Data Dependent Distances for Unsupervised Learning

James M. Murphy
Department of Mathematics
Tufts University

Collaborators

Mauro Maggioni
Johns Hopkins University

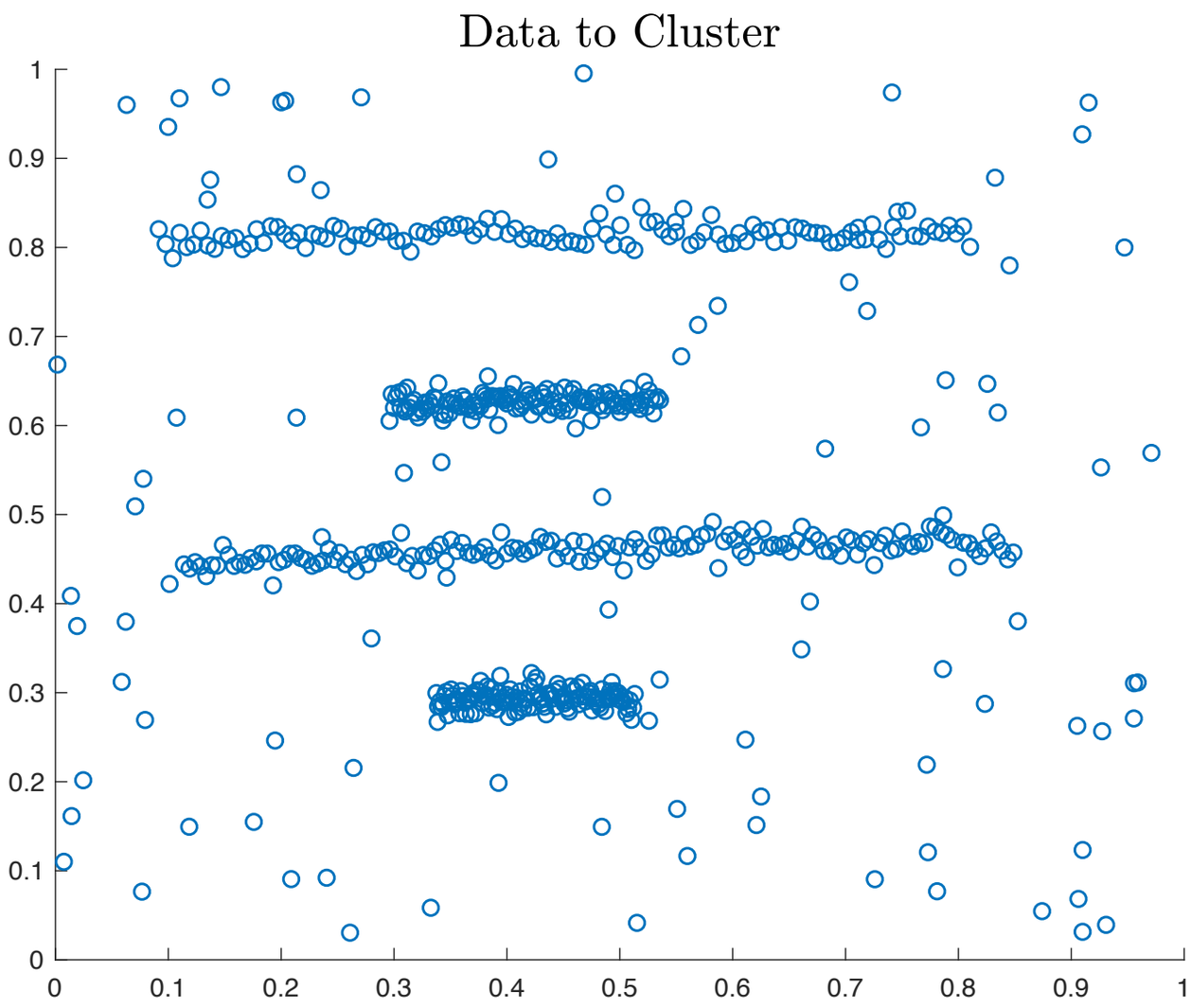
Anna Little
Michigan State University

Tufts UNIVERSITY

Unsupervised Learning

Unsupervised learning: infer structure from data without access to *training data*, i.e. examples belonging to particular classes.

Clustering: unsupervised learning in which the goal is to label points as belonging to a given class.



$$x_1, \ldots, x_n \overset{iid}{\sim} \mu = \sum_{k=1}^{K} w_k \mu_k + \tilde{\mu}, \quad \sum_{k=1}^{K} w_k = 1$$

Labeling: Which x_j were generated from μ_k?
Number of Clusters: Can we estimate K?

Spectral Clustering I

Idea: embed data into a lower-dimensional space in a structure preserving way.

Input: $x_1, \ldots, x_n \subset \mathbb{R}^D$

Step 1: Build a *weight matrix*

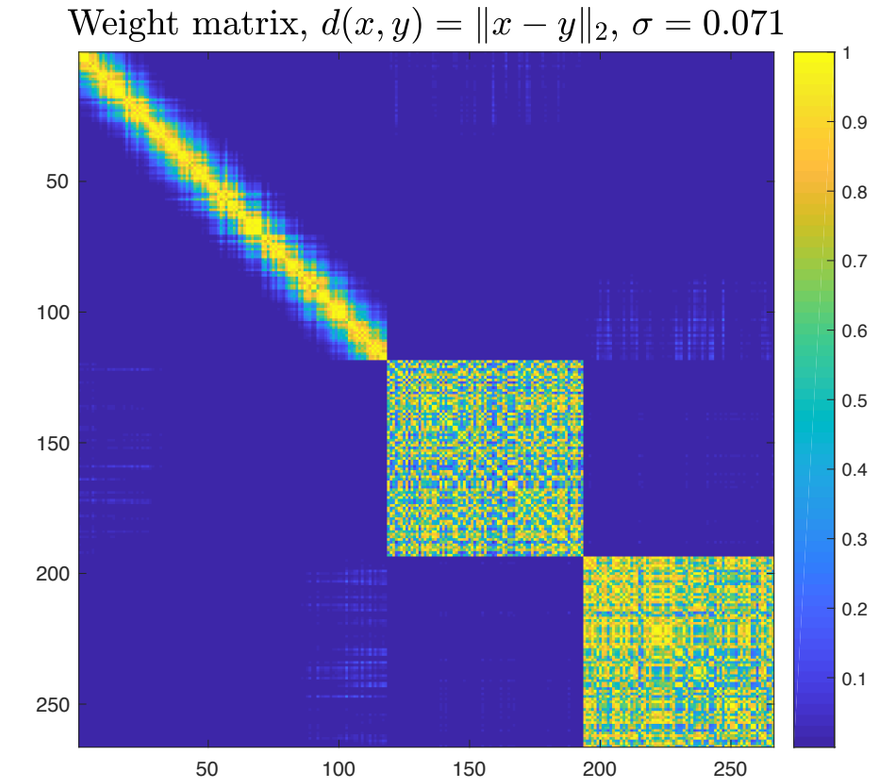
$$W_{ij} = e^{-d(x_i, x_j)^2/\sigma^2}$$

for some metric $d(\cdot, \cdot)$ and σ.

Step 2: Compute the *(graph) Laplacian*

$$L = I - D^{-\frac{1}{2}} W D^{-\frac{1}{2}}$$

$$D_{ii} = \sum_{j=1}^{n} W_{ij}; D_{ij} = 0, i \neq j.$$



U. Von Luxburg. "A tutorial on spectral clustering". *Statistics and Computing*. 2007. 17(4):395-416.

Tufts UNIVERSITY

Spectral Clustering II

Step 3: Compute eigenvalues of L

$$0 \leq \lambda_1 \leq \lambda_2 \leq ... \leq \lambda_n$$

and associated eigenvectors

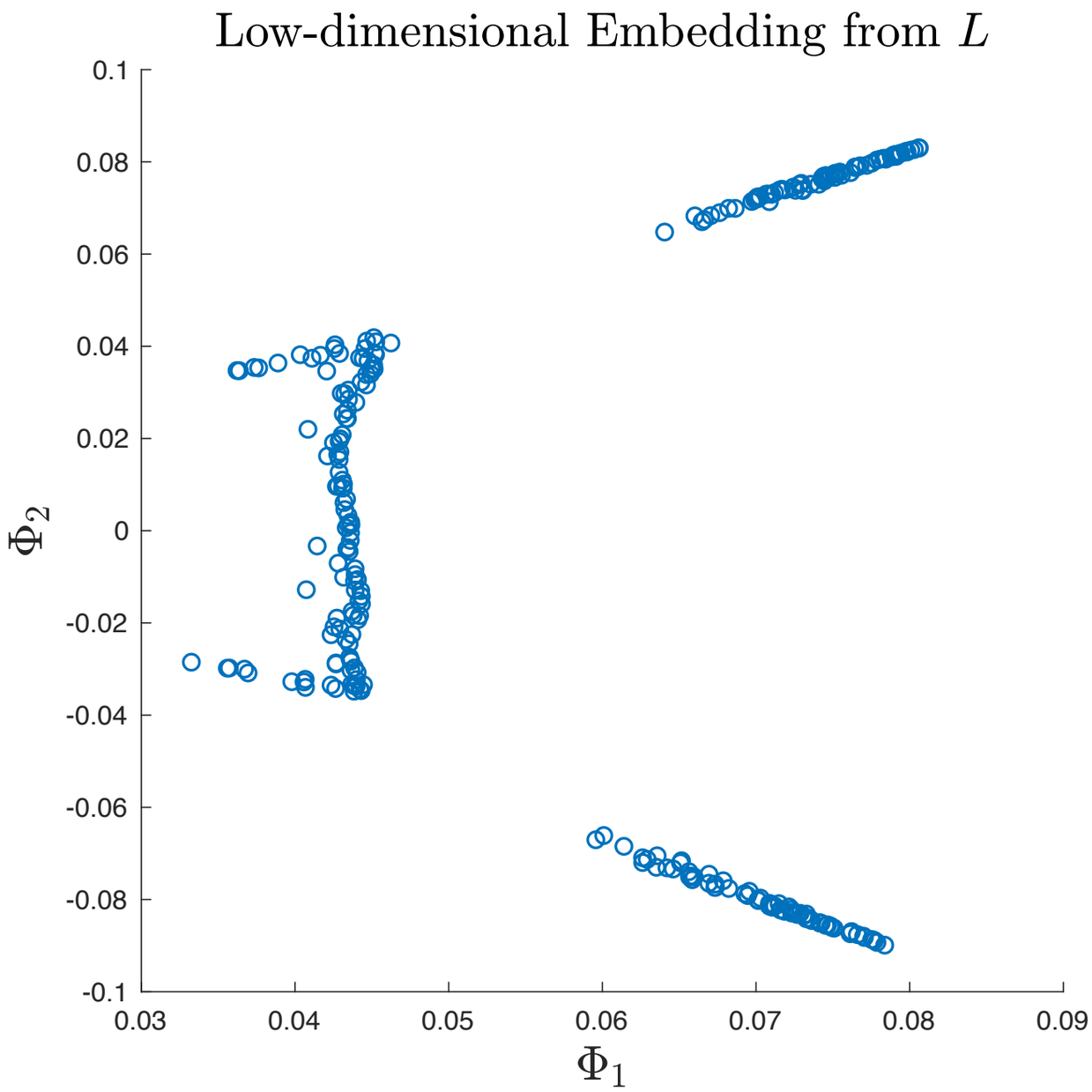
$$\Phi_1, ..., \Phi_n.$$

Step 4: Embed the data as

$$x_i \mapsto (\Phi_1(x_i), \ldots, \Phi_K(x_i))$$

then run K-means. Note

$$\Phi_j(x_i) := \Phi_j(i).$$



Parameter Problems:

- Dependence on parameters σ, K .
- Heuristic: $K \approx \arg\max_k \lambda_{k+1} - \lambda_k$

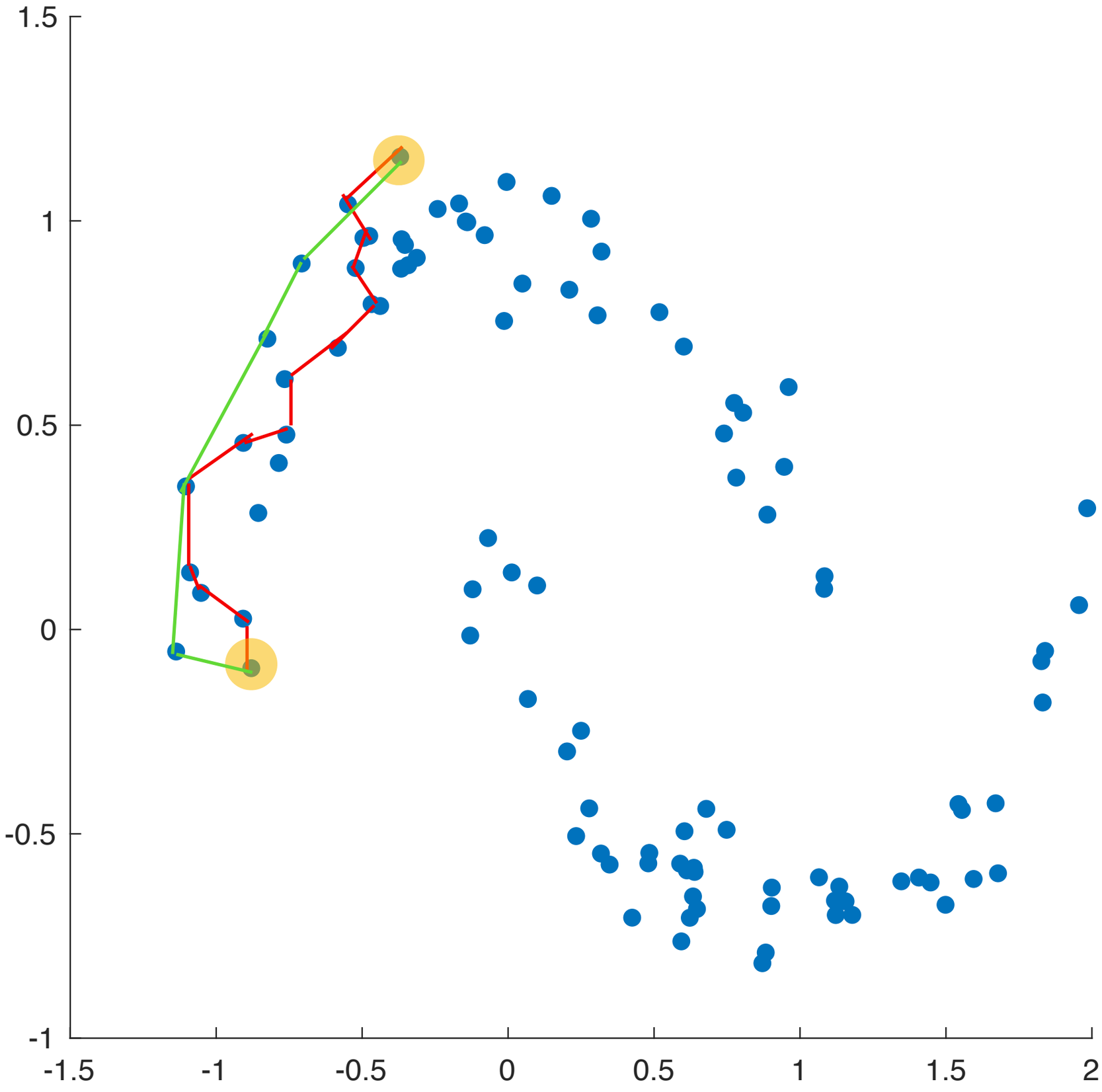
Ultrametric Path Distances

Definition. *For a discrete set $X = \{x_i\}_{i=1}^n \subset \mathbb{R}^D$, let $\mathcal{G}$ be the graph on X with edges given by the Euclidean distance between points. For $x_i, x_s \in X$, let $\mathcal{P}(x_i, x_s)$ denote the space of paths connecting x_i, x_s in $\mathcal{G}$. The* ***longest leg path distance (LLPD)*** *between x_i, x_s is:*

$$d_{\ell\ell}(x_i, x_s) = \min_{\{y_j\}_{j=1}^L \in \mathcal{P}(x_i, x_s)} \max_{j=1,2,\ldots,L-1} \|y_{j+1} - y_j\|_2,$$

- The distance between points x, y is the minimum over all paths between x, y of the longest edge in the path.
- Depending on the data X, this distance changes!
- We are re-shaping the unit ball to respect the geometry of the data.
- $\mathcal{G}$ could be a complete graph (all points connected to all points) or a connected NN graph.
- Looks hard to compute. We will present a fast approximation algorithm.

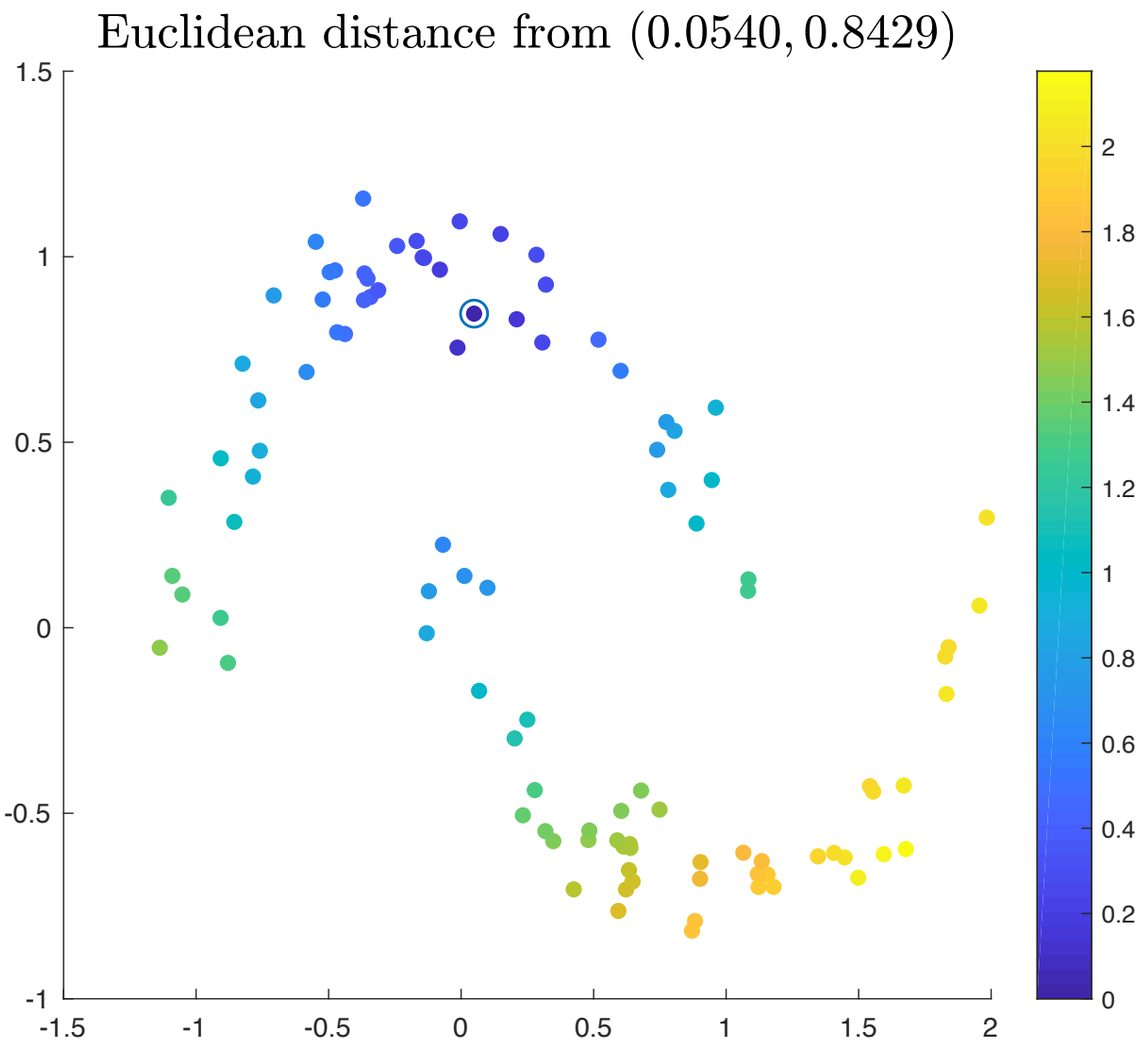
LLPD Visualization

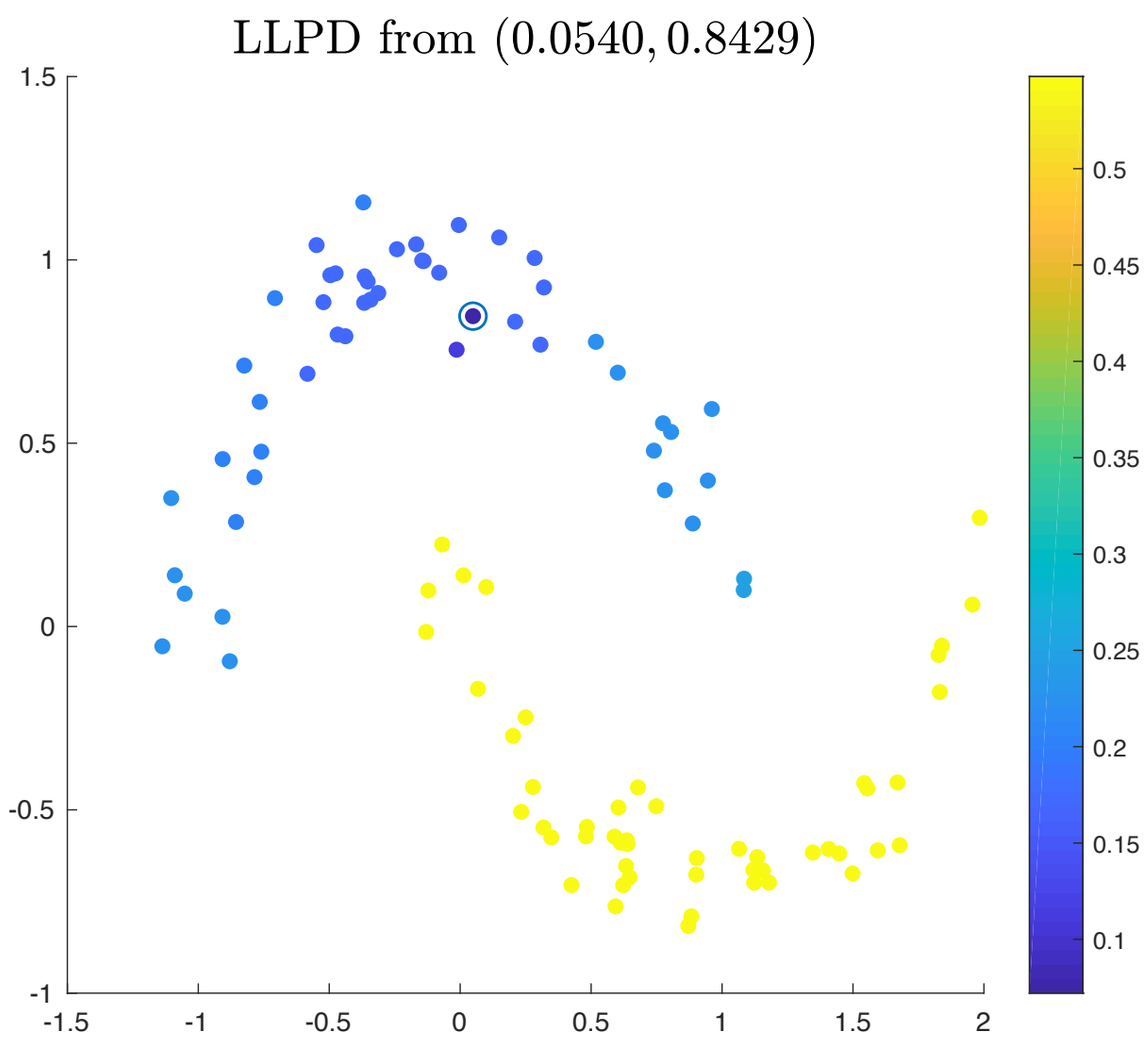


The green and red paths both connect the specified points, but the red path has smaller maximal edge length.

Tufts UNIVERSITY

Euclidean Distance versus LLPD





Low Dimensional, Large Noise (LDLN) Model

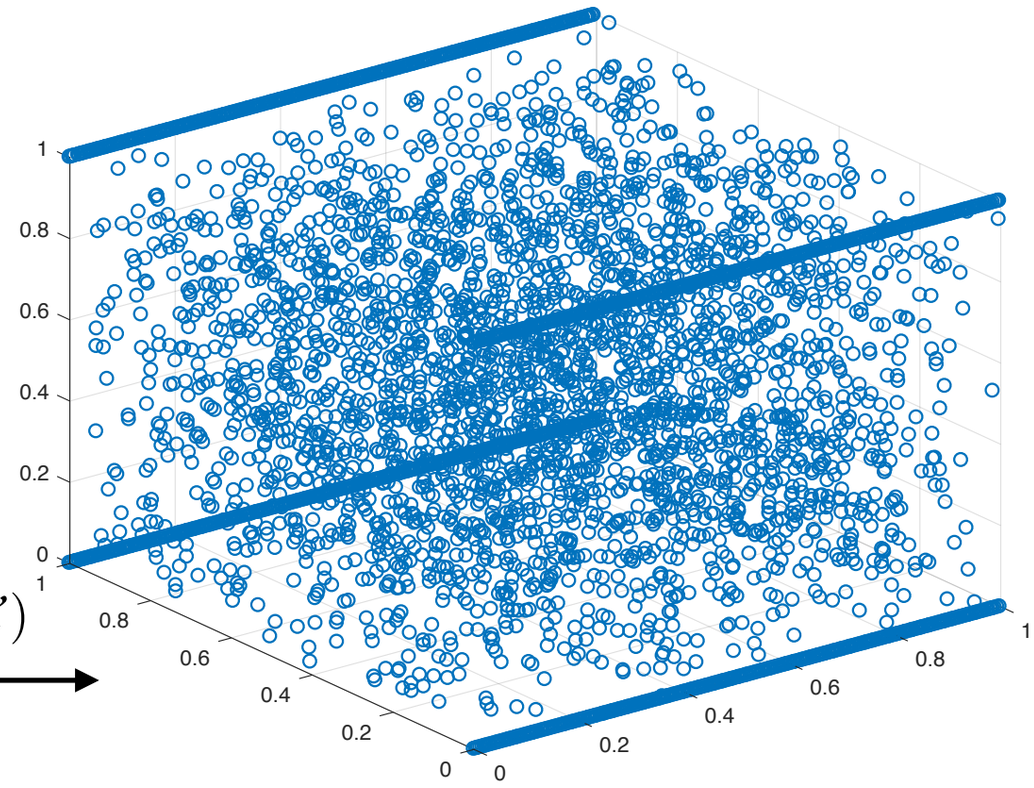
Definition. *A set $S \subset \mathbb{R}^D$ is an element of $\mathcal{S}_d(\kappa, \epsilon_0)$ for some $\kappa \geq 1$ if it has finite d-dimensional Hausdorff measure, denoted by $\mathcal{H}^d$, is connected, and for some $\epsilon_0 > 0$, it satisfies the following geometric condition:*

$$\forall x \in S, \quad \forall \epsilon \in (0, \epsilon_0), \qquad \kappa^{-1}\epsilon^d \leq \frac{\mathcal{H}^d(S \cap B_\epsilon(x))}{\mathcal{H}^d(B_1(0))} \leq \kappa\epsilon^d \,.$$

Low-dimensional

$$\mathcal{X}_1, \ldots, \mathcal{X}_K \subset \mathcal{X} \subset \mathbb{R}^D$$
$$\mathcal{X}_1, \ldots, \mathcal{X}_K \in \mathcal{S}_d(\kappa, \epsilon_0)$$
$$\delta = \min_{k \neq k'} \operatorname{dist}(\mathcal{X}_k, \mathcal{X}_{k'})$$

n_i i.i.d. draws from Unif$(\mathcal{X}_i)$



Large noise

$$\tilde{\mathcal{X}} = \mathcal{X} \setminus (\mathcal{X}_1 \cup \ldots \cup \mathcal{X}_K)$$

$\tilde{n}$ i.i.d. draws from Unif$(\tilde{\mathcal{X}})$

$$n = n_1 + \ldots + n_K + \tilde{n}$$
$$n_{\min} = \min_{1 \leq k \leq K} n_k$$

Tufts UNIVERSITY

Nearest Neighbors in LLPD and Denoising

- In the LDLN model, points within clusters all have comparable distances, and points from different clusters are well separated.

- We denoise points by removing all points whose distance to their $k_{\text{nse}}{}^{th}$ nearest neighbor exceeds some threshold θ .

- k_{nse}, θ are parameters.

- This analysis, based on percolation theory, proves the weight matrix is nearly block constant.

Performance Guarantees

Theorem. *(Little, Maggioni, **M.**) Under the LDLN data model and assumptions, suppose that the cardinality $\tilde{n}$ of the noise set is such that*

$$\tilde{n} \leq \left(\frac{C_2}{C_1}\right)^{\frac{k_{nse}D}{k_{nse}+1}} n_{min}^{\frac{D}{d+1}\left(\frac{k_{nse}}{k_{nse}+1}\right)}.$$

Let $f_\sigma(x) = e^{-x^2/\sigma^2}$ be the Gaussian kernel and assume $k_{nse} = O(1)$ and that $\frac{\max_i n_i}{n_{min}} = O(1)$. Let n_{min} be sufficiently large enough and let θ, σ satisfy

$$C_1 n_{min}^{-\frac{1}{d+1}} \leq \theta \leq C_2 \tilde{n}^{-\left(\frac{k_{nse}+1}{k_{nse}}\right)\frac{1}{D}} \tag{1}$$

$$C_3\theta \leq \sigma \leq C_4\delta \tag{2}$$

Let L be the LLPD Laplacian defined on the denoised data X_N, that is, $L = I - D^{-\frac{1}{2}}WD^{-\frac{1}{2}}$, where $W_{ij} = f_\sigma(\rho_{\ell\ell}(x_i, x_j))$. Then with high probability:

(i) the largest gap in the eigenvalues of L is $\lambda_{K+1} - \lambda_K$.

(ii) spectral clustering with L with K principal eigenvectors achieves perfect accuracy on X_N.

The constants $\{C_i\}_{i=1}^4$ depend on the geometric quantities but do not depend on $n_1, \ldots, n_K, \tilde{n}, \theta, \sigma$.

A. Little, M. Maggioni, and J.M. Murphy. "Path-Based Spectral Clustering: Guarantees, Robustness to Outliers, and Fast Algorithms". ArXic Preprint. 2017

Numerical Implementation

- Recall computation appears hard, since space of paths is large.
- We propose an efficient *approximation* scheme, quasilinear in n .
- Computing the first K eigenvectors of the LLPD Laplacian is

$$O(n(k_1 C_{\mathrm{NN}} + m(k_1 \vee \log(n) \vee K^2)))$$

  - k_1 is the number of neighbors in original graph.
  - m is related to accuracy of approximation.
  - C_{NN} is the cost of a Euclidean nearest neighbor query.

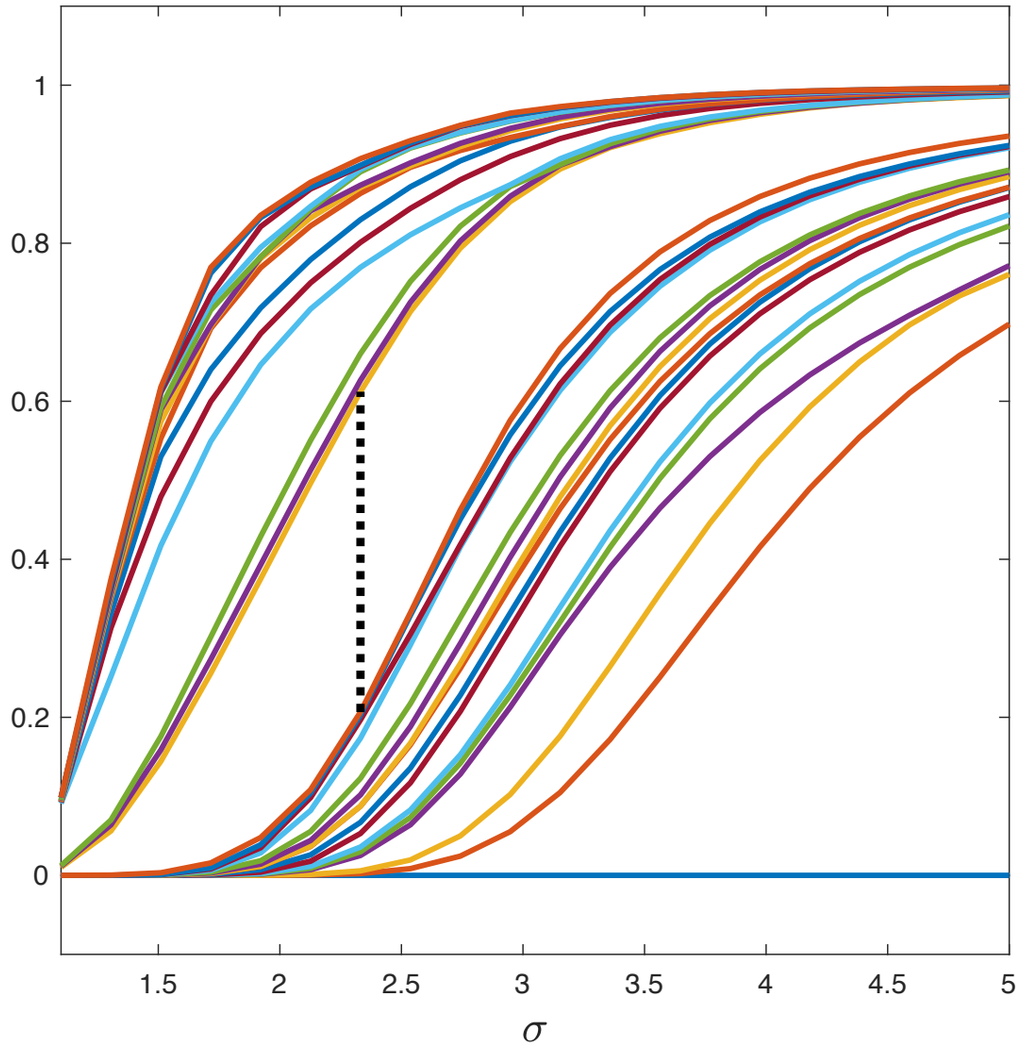
$$O(C^d D n \log(n))$$

Big data regime ($n = 10^8$ takes a few minutes!)

Columbia Object Image Library (COIL)

COIL 16 Classes

Multiscale Eigenvalues for LLPD SC



- 16 classes, ambient dimensionality 1024, about 100 samples per class.

- LLPD spectral clustering achieve 99+% accuracy, and correctly identifies that there are 16 classes.

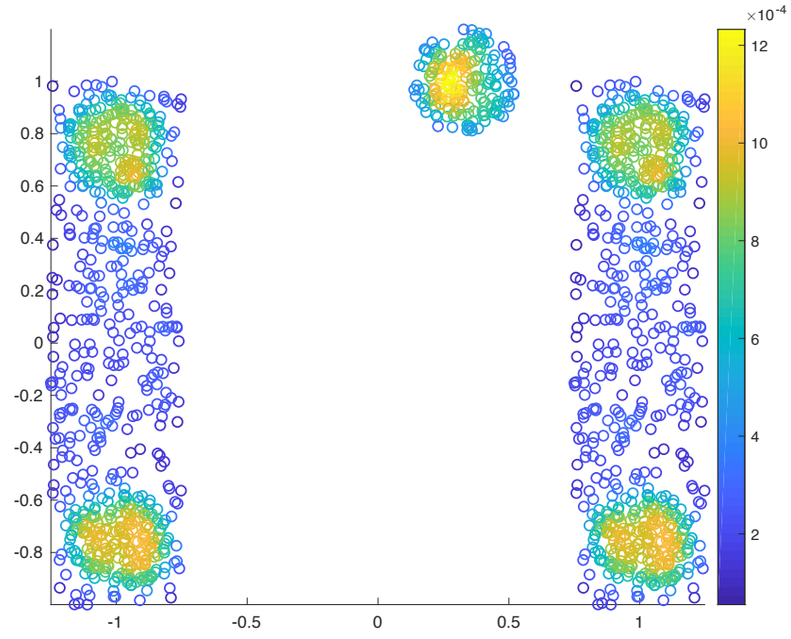
Tufts UNIVERSITY

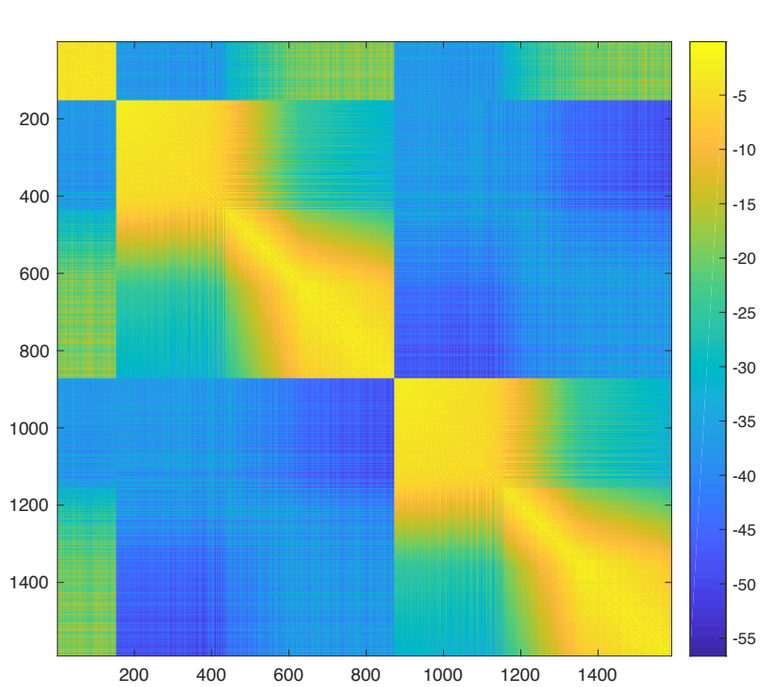
Incorporating Nonlinear Geometry

Learn nonlinear geometry with a diffusion process

$$P_{ij} = \frac{W_{ij}}{\sum_{\ell=1}^{n} W_{i\ell}}$$

$$W_{ij} = \begin{cases} e^{-\frac{\|x_i - x_j\|_2^2}{\sigma}}, & x_i \in NN_k(x_j), \\ 0, & \text{else.} \end{cases}$$





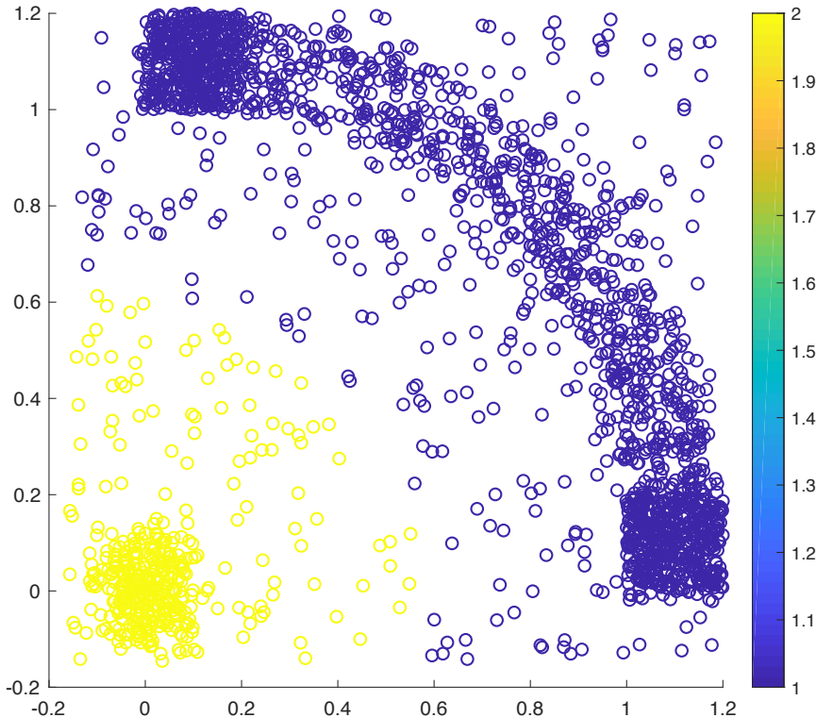
Diffusion Distances:

$$d_t(x_i, x_j) = \sum_{\ell=1}^{n} (P_{i\ell}^t - P_{j\ell}^t)^2 \frac{\mu_\ell}{\pi_\ell}$$

Spectral Formulation

$$d_t(x_i, x_j) = \sum_{\ell=1}^{n} \lambda_\ell^{2t} (\Phi_\ell(i) - \Phi_\ell(j))^2$$

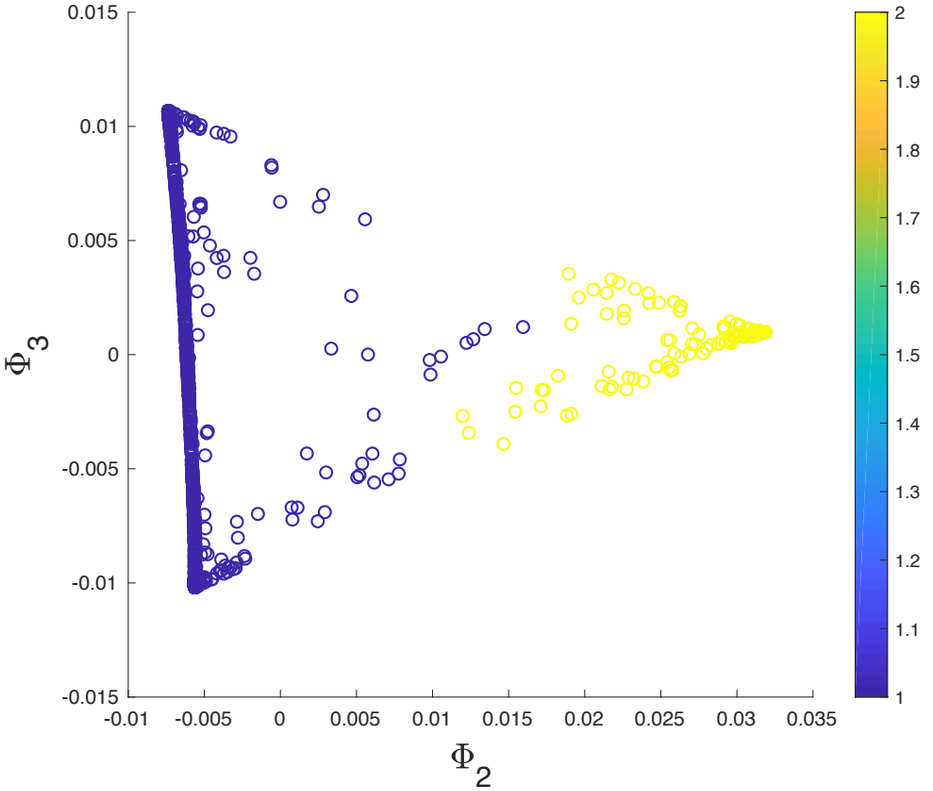
$\{(\lambda_i, \Phi_i)\}_{i=1}^n$ Spectral decomposition of P



Convert to

(Φ_2, Φ_3)

Coordinates



Tufts UNIVERSITY

Learning by Unsupervised Nonlinear Diffusion (LUND)

1.) **Compute empirical density:**

$$p_0(x_i) = \sum_{x_j \in NN_k(x_i)} e^{\frac{-\|x_i - x_j\|_2^2}{\sigma^2}}$$

$$p(x_i) = p_0(x_i) / \sum_{j=1}^{n} p_0(x_j)$$

2.) **Find points that are d_t -far from higher density points:**

$$\tilde{\rho}_t(x_i) = \begin{cases} \min\limits_{\{p(x_j) \geq p(x_i)\}} d_t(x_i, x_j), & x_i \neq \arg\max\limits_i p(x_i), \\ \max\limits_{x_j} d_t(x_i, x_j), & x_i = \arg\max\limits_i p(x_i). \end{cases}$$

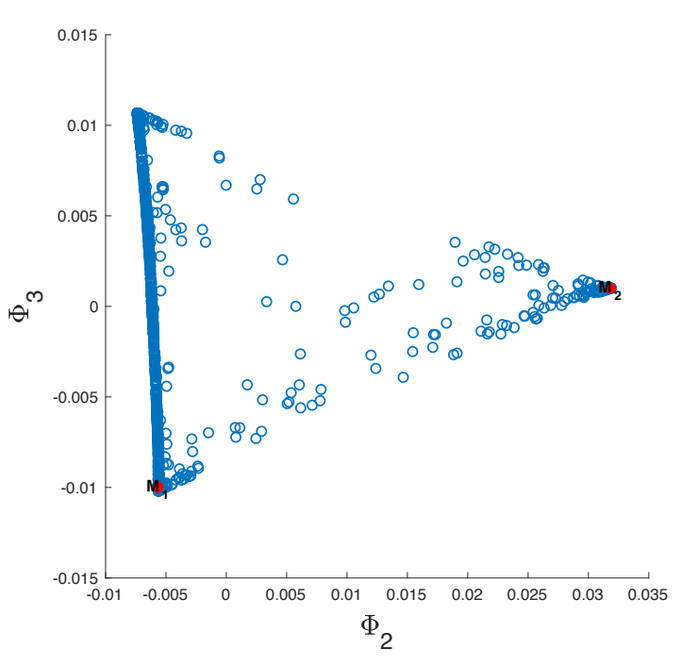
$$\rho_t(x_i) = \tilde{\rho}_t(x_i) / \max_{x_j} \tilde{\rho}_t(x_j)$$

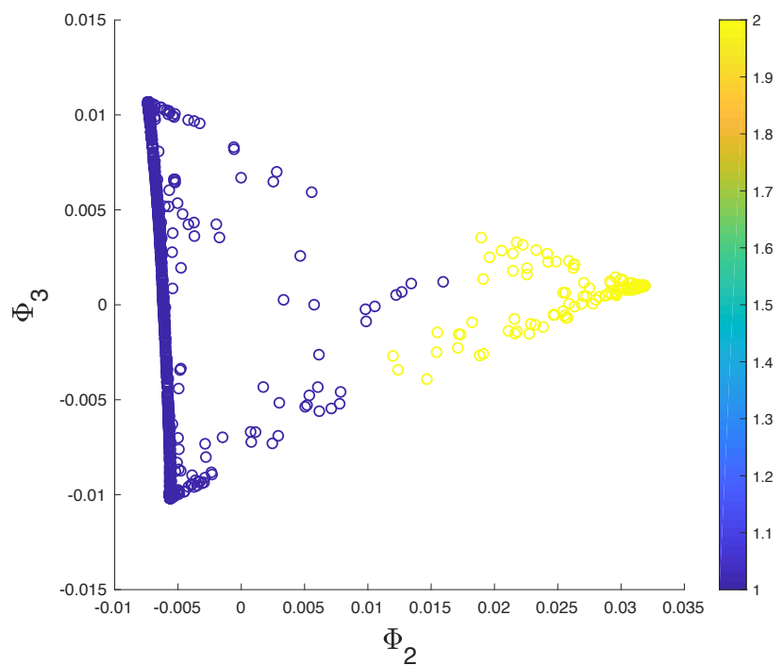
3.) **Estimate modes as maximizers of:**

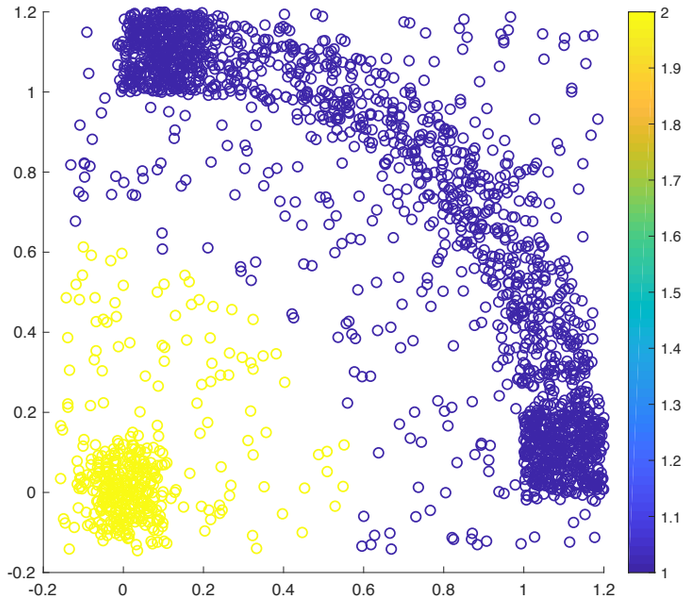
$$\boxed{\mathcal{D}_t(x_i) = p(x_i)\rho_t(x_i)}$$

Learning by Unsupervised Nonlinear Diffusion (LUND)

Assign all points the same label as their d_t-nearest neighbor of higher density.







With fast nearest-neighbor look-ups, complexity is $O(n \log(n) D C^d)$

D — ambient dimension

d — intrinsic dimension

n — number of data points

Tufts UNIVERSITY

Mathematical Guarantees

Let $X = \bigcup_{k=1}^{K} X_k$ be the latent clusters in the data.

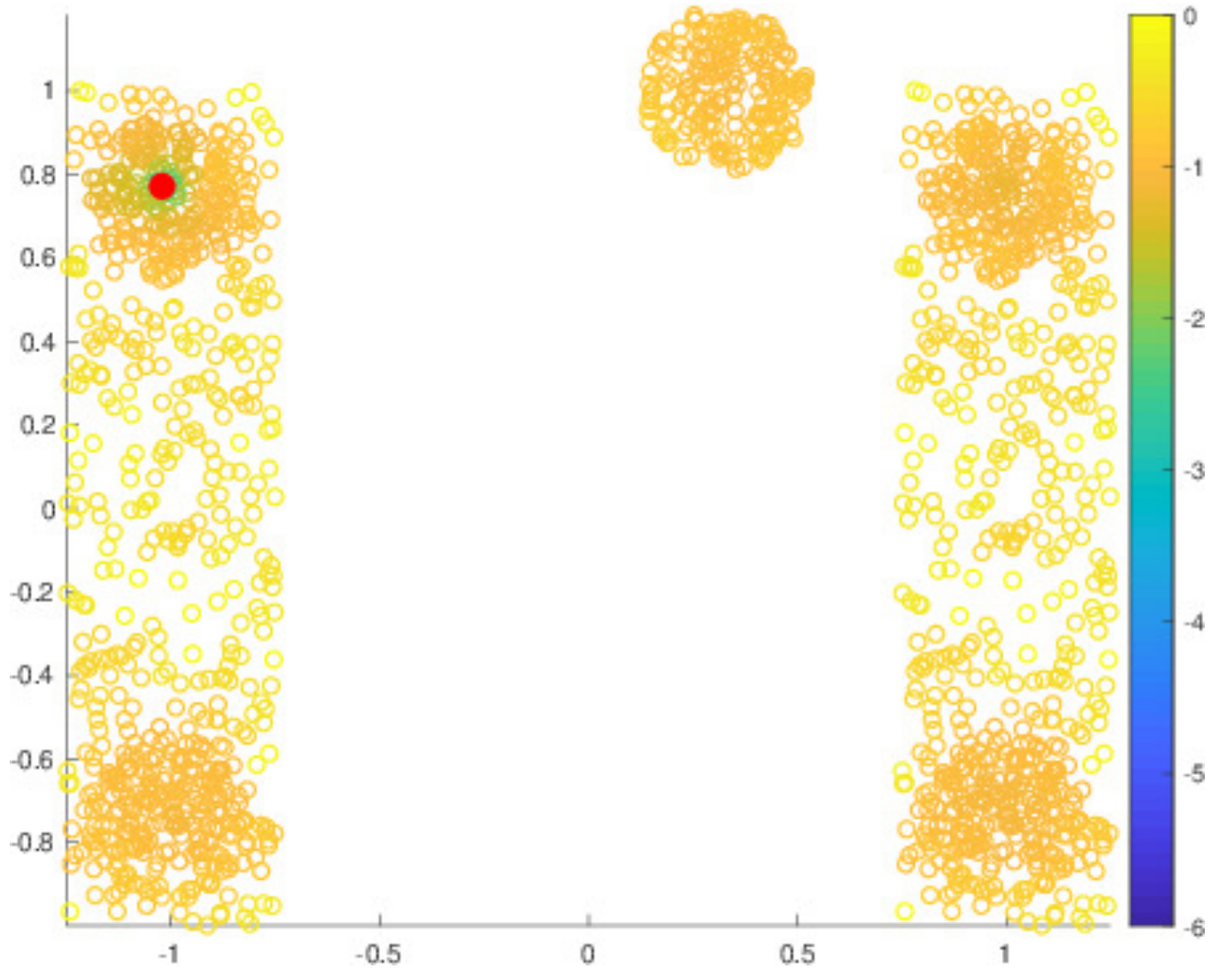
$$D_{\text{in}} = \max_{k} \max_{x,y \in X_k} d_t(x,y), \quad D_{\text{btw}} = \min_{k \neq k'} \min_{x \in X_k, y \in X_{k'}} d_t(x,y)\,.$$

Theorem. *(Maggioni, **M.**) Let $X = \bigcup_{k=1}^{K} X_k$ and let $\mathbf{P}$ be a corresponding Markov transition matrix on X, inducing diffusion distances $\{D_t\}_{t \geq 0}$. Then there exist constants $\{C_i\}_{i=1}^{5} \geq 0$ such that the following holds: for any $\epsilon > 0$, and for any t satisfying $C_1 \ln\left(\frac{C_2}{\epsilon}\right) < t < C_3 \epsilon$, we have*

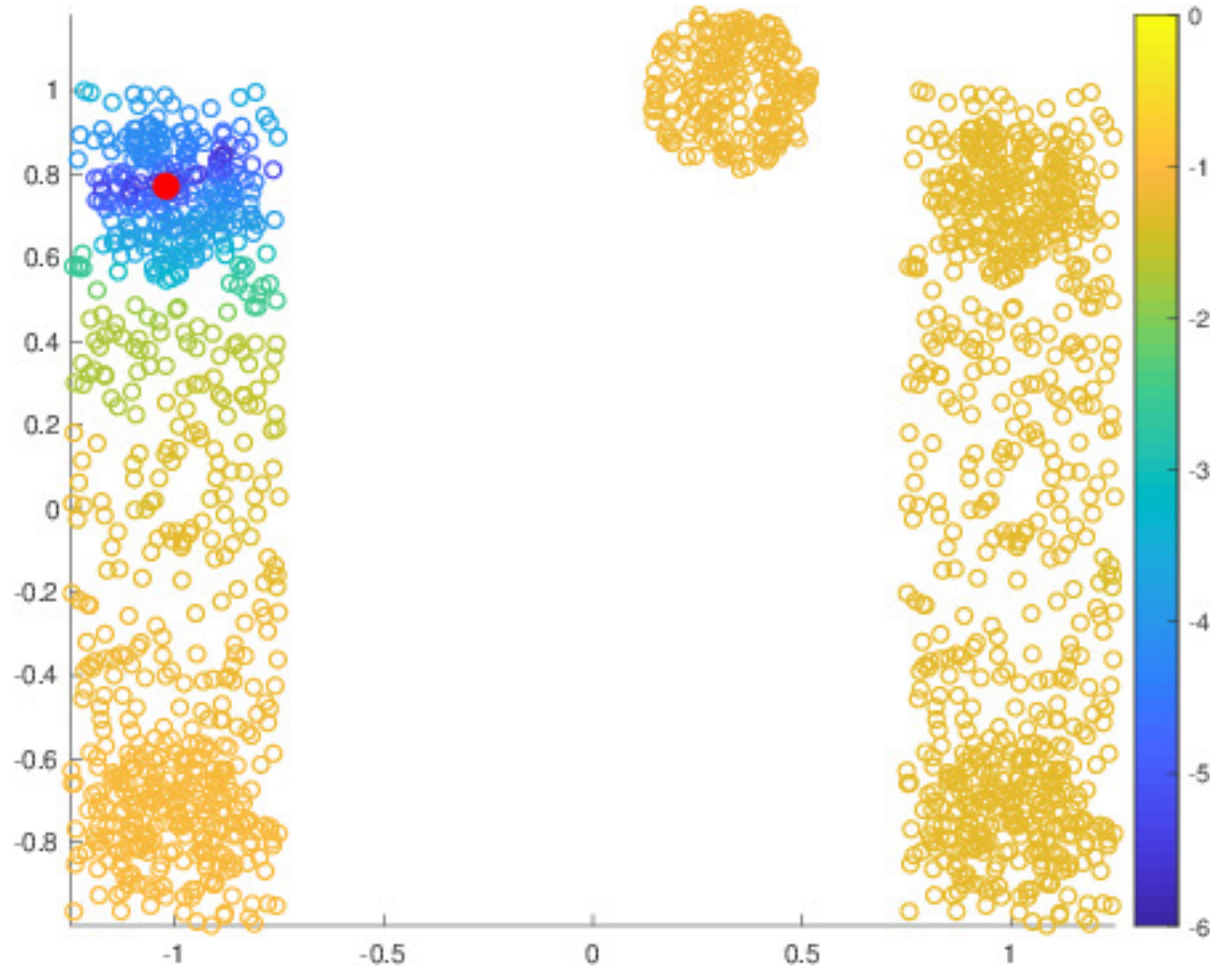
$$D_t^{in} \leq C_4 \epsilon, \quad D_t^{btw} \geq C_5 - C_4 \epsilon.$$

The constants $\{C_i\}_{i=1}^{5}$ depend on the data. More separation between clusters and cohesion within cluster lead to better constants.

Multiscale Equilibria I



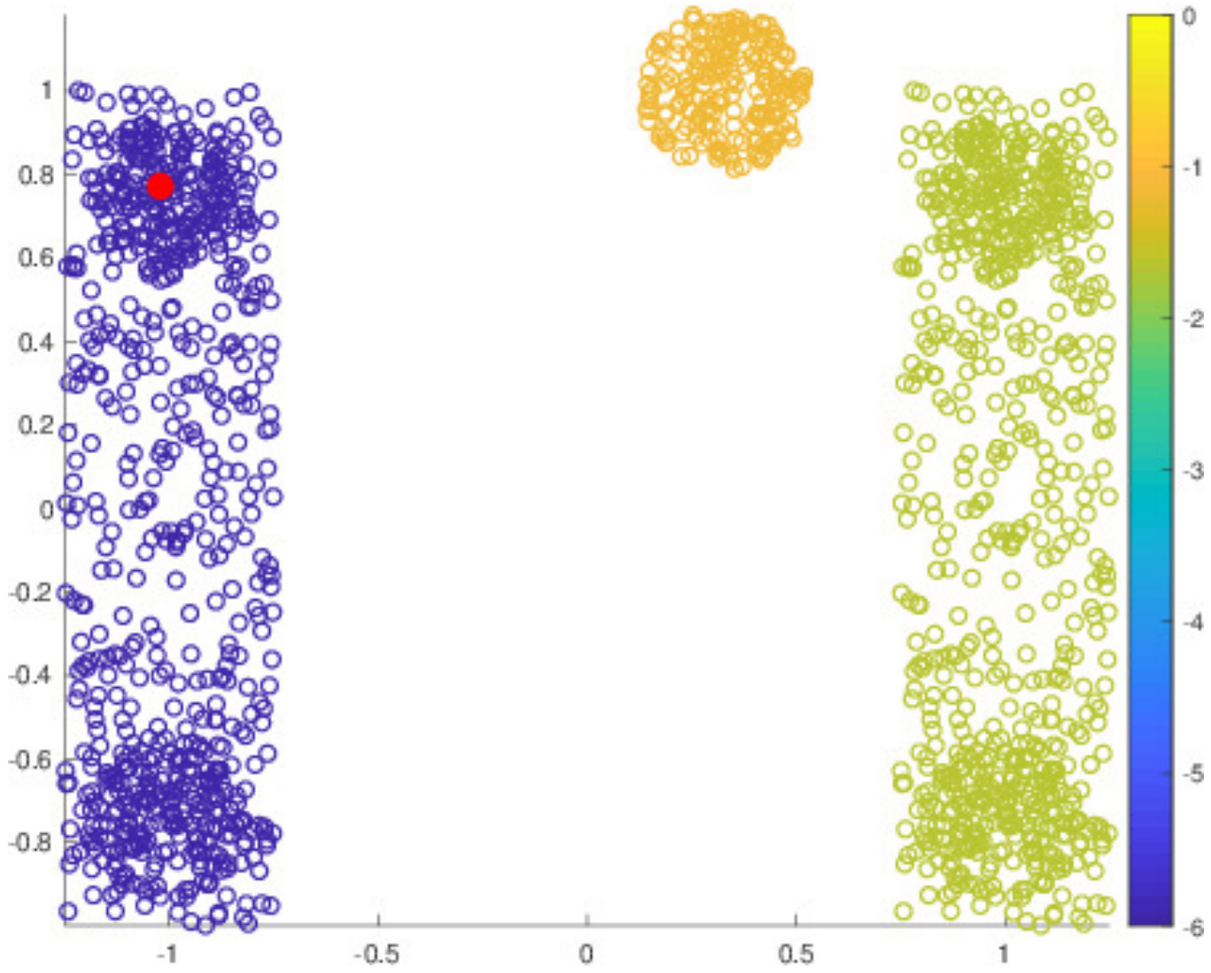
$t = 0$



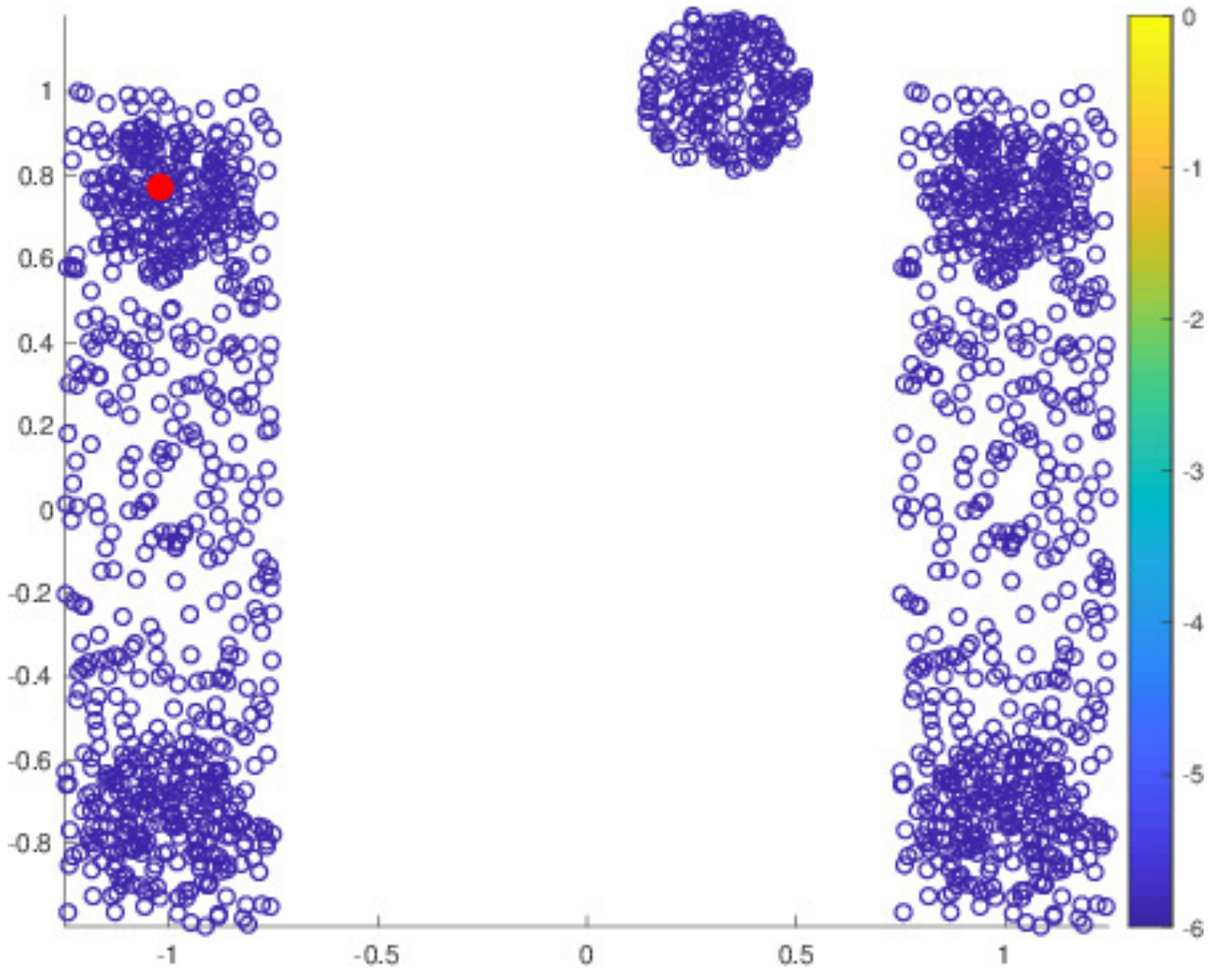
$t = 10^2$

Diffusion distances from red point in log scale:
small times lead to local mixing.

Multiscale Equilibria II



$t = 10^8$



$t = 10^{16}$

Diffusion distances from red point in log scale: as time increases, mesoscopic equilibria, then global equilibrium is reached.

Tufts UNIVERSITY

Mathematical Guarantees

$$M = \{p(x) \mid \exists k \text{ such that } x = \text{argmax}_{y \in X_k} p(y)\}$$

Theorem. *(Maggioni, **M.**) Suppose* $X = \bigcup_{k=1}^{K} X_k$ *as above. The proposed algorithm labels all points accurately, and correctly estimates* K*, provided that*

$$\frac{D_{in}}{D_{btw}} < \frac{\min(M)}{\max(M)}.$$

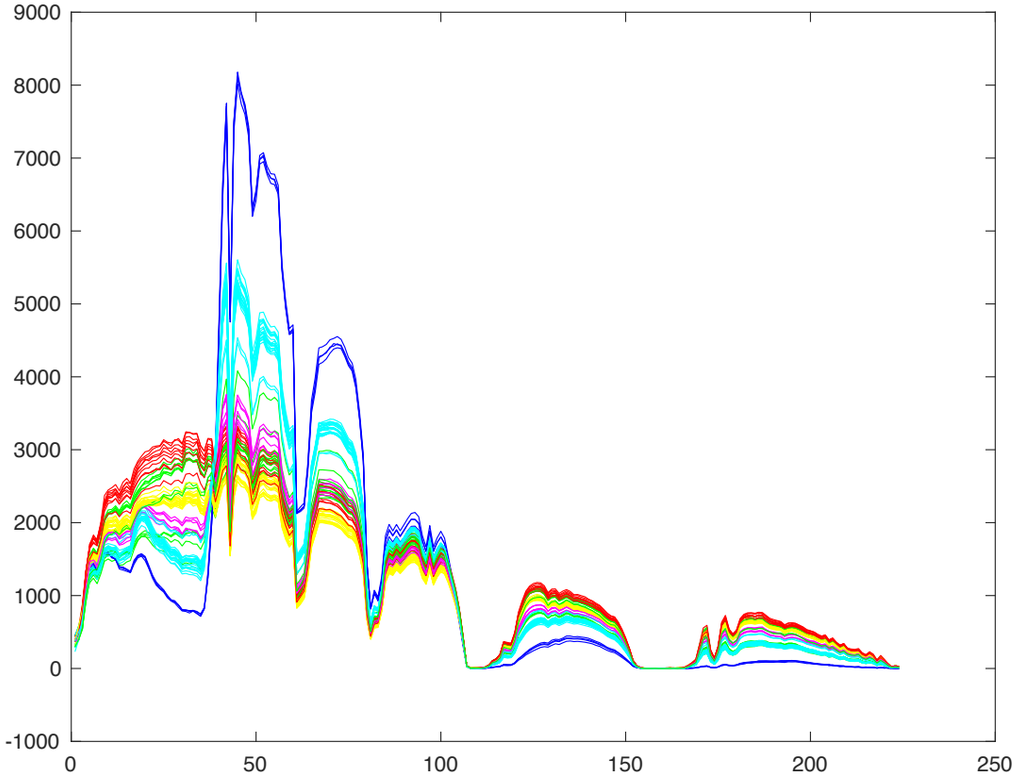
The more well-separated and internally cohesive the clusters are, the greater time range in which accuracy is assured.

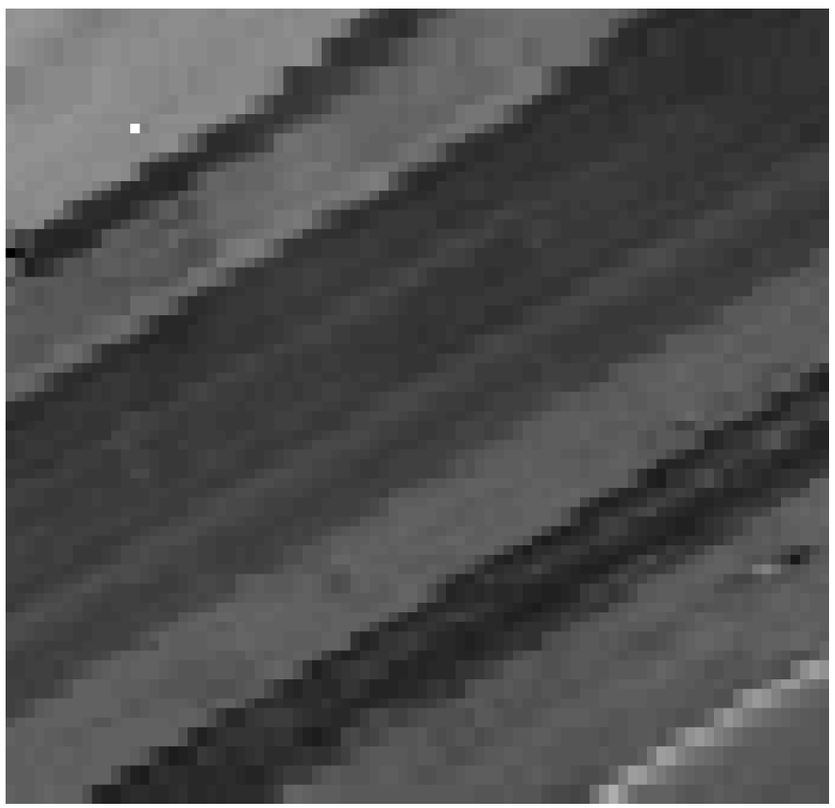
Proofs base on analysis of Markov matrices in relationship to near reducibility and mixing times.

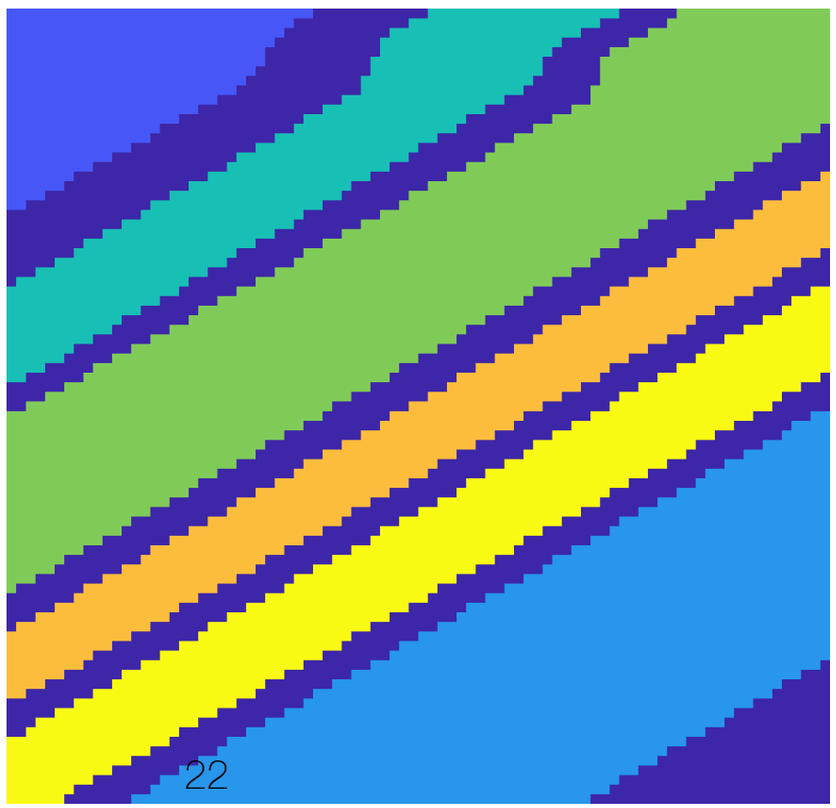
HSI Clustering

Salinas A HSI: $D = 220,\ n = 7138$

Six classes, substantial within-class variation

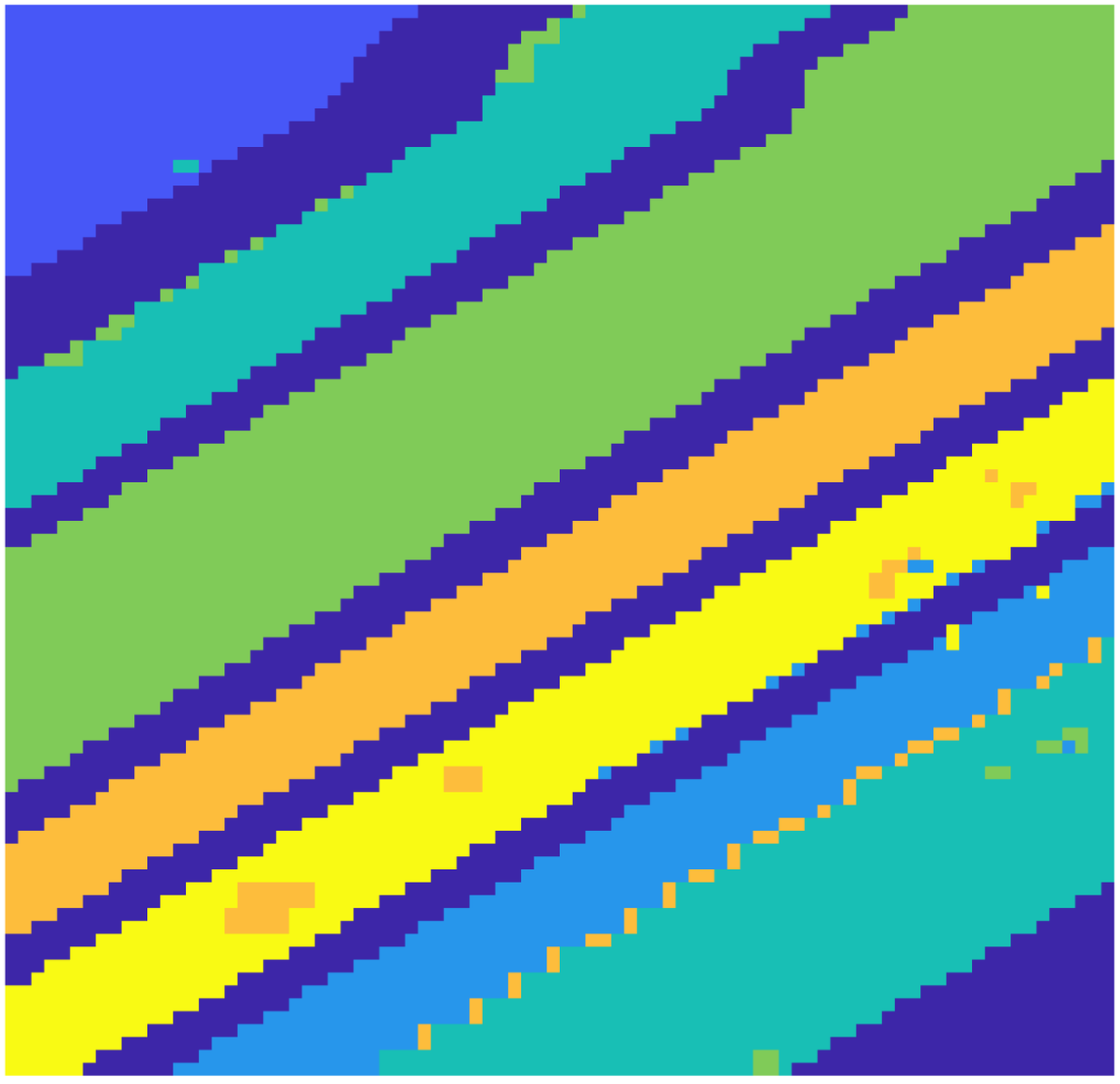




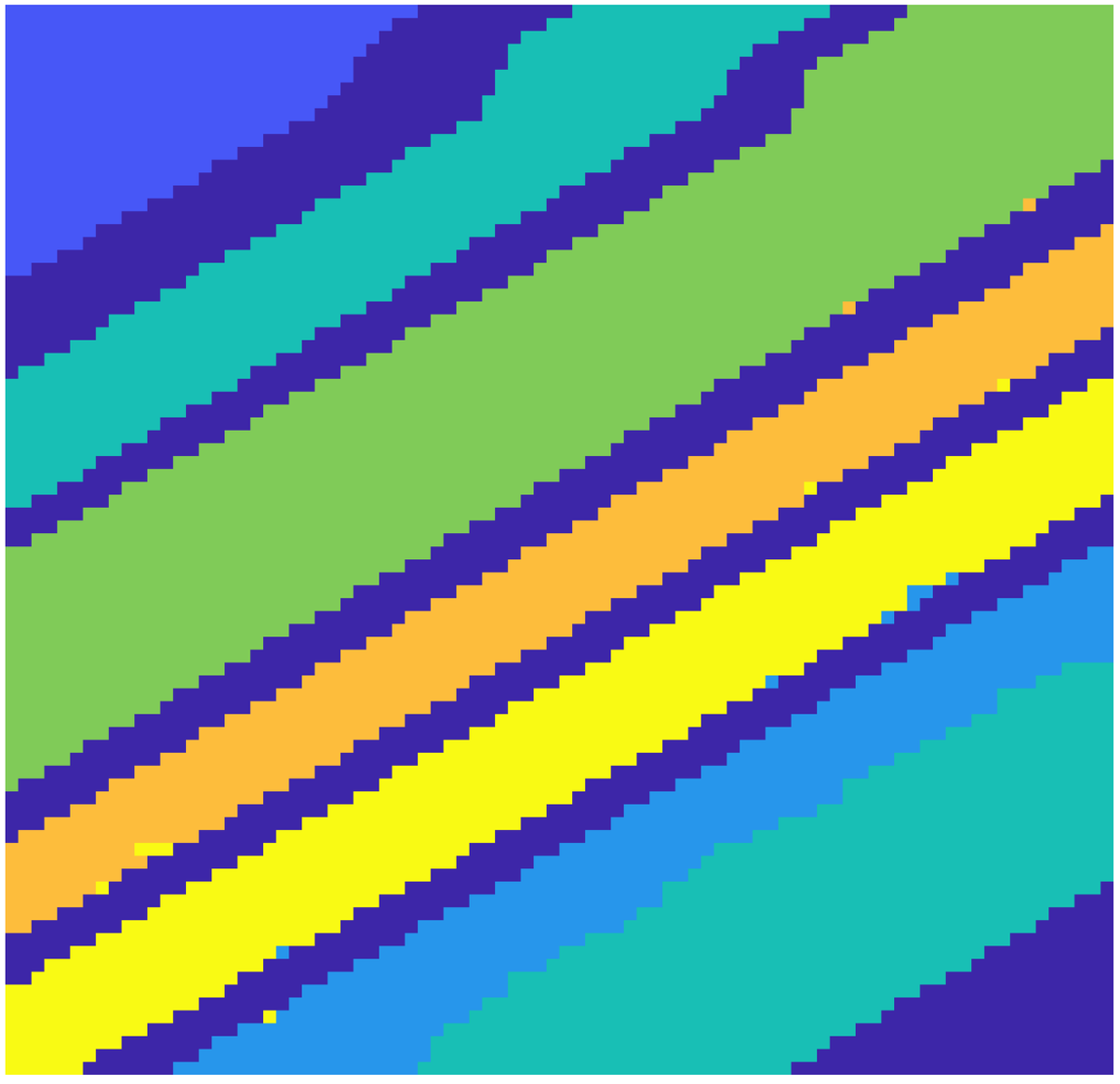


Tufts UNIVERSITY

Spatial Regularization



LUND

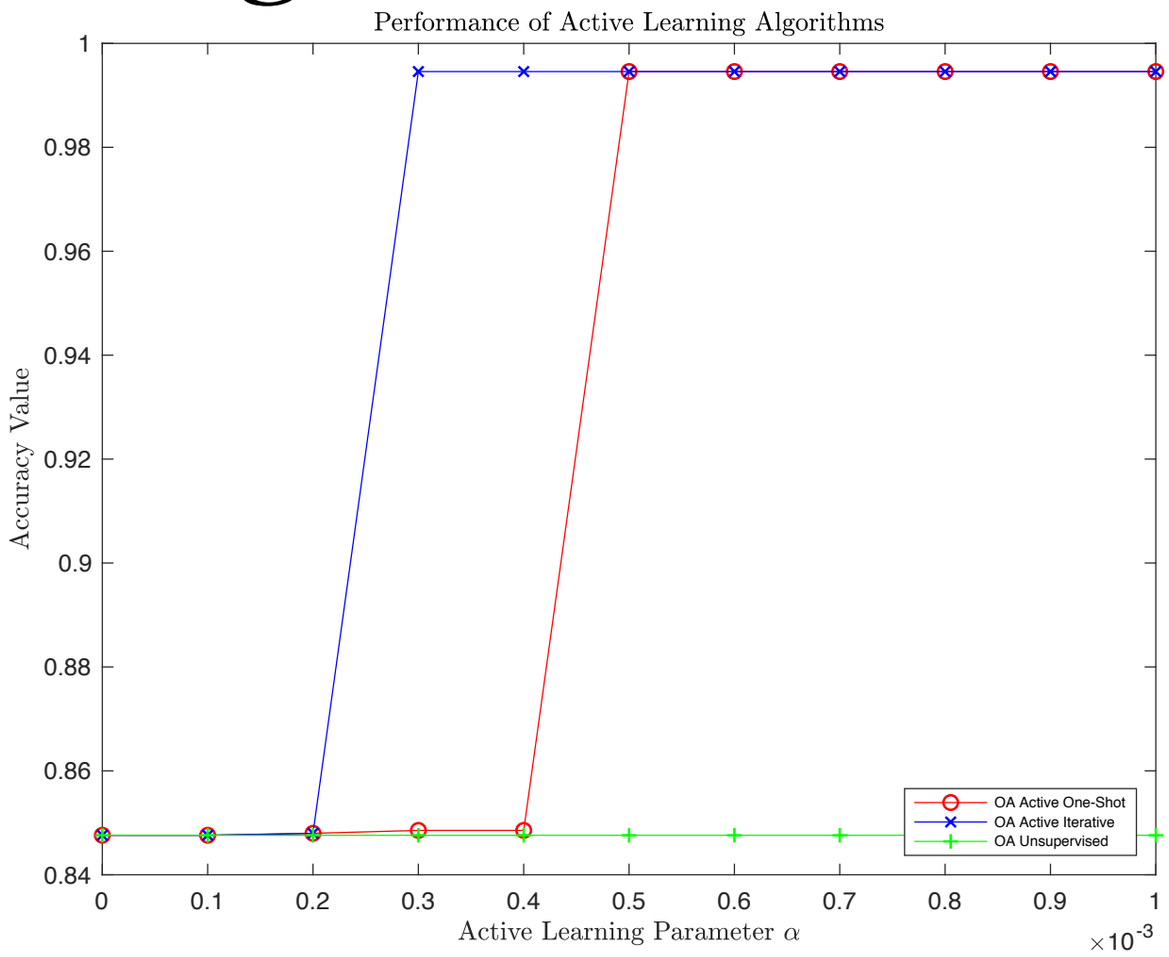


LUND+Spatial Regularization

Active Learning

Active learning: We can ask for O(1) labels…who to query?

Let $x^*_{n_1}, x^*_{n_2}$ be the two labeled points d_t-nearest to x_n.



Query for labels the minimizers of

$$F_t(x_n) = |d_t(x_n, x^*_{n_1}) - d_t(x_n, x^*_{n_2})|$$

Adding $O(1)$ training labels to perfect accuracy!

Tufts UNIVERSITY

Summary and Broad Future Directions

- The *low-dimensional, high noise* model allows to prove performance guarantees for clustering algorithms.

- *Efficient algorithms* for data-dependent metrics allow to handle large numbers of data points in high dimensions.

- Applicable to *real-world data*, including image datasets, remotely sensing signals,…

- Future:

- **Mathematics**: Investigate problems in discrete-to-continuum limits, multiscale hierarchies, metrics for graph construction, directed graphs…
- **Machine Learning:** Theoretical models for active learning, alternative diffusion constructions
- **Data Science and Interdisciplinary Collaboration:** Develop algorithms and software for application to real data: remote sensing, medical signals, social media networks…

References

- Little, A., M. Maggioni and J.M. Murphy. "Path-Based Spectral Clustering: Guarantees, Robustness to Outliers, and Fast Algorithms." arXiv:1712.06206. 2017.

- Maggioni, M. and J.M. Murphy. "Clustering by Unsupervised Geometric Learning of Modes." arXiv:1810.06702. 2018.

- Murphy, J.M. and M. Maggioni. "Unsupervised Clustering and Active Learning of Hyperspectral Images with Nonlinear Diffusion." *IEEE Transactions on Geoscience and Remote Sensing*. volume 57(3), p. 1829-145. 2019.

- Murphy, J.M., and M. Maggioni. "Spectral-Spatial Diffusion Geometry for Hyperspectral Image Clustering." arXiv:1902.05402. 2019.

- Murphy, J.M. and M. Maggioni. "Diffusion geometric methods for fusion of remotely sensed data." *SPIE Defense+Security*. 2018.

- Murphy, J.M. and M. Maggioni. "Iterative Active Learning with Diffusion Geometry for Hyperspectral Images." *IEEE Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS)*. 2018.

Code and Contact Information

Code: https://jmurphy.math.tufts.edu/Code/

Contact: jm.murphy@tufts.edu

Thanks for Your Attention!

Tufts UNIVERSITY