\{V_j = v_j\}) = \rho(x_i)^k e^{-\rho(x_i)vol(B_{r_{ik}}(x_i))}$

- **Local model of NN distances depending on ρ**

Estimating the density

By max likelihood estimate ρ and error Δρ

$$\rho(x_i) = \frac{k}{\omega_d r_{ik}^d}, \quad \Delta\rho(x_i) = \frac{\sqrt{k}}{\omega_d r_{ik}^d}$$

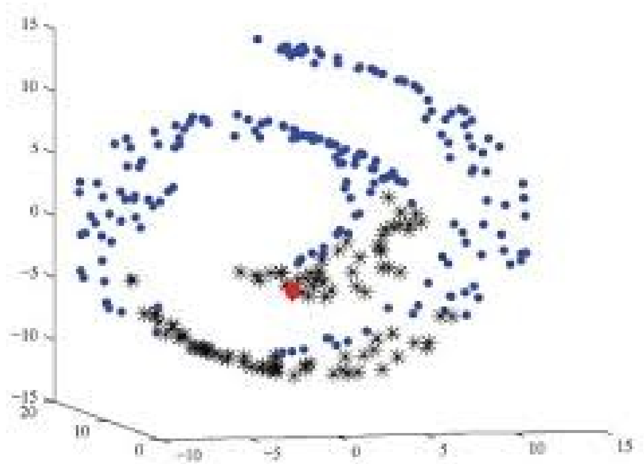
Two problems:

- 1) **what is right *k***?
- 2) **what is right *d***?

What is right *k*? **increase k until local model fails** [Rodriguez et al., JCTC 2017]

What is right *d*? **Intrinsic dimension**

- The data lie on d-dimensional hypersurface
- ρ should be evaluated on this hypersurface



ID estimation: TWO-NN

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

local model for k=2

$$\nu = v_2/v_1$$

distribution of ν is independent of ρ

$$\mathcal{L}(\nu) = \frac{1}{1+\nu^2}$$

$$\mu = r_{i2}/r_{i1} \qquad \nu = \mu^d - 1$$

under local model, distribution of μ depends only on *d*

$$\mathcal{L}(\mu) = de^{-(d+1)\mu}$$

ID can be inferred from the **μ of all points** collectively

This is **independent of the estimates of ρ**
(**assuming ρ is constant over scale of first 2 neighbors**)

ID estimation: heterogeneous case

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

ID may not be uniform in the dataset!

H3) ρ(x) has **support on the union of a finite number K of manifolds**

with different intrinsic dimensions $d_1 \ldots d_k$

Mixture model $$\rho(x) = \sum_{k=1}^{K} p_k \rho(x|k)$$

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

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 $\mathbf{d} = d_1 \ldots d_K, \quad \mathbf{p} = p_1 \ldots p_K$

ID estimation: heterogeneous case

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

ID may not be uniform in the dataset!

H3) ρ(x) has **support on the union of a finite number K of manifolds**

with different intrinsic dimensions $d_1 \dots d_k$

Mixture model $$\rho(x) = \sum_{k=1}^{K} p_k \rho(x|k)$$

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

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

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

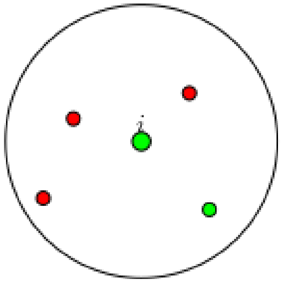
Then we can again estimate $\mathbf{d} = d_1 \dots d_K, \quad \mathbf{p} = p_1 \dots p_K$

ID estimation: heterogeneous case

- Introduce **latent variables** (manifold membership of each point) $\mathbf{Z} = Z_1, \ldots, Z_N$
- Likelihood with latent variables $\mathcal{L}(\boldsymbol{\mu}|\mathbf{d},\mathbf{p},\mathbf{Z}) = \prod_{i=1}^{N} p_{Z_i} d_{Z_i} \mu_i^{-d_{Z_i}-1}$
- Estimate jointly $\mathbf{d}, \mathbf{p}, \mathbf{Z}$
- *K* fixed by trying increasing values in $[1, K_{max}]$ and performing a model selection test e.g. likelihood ratio test

Problem:

- Pareto distributions with different *d* are highly overlapping
- **estimation of manifold membership fails**
- diagnostic: **neighboring points have different Z**



Hidalgo

- **H4)** the first *q* neighbors of a point mostly belong to the same manifold

Probabilistic requirement on the Z of neighboring points

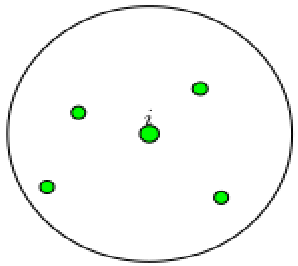
Z: be probability that a neighbour of *i* belong to same manifold as *i*

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

n_i^{out} # neighbors with diffferent Z

additional term in the likelihood

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

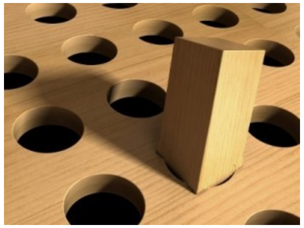


$\zeta > \frac{1}{2}$ Controls the degree of uniformity

Two types of data analysis

Confirmatory analyis

- Starts from assumed *model* for the data, given a priori
- Uses *statistics* to verify whether the data fit the assumed model
- Can be rigid: fail to exploit richness of the data



Exploratory analyis

- Procedures (algorithms) to find structure in the data
- Often, *no formal evaluation* of the results
- Danger of falling into magical thinking (seeing structures that are not there)

A possible compromise

- We started with E.D.A. method (density peak clustering) with no statistical validation of results
- for statistical validation, some assumption on the data was needed
- we introduced *minimal assumptions* on the data, allowing to maintain high flexibility
- as a result, we developed complex procedure to reconstruct probability density, its intrinsic dimension and peaks in high dimensional space

Acknowledgments

Alessandro Laio

Maria d’Errico

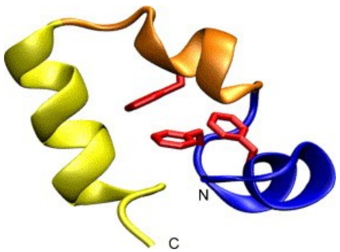
Elena Facco

Alex Rodriguez

Antonietta Mira

Hidalgo

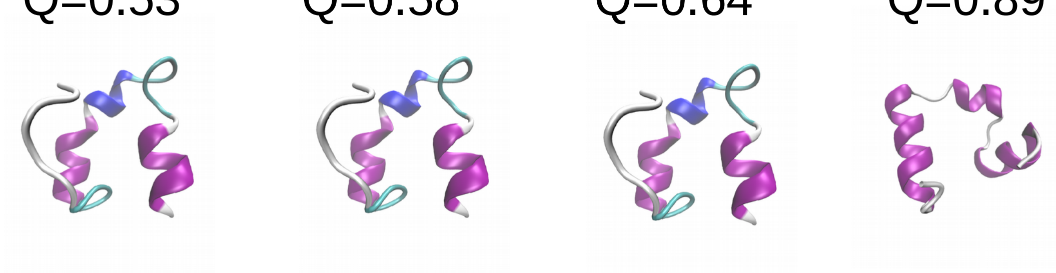
Example: molecular dynamics



- 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!