Density-Driven Path Metrics: Graphs, Manifolds, and Data

James M. Murphy
Department of Mathematics

September 8, 2022

Tufts
UNIVERSITY

Collaborators

A. Little, Utah

M. Maggioni, JHU

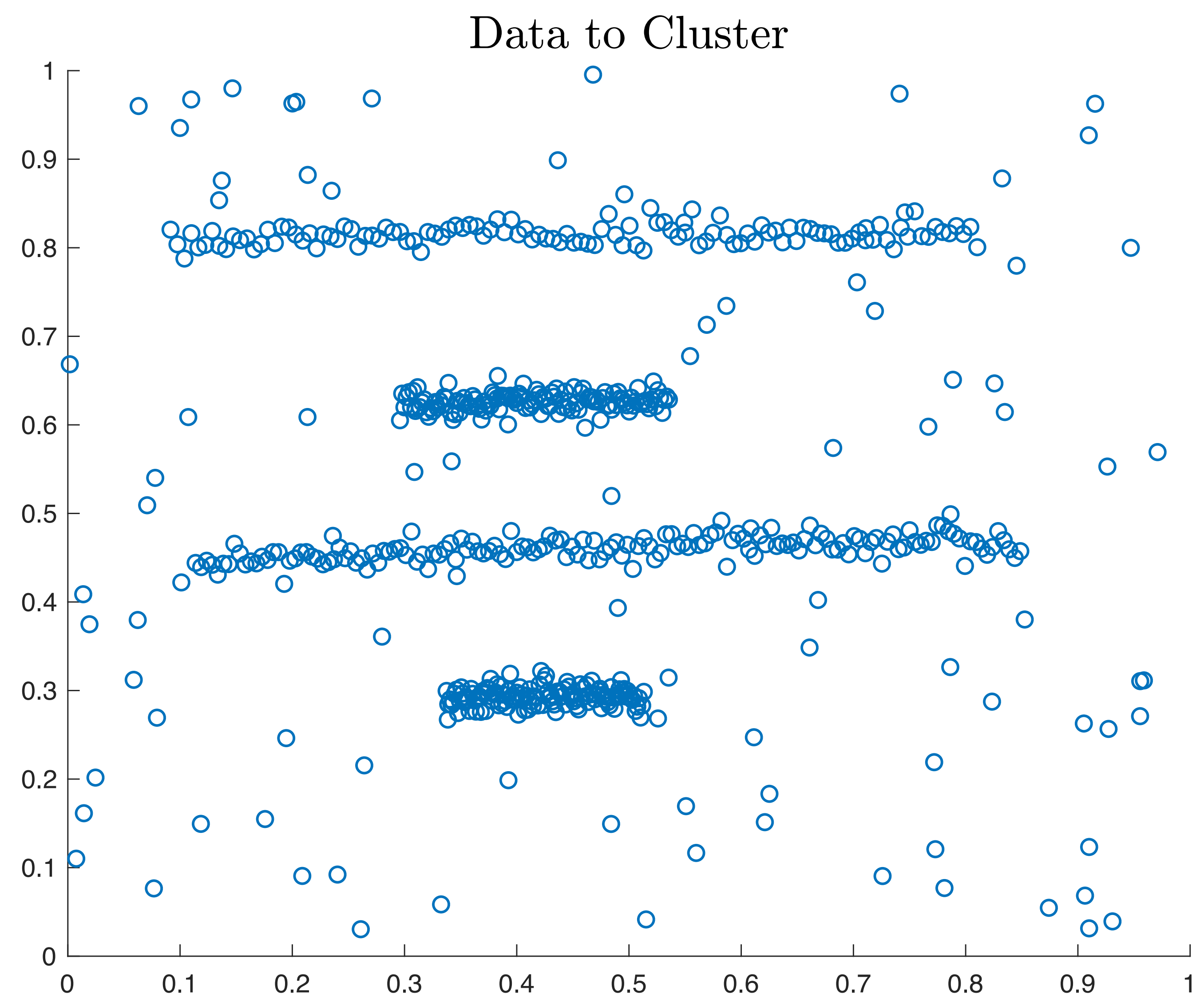
D. McKenzie, Mines

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.



$$x_1, \ldots, x_n \overset{i.i.d.}{\sim} \mu = \sum_{k=1}^{K} w_k \mu_k + w_0 \tilde{\mu}, \; \sum_{k=0}^{K} w_k = 1$$

Labeling: Which x_j were generated from μ_k?

Number of Clusters: Can we estimate K?

Tufts UNIVERSITY

Standard Method: K-Means

- **Idea**: find K centroids, then assign each point to its nearest centroid.
- Empirically good for same sized, spherical clusters.
- Guaranteed for certain Gaussians.
- Exact solution is NP-Hard to compute.
- Standard implementations involve non-convex optimization.
- Need to know K.

$$C^* = \underset{C=\{C_k\}_{k=1}^{K}}{\arg\min} \sum_{k=1}^{K} \sum_{x \in C_k} \|x - \bar{x}_k\|_2^2$$

Tufts UNIVERSITY

Standard Method: K-Means

- **Idea**: find K centroids, then assign each point to its nearest centroid.

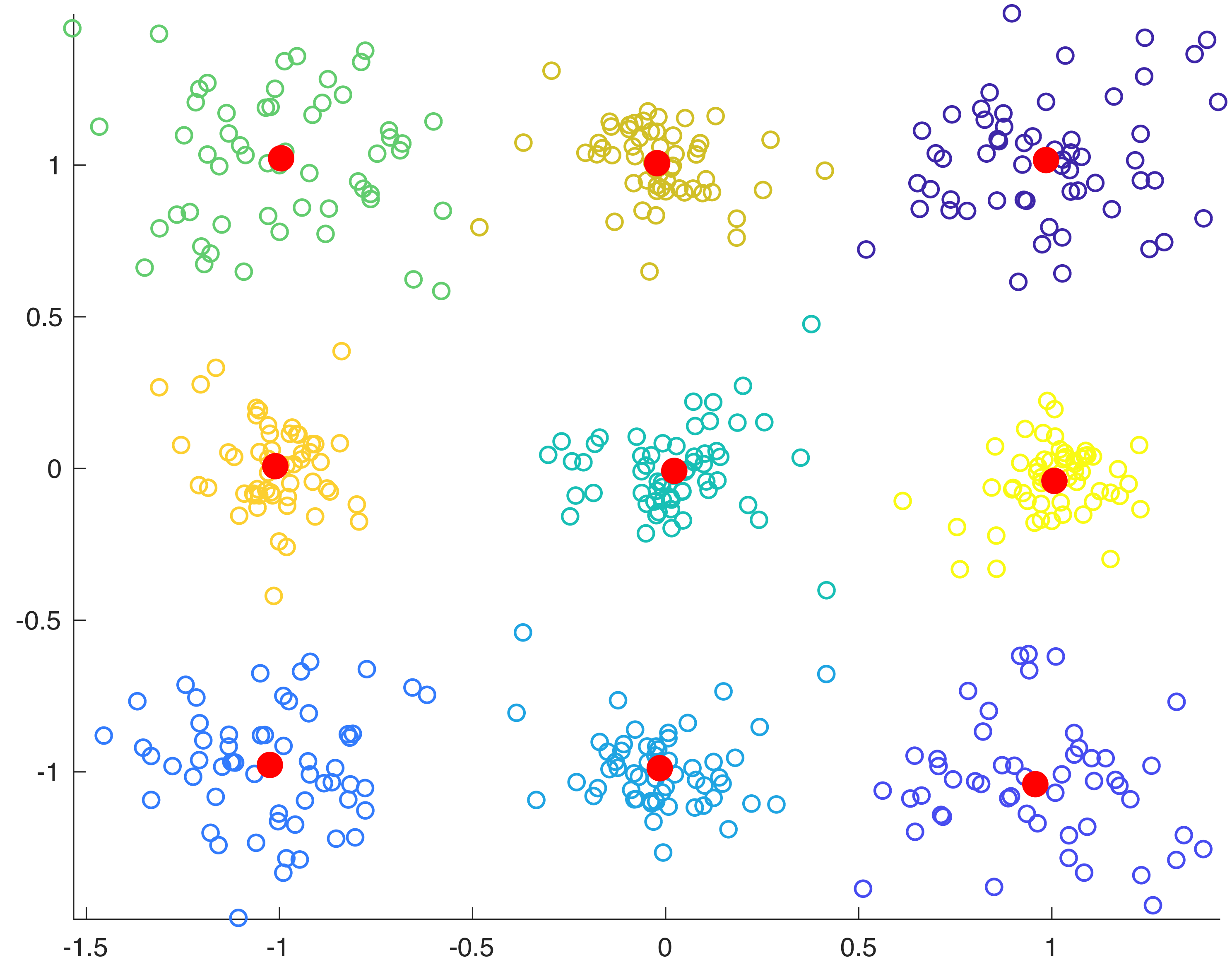
- Empirically good for same sized, spherical clusters.

- Guaranteed for certain Gaussians.

- Exact solution is NP-Hard to compute.

- Standard implementations involve non-convex optimization.

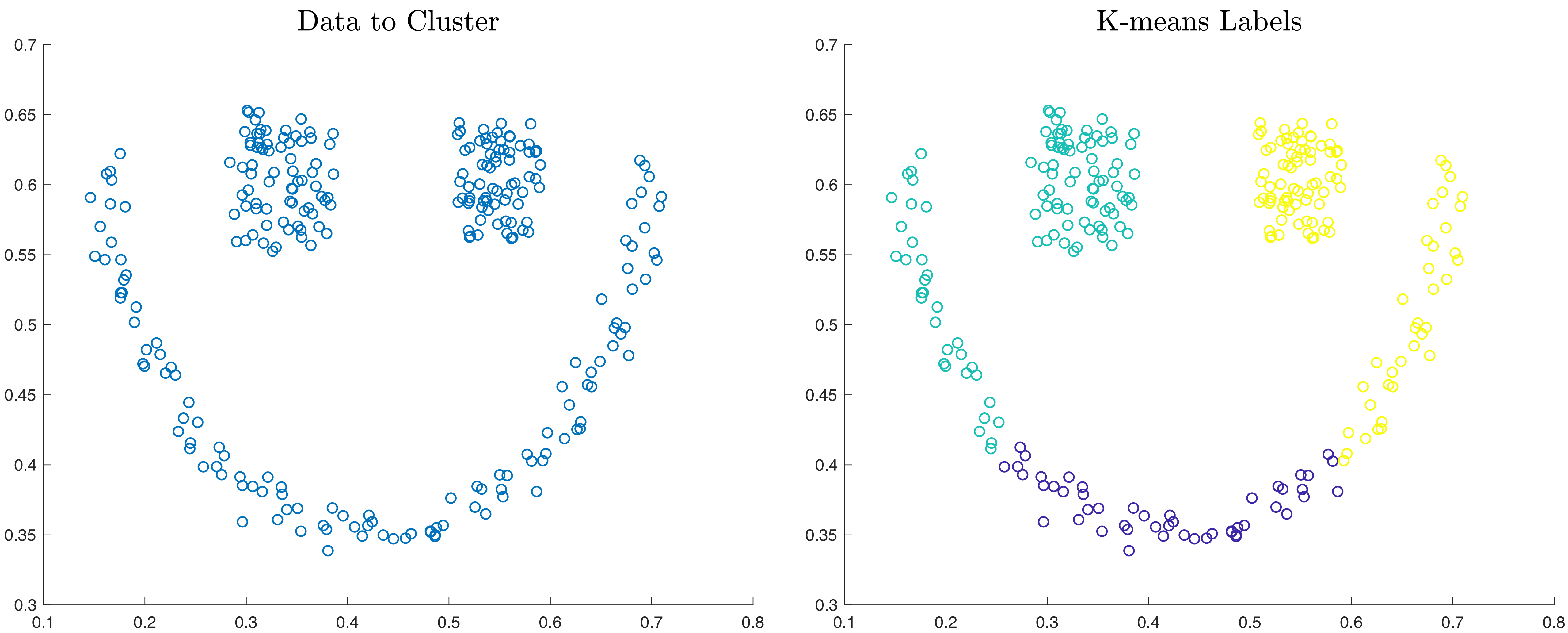
- Need to know K.



$$C^* = \underset{C=\{C_k\}_{k=1}^K}{\arg\min} \sum_{k=1}^{K} \sum_{x \in C_k} \|x - \bar{x}_k\|_2^2$$

Tufts UNIVERSITY

K-Means Often Fails



Problem: Some clusters are non-spherical!

Tufts UNIVERSITY

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*

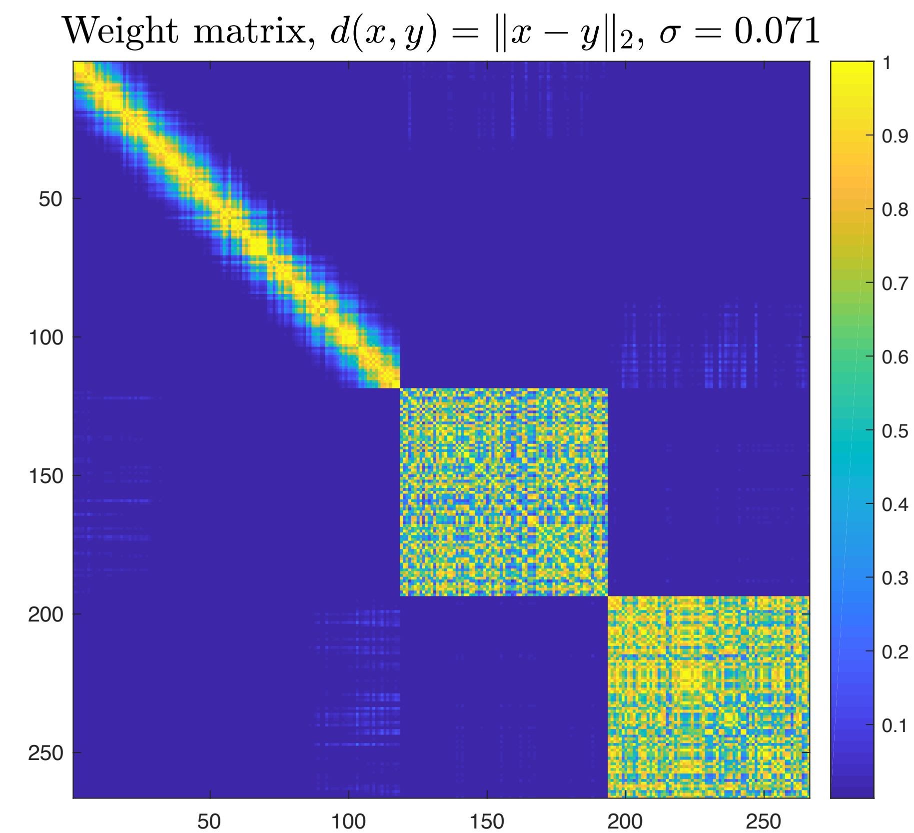
$$W_{ij} = e^{-d(x_i,x_j)^2/\sigma^2}$$

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

Step 2: Compute the *(graph) Laplacian*

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

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



Spectral Clustering II

Step 3: Compute eigenvalues of L

$$0 \leq \lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n$$

and associated eigenvectors

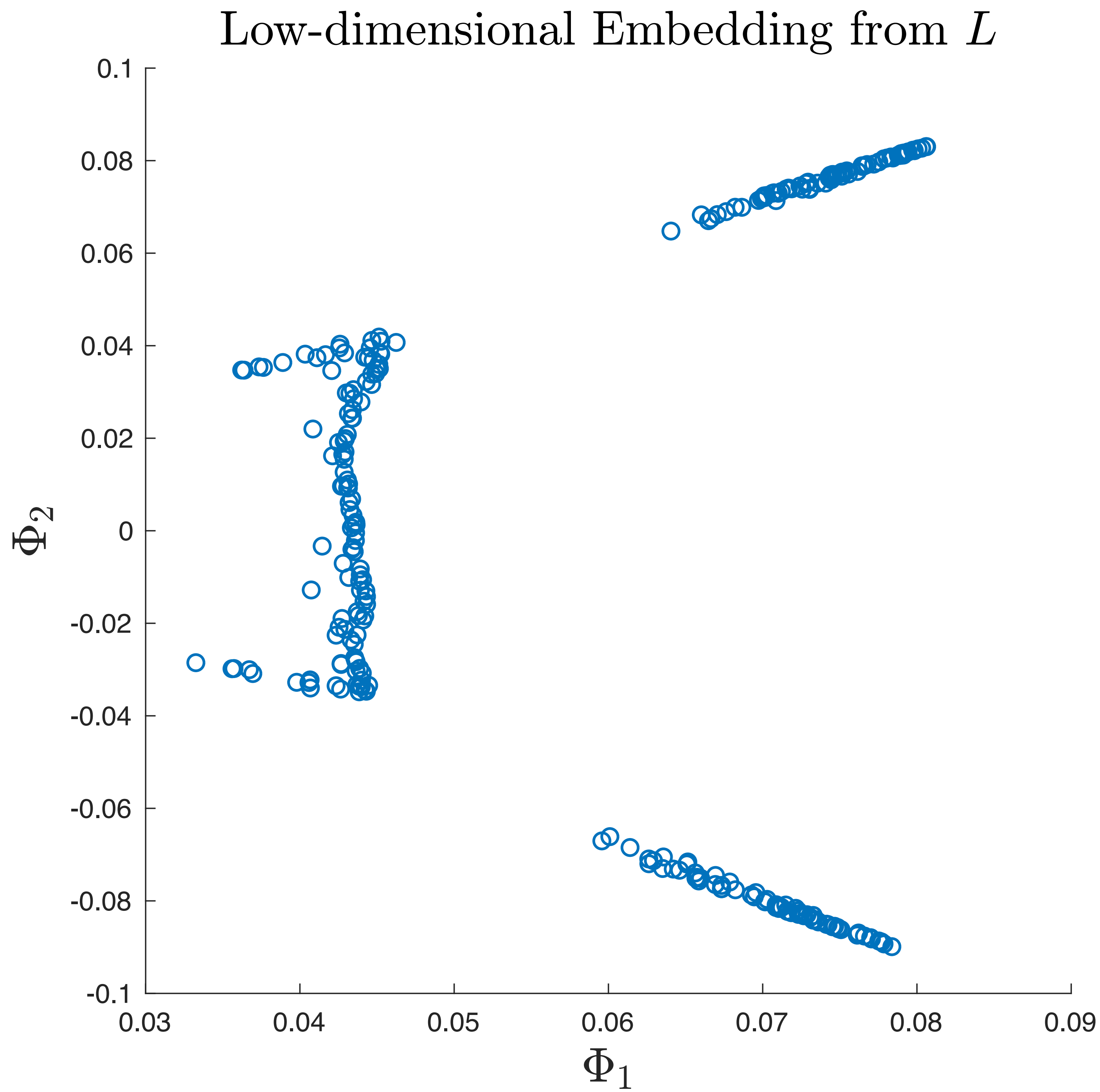
$$\Phi_1, \ldots, \Phi_n.$$

Step 4: Embed the data as

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

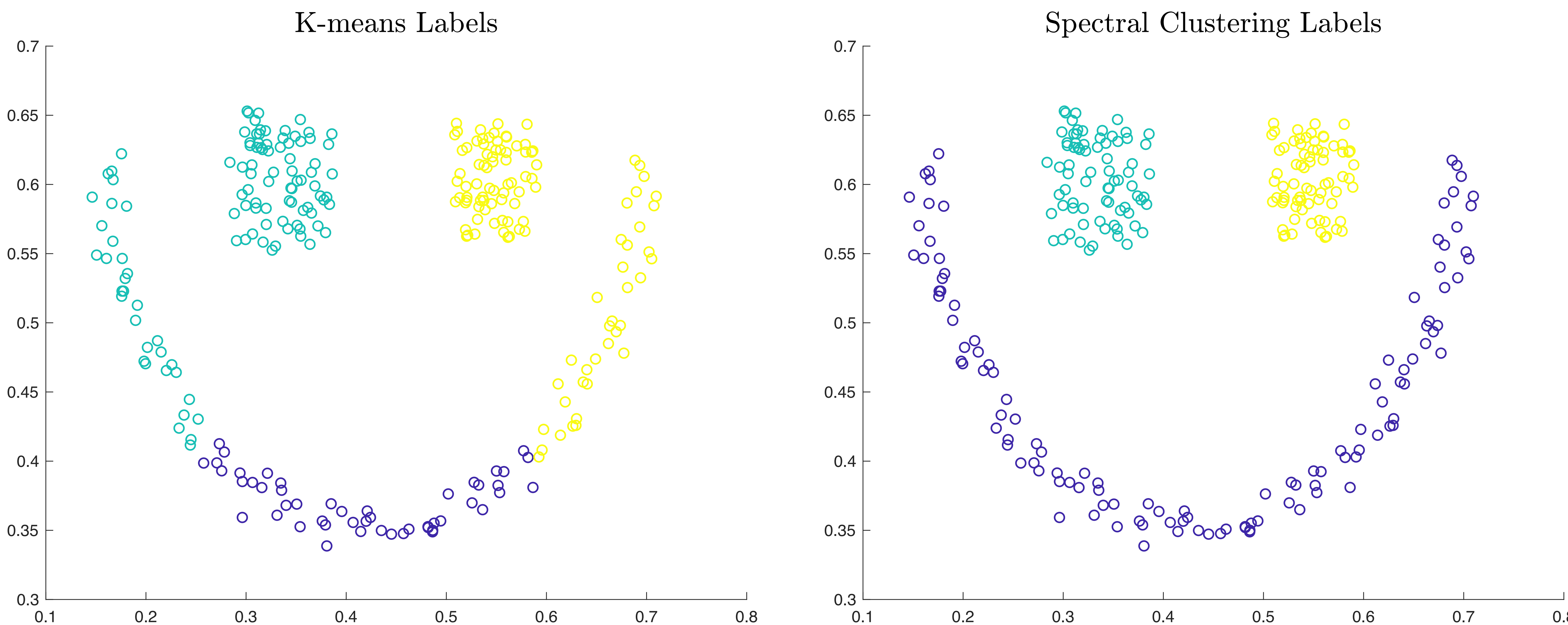
then run K-means. Note

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



Tufts UNIVERSITY

K-Means v. Spectral Clustering



- Spectral clustering (with a "good" σ) succeeds where K-means fails!

- Theoretical estimates are limited, particularly for estimating the number of clusters. Common heuristic: $K \approx \arg\max_k \lambda_{k+1} - \lambda_k$.

Tufts UNIVERSITY

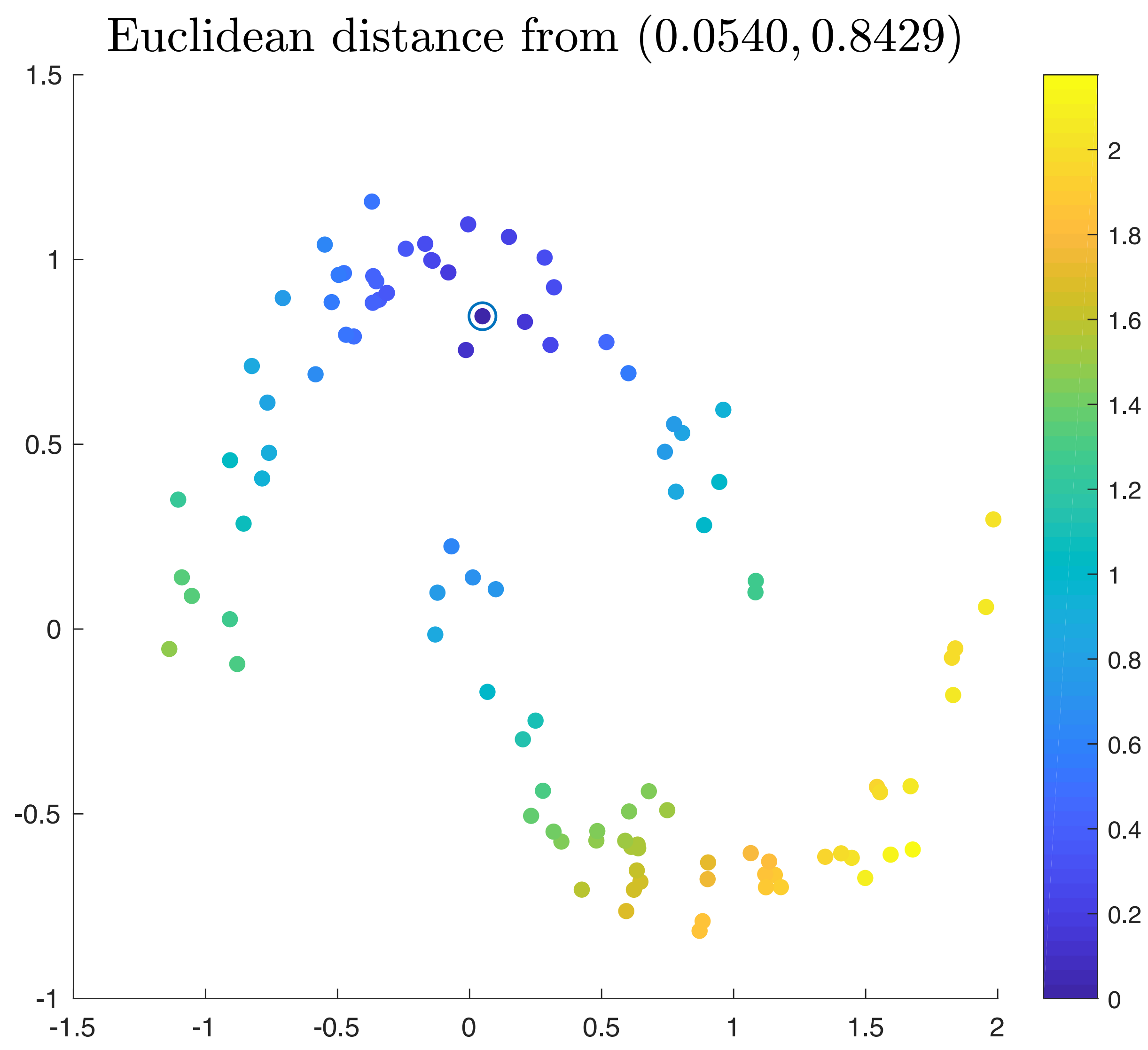
Data-Dependent LLPD Metric

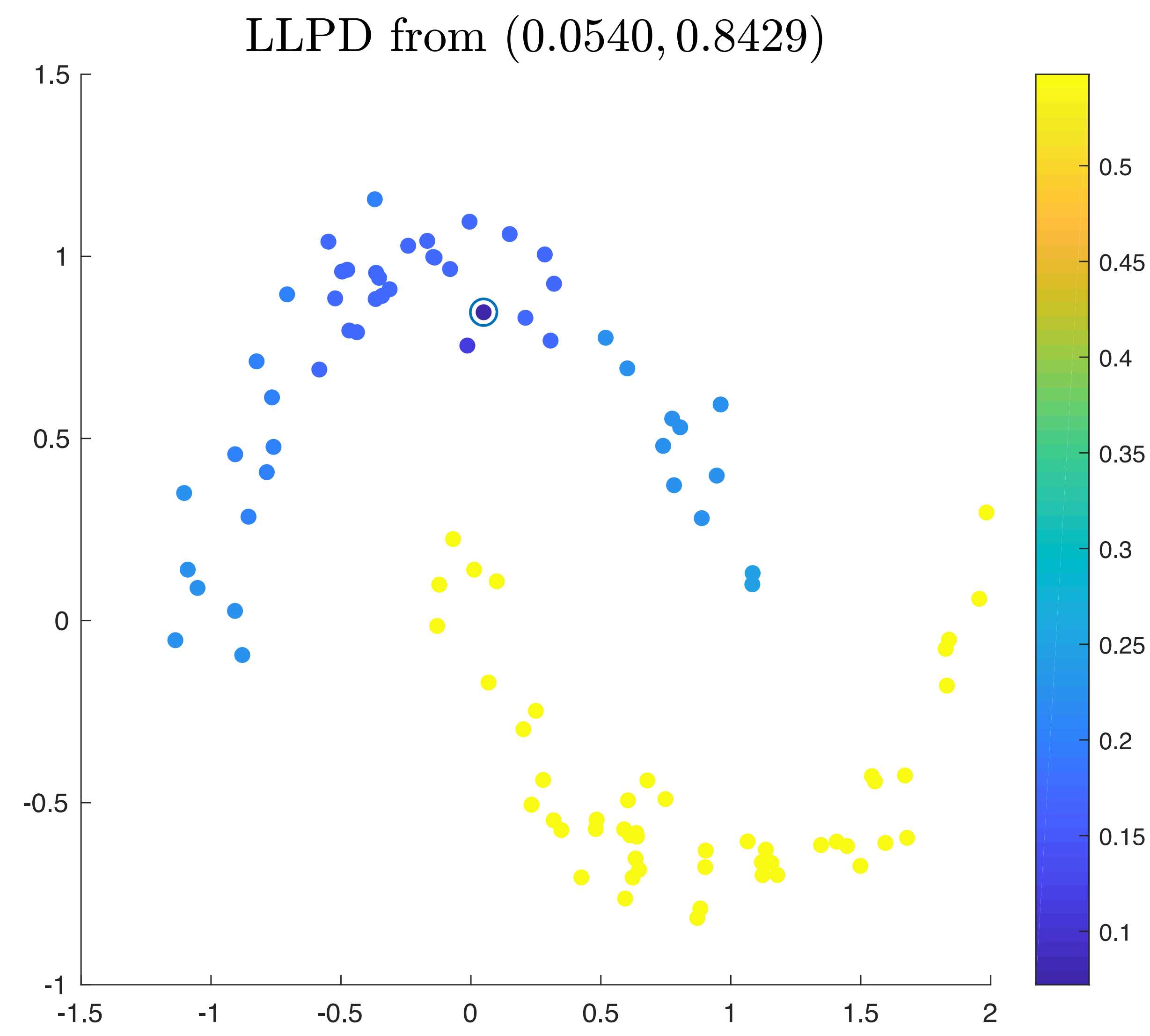
Definition. *For a discrete set* $X = \{x_i\}_{i=1}^n \subset \mathbb{R}^D$*, let* $\mathcal{G}$ *be the graph on* X *with edges given by the Euclidean distance between points. For* $x_i, x_s \in X$*, let* $\mathcal{P}(x_i, x_s)$ *denote the space of paths connecting* x_i, x_s *in* $\mathcal{G}$*. The*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

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

- 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!
- $\mathcal{G}$ could be a complete graph (all points connected to all points) or a connected NN graph.
- Ultrametric structure is compatible with fast matrix-vector multipliers.

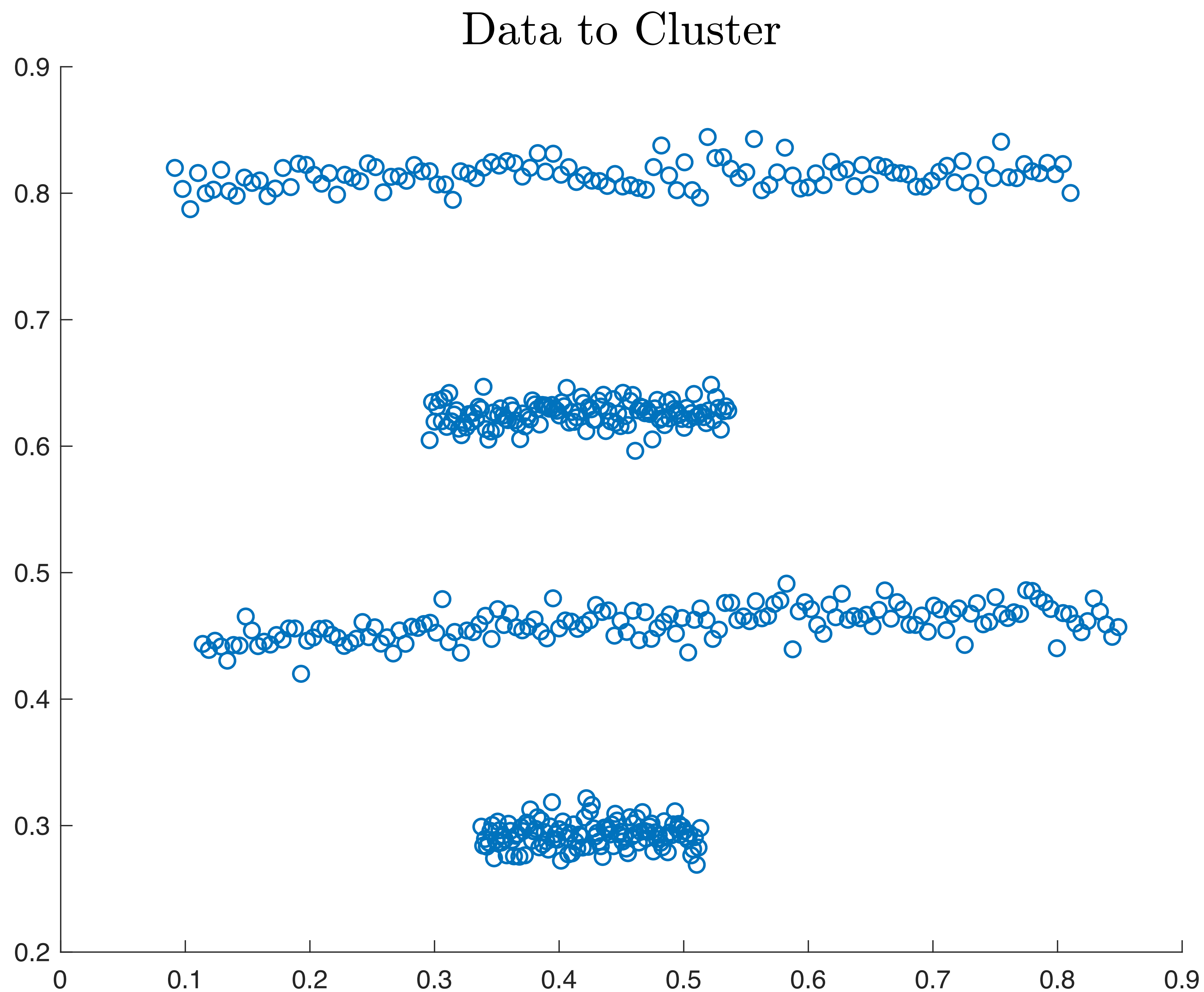
Euclidean Distance versus LLPD





Tufts UNIVERSITY

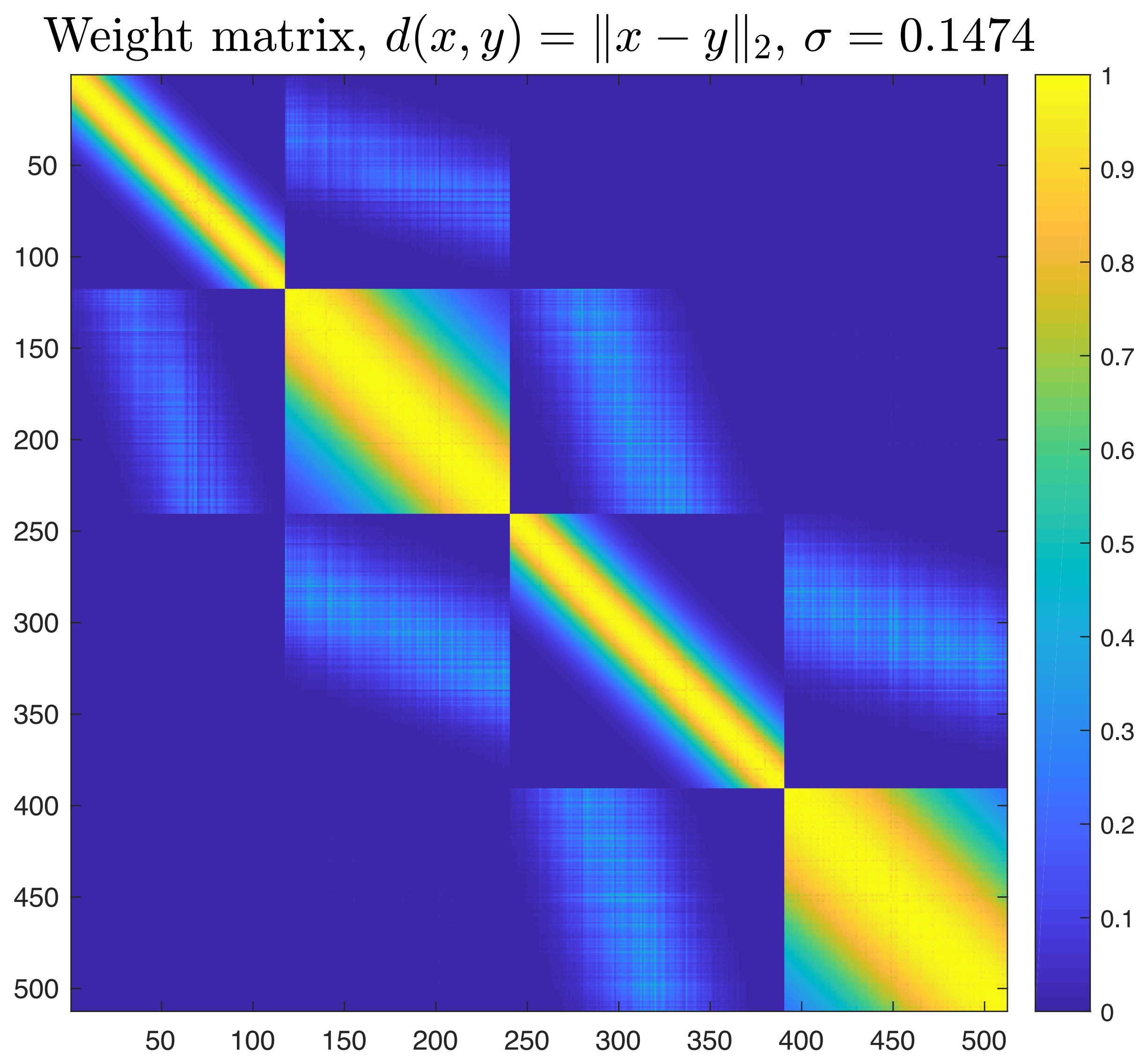
Data Well-Suited for LLPD

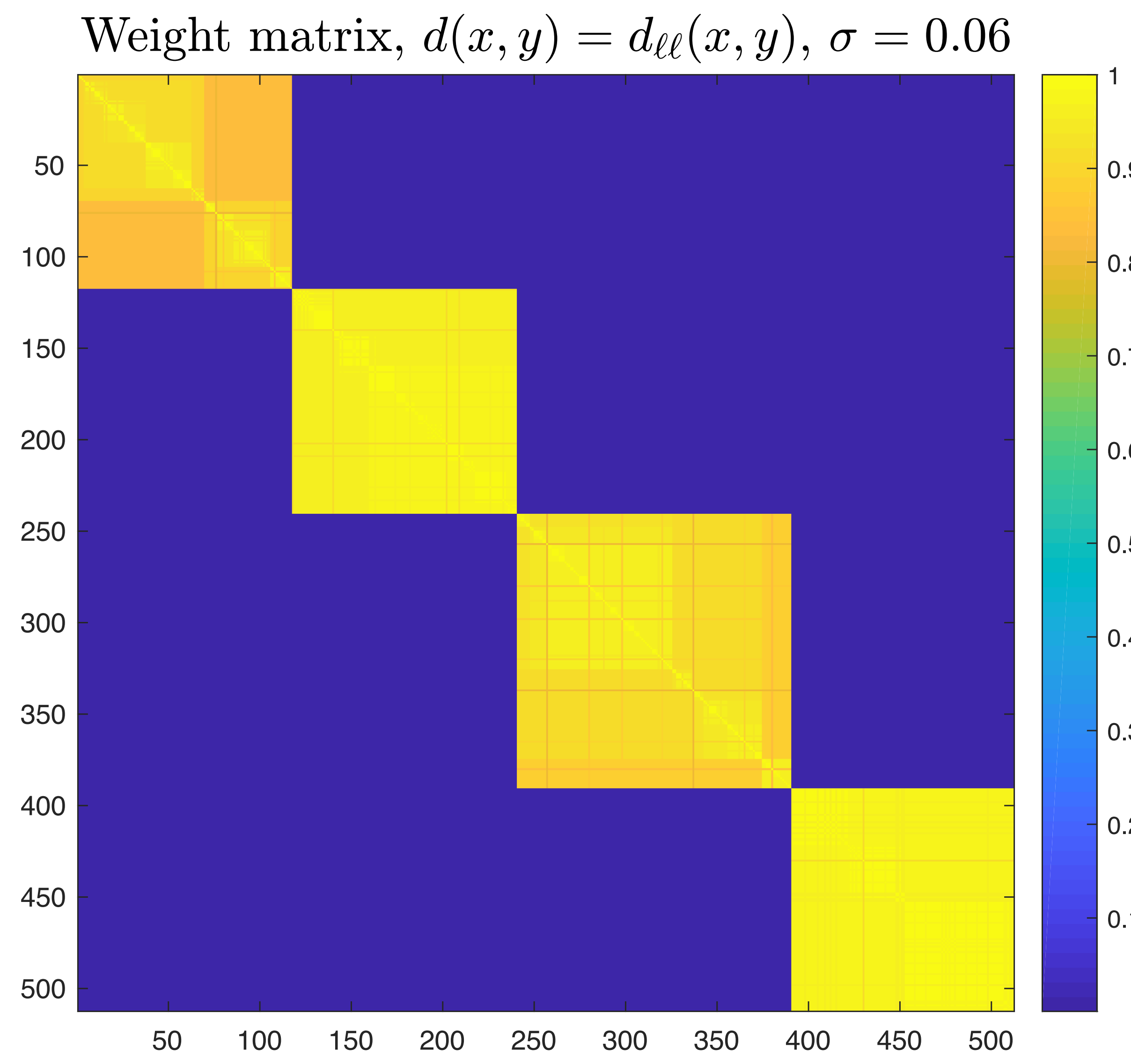


Tufts UNIVERSITY

LLPD Weight Matrix

- For our simple "four lines" data, there is a big difference between Euclidean distance (data independent) and LLPD (data dependent).
- The LLPD weight matrix has block-constant structure.





Tufts UNIVERSITY

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

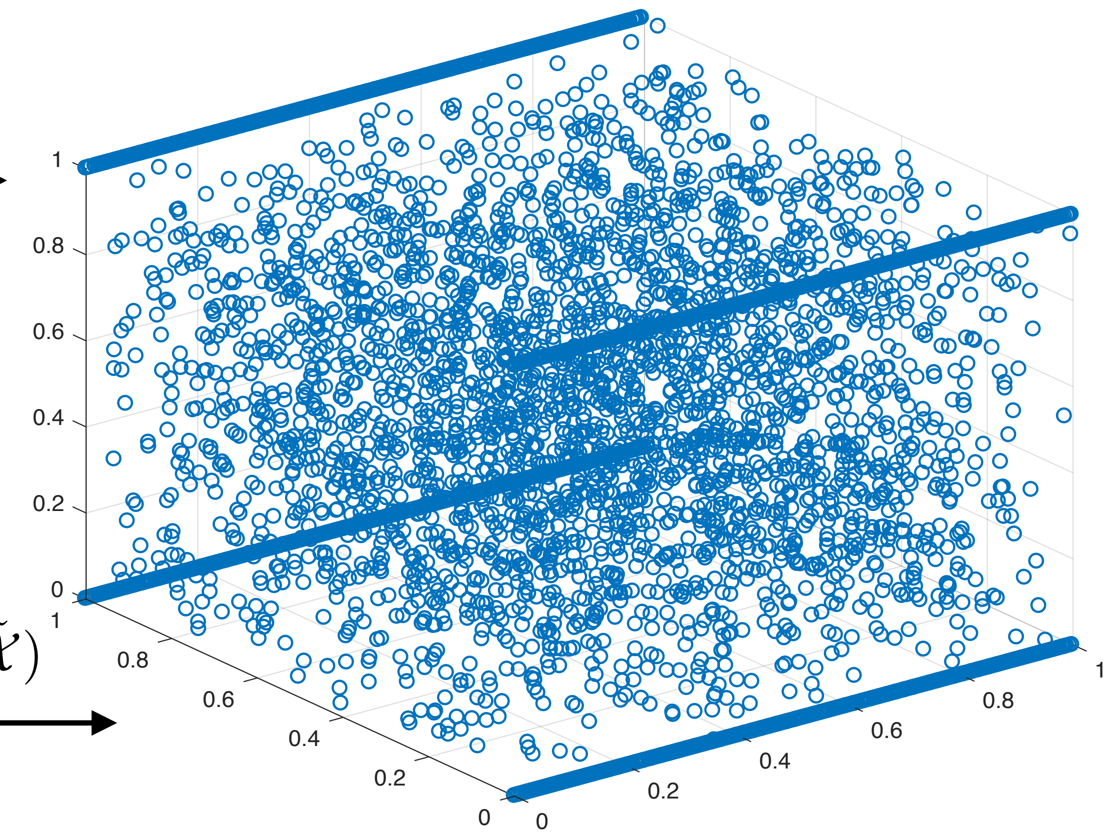
$$\mathcal{X}_1, \ldots, \mathcal{X}_K \subset \mathcal{X} \subset \mathbb{R}^D$$
$$\mathcal{X}_1, \ldots, \mathcal{X}_K \in \mathcal{S}_d(\kappa, \epsilon_0)$$
$$\delta = \min_{k \neq k'} \mathrm{dist}(\mathcal{X}_k, \mathcal{X}_{k'})$$

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

Large noise

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

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



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

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 ${k_{\mathrm{nse}}}^{th}$ 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.

Tufts UNIVERSITY

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 $\frac{\min_i n_i}{n_{max}} = O(1)$. If n_{min} is large enough and θ, σ satisfy

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

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

then with high probability the graph Laplacian L on the denoised LDLN data X_N satisfies:

(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.

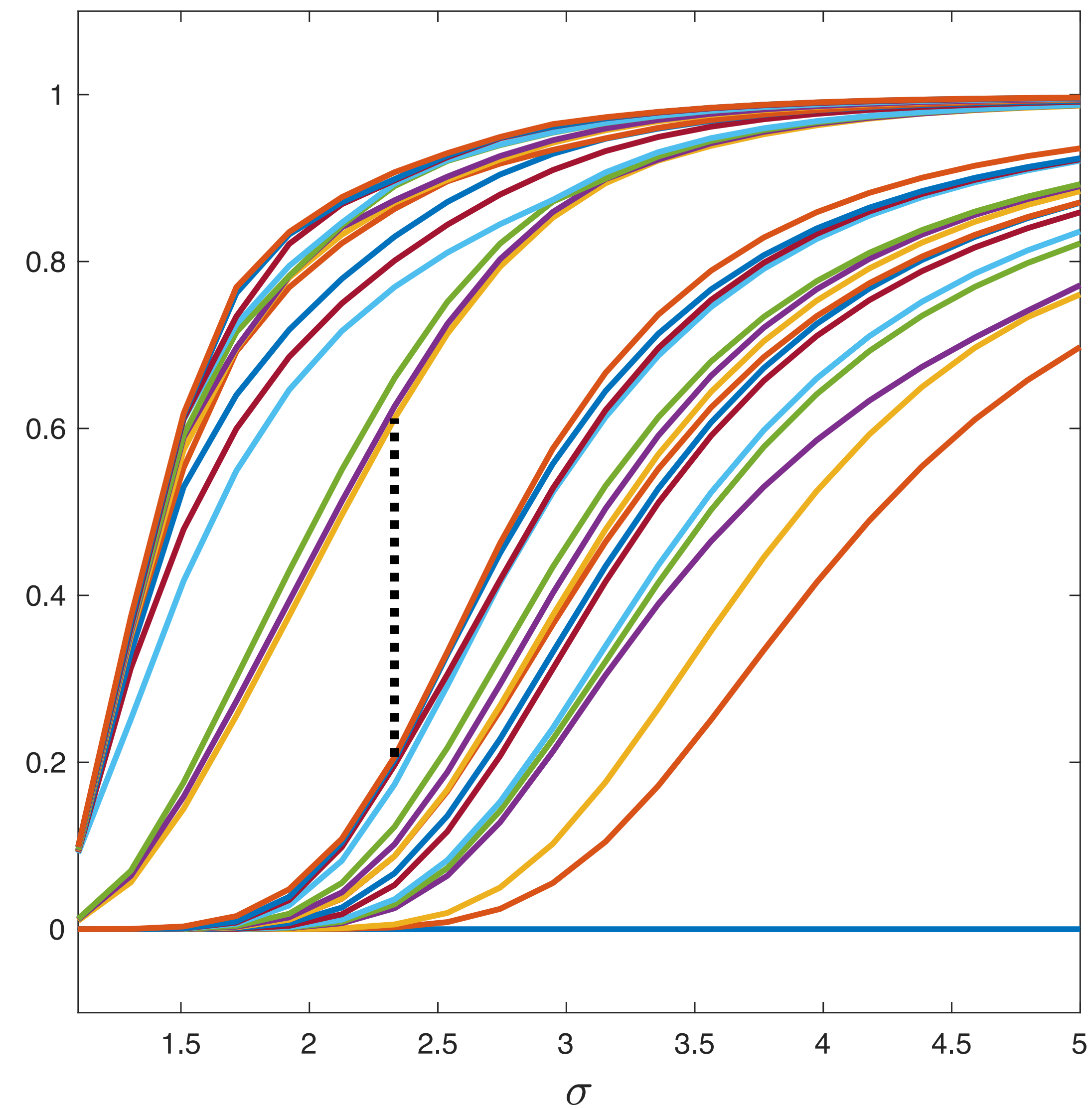
The constants $\{C_i\}_{I=1}^4$ depend on geometric quantities but do not depend on $n_1, \ldots, n_K, \tilde{n}, \theta, \sigma$.

Tufts UNIVERSITY

Application: Image Clustering

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.

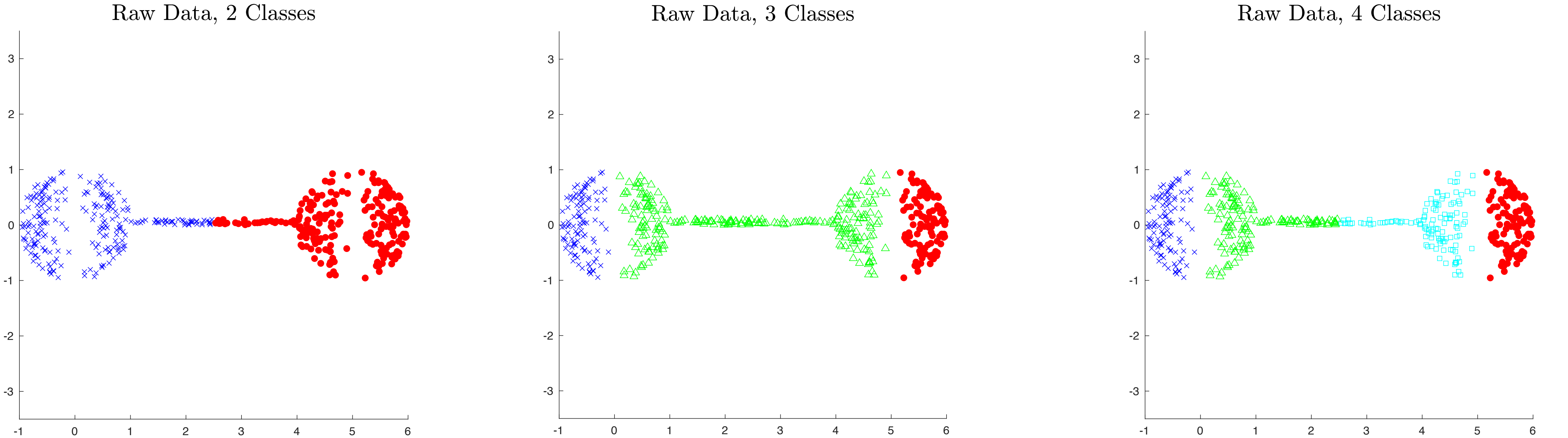
Tufts UNIVERSITY

Interpolating Between Geometry and Density

Definition. *For* $p \in [1, \infty)$ *and for* $x, y \in \mathcal{X}$, *the* (discrete) p-Fermat distance *from* x *to* y *is:*

$$\ell_p(x,y) = \min_{\pi=\{x_{i_j}\}_{j=1}^{T}} \left(\sum_{j=1}^{T-1} \|x_{i_j} - x_{i_{j+1}}\|^p \right)^{\frac{1}{p}},$$

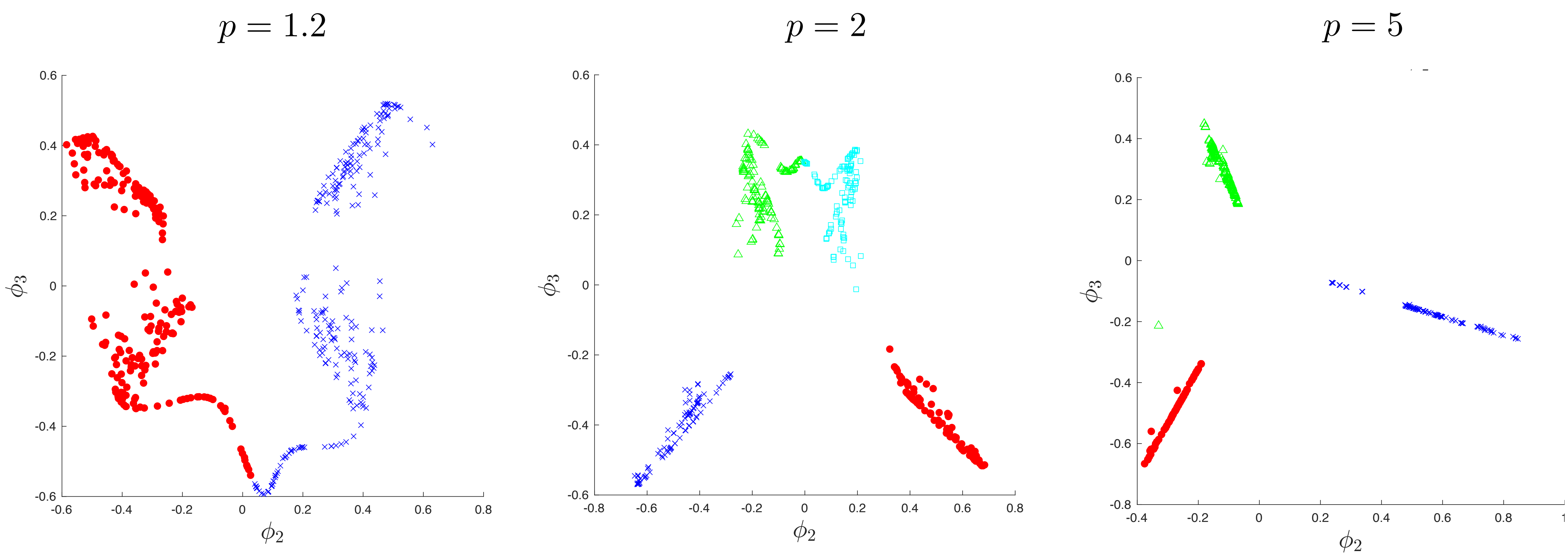
where π *is a path of points in* $\mathcal{X}$ *with* $x_{i_1} = x$ *and* $x_{i_T} = y$ *and* $\|\cdot\|$ *is the Euclidean norm.*



How to balance density and geometry when both are salient?

Tufts UNIVERSITY

Role of p



- As p changes, the embedding changes.
- Small p emphasizes geometry (cutting along the bottleneck).
- Large p emphasizes density (close to LLPD)

Fast Algorithms for Fermat Distances

- One can compute Fermat distances in quite general settings very fast, at least when $p \gg 1$.

Theorem. *(Little, McKenzie,* ***M.****) Let* $\mathcal{M}$ *be a compact, d-dimensional manifold with positive reach. Let* $\mathcal{X} = \{x_i\}_{i=1}^n$ *be drawn i.i.d. from* $\mathcal{M}$ *according to a probability distribution with continuous density* f *satisfying* $0 < f_{\min} \leq f(x) \leq f_{\max}$ *for all* $x \in \mathcal{M}$*. For* $p > 1$ *and* n *sufficiently large, Fermat distances computed using (i) a complete Euclidean distances graph and (ii) a Euclidean* k*-nearest neighbors graph are the same with probability at least* $1 - 1/n$ *if*

$$k \gtrsim \left[\frac{f_{\max}}{f_{\min}}\right] \left[\frac{4}{4^{1-1/p} - 1}\right]^{d/2} \log(n). \tag{1}$$

- Implicit constant in (1) depends on manifold reach and curvature.

Tufts UNIVERSITY

Continuum Formulation

Definition. *Let* $(\mathcal{M}, g)$ *be a compact, d-dimensional Riemannian manifold and* f *a continuous density function on* $\mathcal{M}$ *that is lower bounded away from zero (i.e.* $f_{\min} := \min_{x \in \mathcal{M}} f(x) > 0$ *on* $\mathcal{M}$*). For* $p \in [1, \infty)$ *and* $x, y \in \mathcal{M}$*, the* (continuum) p-Fermat distance *from* x *to* y *is:*

$$\mathcal{L}_p(x, y) = \left(\inf_{\gamma} \int_0^1 \frac{1}{f(\gamma(t))^{\frac{p-1}{d}}} \sqrt{g\left(\gamma'(t), \gamma'(t)\right)} dt \right)^{\frac{1}{p}}, \qquad (1)$$

where $\gamma : [0, 1] \to \mathcal{M}$ *is a* $\mathcal{C}^1$ *path with* $\gamma(0) = x, \gamma(1) = y$*.*

- Let $\mathscr{D}(x, y)$ be the geodesic on the manifold
- Let $\mathscr{D}_{f,\text{Euc}}(x, y) = \dfrac{\|x - y\|}{(f(x)f(y))^{\frac{p-1}{2d}}}$

be a density-based stretch of Euclidean distance.

Local Equivalence

Theorem. *(Little, McKenzie, **M.**) Assume $\mathcal{M}$ is sufficiently regular and that f is a bounded $\mathfrak{L}$-Lipschitz density function on $\mathcal{M}$ with $f_{min} > 0$. Let $\epsilon > 0$. Then there exist constants $\epsilon_0, C_1, C_2, C_3$ depending only on the geometry of $\mathcal{M}$, f_{min}, $\mathfrak{L}$, p, and d such that for all $x, y \in \mathcal{M}$ such that $\mathscr{D}(x,y) \leq \epsilon_0$ and $\|x-y\| \leq \epsilon$,*

$$|\mathcal{L}_p(x,y) - \mathscr{D}_{f,Euc}^{1/p}(x,y)| \leq C_1 \epsilon^{1+\frac{1}{p}} + C_2 \epsilon^{2+\frac{1}{p}} + O(\epsilon^{3+\frac{1}{p}}).$$

- This gives an opening to developing a discrete-to-continuunm limit theory for graph operators constructed with Fermat distances which reveal how p balances density with geometric structure.

- Ongoing work making this precise.

Tufts UNIVERSITY

References & Support

- Little, Maggioni, and **Murphy**. "Path-Based Spectral Clustering: Guarantees, Robustness to Outliers, and Fast Algorithms." *Journal of Machine Learning Research*. 2020.

- Little, McKenzie, and **Murphy**. "Balancing Geometry and Density: Path Distances on High-Dimensional Data." *SIAM Journal on the Mathematics of Data Science*. 2022.

The Camille & Henry Dreyfus Foundation

NSF

DMS 1912737
DMS 1924513

Tufts University

DISC

Code and Contact Information

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

Contact: jm.murphy@tufts.edu

Thanks for Your Attention!