Balancing Geometry and Density with Fermat Metrics: Graphs, Clustering, and Applications

James M. Murphy

The 15th AIMS Conference
July 9, 2026

Tufts
UNIVERSITY

Learning in High Dimensions is Hard

- High-dimensional problems (i.e., many variables relative to number of observations) are hard.

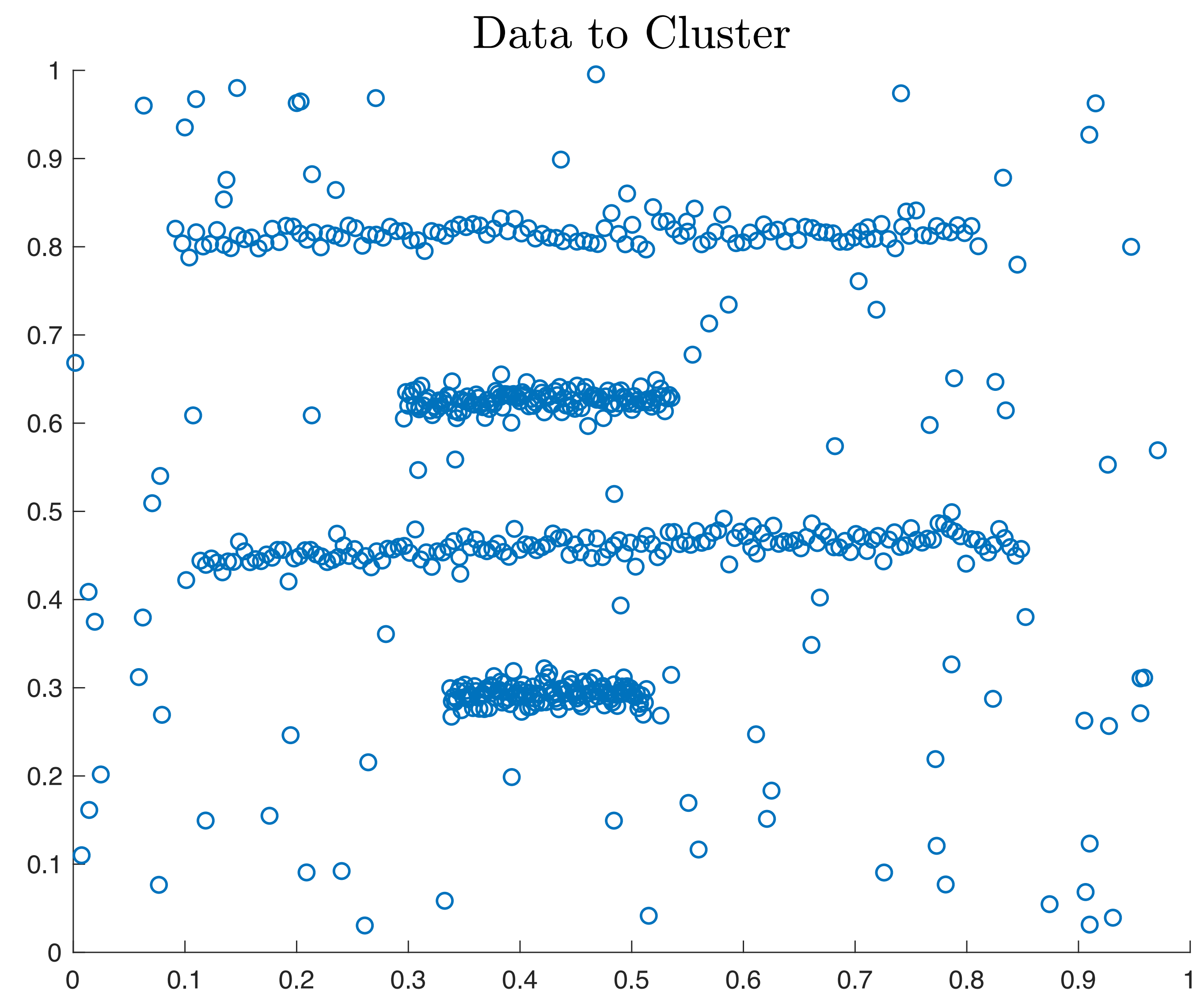
- The *curse of dimensionality* dooms inference in the absence of structural assumptions on the data.

- *Manifold Hypothesis*: data lies near low-dimensional subspace or manifold.

Unsupervised Learning

Unsupervised learning: infer structure from data without access to training data.

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 (e.g., Gaussians).

- Need to know K.

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

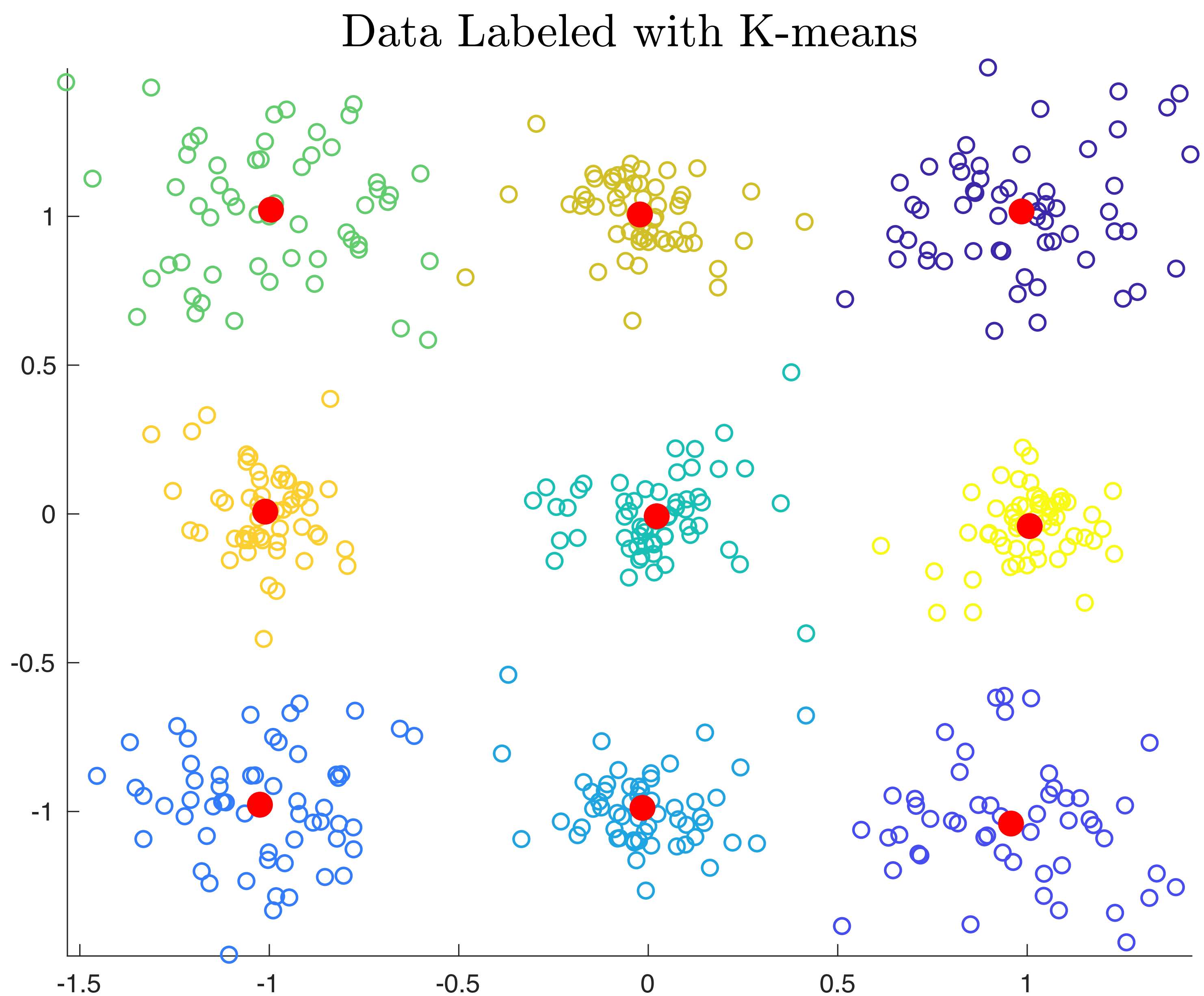
Tufts UNIVERSITY

Standard Method: K-Means

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

- Empirically good for same sized, spherical clusters (e.g., Gaussians).

- 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

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*

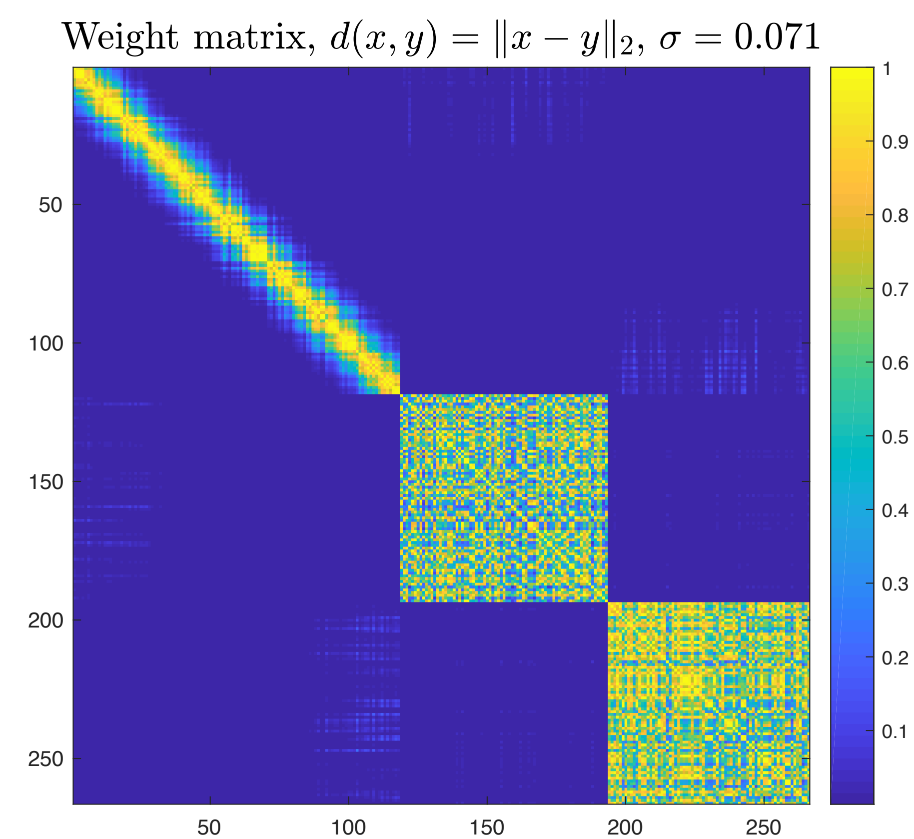
$$W_{ij} = e^{-d(x_i, x_j)^2/\sigma^2}$$

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

Step 2: Compute the *(graph) Laplacian*

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

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



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

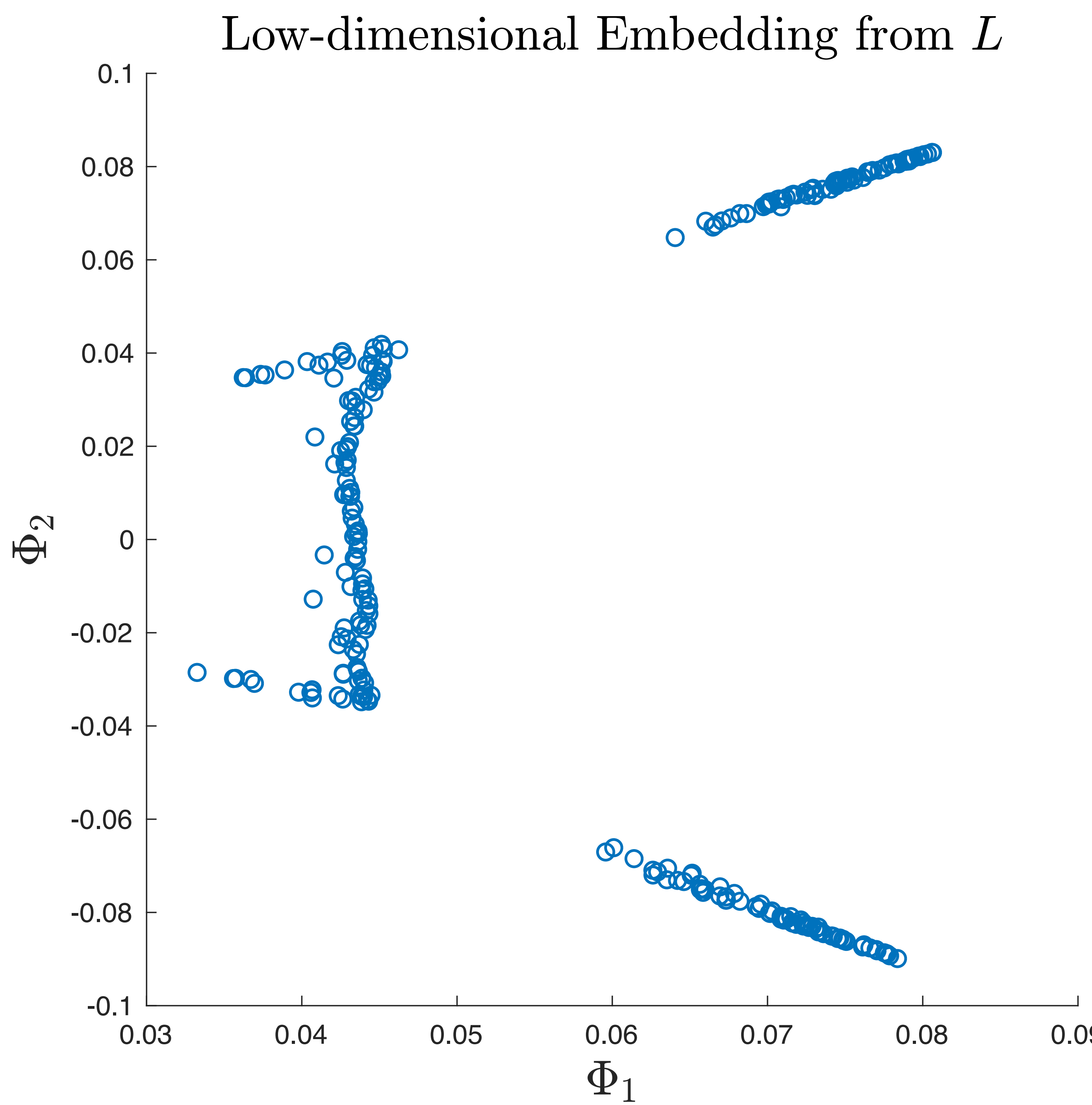
$$\Phi_1, ..., \Phi_n.$$

Step 4: Embed the data as

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

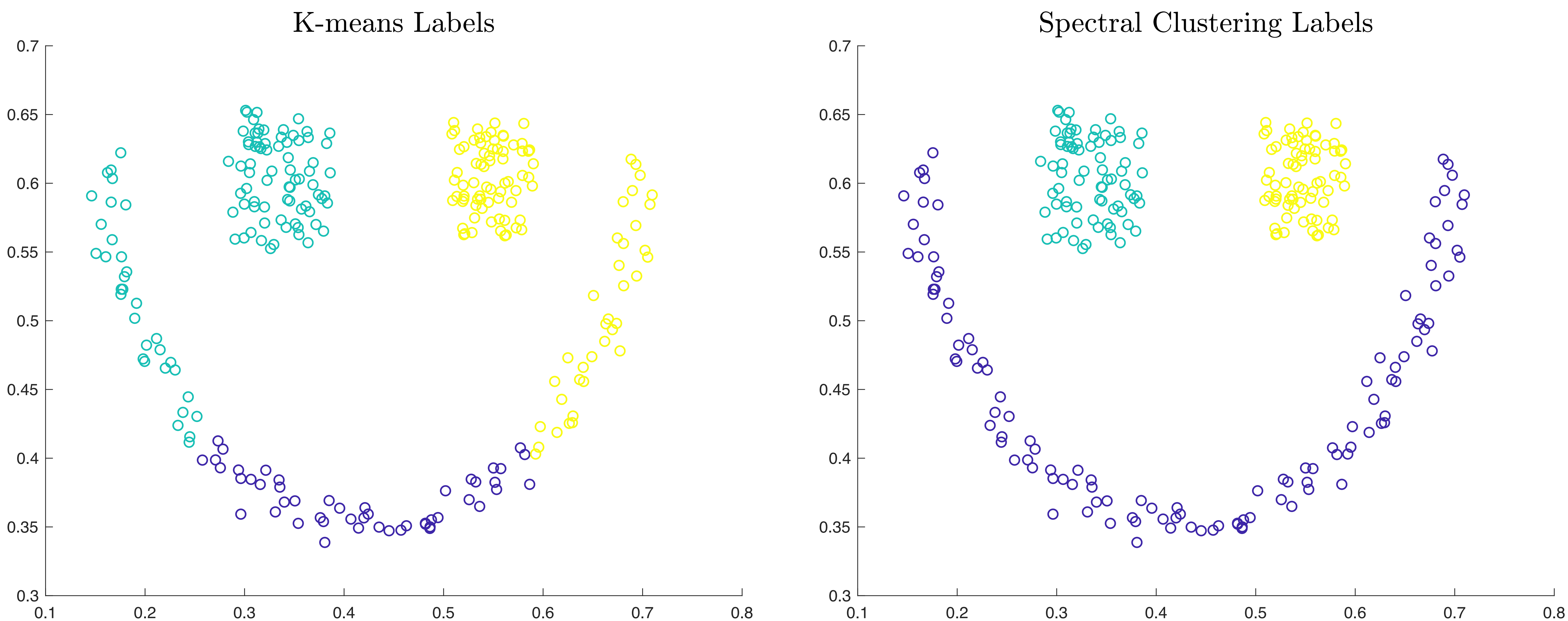
then run K-means. Note

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



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

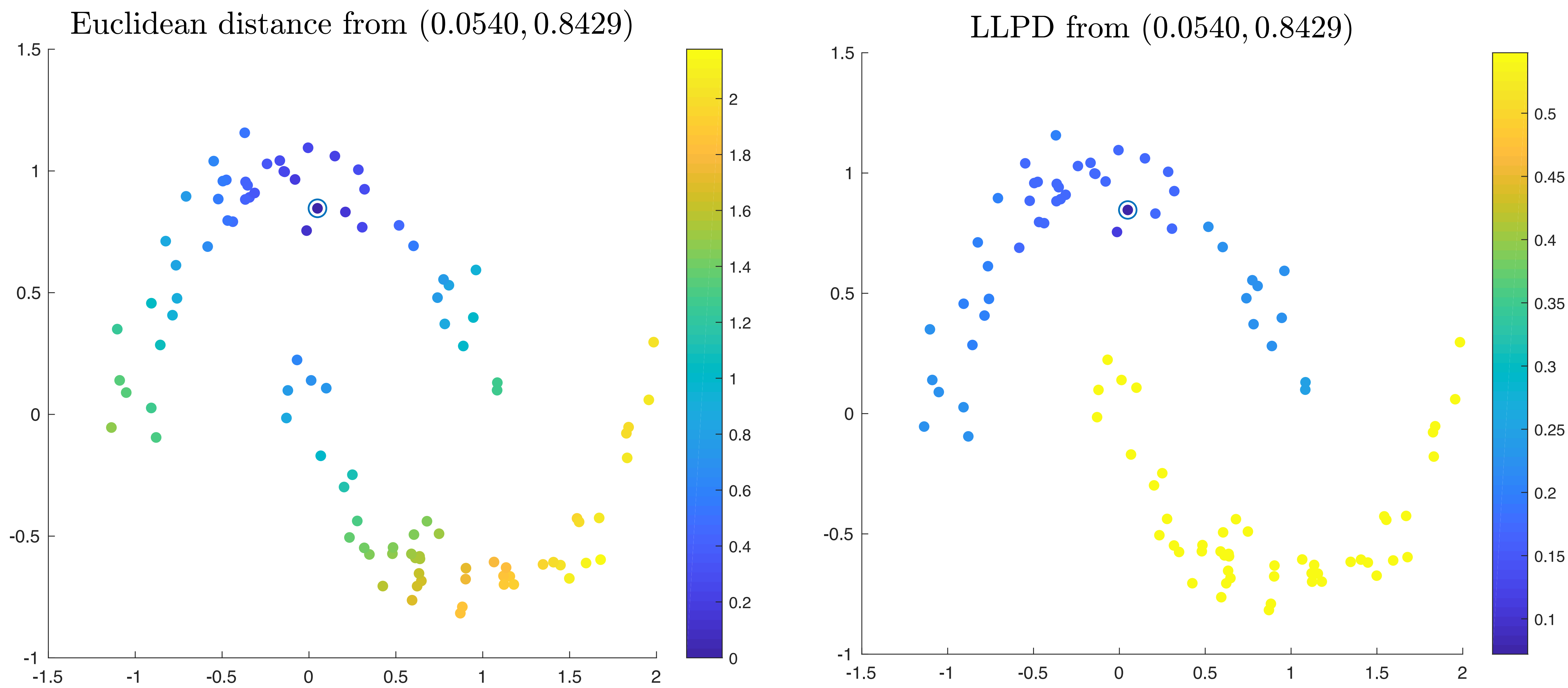
Longest Leg Path Distance

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:*

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

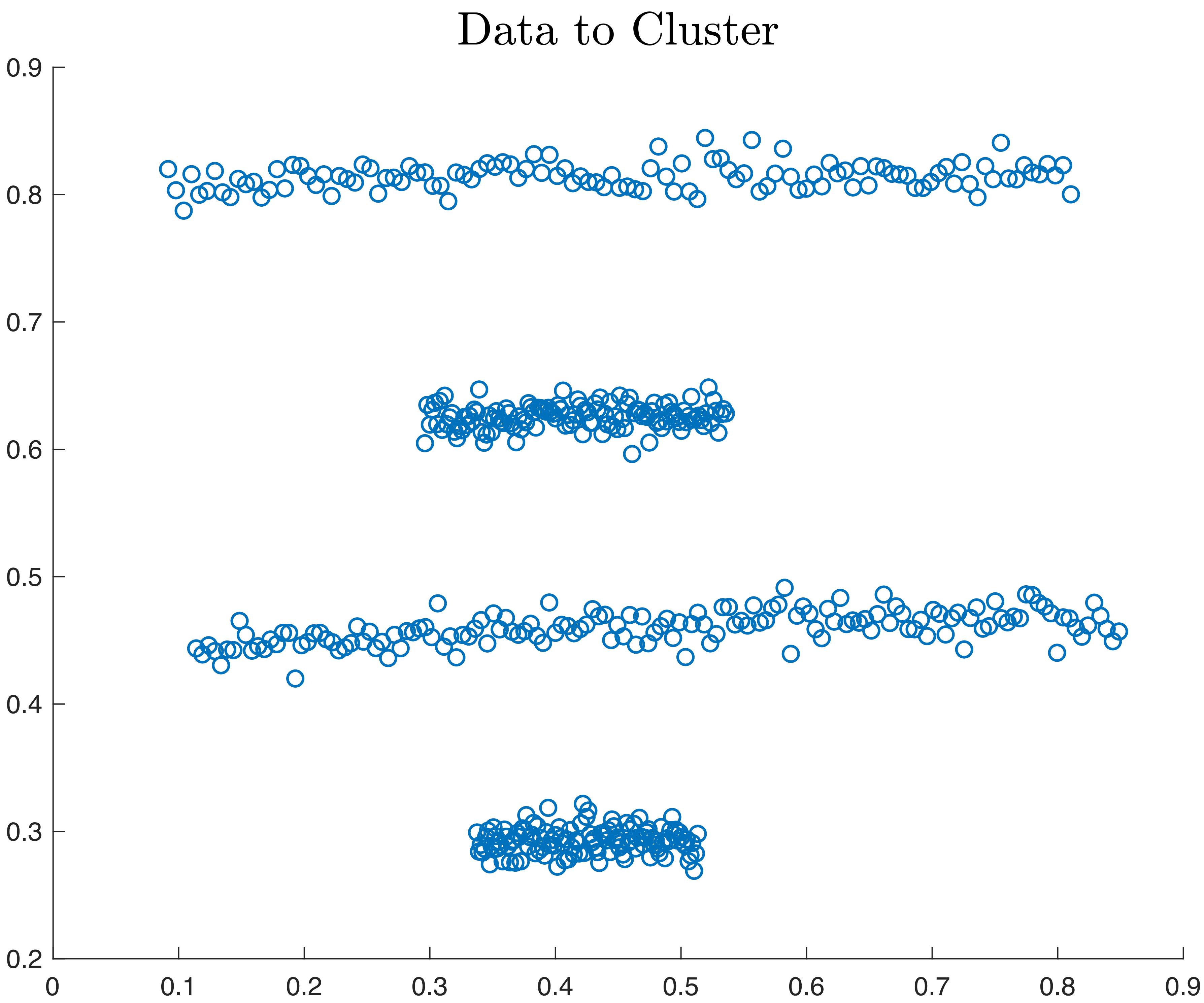
- The distance between points x, y is the minimum over all paths between x, y of the longest edge in the path.
- Depending on the data X, this distance changes!
- $\mathcal{G}$ could be a complete graph (all points connected to all points) or a connected NN graph.
- Looks hard to compute. We have a fast approximation, so this turns out not to be an obstacle.

Euclidean Distance versus LLPD



Tufts UNIVERSITY

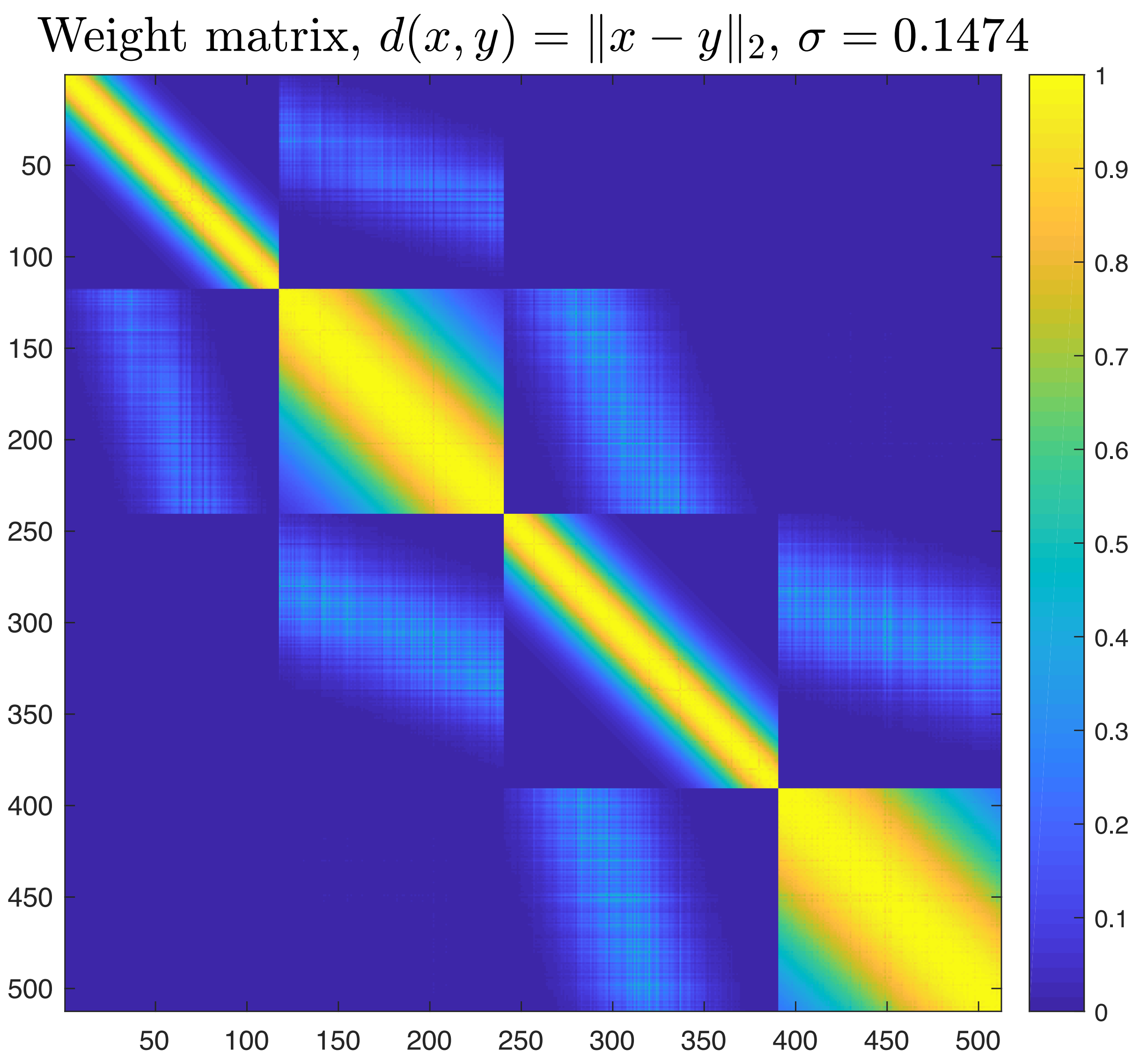
Data Well-Suited for LLPD

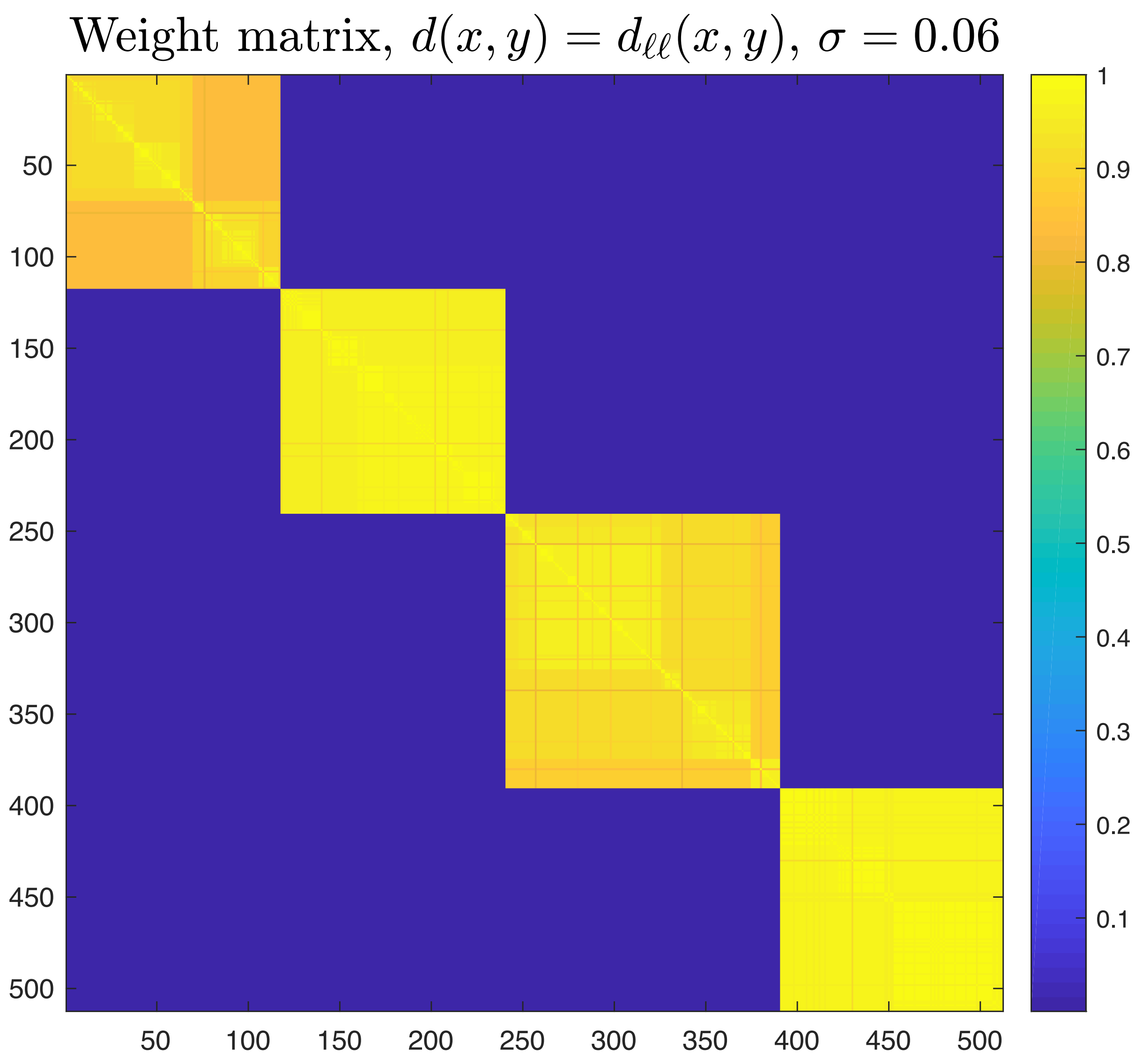


Tufts UNIVERSITY

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

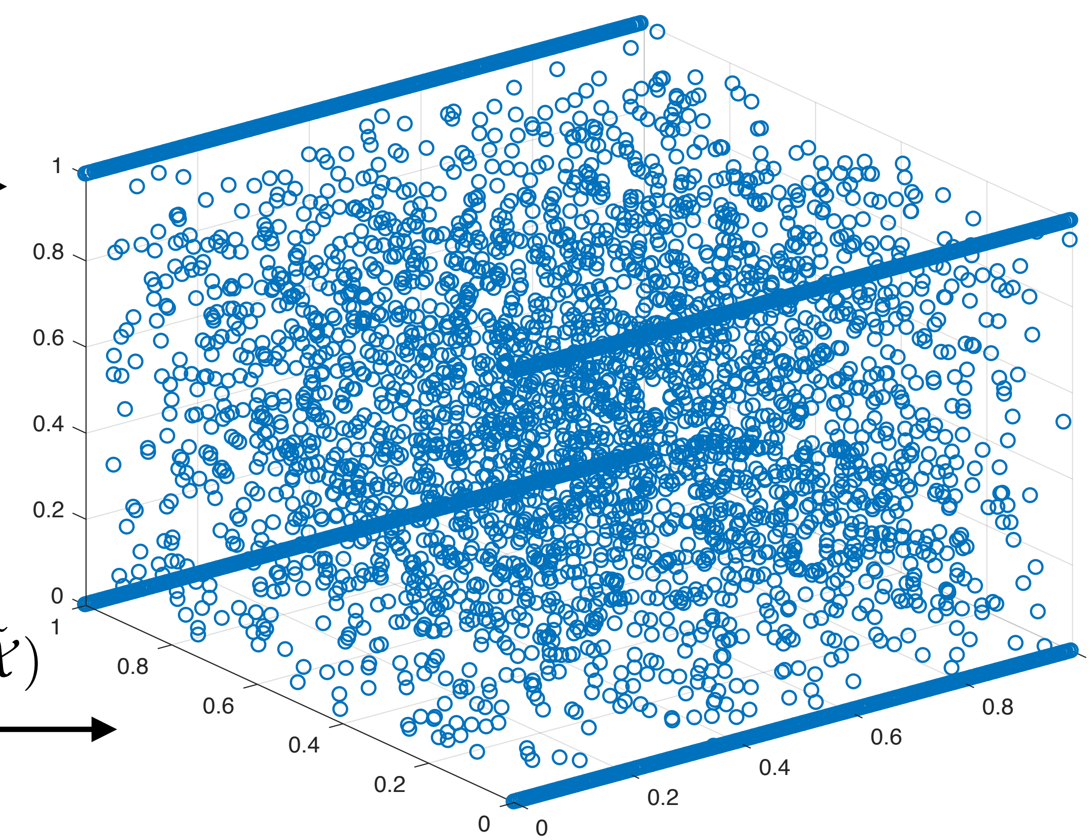
$$\boxed{\begin{aligned} &\mathcal{X}_1, \ldots, \mathcal{X}_K \subset \mathcal{X} \subset \mathbb{R}^D \\ &\mathcal{X}_1, \ldots, \mathcal{X}_K \in \mathcal{S}_d(\kappa, \epsilon_0) \\ &\delta = \min_{k \neq k'} \operatorname{dist}(\mathcal{X}_k, \mathcal{X}_{k'}) \end{aligned}}$$

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

Large noise

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

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



$$n = n_1 + \ldots + n_K + \tilde{n}$$

$$n_{\min} = \min_{1 \leq k \leq K} n_k$$

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.

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* $n_1, \ldots, n_K, \tilde{n}, \theta, \sigma$.

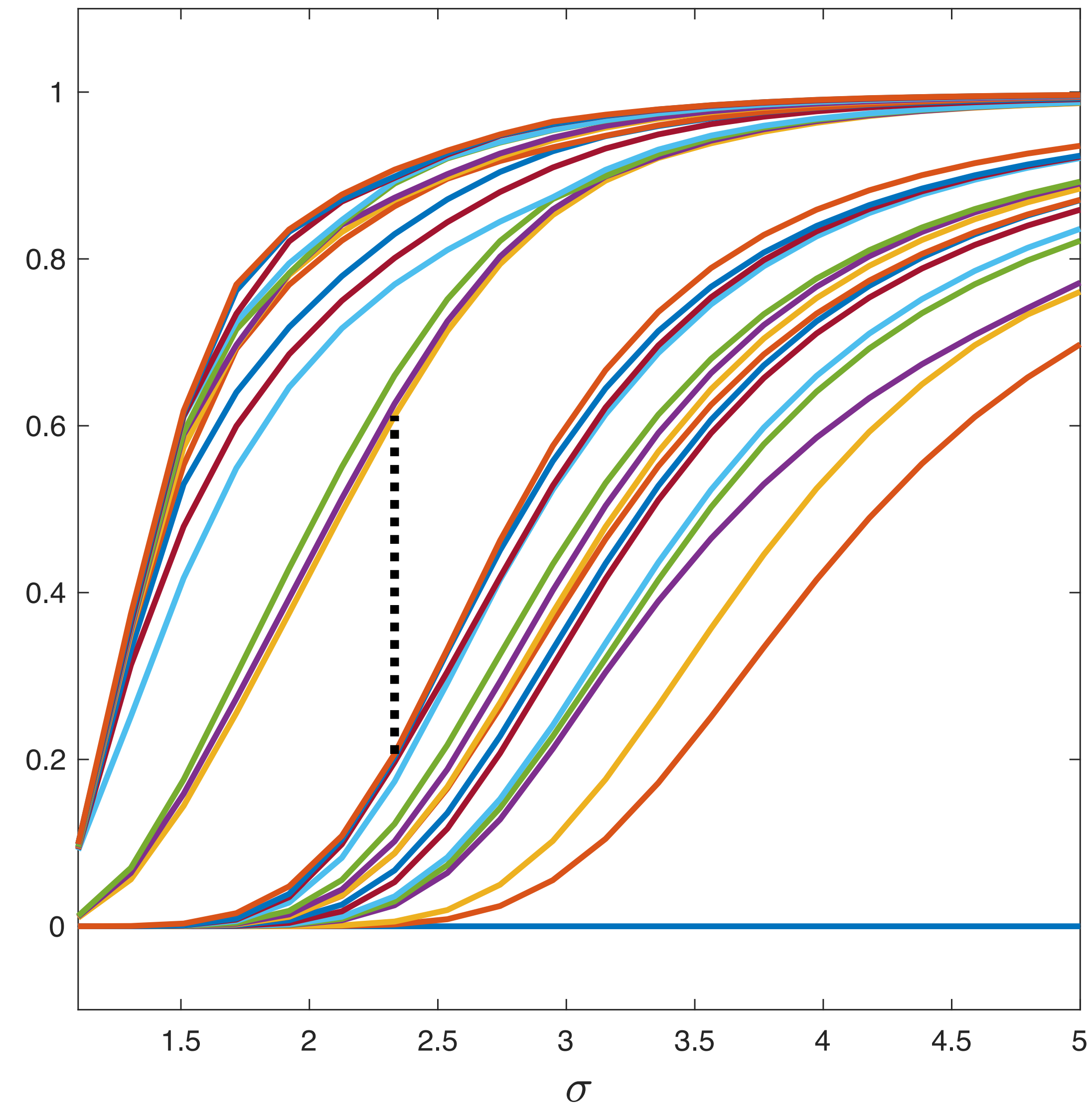
Little, Maggioni, Murphy
“Path-Based Spectral Clustering: Guarantees, Robustness to Outliers, and Fast Algorithms"
Journal of Machine Learning Research, 2020

Tufts
UNIVERSITY

Clustering Collection of Images

COIL 16 Classes

Multiscale Eigenvalues for LLPD SC



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

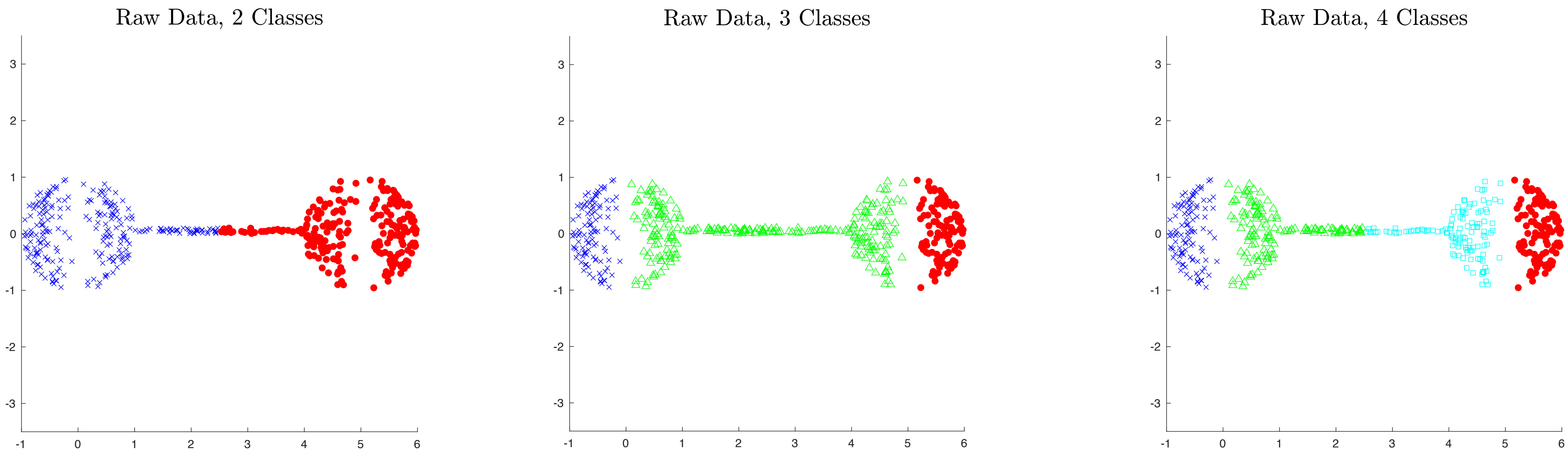
- LLPD spectral clustering achieve 99+% accuracy, and correctly identifies that there are 16 classes.

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:*

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

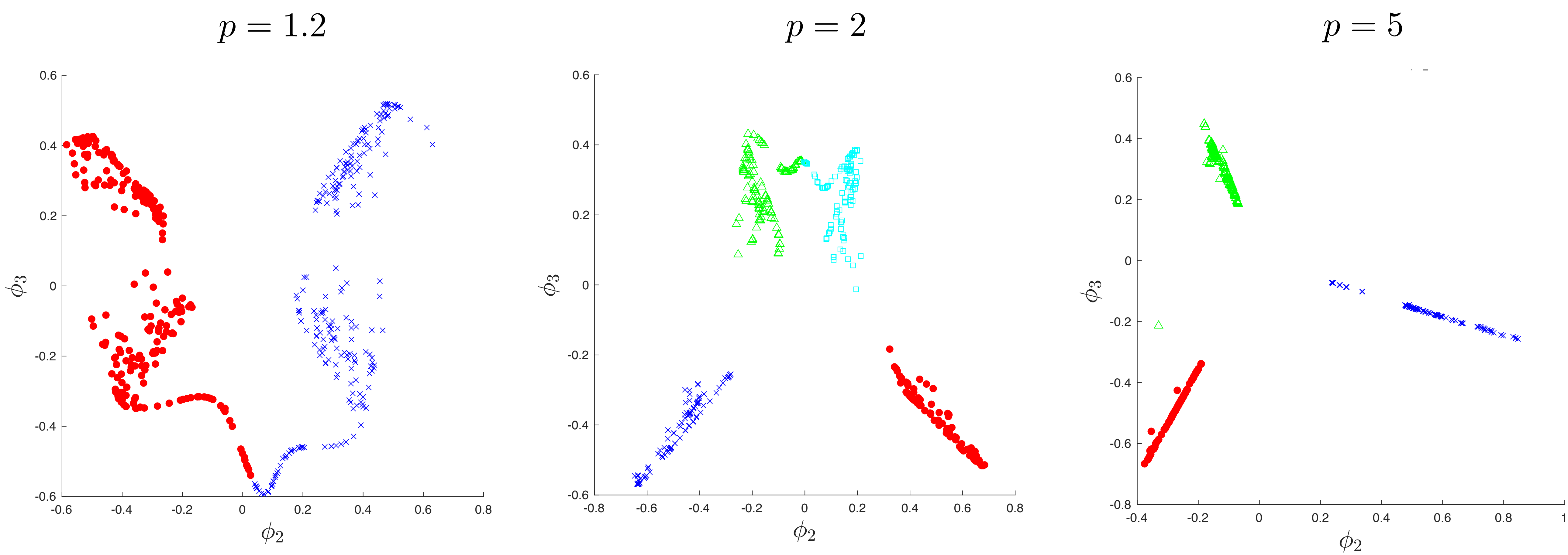
where π is a path of points in $\mathcal{X}$ with $x_{i_1} = x$ and $x_{i_T} = y$ and $\|\cdot\|$ is the Euclidean norm.



How to balance density and geometry when both matter?

Tufts UNIVERSITY

Role of p



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

Fast Algorithm for Fermat Distances

- When data lies near a low-dimensional manifold, one can compute approximate Fermat distances very fast.

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*

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

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

Segmenting Images with Fermat Laplacian

Raw Image

Euclidean Laplacian

Fermat Laplacian

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:*

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

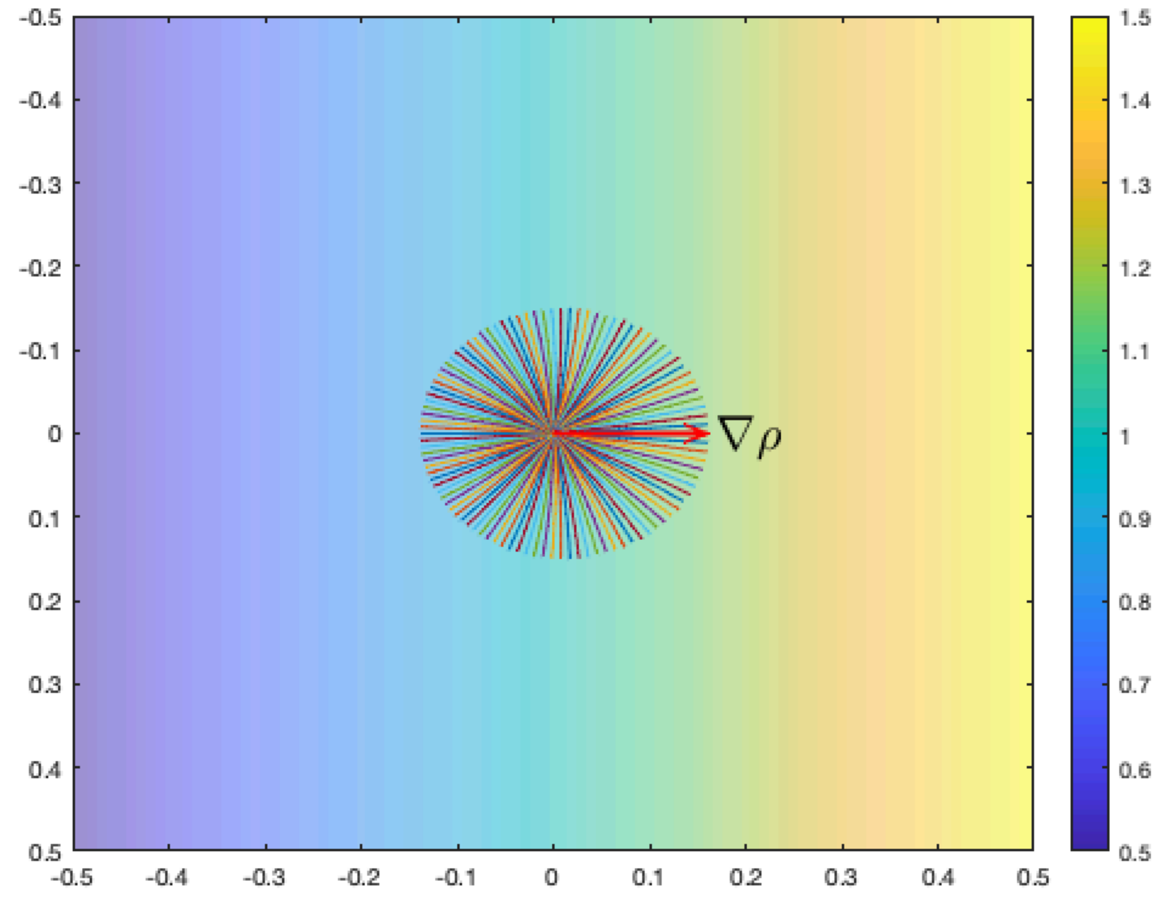
where $\gamma : [0,1] \to \mathcal{M}$ *is a* $\mathcal{C}^1$ *path with* $\gamma(0) = x, \gamma(1) = y$*.*

- One can show convergence of empirical Fermat distances to continuum Fermat distances.
- This can be combined with operator convergence arguments to prove the spectra of the (random) empirical Fermat graph Laplacian converge to those of a continuum operator.

Continuum Limit of Fermat Laplacians

Theorem. *(García Trillos, Little, McKenzie,* ***M.****) Let* μ *be an absolutely continuous measure supported on a smooth, compact Riemannnian manifold* $\mathcal{M}$ *of dimension* d*. Let* ψ_2^n *be the (random) discrete Laplacian eigenvector with second smallest eigenvalue associated to the Fermat graph Laplacian* L_n *constructed with* n *random samples from* μ*. Let* f *be the eigenfunction of a corresponding continuum operator on* $L^2(\mathcal{M}, \mu)$ *with second smallest eigenvalue. Then with high probability,* $\|f - \psi_2^n\|_{L^2(\mu)}^2 \lesssim n^{-\frac{1}{3d}}$.

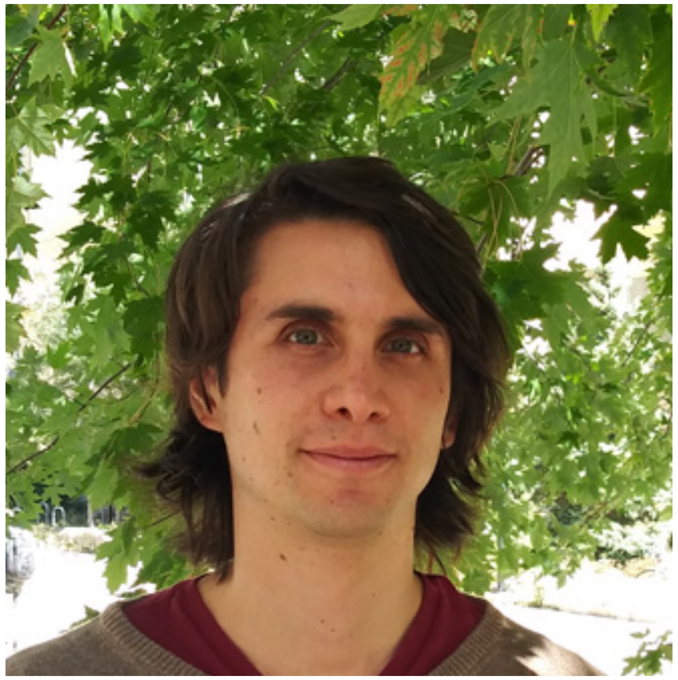
- The continuum operator shows a drift given by the density, whose impact increases with p.
- The rate is slower than with Euclidean distances($\approx n^{-1/d}$) due to the slow convergence of empirical Fermat distances to their population version.



García Trillos, Little, Mckenzie, Murphy
"Fermat Distances: Metric Approximation, Spectral Convergence, and Clustering Algorithms"
Journal of Machine Learning Research, 2024

Tufts UNIVERSITY

Collaborators



N. García Trillos (Wisconsin Statistics)

A. Little (Utah Math)

M. Maggioni, (JHU Applied Math)

D. McKenzie (Mines Applied Math)

Tufts UNIVERSITY

Support

DMS 1912737
DMS 1924513
DMS 2309519
DMS 2318894

NSF

NATIONAL GEOSPATIAL-INTELLIGENCE AGENCY
UNITED STATES OF AMERICA

DISC

The Camille & Henry Dreyfus Foundation

Tufts
UNIVERSITY

Papers

García Trillos, Little, Mckenzie, Murphy
"Fermat Distances: Metric Approximation, Spectral Convergence, and Clustering Algorithms"
Journal of Machine Learning Research, 2024

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

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

Code and Contact Information

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

Contact: jm.murphy@tufts.edu

Thanks for Your Attention!

Tufts UNIVERSITY