

Breaking the Curse of Dimensionality: Graphs, Probability Measures, and Data

James M. Murphy

Brown LCDS
October 20, 2025



Support



DMS 1912737
DMS 1924513
DMS 2309519
DMS 2318894



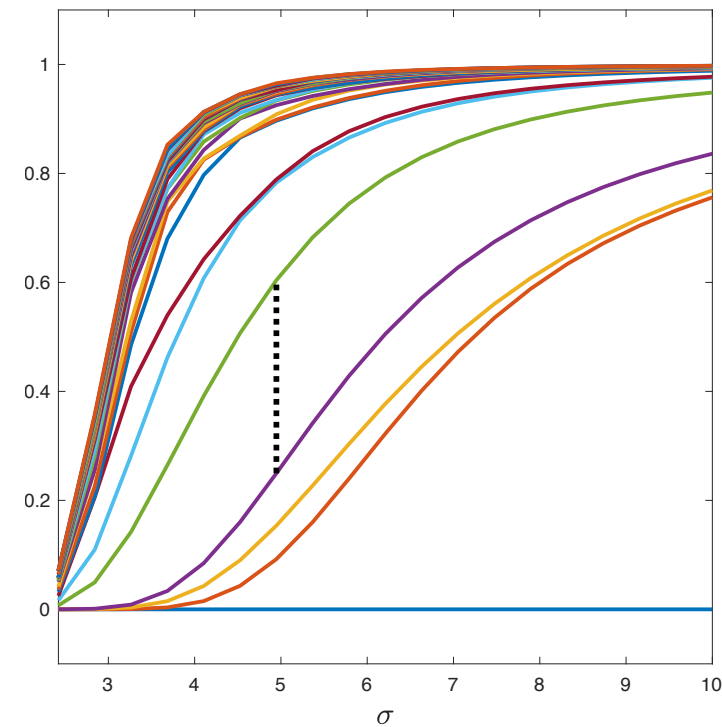
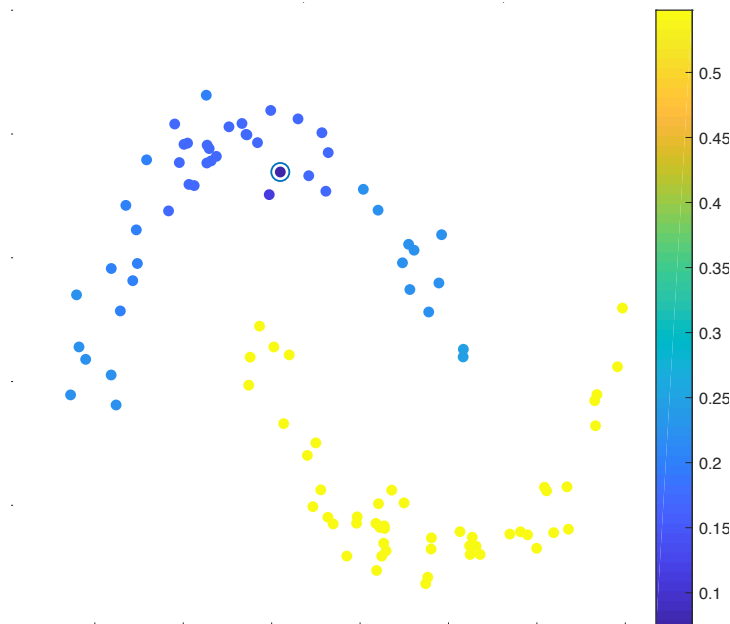
THE CAMILLE & HENRY DREYFUS FOUNDATION

Learning in High Dimensions is Hard

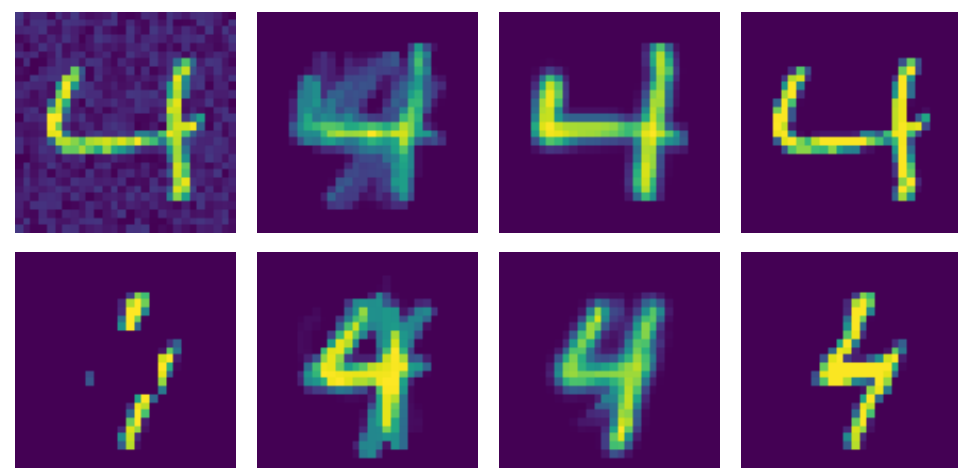
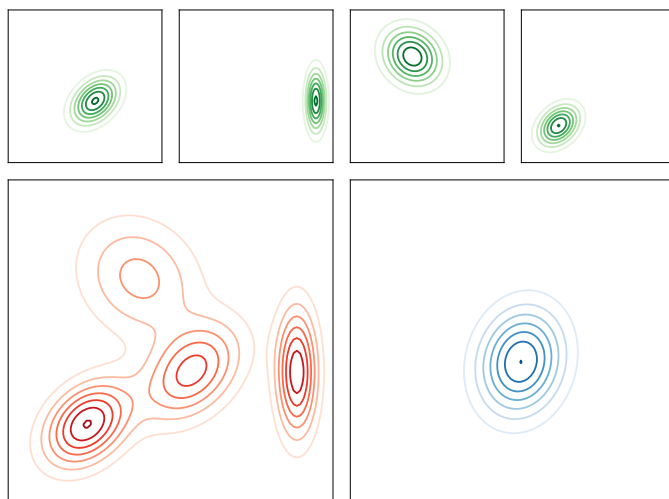
- High-dimensional problems (i.e., many variables relative to number of observations) are hard.
- The *curse of dimensionality* dooms inference in the absence of structural assumptions on the data.
- *Manifold Hypothesis*: data lies near low-dimensional subspace or manifold.

Outline

- Part I: Unsupervised Learning with Data-Driven Metrics

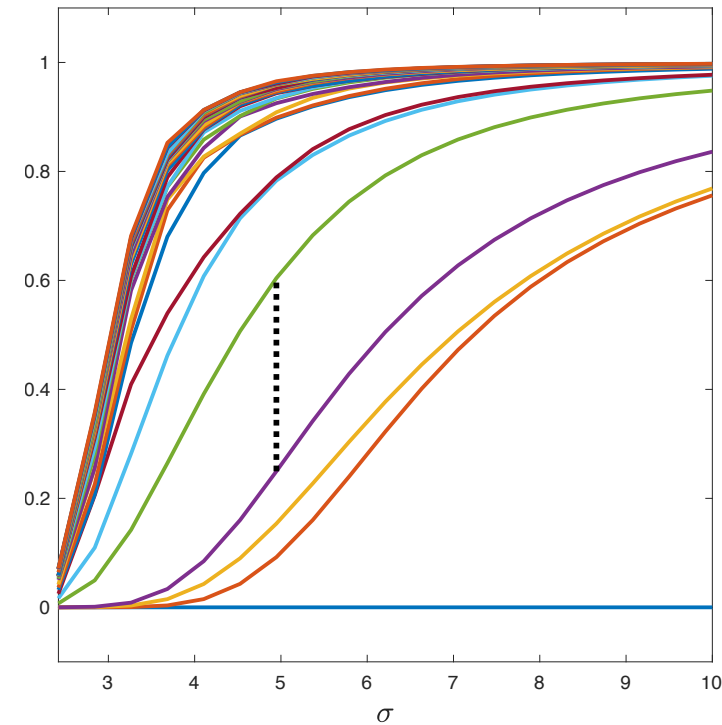
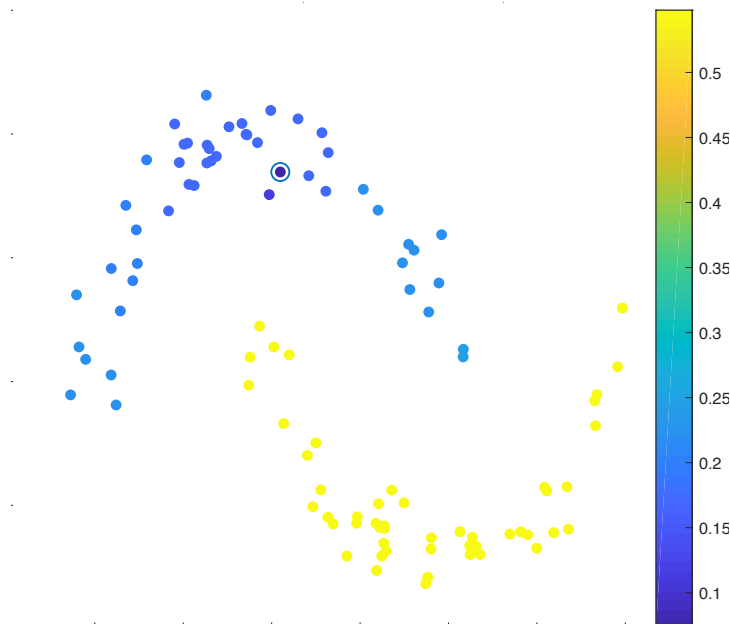


- Part II: Representation Learning in Wasserstein Space

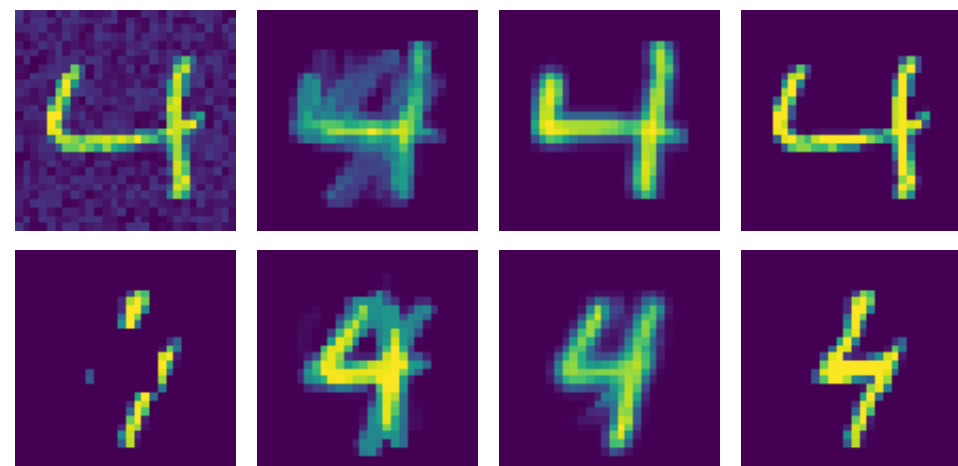
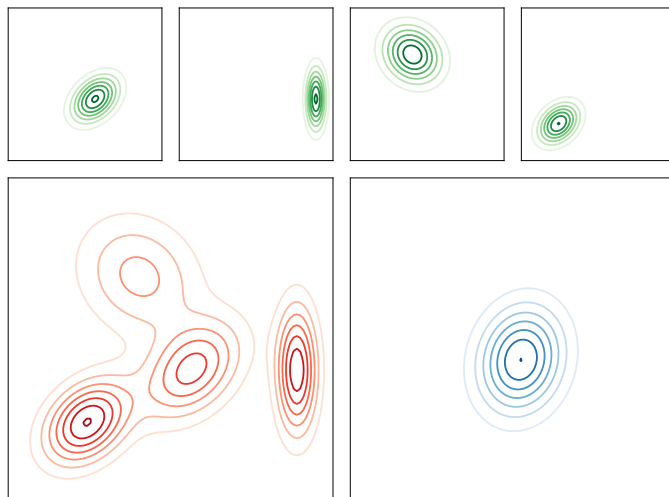


Outline

- Part I: Unsupervised Learning with Data-Driven Metrics



- Part II: Representation Learning in Wasserstein Space



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.



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

Labeling: Which x_j were generated from μ_k ?

Number of Clusters: Can we estimate K ?

Standard Method: K-Means

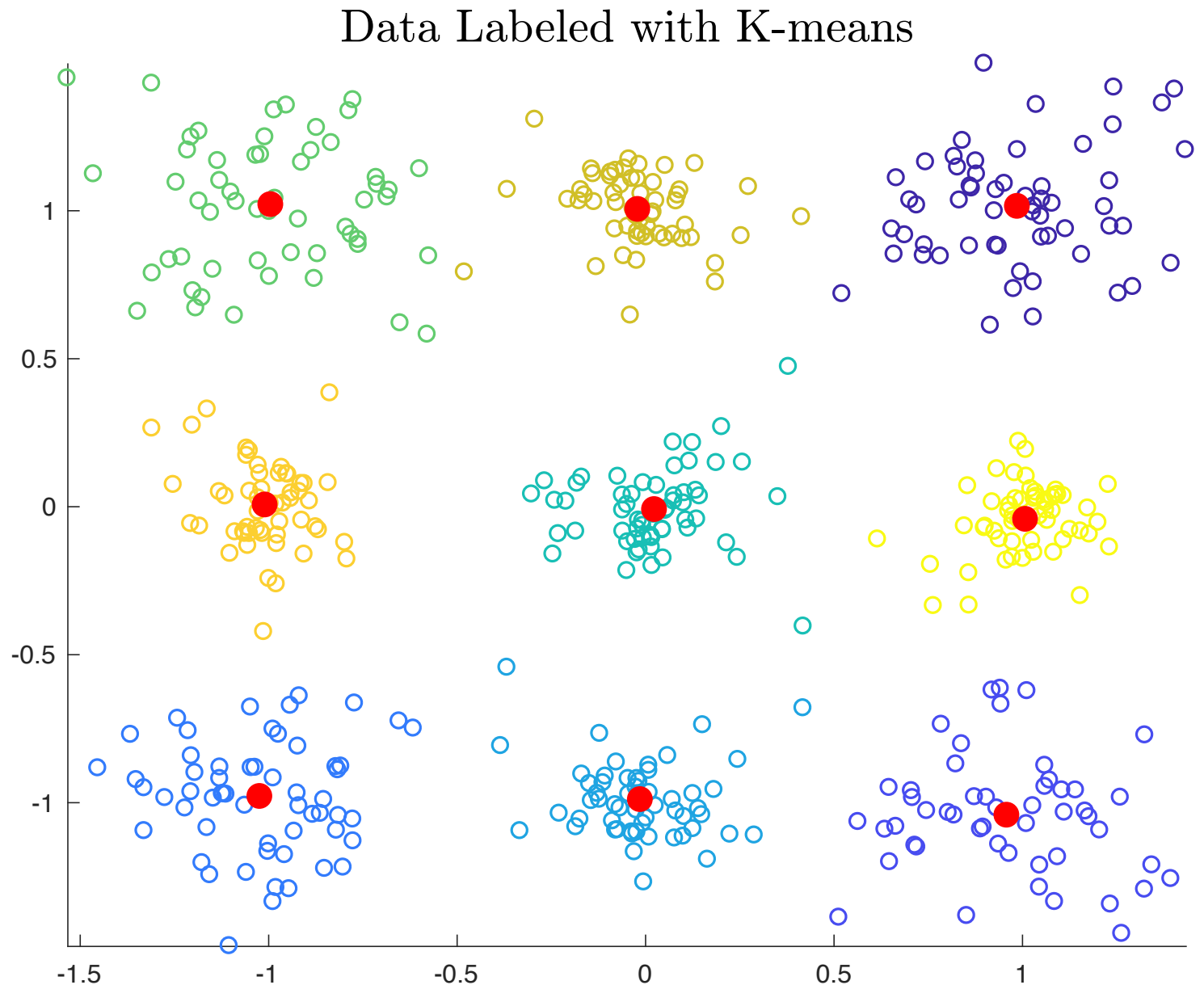
- **Idea:** find K centroids, then assign each point to its nearest centroid.
- Empirically good for same sized, spherical clusters (e.g., Gaussians).
- Need to know K .



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

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 .



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

Spectral Clustering I

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

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

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 .

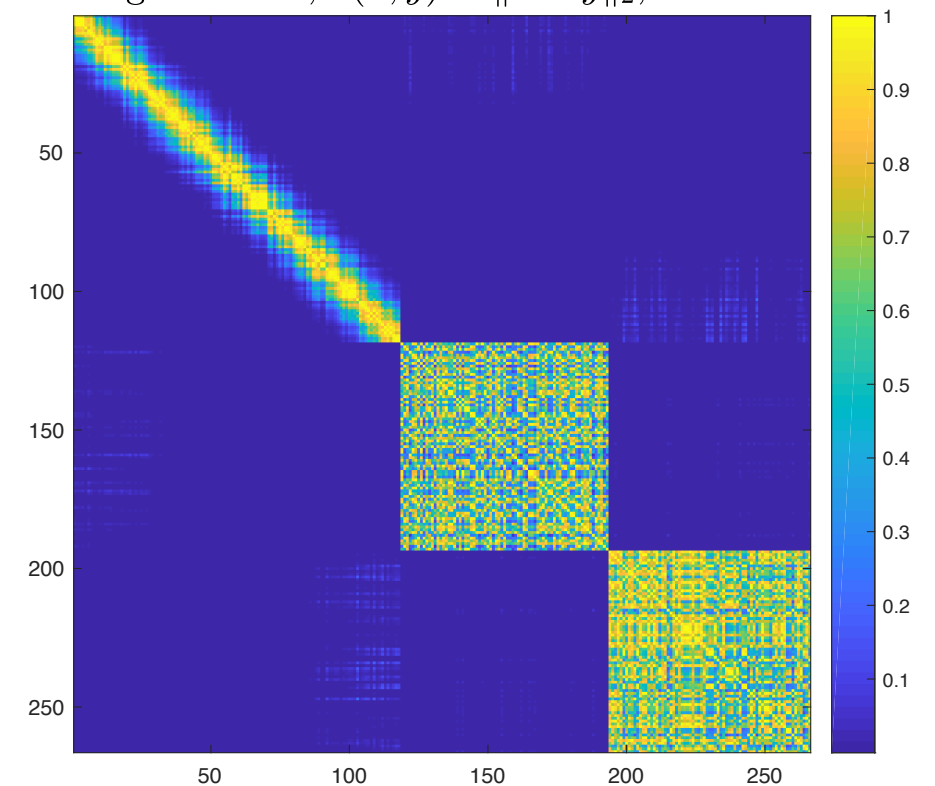
Step 2: Compute the (*graph*) *Laplacian*

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

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



Weight matrix, $d(x, y) = \|x - y\|_2$, $\sigma = 0.071$



Spectral Clustering II

Step 3: Compute eigenvalues of L

$$0 \leq \lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_n$$

and associated eigenvectors

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

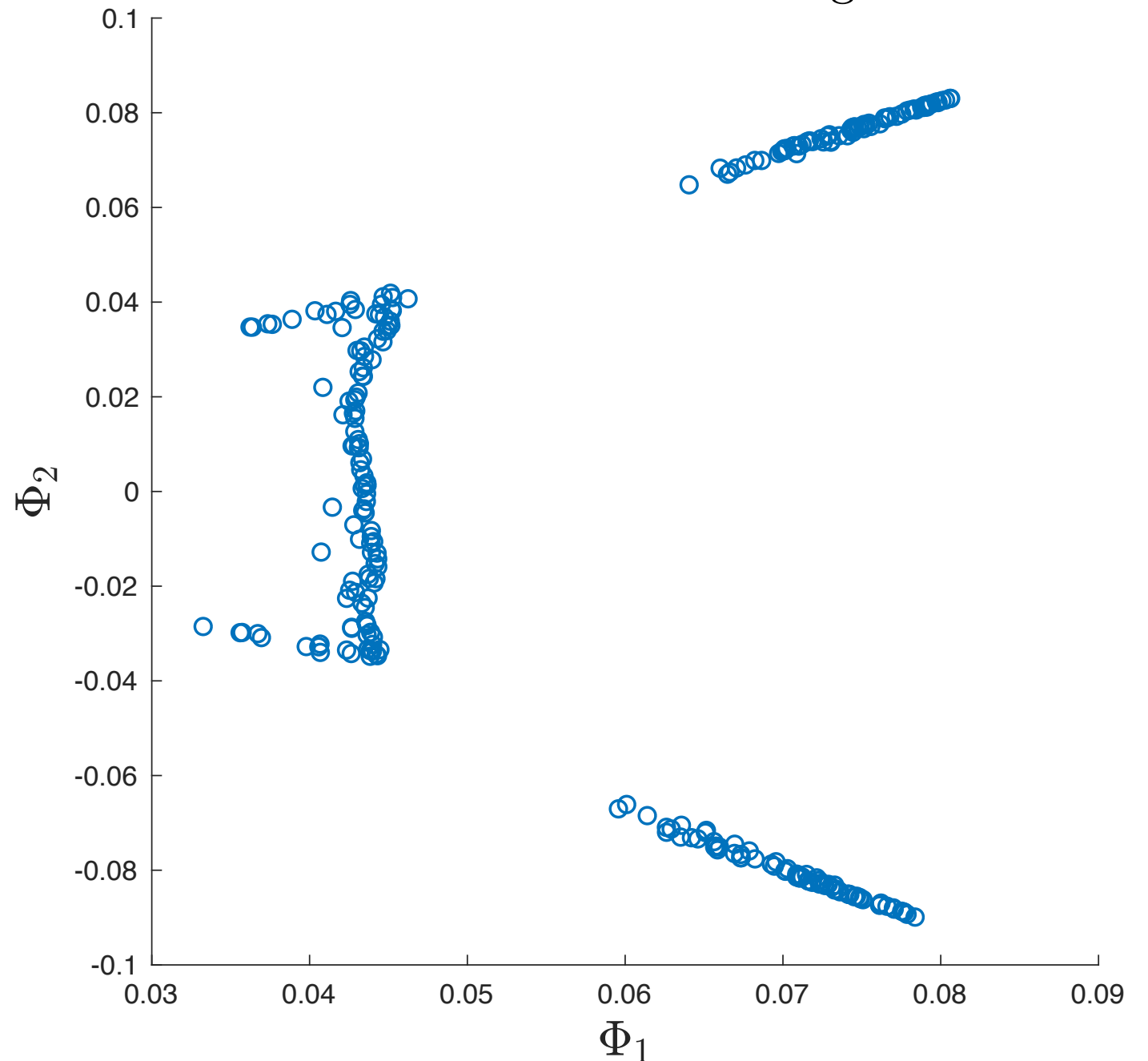
Step 4: Embed the data as

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

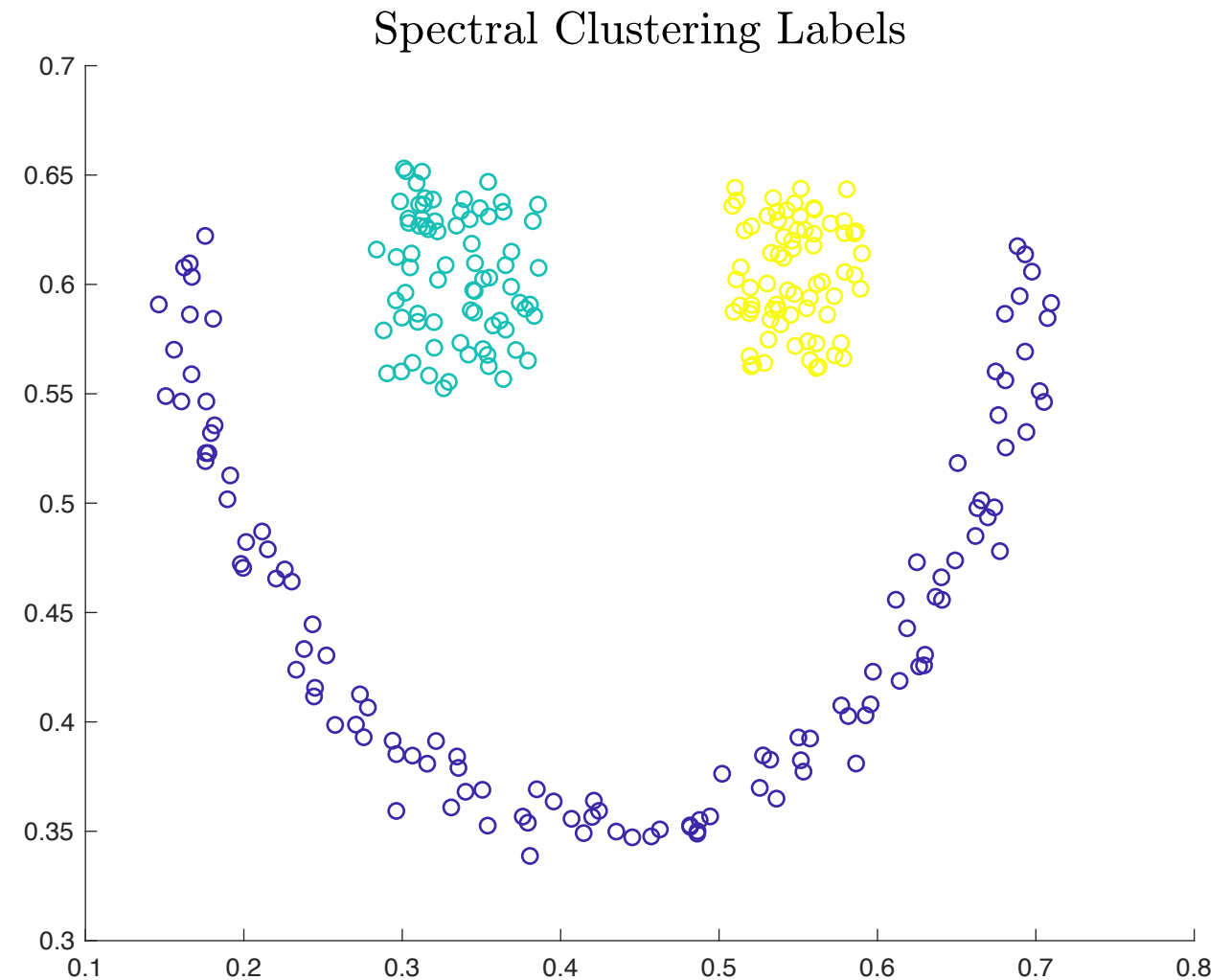
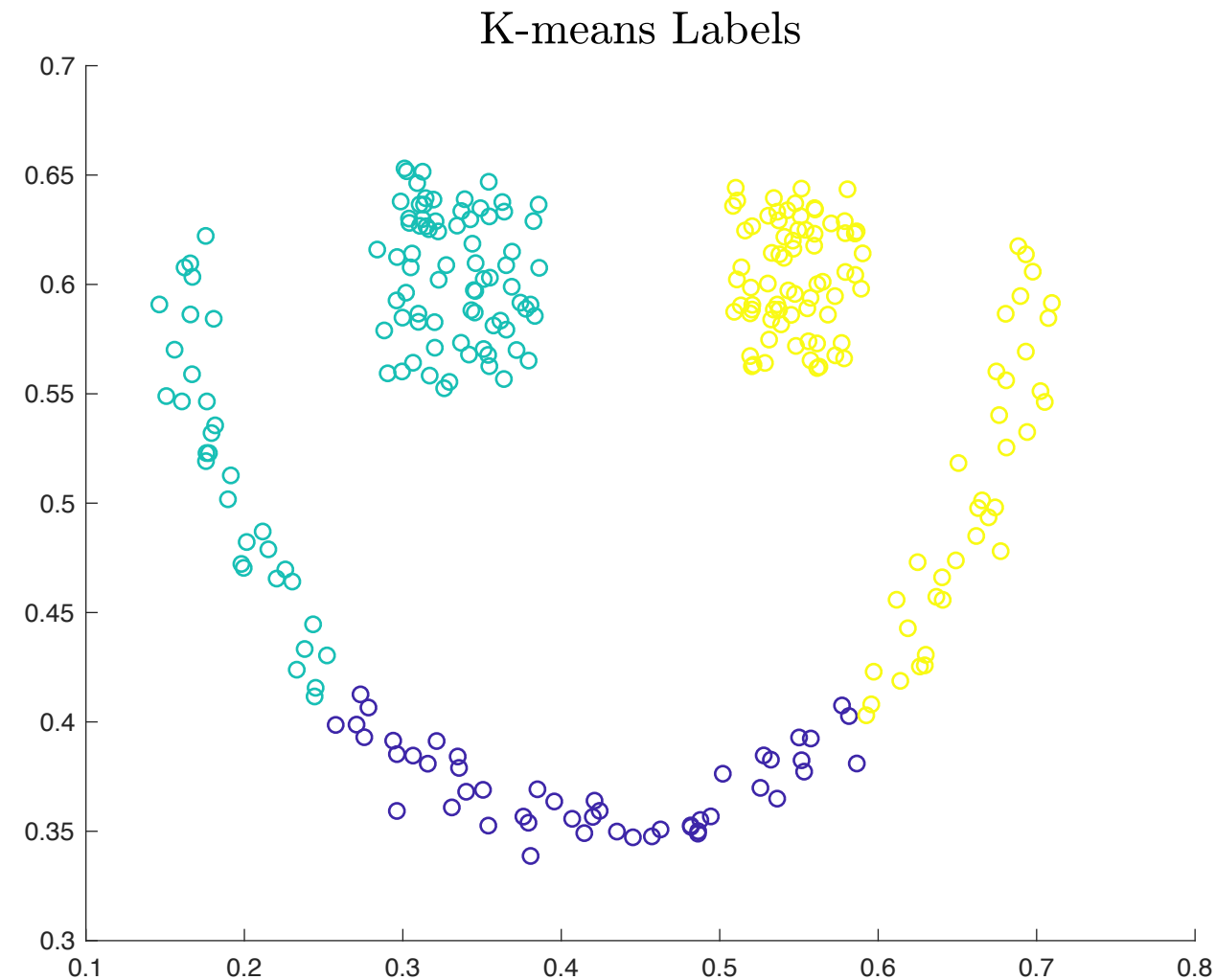
then run K-means. Note

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

Low-dimensional Embedding from L



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: $K \approx \arg \max_k \lambda_{k+1} - \lambda_k$.

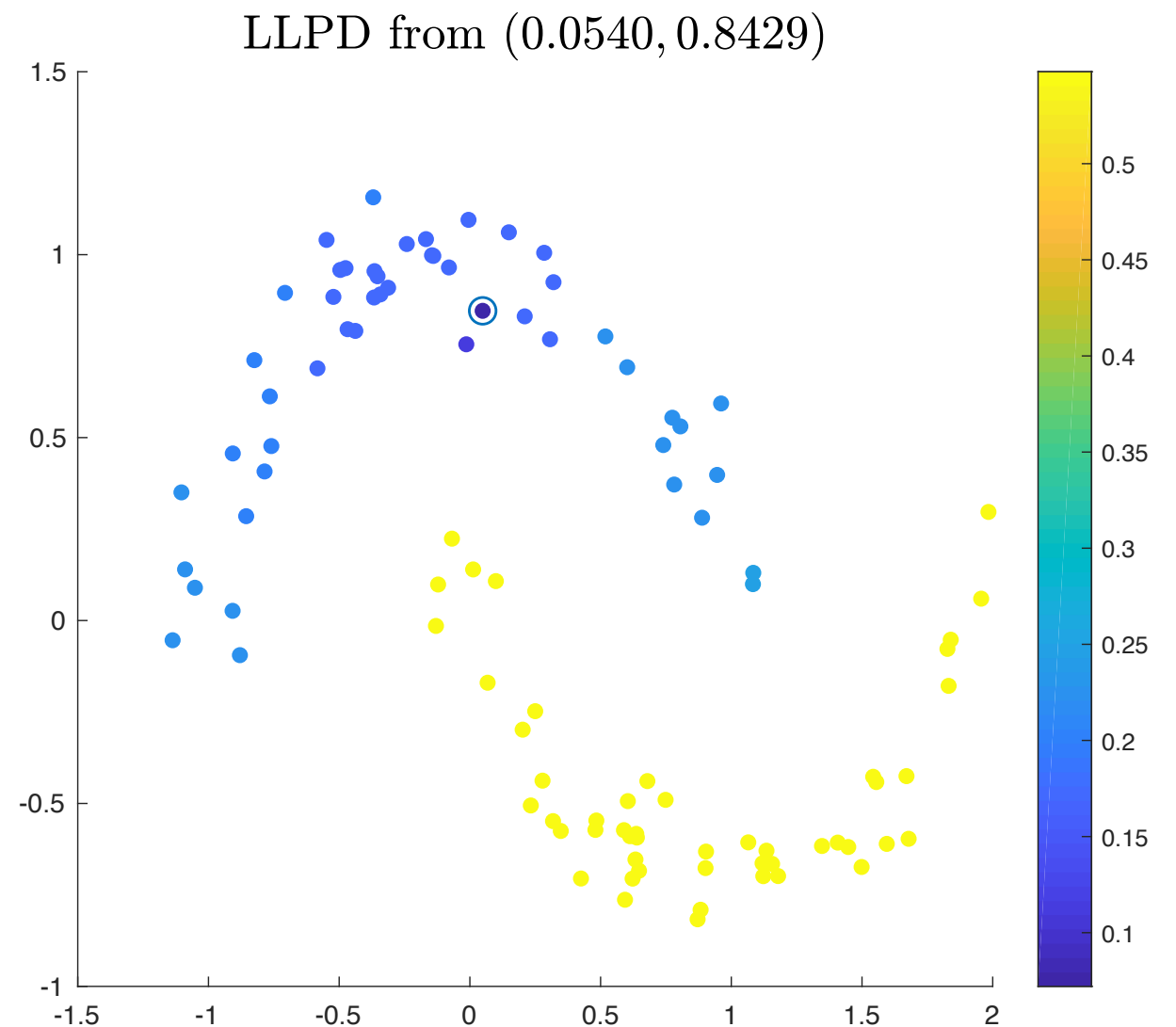
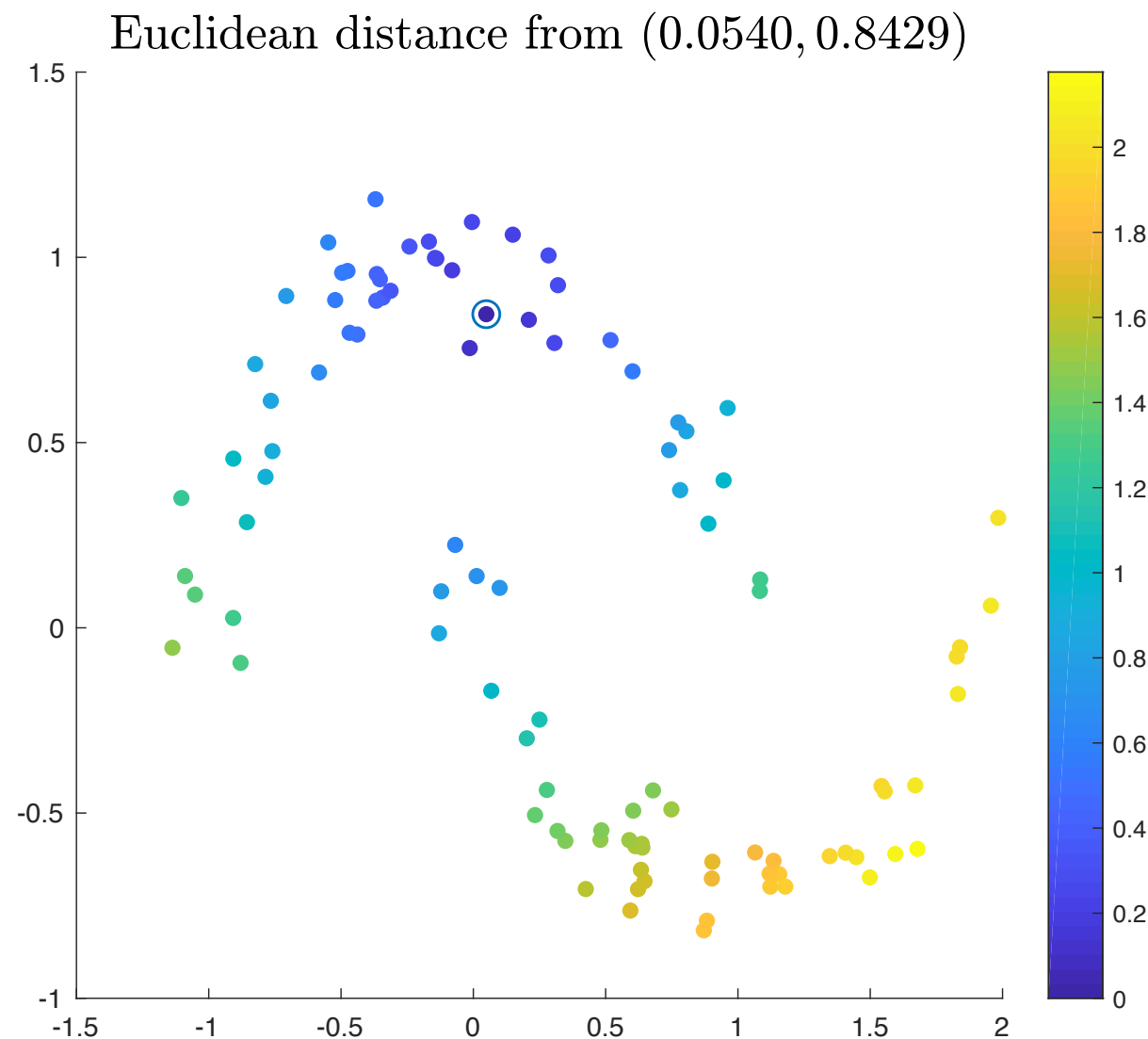
Longest Leg Path Distance

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:

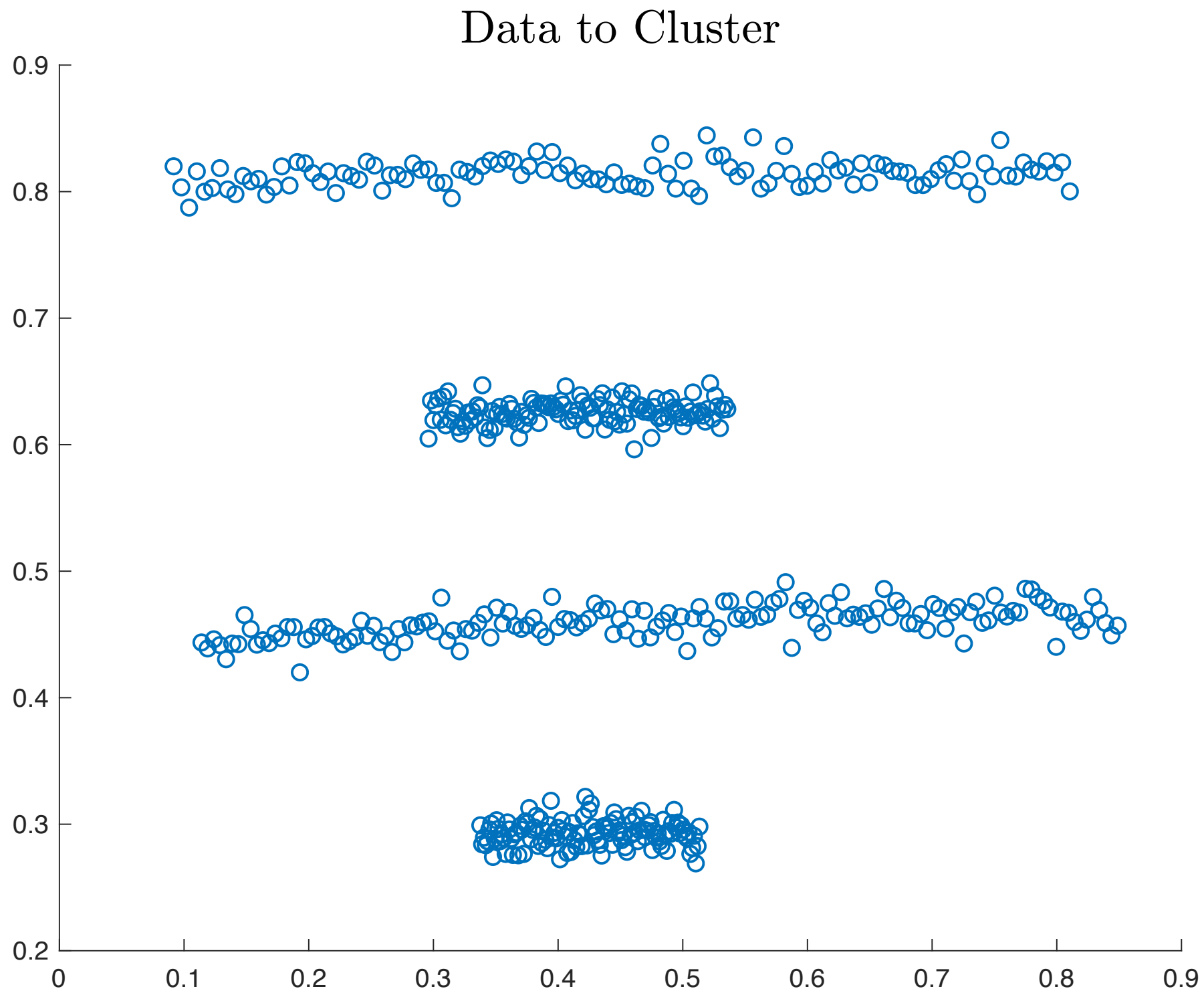
$$d_{\ell\ell}(x_i, x_s) = \min_{\{y_j\}_{j=1}^L \in \mathcal{P}(x_i, x_s)} \max_{j=1,2,\dots,L-1} \|y_{j+1} - y_j\|_2,$$

- 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



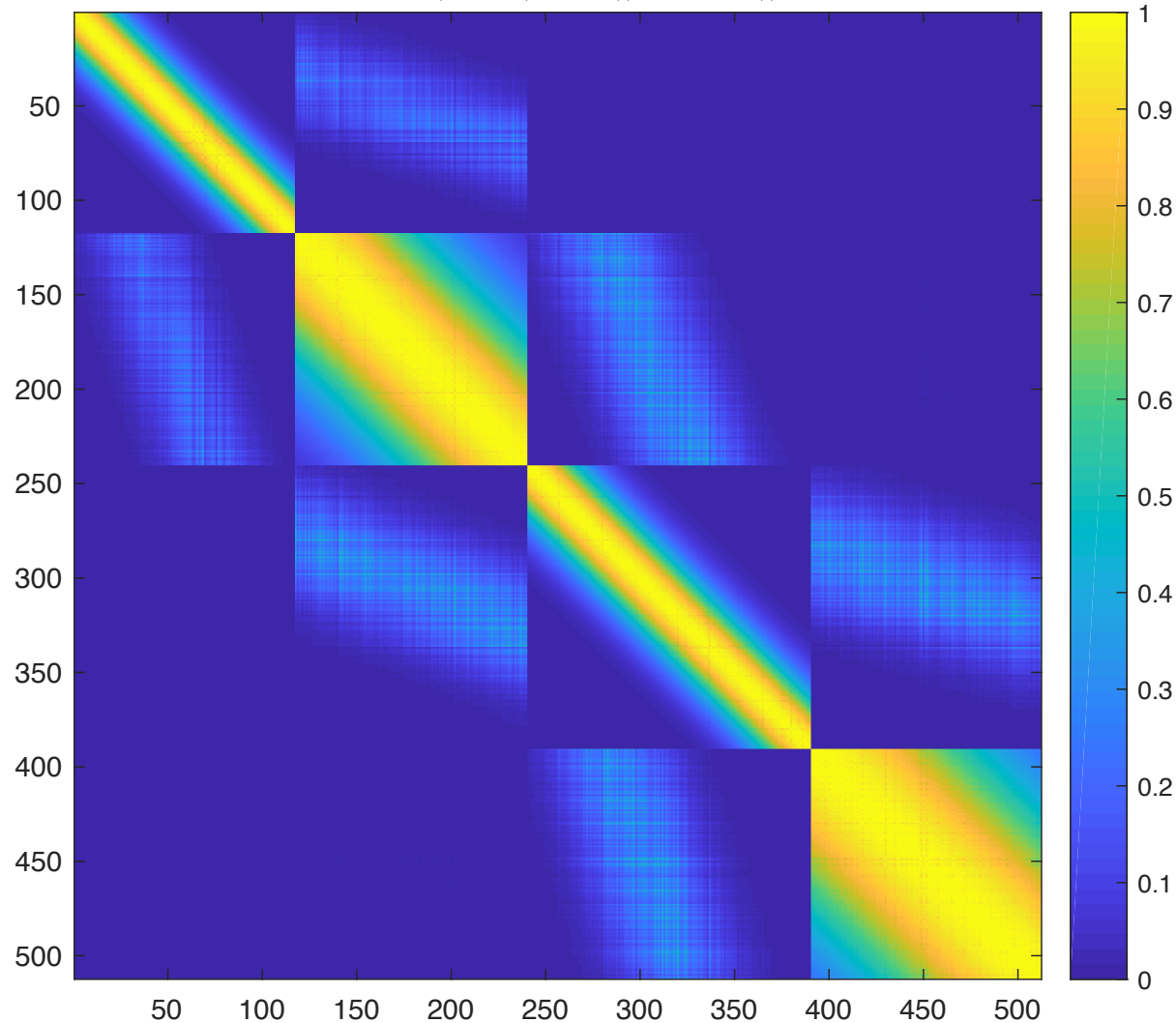
Data Well-Suited for LLPD



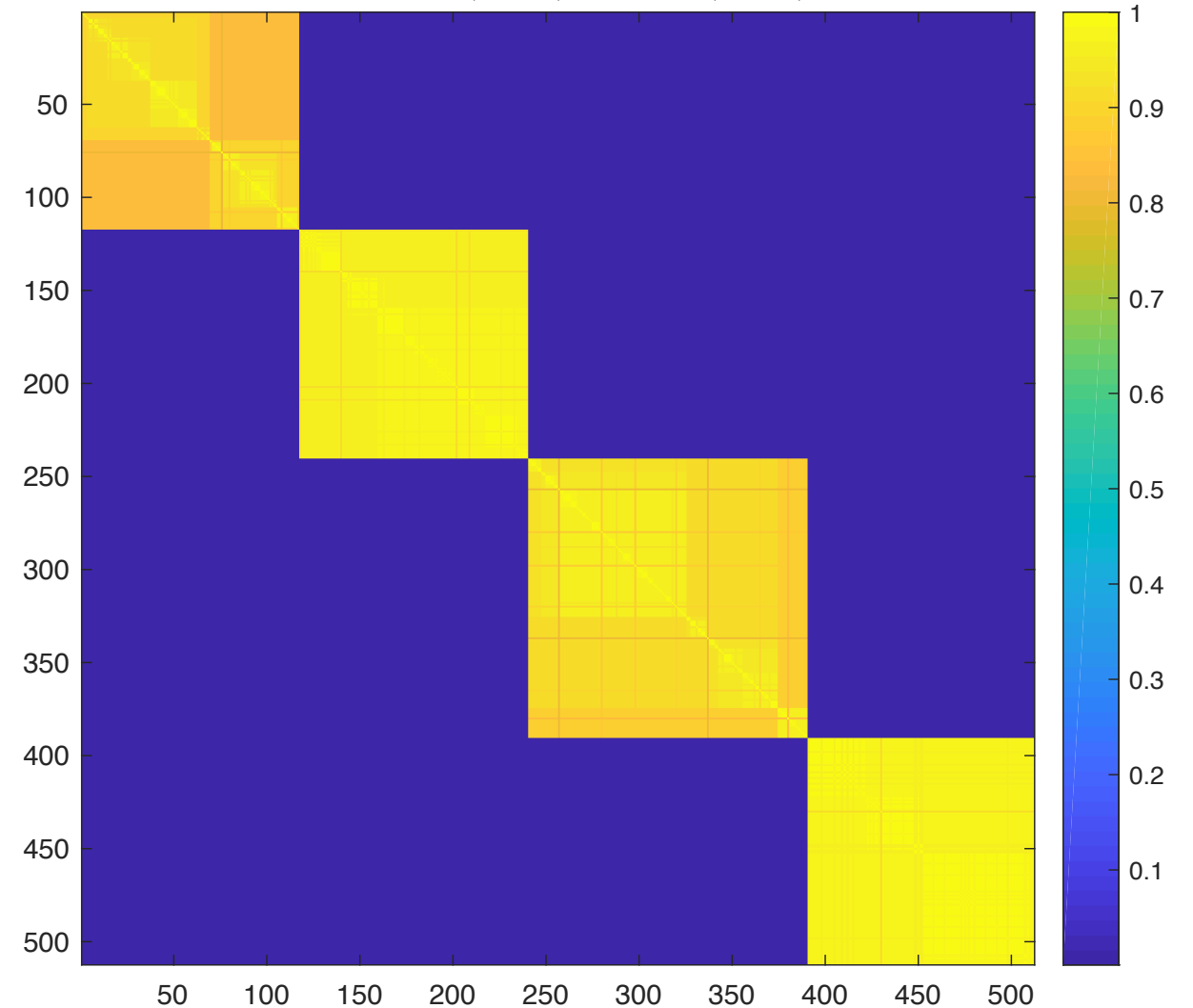
LLPD Weight Matrix

- 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.

Weight matrix, $d(x, y) = \|x - y\|_2$, $\sigma = 0.1474$



Weight matrix, $d(x, y) = d_{\ell\ell}(x, y)$, $\sigma = 0.06$



Low Dimensional, Large Noise (LDLNN) 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:

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

Low-dimensional

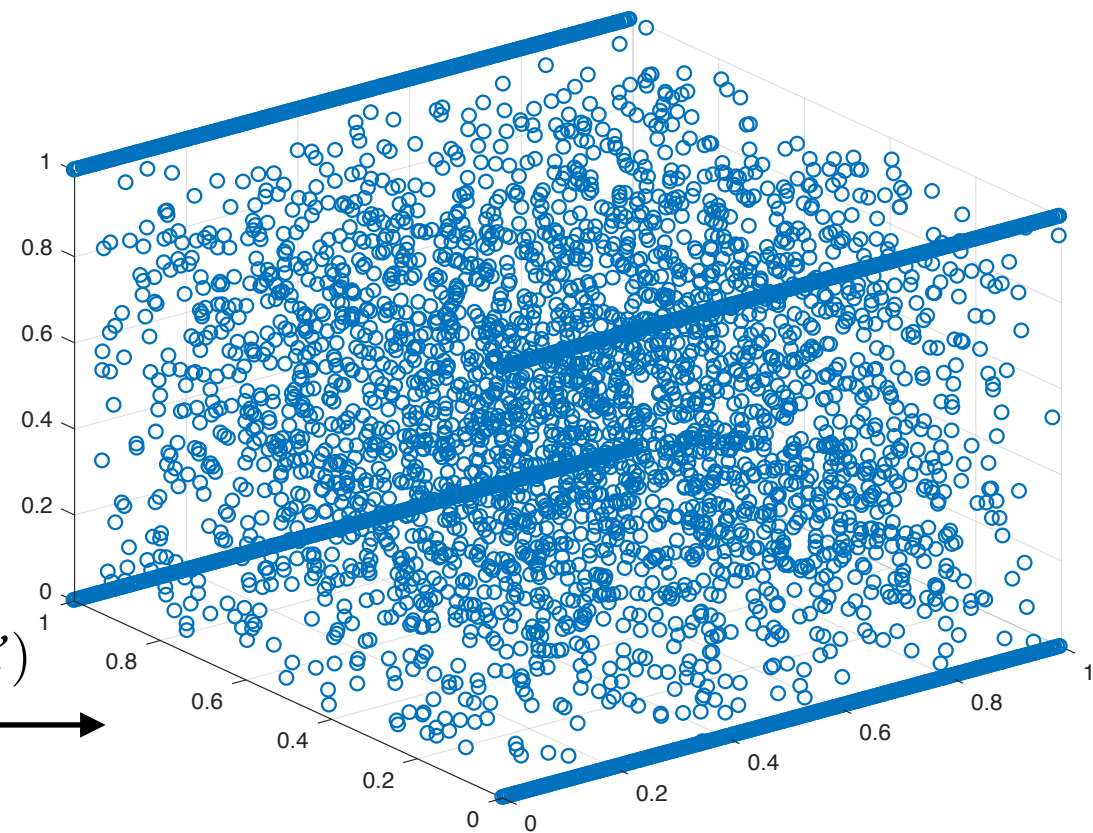
$$\begin{aligned} \mathcal{X}_1, \dots, \mathcal{X}_K &\subset \mathcal{X} \subset \mathbb{R}^D \\ \mathcal{X}_1, \dots, \mathcal{X}_K &\in \mathcal{S}_d(\kappa, \epsilon_0) \\ \delta &= \min_{k \neq k'} \text{dist}(\mathcal{X}_k, \mathcal{X}_{k'}) \end{aligned}$$

Large noise

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

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

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



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

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

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

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

$$C_3 \theta \leq \sigma \leq C_4 \delta \quad (2)$$

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 .

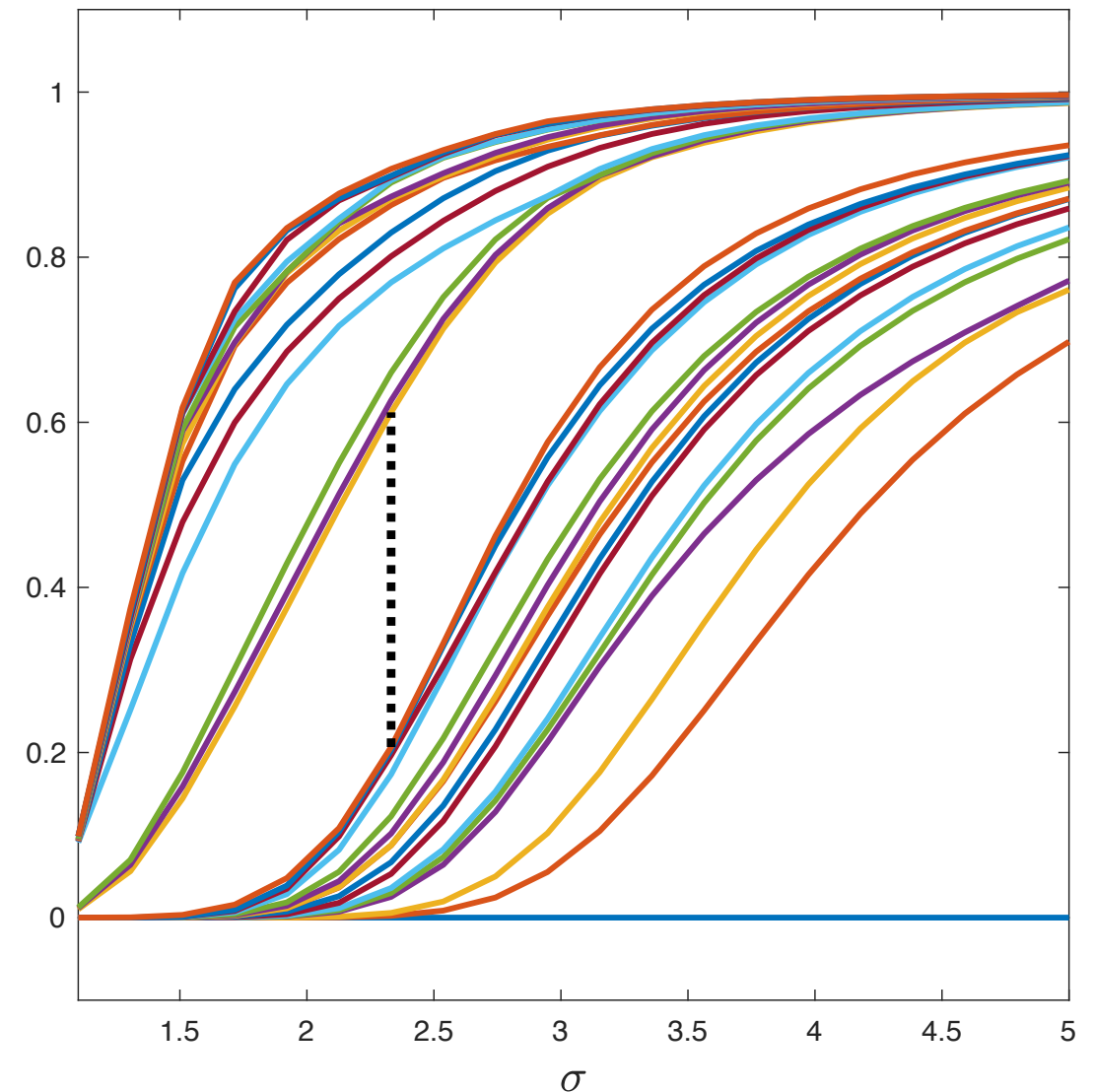
The constants $\{C_i\}_{i=1}^4$ depend on geometric quantities but do not depend on $n_1, \dots, n_K, \tilde{n}, \theta, \sigma$.

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.

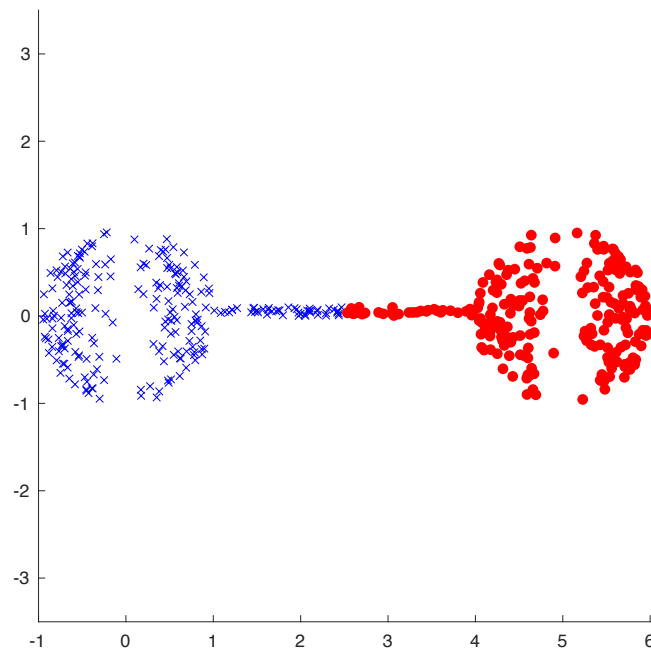
Interpolating Between Geometry and Density

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

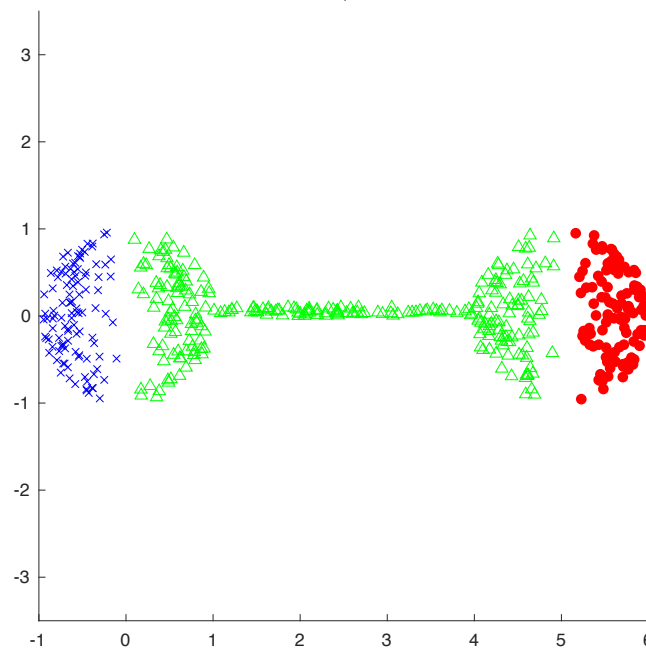
$$\ell_p(x, y) = \min_{\pi = \{x_{i_j}\}_{j=1}^T} \left(\sum_{j=1}^{T-1} \|x_{i_j} - x_{i_{j+1}}\|^p \right)^{\frac{1}{p}},$$

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.

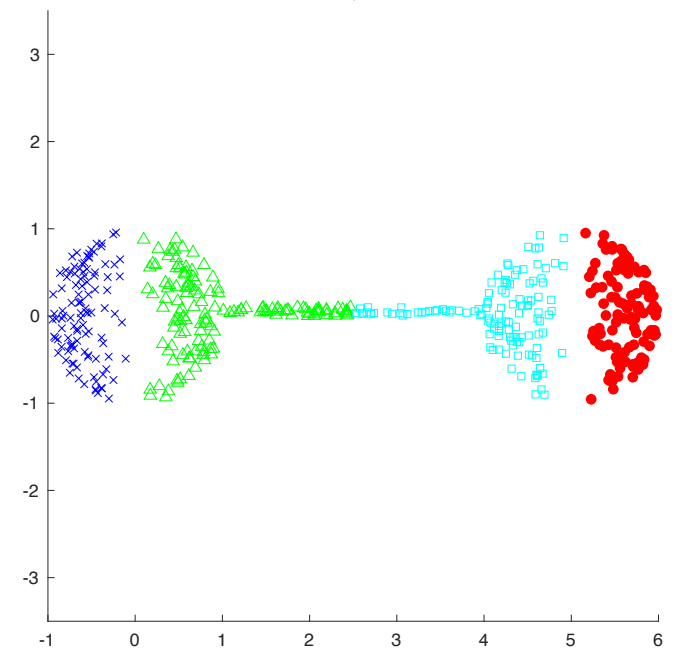
Raw Data, 2 Classes



Raw Data, 3 Classes



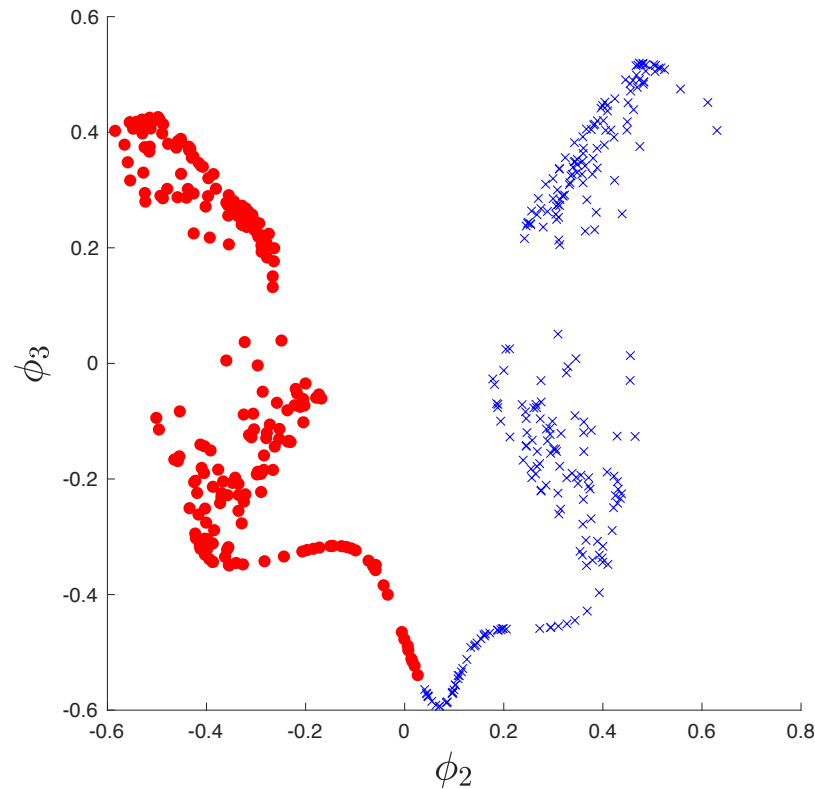
Raw Data, 4 Classes



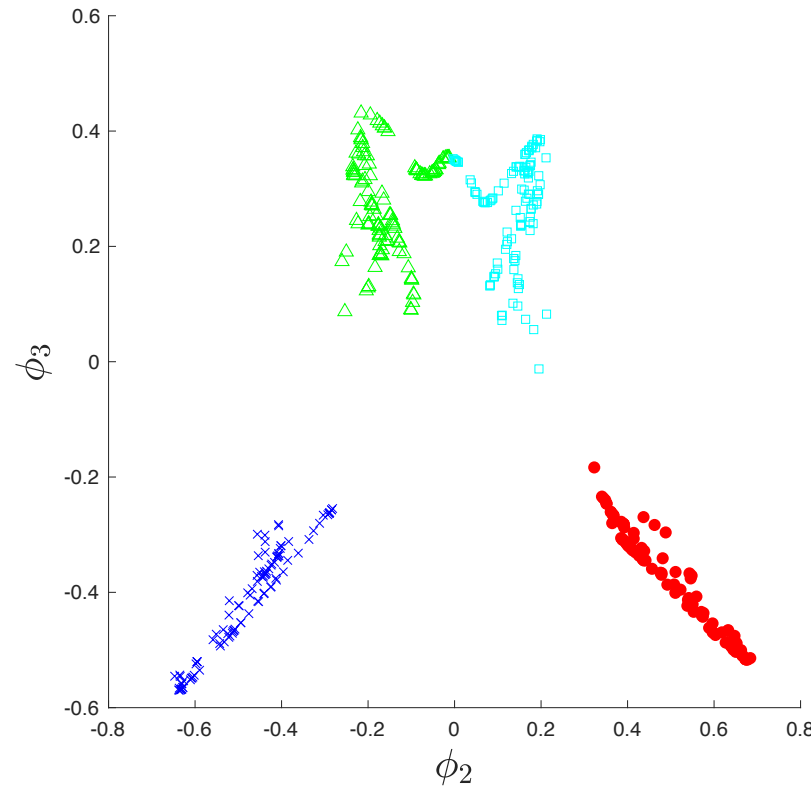
How to balance density and geometry when both matter?

Role of p

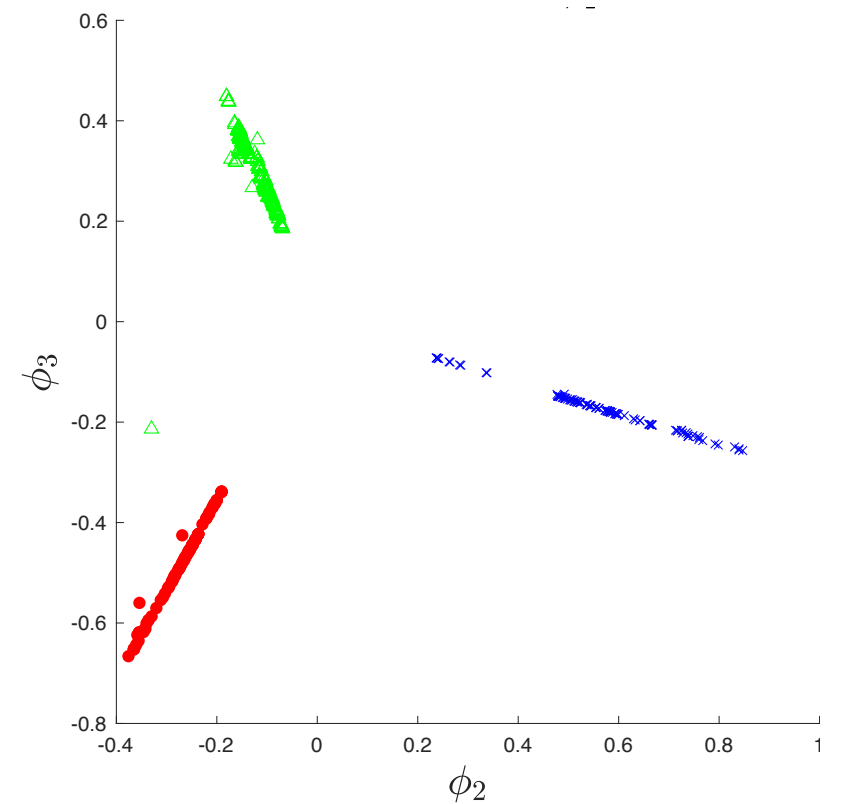
$p = 1.2$



$p = 2$



$p = 5$



- 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*

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

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:

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

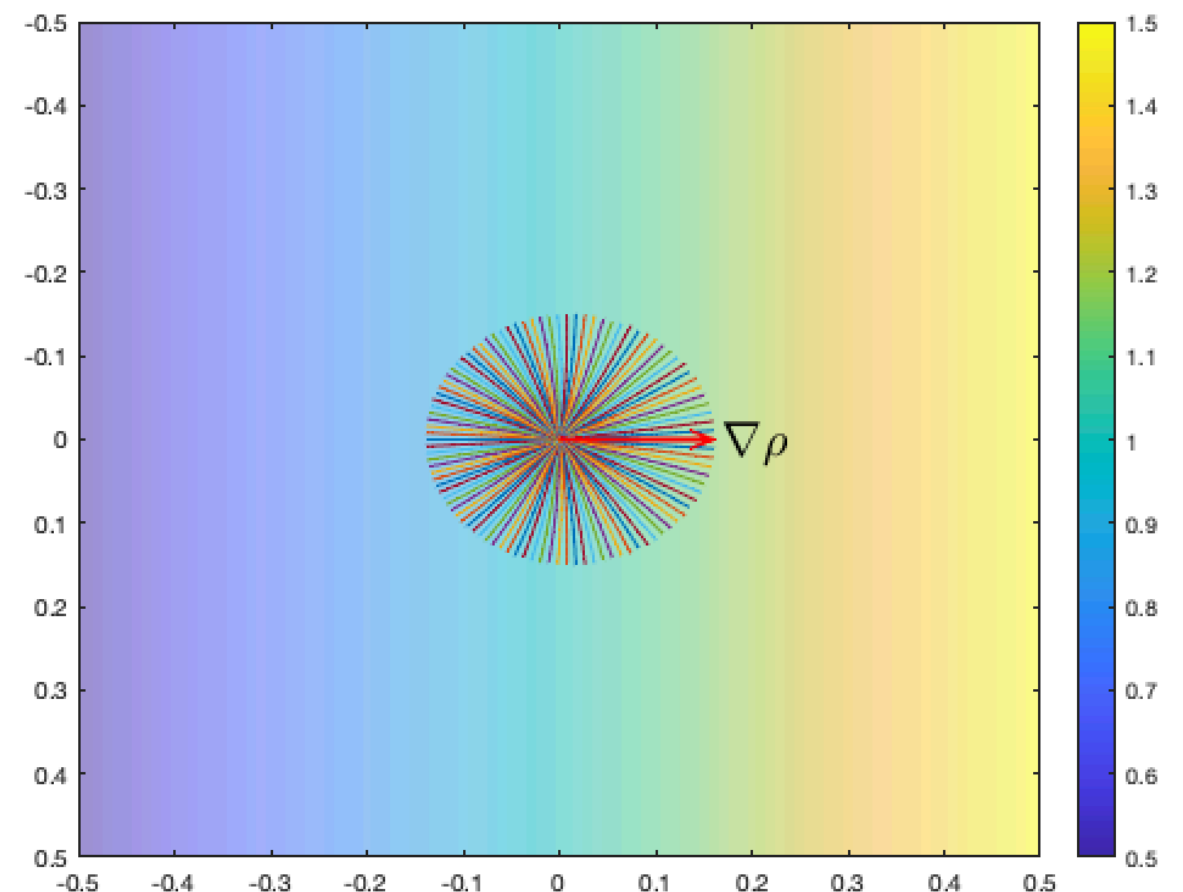
where $\gamma : [0, 1] \rightarrow \mathcal{M}$ is a $\mathcal{C}^1$ path with $\gamma(0) = x, \gamma(1) = y$.

- 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 Riemannian 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.



Collaborators



N. García Trillos (Wisconsin Statistics)



A. Little (Utah Math)



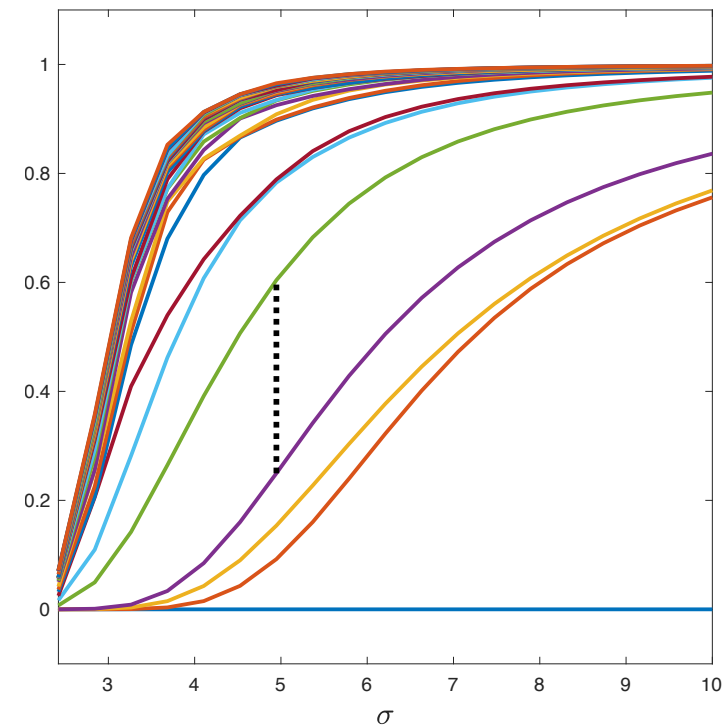
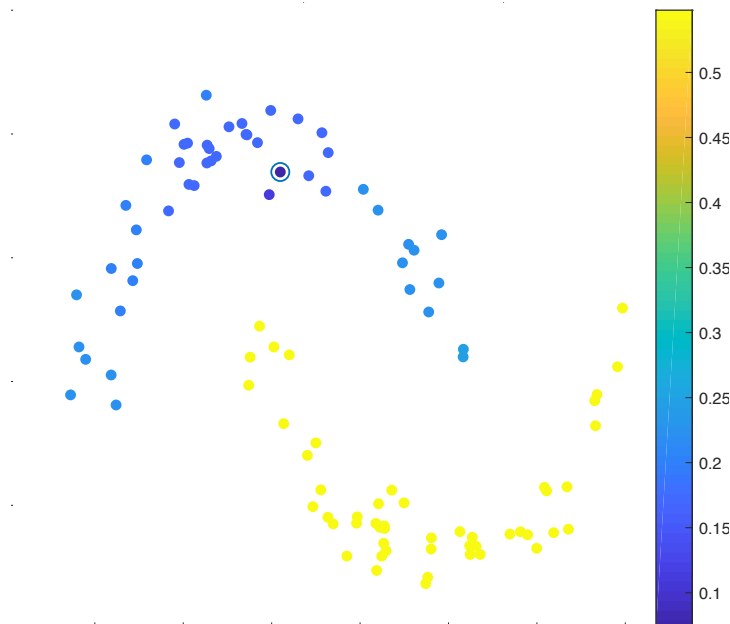
M. Maggioni, (JHU Applied Math)



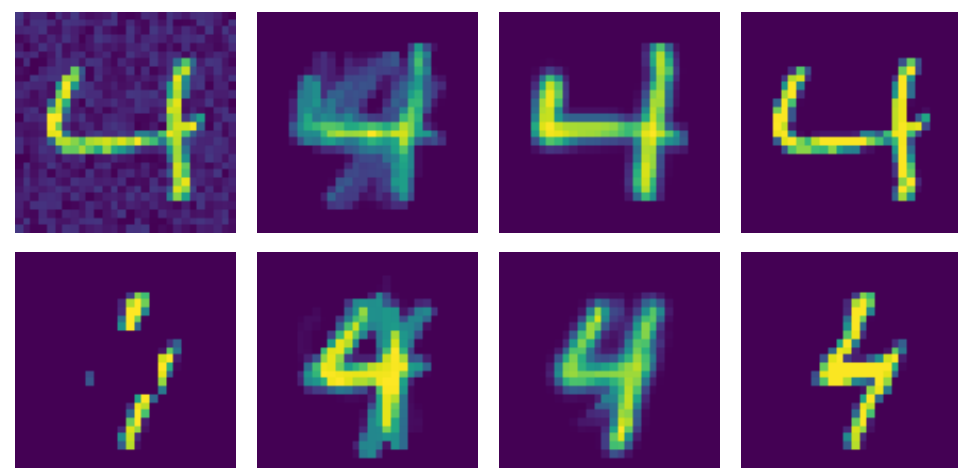
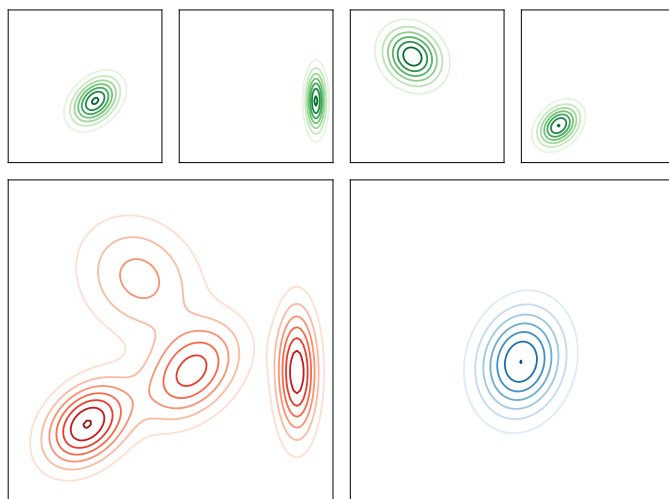
D. McKenzie (Mines Applied Math)

Outline

- Part I: Unsupervised Learning with Data-Driven Metrics

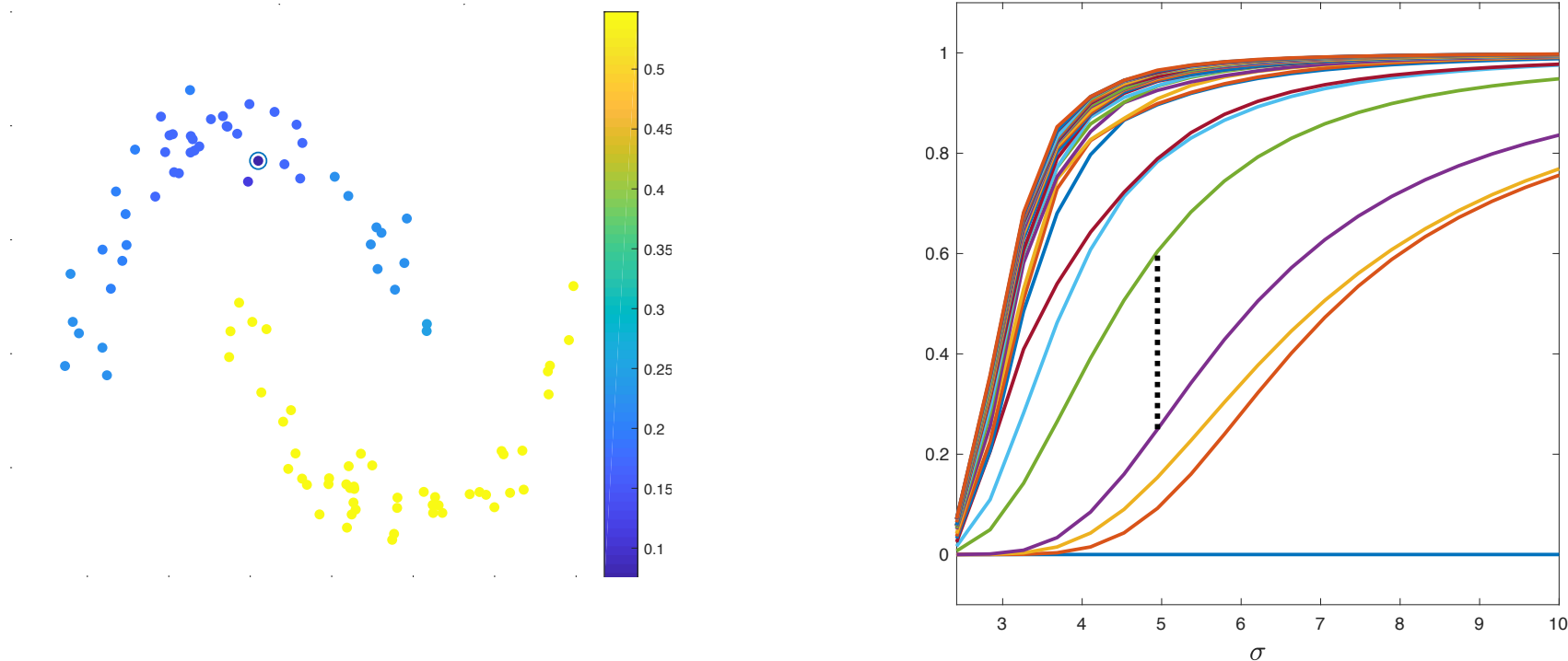


- Part II: Representation Learning in Wasserstein Space

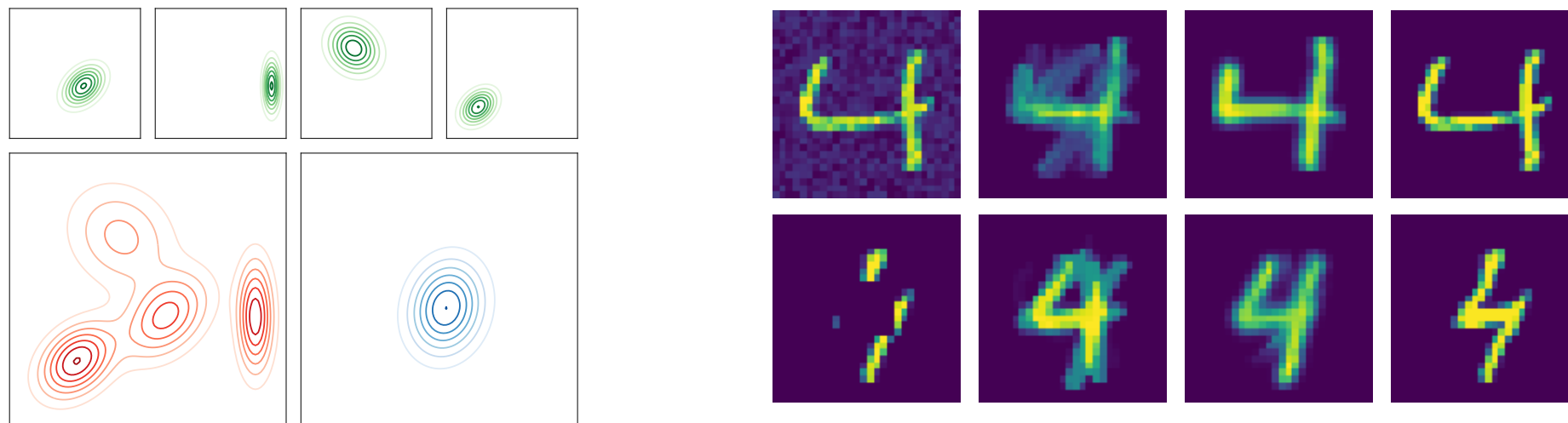


Outline

- Part I: Unsupervised Learning with Data-Driven Metrics

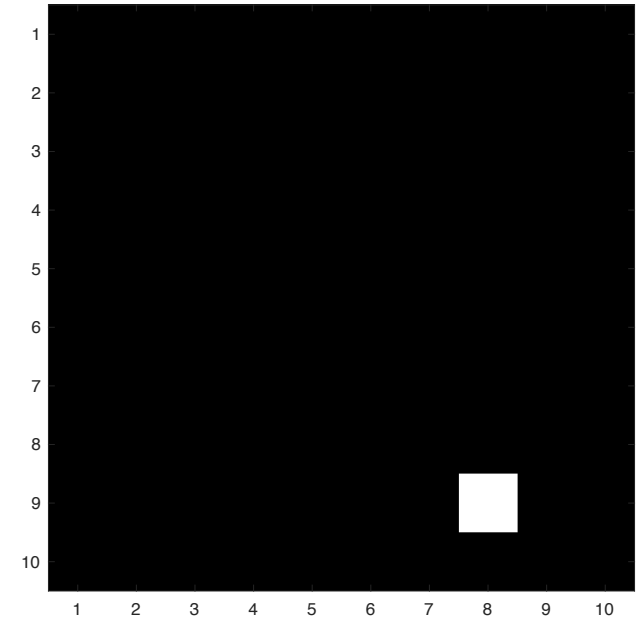
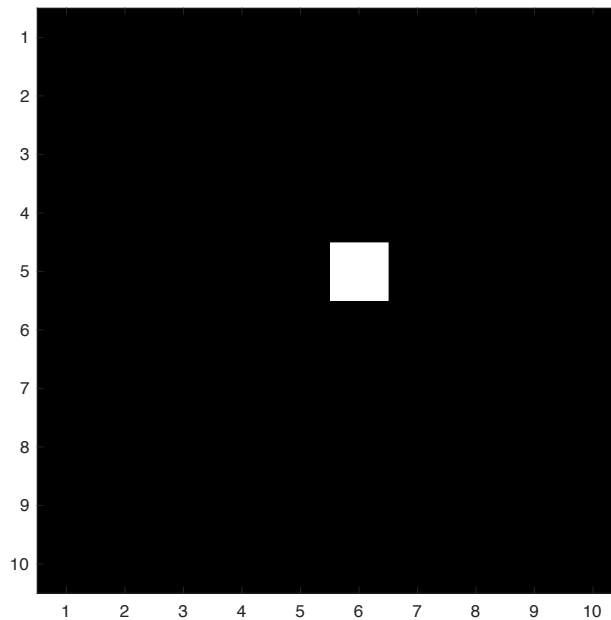
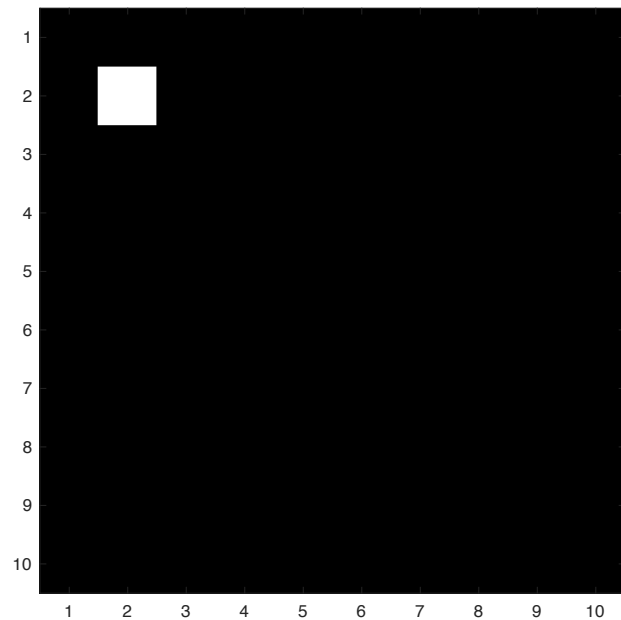


- Part II: Representation Learning in Wasserstein Space



Beyond the Euclidean

- Manifold learning methods based on local Euclidean distances may be insufficient to capture the geometry of certain data.
- **Toy Example:** black and white images with single white pixel:



- Everything is equally far in Euclidean distance, and therefore in any graph metric.
- Need to capture the distance *between the support of these images*.

Data as Measures & 2-Wasserstein Space

- Let $\mathcal{P}_{2,\text{ac}}(\mathbb{R}^d)$ denote the space of absolutely continuous measures (i.e., having density with respect to the Lebesgue measure) with finite second moment.
- For $\mu, \nu \in \mathcal{P}_{2,\text{ac}}(\mathbb{R}^d)$, the 2-Wasserstein metric is

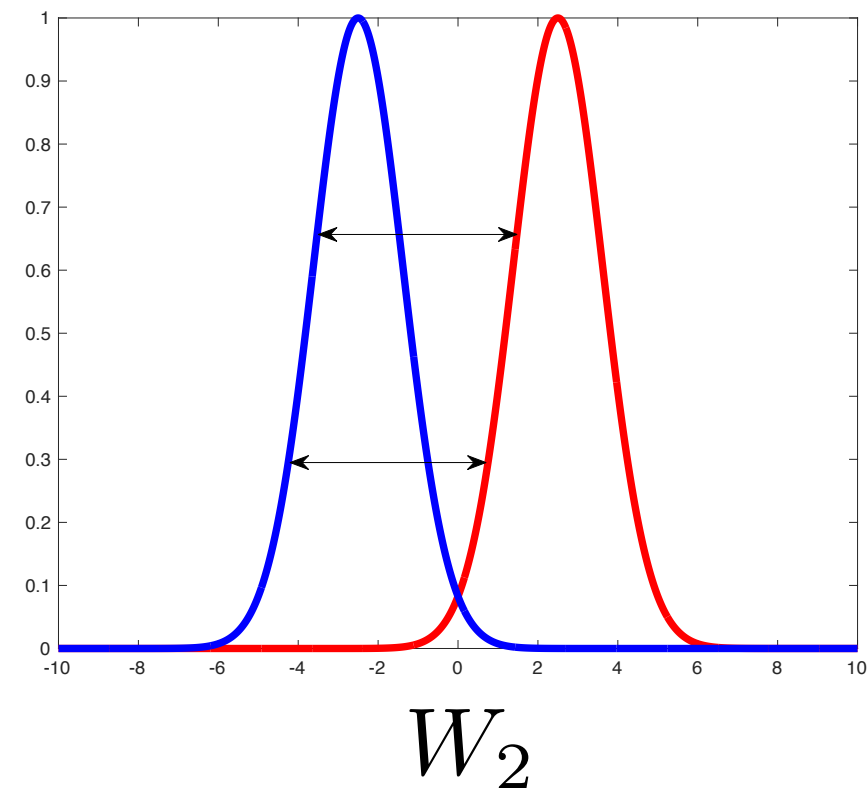
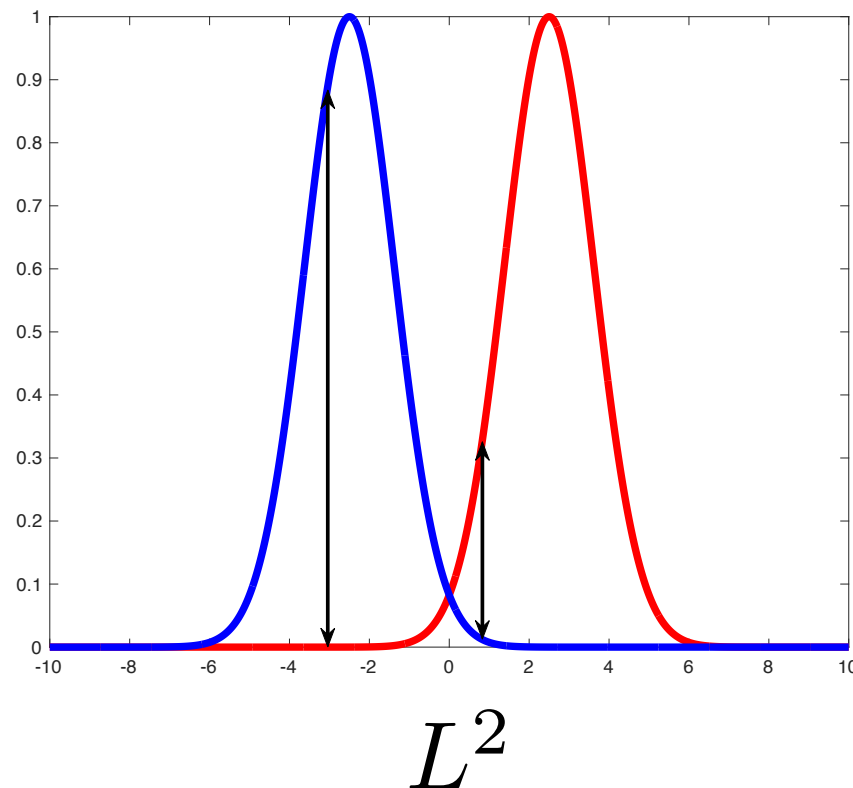
$$W_2^2(\mu, \nu) = \min_{T \# \mu = \nu} \int_{\mathbb{R}^d} \|T(x) - x\|_2^2 d\mu(x)$$

where the minimization is over all maps $T : \mathbb{R}^d \rightarrow \mathbb{R}^d$ that pushforward μ onto ν :

$$T \# \mu = \nu \iff \nu[B] = \mu[T^{-1}(B)] \text{ for all Borel sets } B.$$

Optimal Transport Maps

- Pushforwards transfer mass from one distribution to another.



- The T^* realizing $W_2^2(\mu, \nu) = \int_{\mathbb{R}^d} ||T^*(x) - x||_2^2 d\mu(x)$

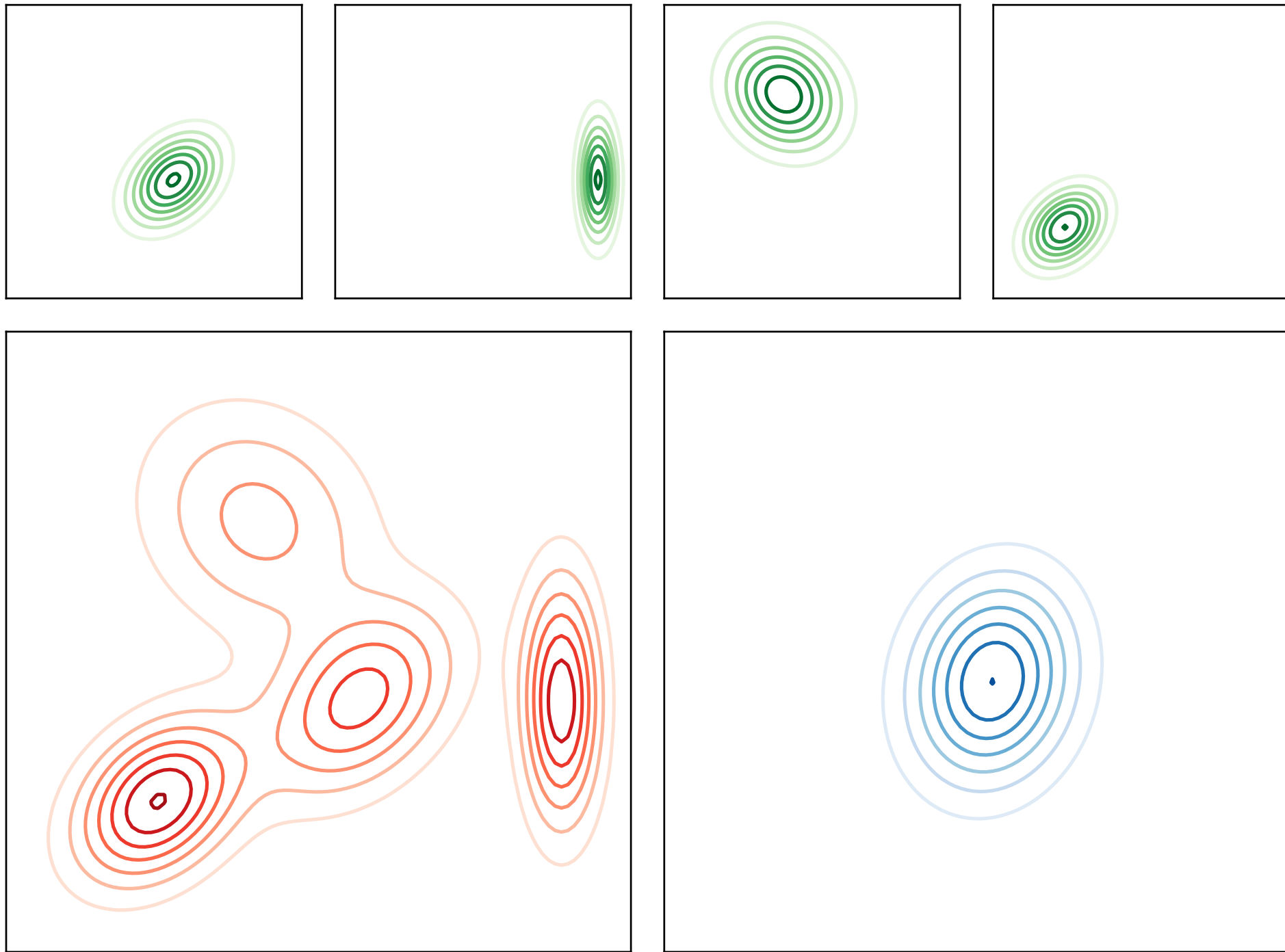
is the *optimal transport map*.

Averaging in $\mathcal{W}_2$: Barycenters

- Let $\Delta^p = \left\{ \lambda = (\lambda_1, \dots, \lambda_p) \in \mathbb{R}^p : \lambda_i \geq 0, \sum_{i=1}^p \lambda_i = 1 \right\}$.
- For measures $\{\mu_i\}_{i=1}^p \subset \mathcal{P}_{2,\text{ac}}(\mathbb{R}^d)$ and coordinates $\lambda \in \Delta^p$, define the *2-Wasserstein barycenter* as

$$\nu_\lambda = \arg \min_{\nu \in \mathcal{P}_{2,\text{ac}}(\mathbb{R}^d)} \frac{1}{2} \sum_{i=1}^p \lambda_i W_2^2(\nu, \mu_i).$$

Barycenters Preserve Shape



$$\sum_{i=1}^p \lambda_i \mu_i$$

$$\arg \min_{\nu \in \mathcal{P}_{2,\text{ac}}(\mathbb{R}^d)} \frac{1}{2} \sum_{i=1}^p \lambda_i W_2^2(\nu, \mu_i)$$

The *Barycentric Coding Model*

- Let $\text{Bary}(\{\mu_i\}_{i=1}^p) = \{\nu_\lambda : \lambda \in \Delta^p\}$ be the set of all barycenters that can be generated from $\{\mu_i\}_{i=1}^p$.
- We denote by the *barycentric coding model (BCM)* the identification of a measure

$$\mu_0 \in \text{Bary}(\{\mu_i\}_{i=1}^p)$$

with its coordinates $\lambda \in \Delta^p$.

- $\text{Bary}(\{\mu_i\}_{i=1}^p)$ can be thought of as the “span” of the reference measures, but with respect to the geometry of Wasserstein space.

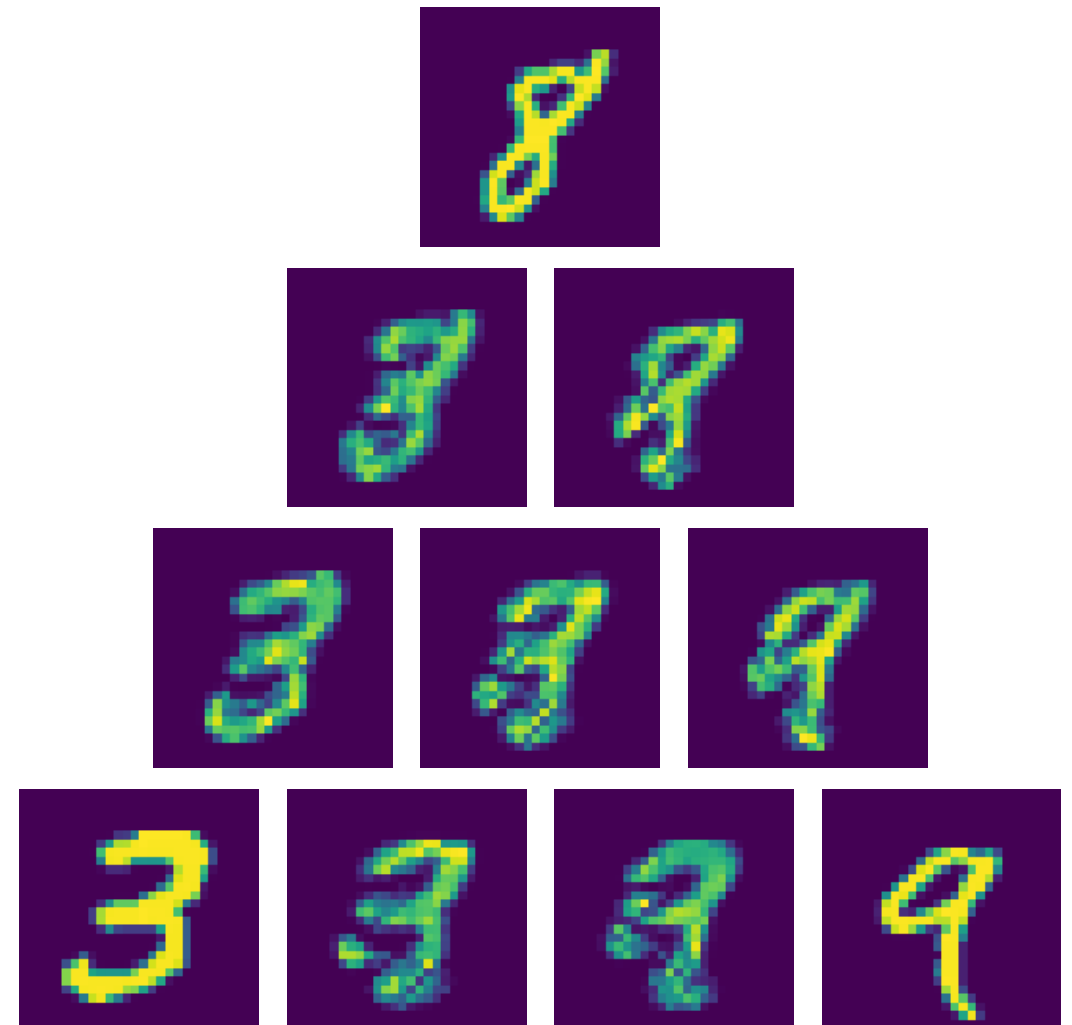
The Analysis Problem

- Given a measure μ_0 and $\{\mu_i\}_{i=1}^p$ reference measures, the *analysis problem* solves

$$\arg \min_{\lambda \in \Delta^p} W_2^2(\mu_0, \nu_\lambda).$$

- If $\mu_0 \in \text{Bary}(\{\mu_i\}_{i=1}^p)$, then:

$$\min_{\lambda \in \Delta^p} W_2^2(\mu_0, \nu_\lambda) = 0.$$



Exact Coefficients in BCM via Quadratic Program

Theorem. (Werenski et al.) Suppose $\{\mu_i\}_{i=0}^p$ are sufficiently regular. Then $\mu_0 \in \text{Bary}(\{\mu_i\}_{i=1}^p)$ if and only if

$$\min_{\lambda \in \Delta^p} \lambda^T A \lambda = 0,$$

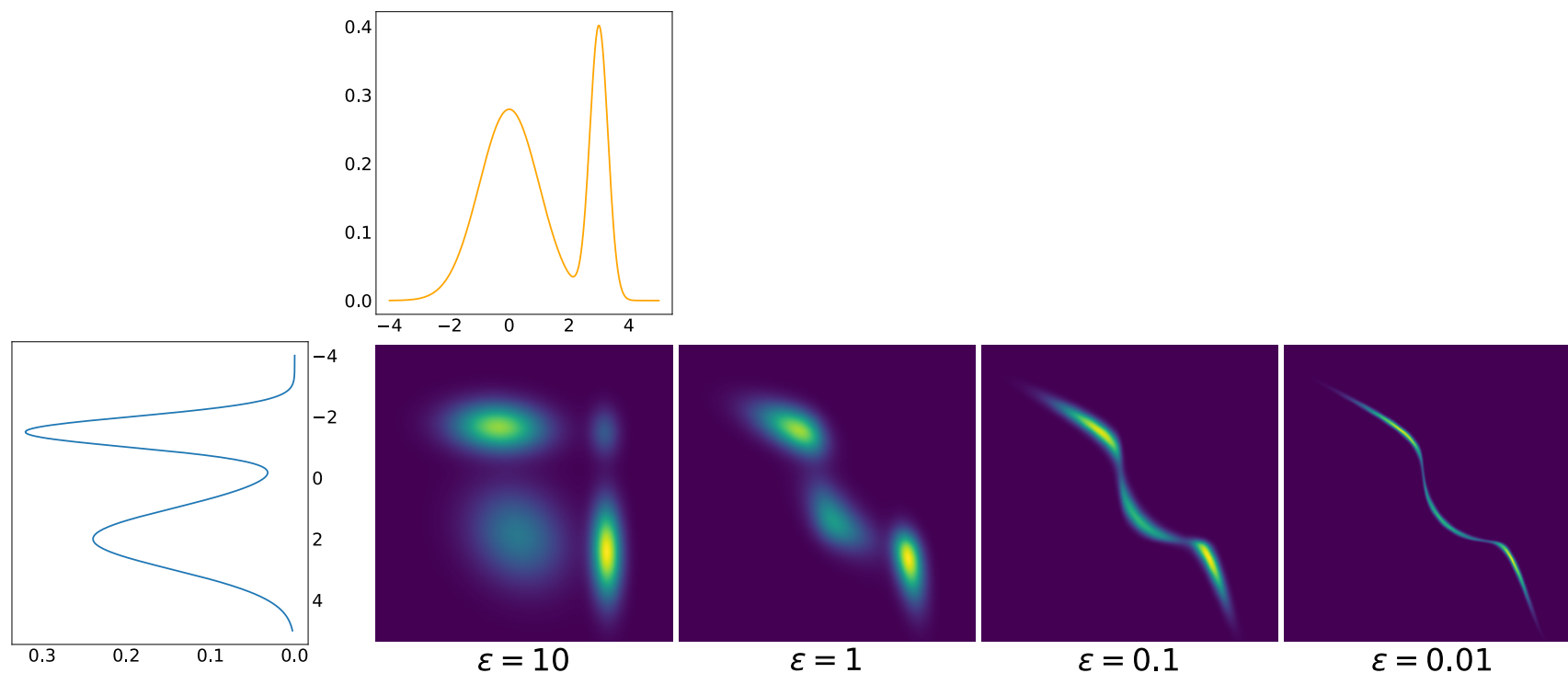
where $A \in \mathbb{R}^{p \times p}$ is given by $A_{ij} = \int_{\mathbb{R}^d} \langle T_i(x) - \text{Id}(x), T_j(x) - \text{Id}(x) \rangle d\mu_0(x)$ for T_i the optimal transport map between μ_0 and μ_i . Furthermore, if the minimum value is 0 and λ_* is an optimal argument, then $\mu_0 = \nu_{\lambda_*}$.

- Requires OT maps between candidate barycenter and reference measures.
- We will estimate these with entropy-regularized optimal transport.

Estimating Transport Maps with Entropy-Regularization

- To estimate OT maps, we use *entropy-regularized optimal transport*.
- We solve the regularized problem:

$$\pi_\epsilon = \arg \min_{\pi \in \Pi(\mu, \nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|^2 d\pi(x, y) + \epsilon \text{KL}(\pi \parallel \mu \otimes \nu)$$



Estimating Transport Maps with Entropy-Regularization

- We then use conditional expectation of π_ϵ as a plug-in estimator for an OT map:

$$T_\epsilon(x) = \mathbb{E}_{(X,Y) \sim \pi_\epsilon} [Y \mid X = x]$$

- The entropy-regularized map has parametric sample complexity; one can also develop a BCM entirely in terms of entropy-regularized maps.

Estimating BCM Coordinates from Data

Algorithm 1 Estimate λ

Input: i.i.d. samples $\{X_1, \dots, X_{2n}\} \sim \mu_0, \{\{Y_1^i, \dots, Y_n^i\} \sim \mu_i : i = 1, \dots, p\}$, regularization parameter $\epsilon > 0$.

for $i = 1, \dots, p$ **do**

Set $M^i \in \mathbb{R}^{n \times n}$ with $M_{jk}^i = \frac{1}{2} \|X_j - Y_k^i\|_2^2$.

Solve for g^i as the optimal g in

$$\max_{f, g \in \mathbb{R}^n} \frac{1}{n} \sum_{j=1}^n f_j + \frac{1}{n} \sum_{k=1}^n g_k - \frac{\epsilon}{n^2} \sum_{j,k} \exp((f_j + g_k - M_{jk}^i)/\epsilon)$$

entropy-regularized
dual formulation
(Kantorovich; Cuturi)

$$\text{Define } \hat{T}_i(x) = \frac{\sum_{i=1}^n Y_i \exp\left(\frac{1}{\epsilon}(g^i(Y_i) - \frac{1}{2}\|x - Y_i\|_2^2)\right)}{\sum_{i=1}^n \exp\left(\frac{1}{\epsilon}(g^i(Y_i) - \frac{1}{2}\|x - Y_i\|_2^2)\right)}.$$

empirical entropic map
(Pooladian, Niles-Weed)

end for

Set $\hat{A} \in \mathbb{R}^{p \times p}$ to be the matrix with entries

$$\hat{A}_{ij} = \frac{1}{n} \sum_{k=n+1}^{2n} \langle \hat{T}_i(X_k) - X_k, \hat{T}_j(X_k) - X_k \rangle$$

Return $\hat{\lambda} = \arg \min_{\lambda \in \Delta^p} \lambda^T \hat{A} \lambda$.

Consistency of Coefficient Estimation

Theorem. (Werenski et al.) Let $i, j \in \{1, \dots, p\}$ and suppose that μ_i, μ_j, μ_0 are supported on bounded domains and that the maps T_i and T_j are sufficiently regular. Let $X_1, \dots, X_{2n} \sim \mu_0, Y_1, \dots, Y_n \sim \mu_i, Z_1, \dots, Z_n \sim \mu_j$. For an appropriately chosen ϵ , let $\hat{T}_i$ and $\hat{T}_j$ be the entropic maps computed using $\{X_i\}_{i=1}^n, \{Y_i\}_{i=1}^n, \{Z_i\}_{i=1}^n$. Then we have

$$\mathbb{E} \left[\left| A_{ij} - \frac{1}{n} \sum_{k=n+1}^{2n} \langle \hat{T}_i(X_k) - X_k, \hat{T}_j(X_k) - X_k \rangle \right| \right] \lesssim \frac{1}{\sqrt{n}} + n^{-\frac{\alpha+1}{4(d'+\alpha+1)}} \sqrt{\log n}$$

where $d' = 2\lceil d/2 \rceil$, and $\alpha \leq 3$ depends on the regularity of optimal maps.

Corollary. (Werenski et al.) Let $\hat{\lambda}$ be the random estimate obtained from Algorithm 1. Suppose that A has an eigenvalue of 0 with multiplicity 1 and that $\lambda_* \in \Delta^p$ realizes $\lambda_*^T A \lambda_* = 0$. Then under the assumptions of the Theorem,

$$\mathbb{E}[\|\hat{\lambda} - \lambda_*\|_2^2] \lesssim \frac{1}{\sqrt{n}} + n^{-\frac{\alpha+1}{4(d'+\alpha+1)}} \sqrt{\log n}.$$

- Convergence rate depends (poorly) on dimensionality and smoothness of OT maps.
- Regularity theory for OT maps relevant.

Learning Reference Measures

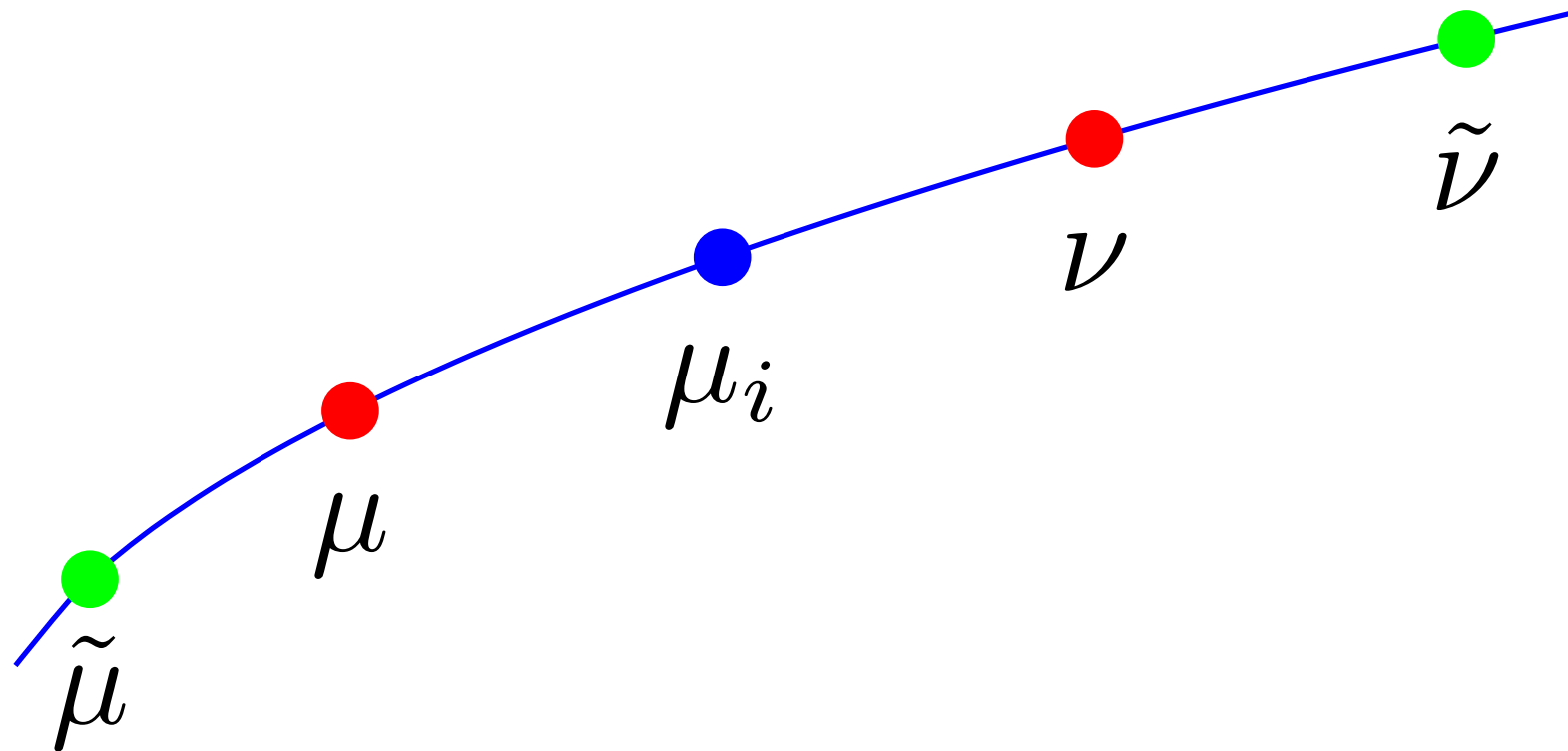
- For a *dictionary* $\mathcal{D} = \{\mathcal{D}_j\}_{j=1}^m \subset \mathcal{P}(\mathbb{R}^d)$ and $\boldsymbol{\lambda} \in \Delta^m$ let
$$\text{Bary}(\mathcal{D}, \boldsymbol{\lambda})$$
denote the barycenter generated from $\mathcal{D}$ with coordinates $\boldsymbol{\lambda}$.

- Given observed data $\{\mu_i\}_{i=1}^n$, *Wasserstein dictionary learning* (*WDL*) solves:

$$(\mathcal{D}^*, \boldsymbol{\Lambda}^*) = \arg \min_{\mathcal{D}, \boldsymbol{\Lambda}} \sum_{i=1}^n W_2^2(\text{Bary}(\mathcal{D}, \boldsymbol{\lambda}_i), \mu_i).$$

WDL Is Ill-Posed

- **Toy Example:** $\{\mu_i\}_{i=1}^n$ live on the Wasserstein geodesic between μ, ν .
- Then any measures $\tilde{\mu}, \tilde{\nu}$ that “extend” this geodesic can also generate $\{\mu_i\}_{i=1}^n$.



Geometric Wasserstein Dictionary Learning (GeoWDL)

$$\mathcal{R}_G(\mathcal{D}, \Lambda) := \sum_{i=1}^n \sum_{j=1}^m (\lambda_i)_j W_2^2(\mathcal{D}_j, \mu_i).$$

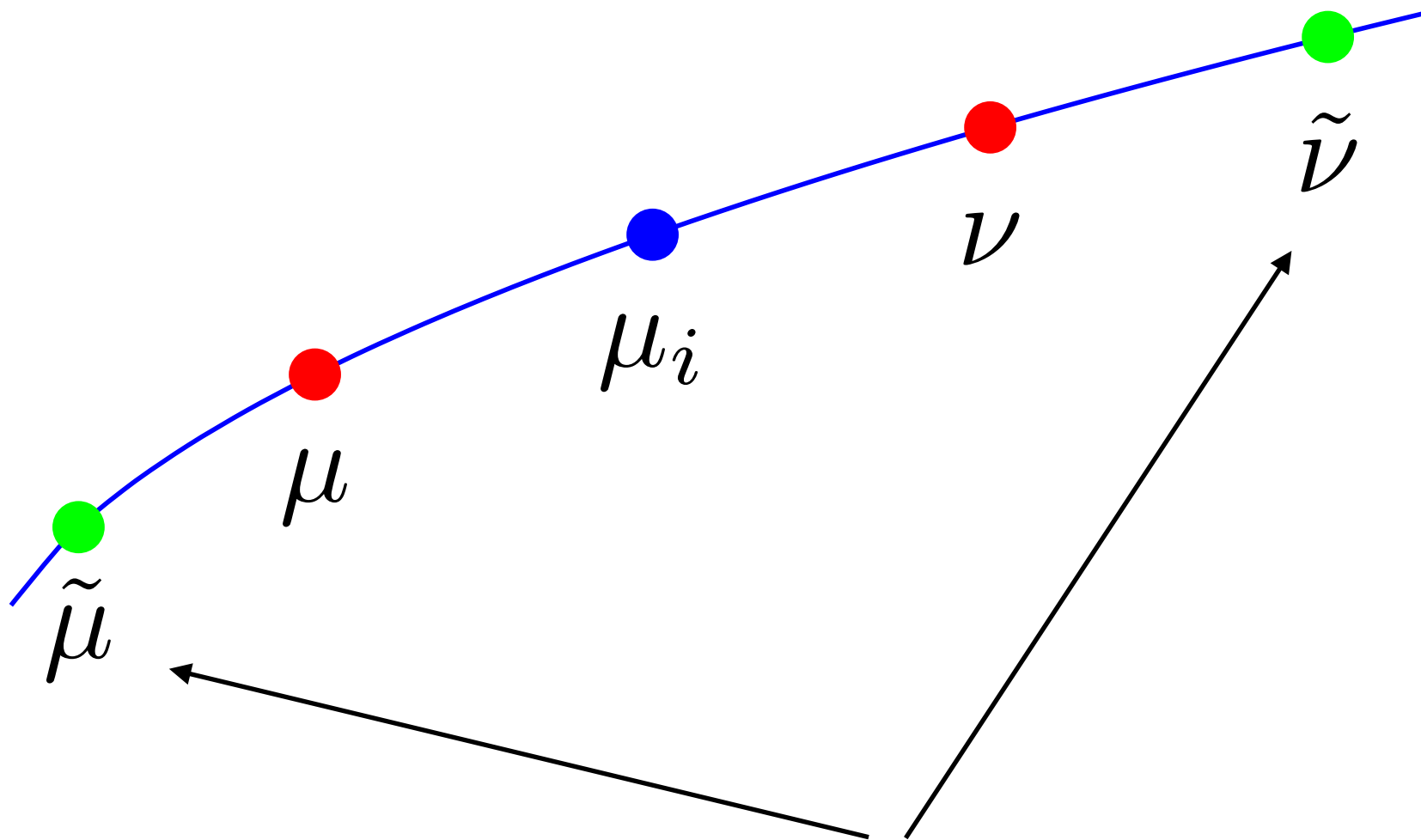
$$\mathcal{G}(\mathcal{D}, \Lambda, \{\mu_i\}_{i=1}^n) := \sum_{i=1}^n W_2^2(\text{Bary}(\mathcal{D}, \lambda_i), \mu_i) + \rho \mathcal{R}_G(\mathcal{D}, \Lambda).$$



- Learn a dictionary that *reconstructs well* using *nearby atoms*; $\rho > 0$.
- GeoWDL: $(\mathcal{D}^*, \Lambda^*) = \arg \min_{\mathcal{D}, \Lambda} \mathcal{G}(\mathcal{D}, \Lambda, \{\mu_i\}_{i=1}^n).$

Provable Learning of Generators

- If data lives on a Wasserstein geodesic, minimizing the geometric regularizer subject to perfect reconstruction learns μ, ν .



Geometric regularizer prevents learning these “extensions.”

Hyperspectral Images (HSI)

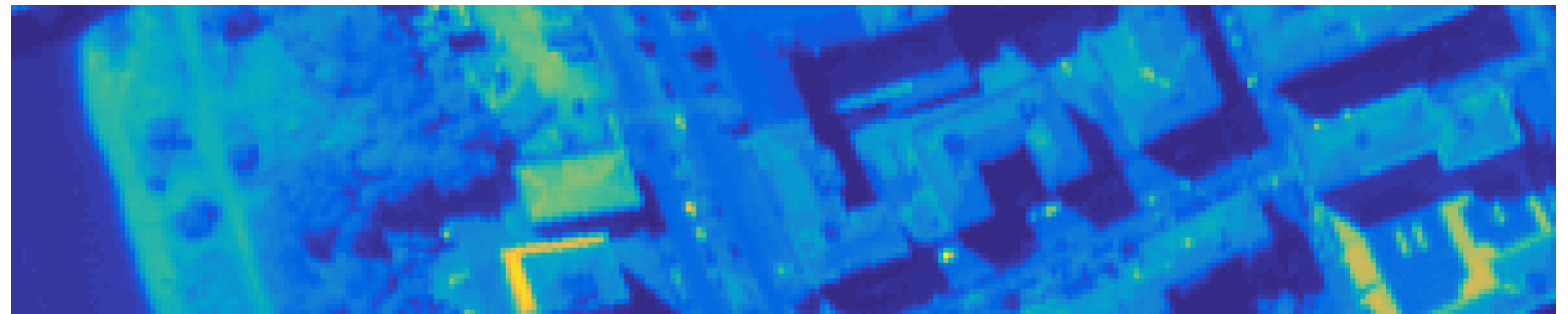
High dimensional images:

$$\begin{array}{ccc} M & \times & N & \times & D \\ & \swarrow & \nearrow & & \uparrow \\ & \text{Spatial} & & & \text{Spectral} \end{array}$$

The spectral bands are localized at certain electromagnetic frequencies, allowing for precise differentiation of materials in scenes

Learning Problems: clustering, anomaly detection, active learning, classification, segmentation, compression...

Applications: land cover change, water quality evaluation, precision agriculture, pollution tracking, defense and security...



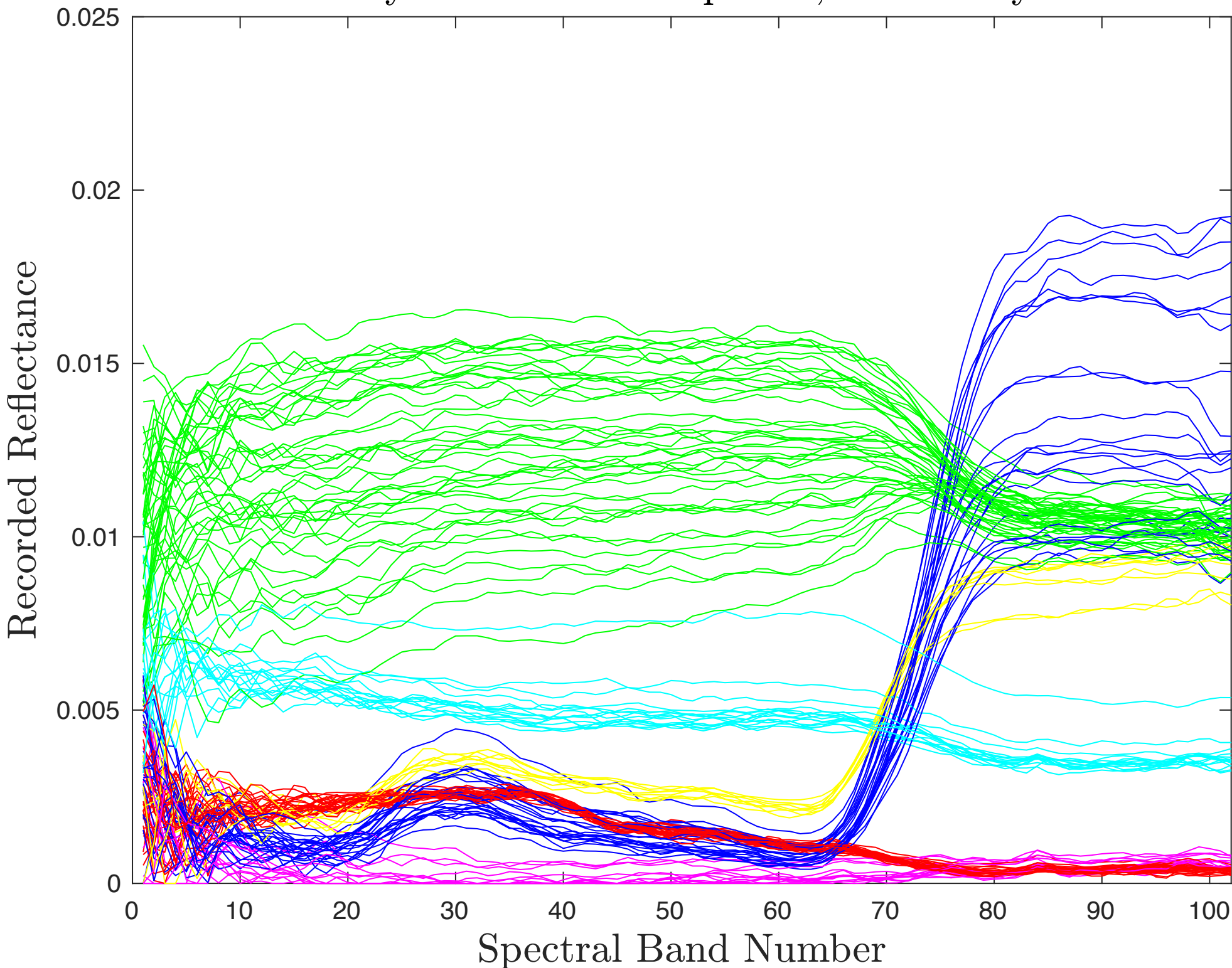
Subset of Pavia dataset, sum of all bands (sum across D)



Ground Truth

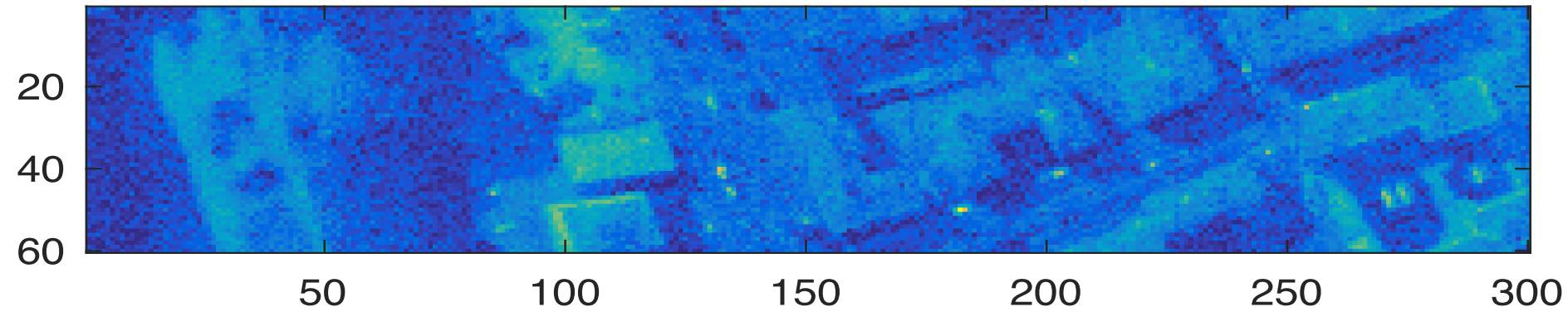
Each Pixel is a High-Dimensional Vector

Randomly Selected Pixel Spectra, Colored by Class

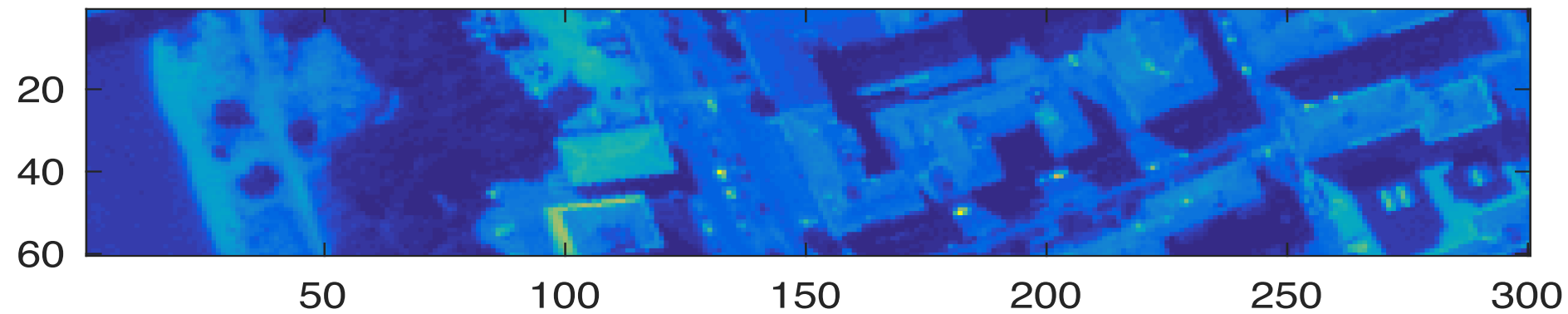


- Large within-class variation.
- Significant between-class overlap in some dimensions.

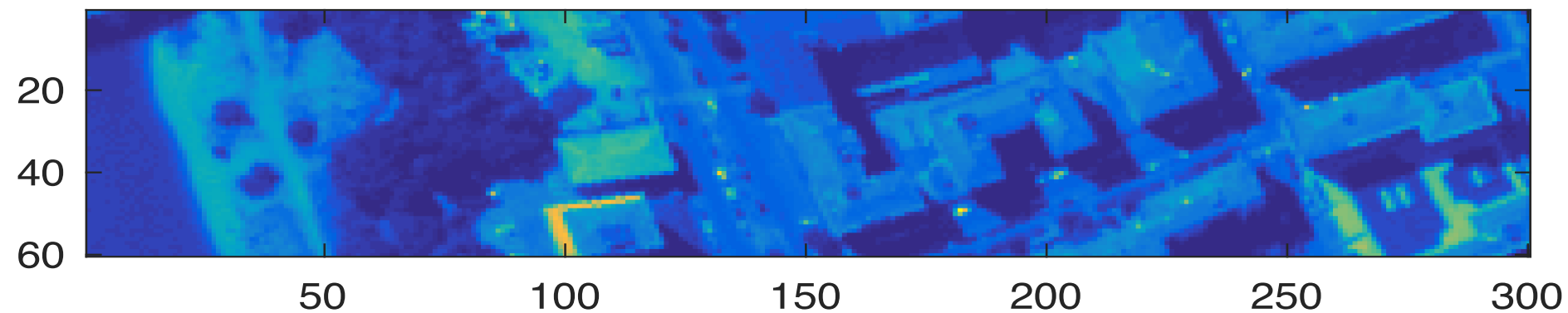
Individual Spectral Bands



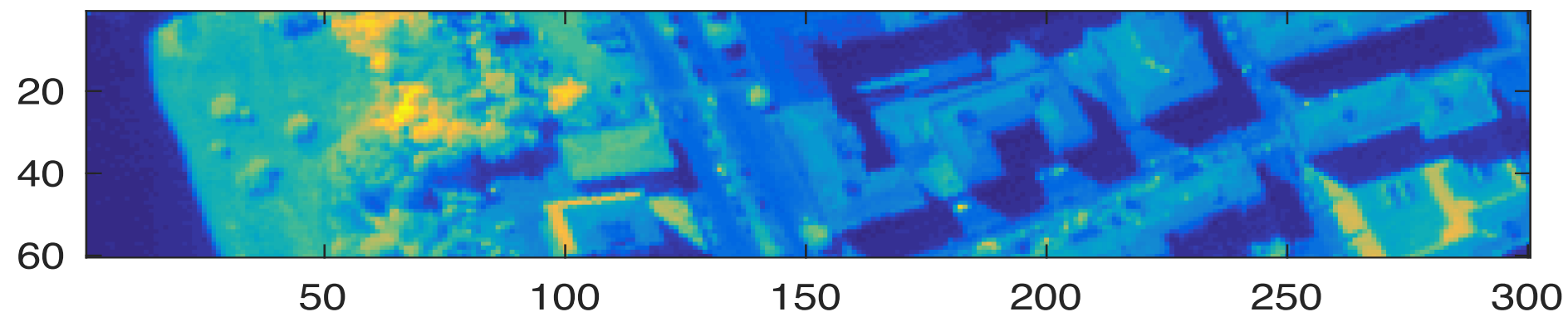
Band 2



Band 20



Band 40

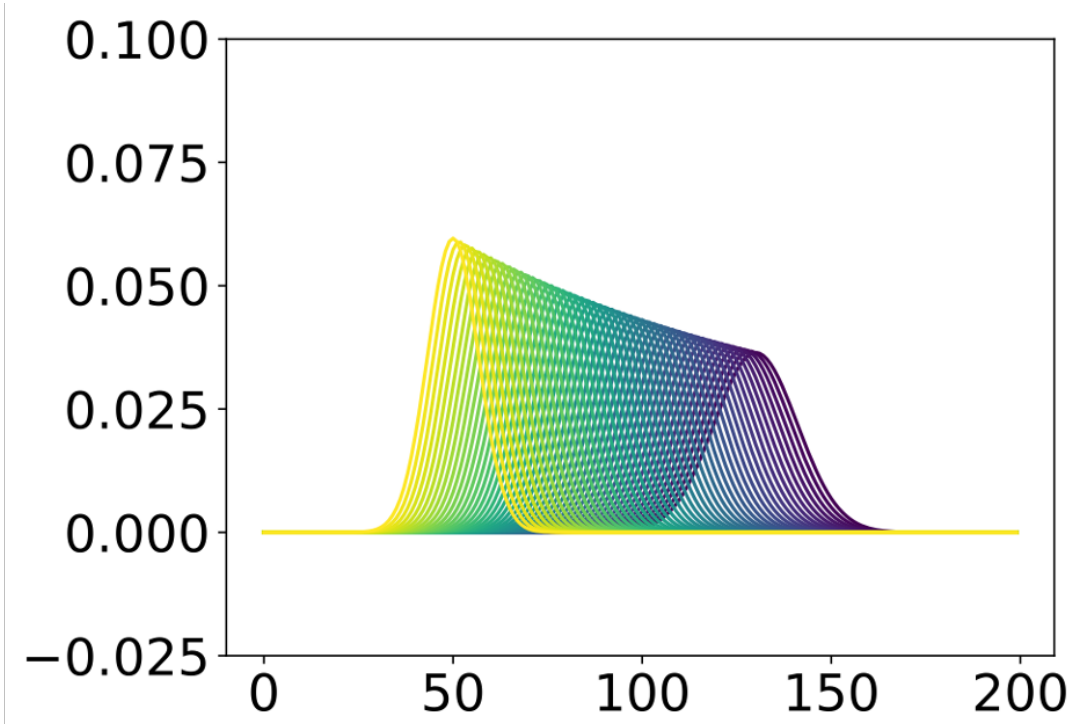
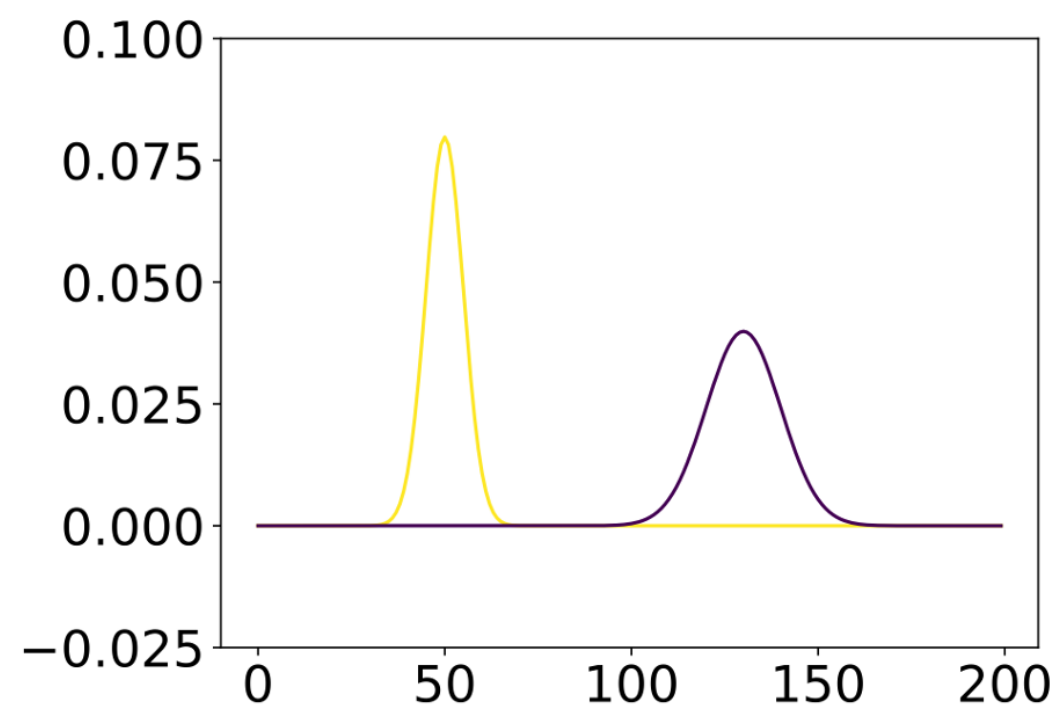


Band 102

Hyperspectral Unmixing

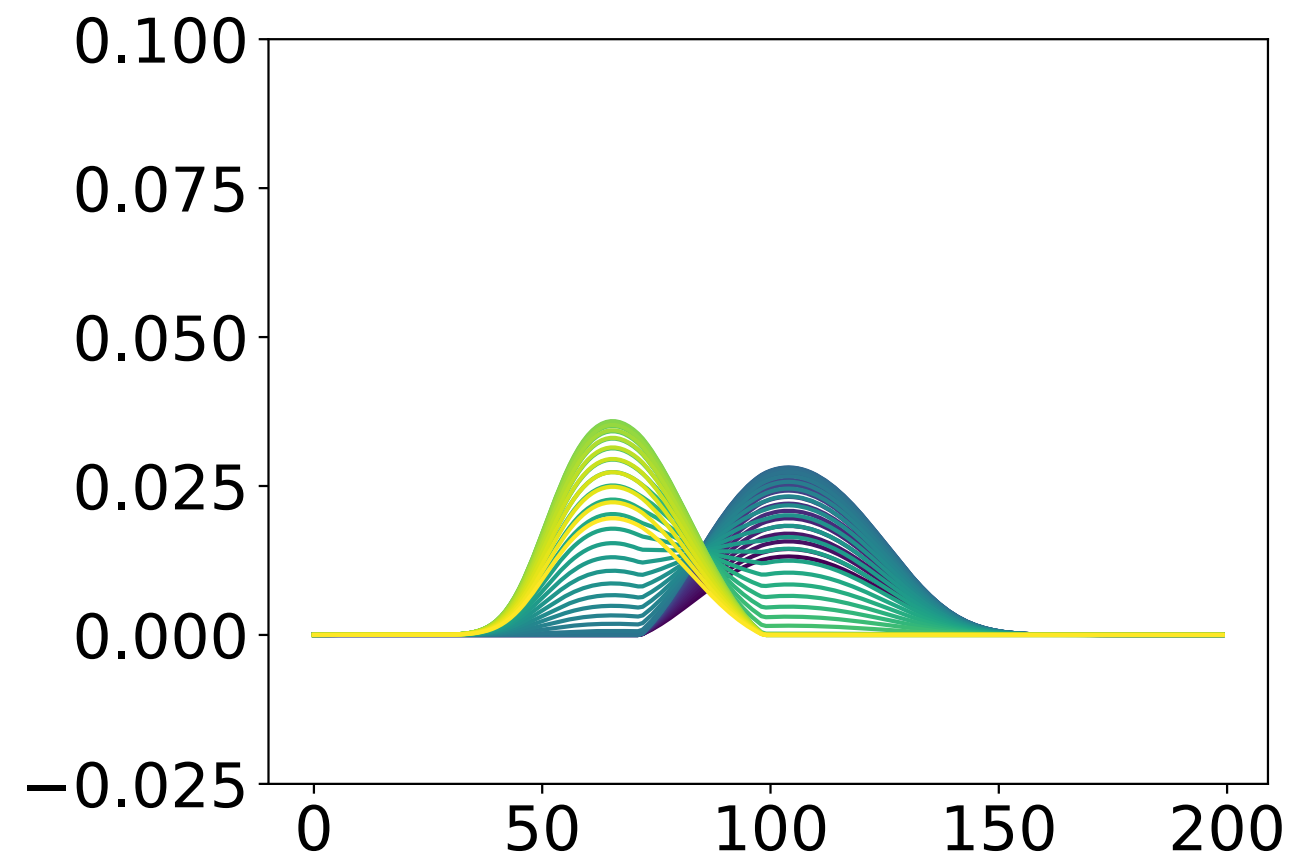
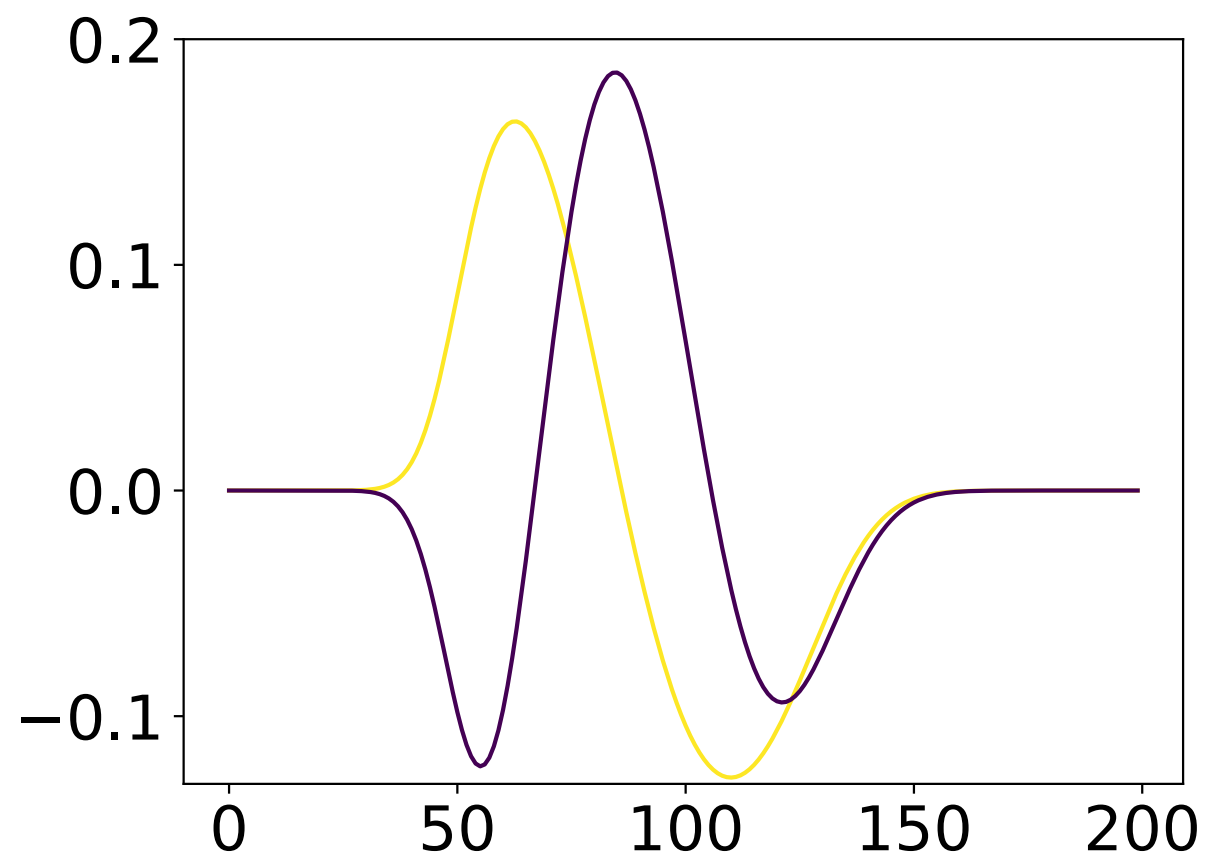
- Spatial resolution can be as high as 30m^2 per pixel. Many materials mixed up in a single pixel.
- *HSI unmixing*: learn “pure” material spectra and decompose each pixel into a combination of these pure spectra (“endmembers”).
- We propose to do this by modeling HSI spectra as probability distributions over reflectance space.

Wasserstein Hyperpsectral Unmixing (WaHU)

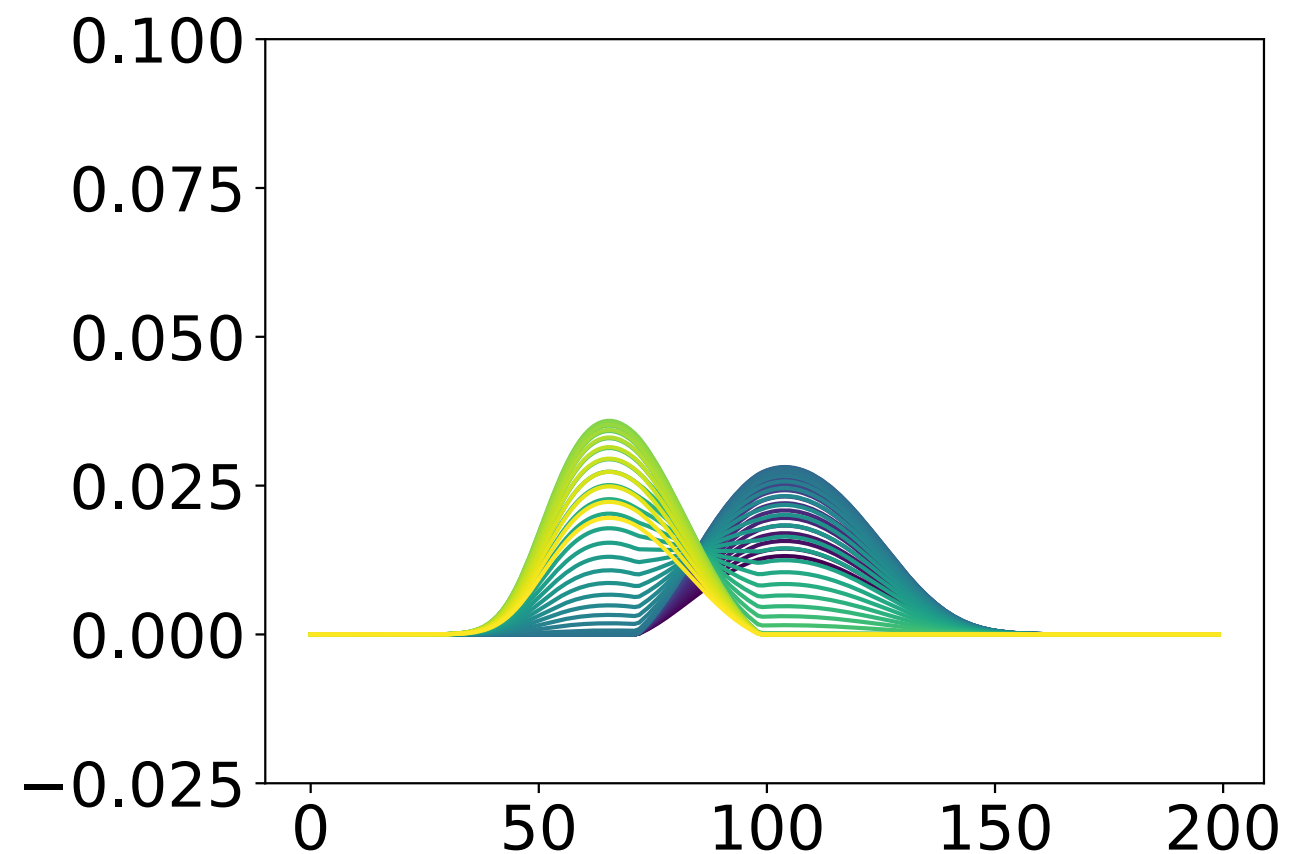
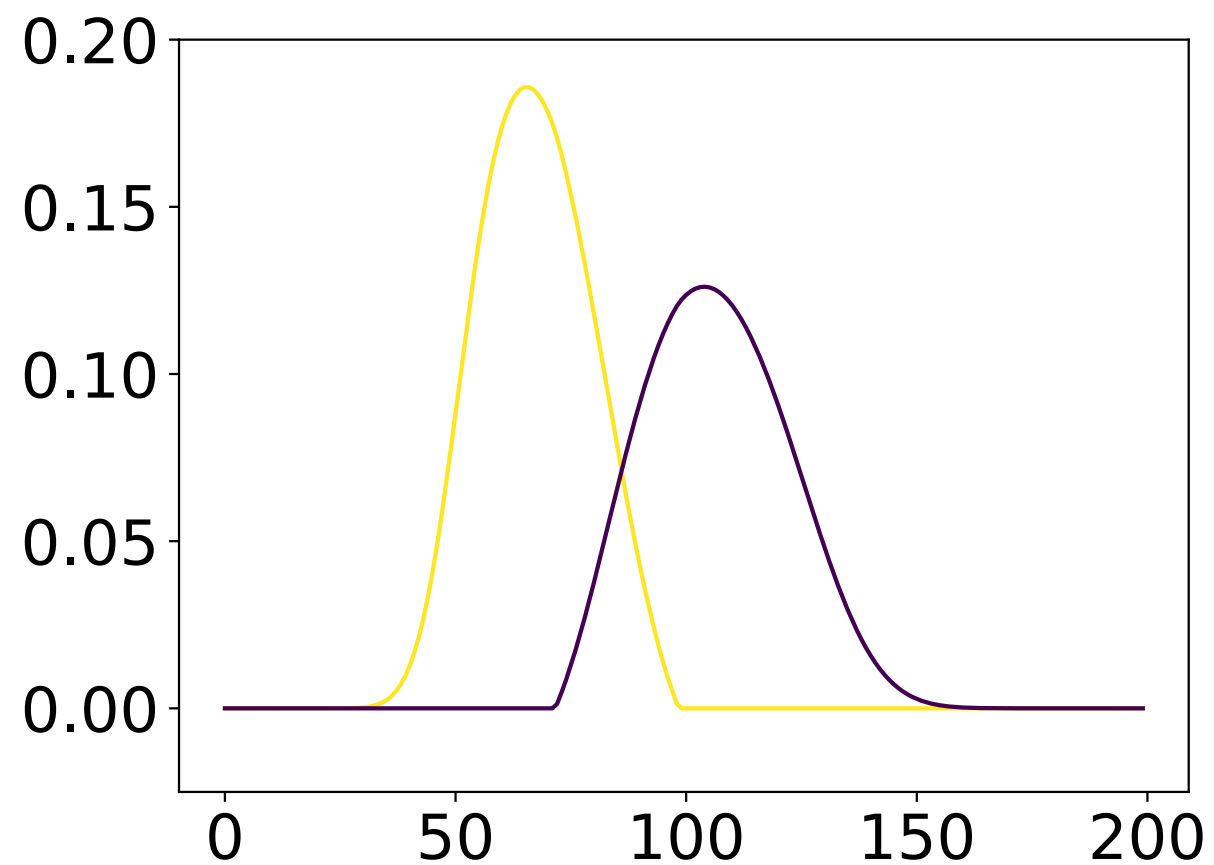


- Can we unmix this synthetic mixture of Gaussians and recover the original atoms?
- Contrast using our geometrically-regularized Wasserstein dictionary learning method with linear methods (PCA, NMF).

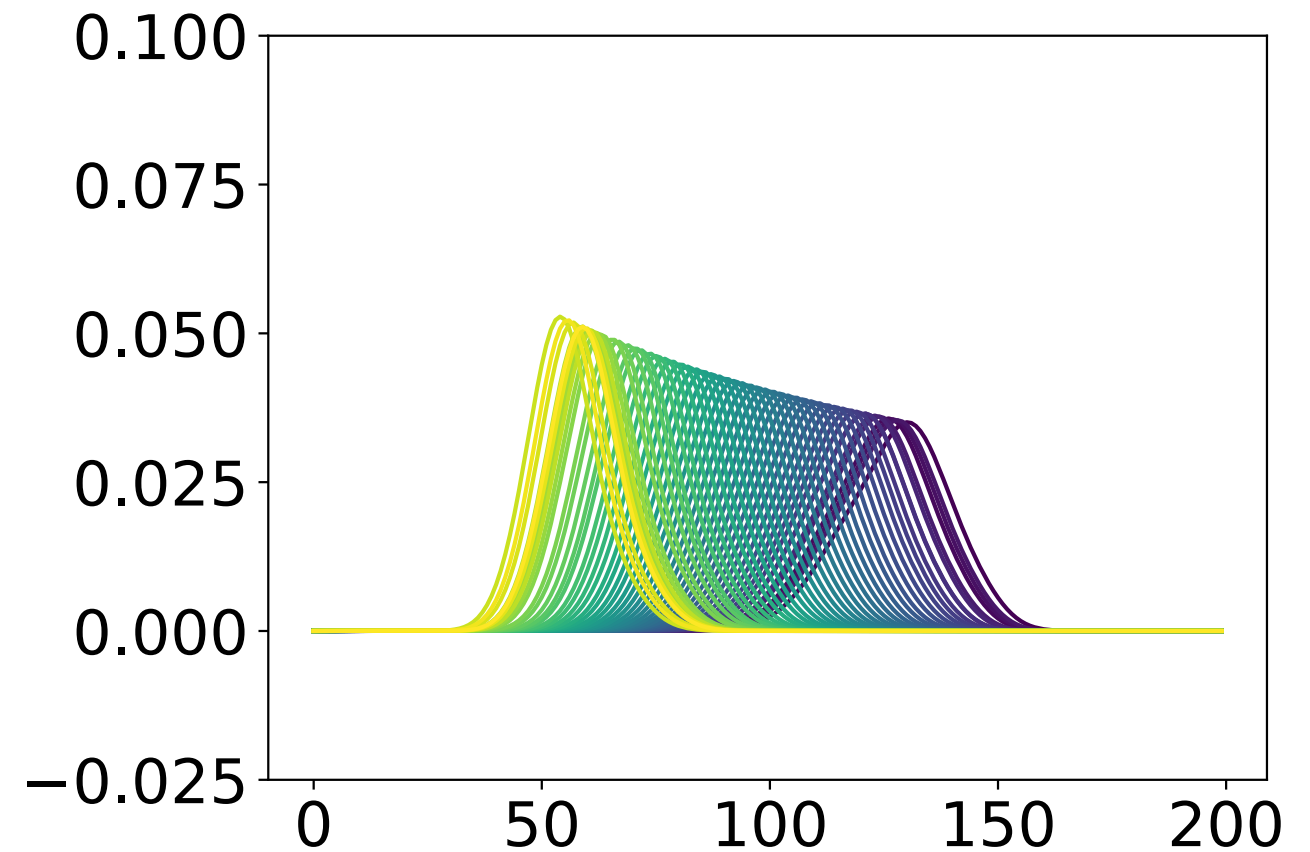
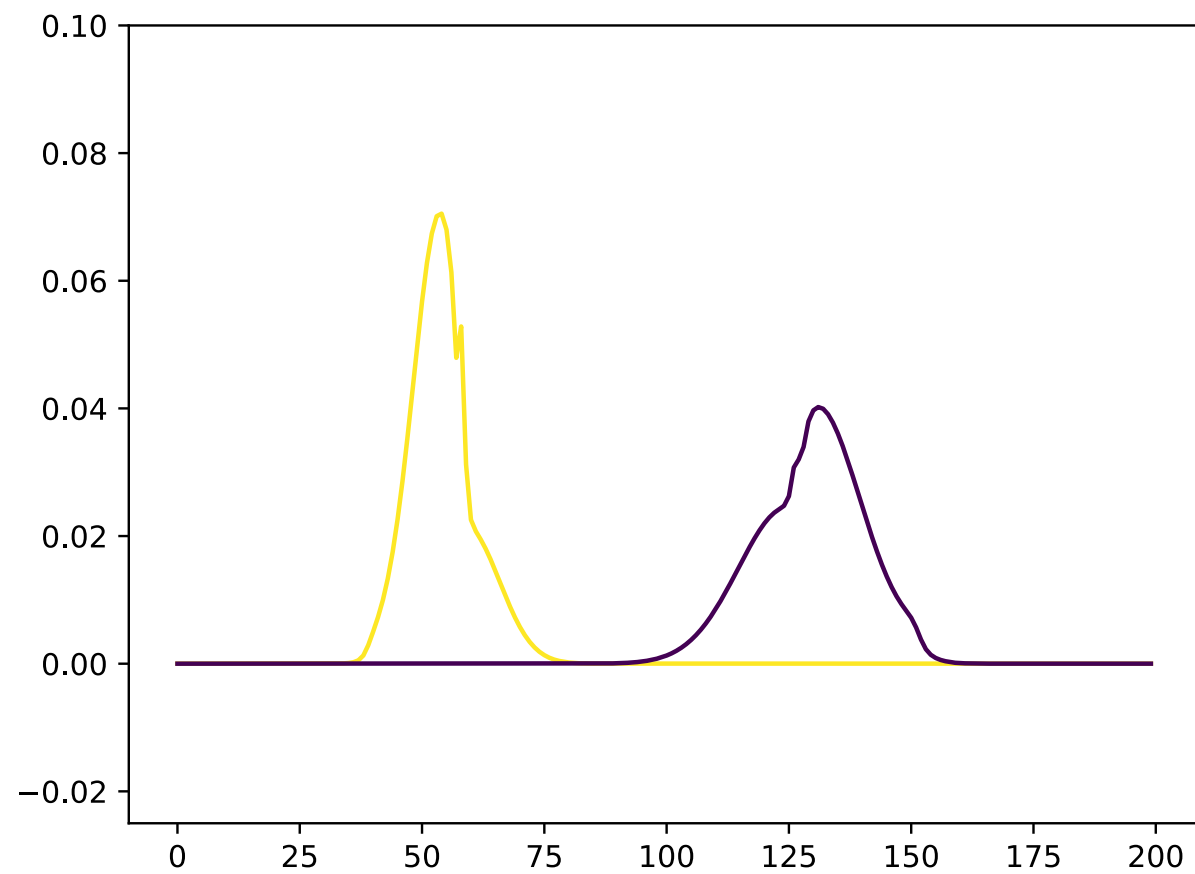
PCA on Synthetic Data



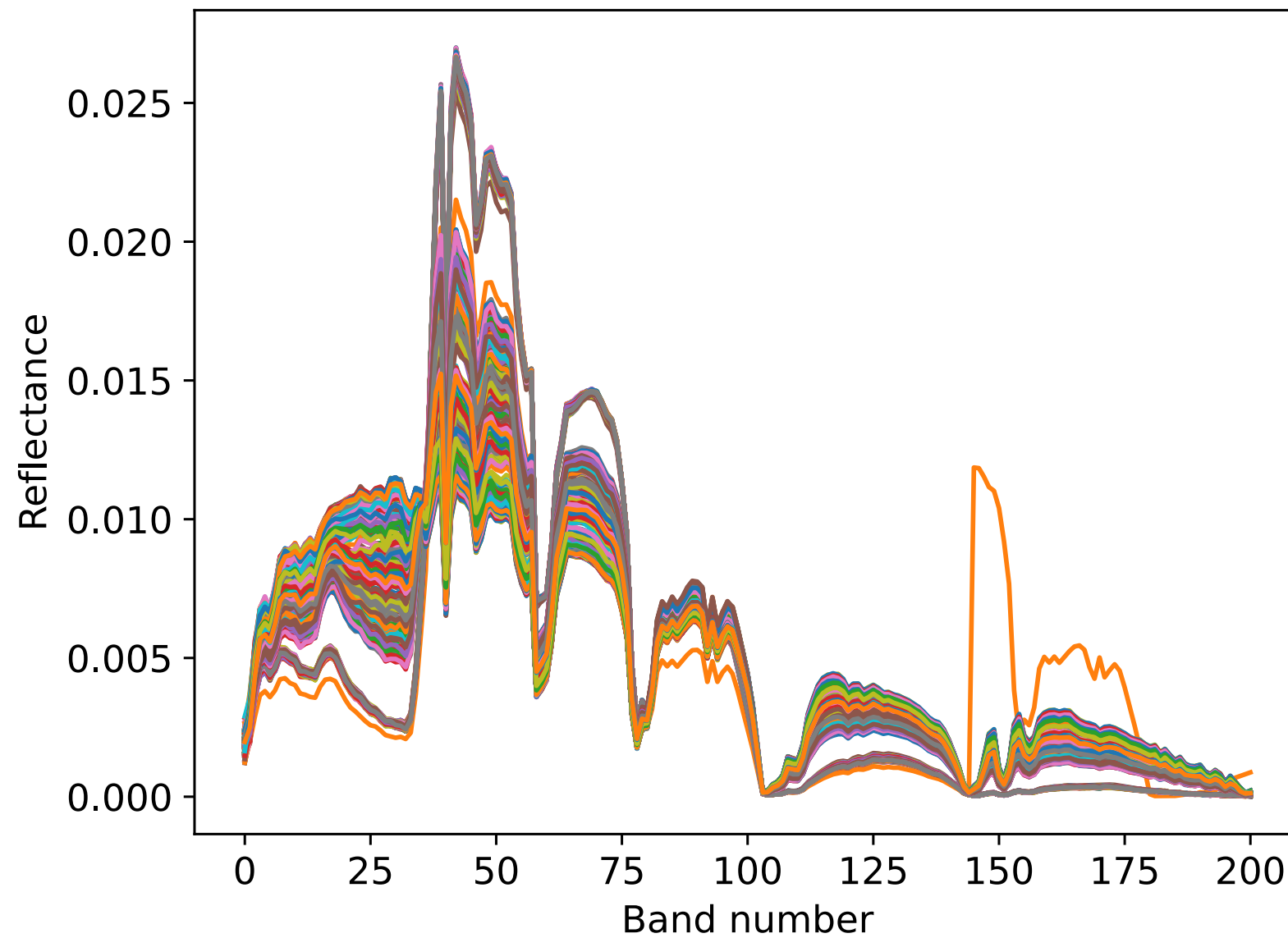
NMF on Synthetic Data



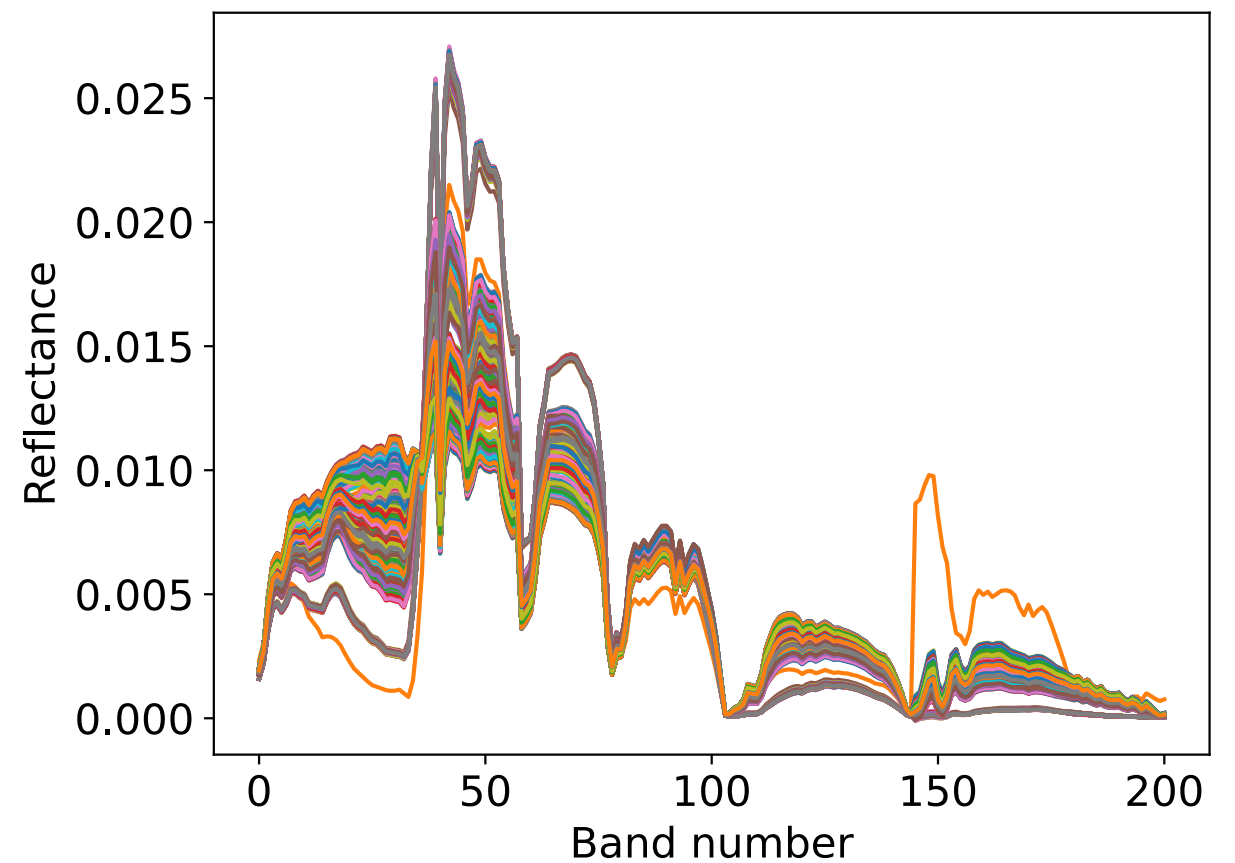
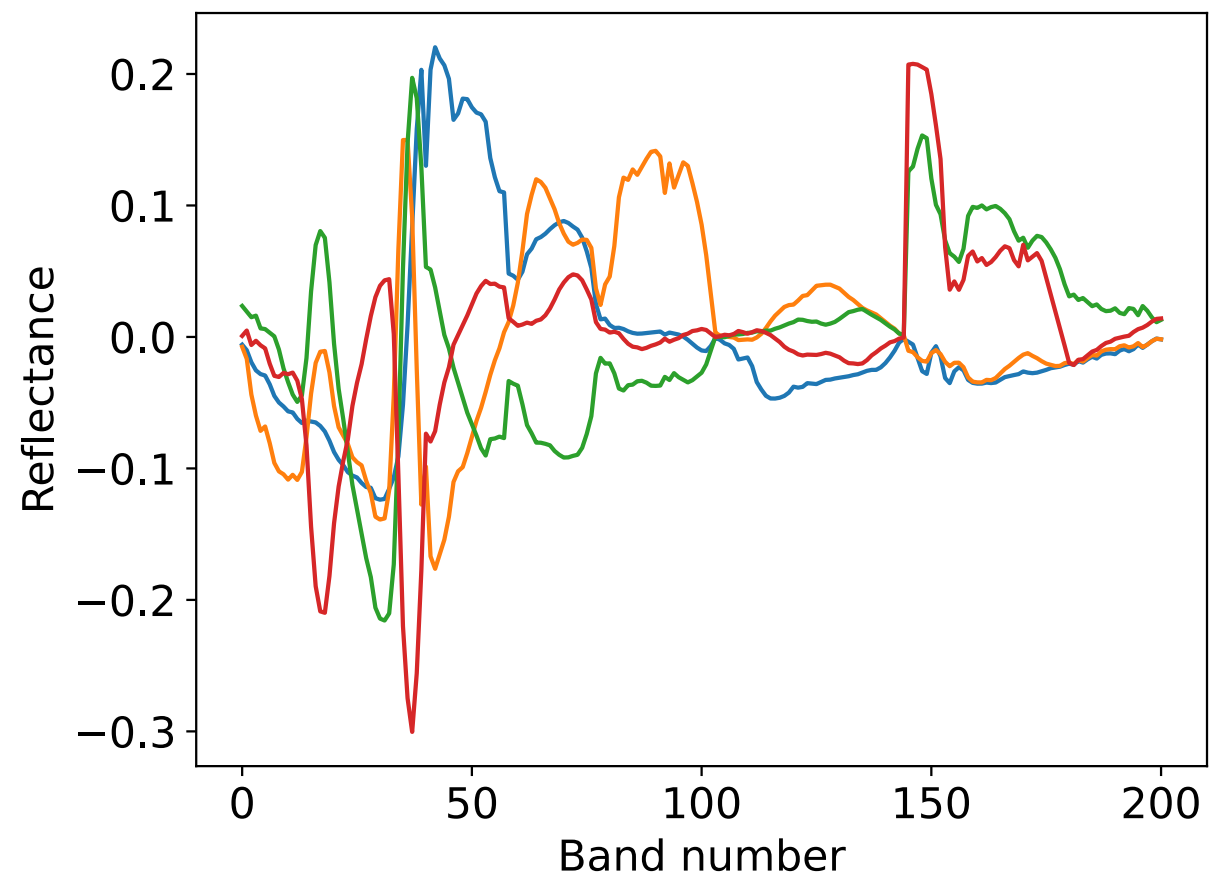
GeoWDL on Synthetic Data



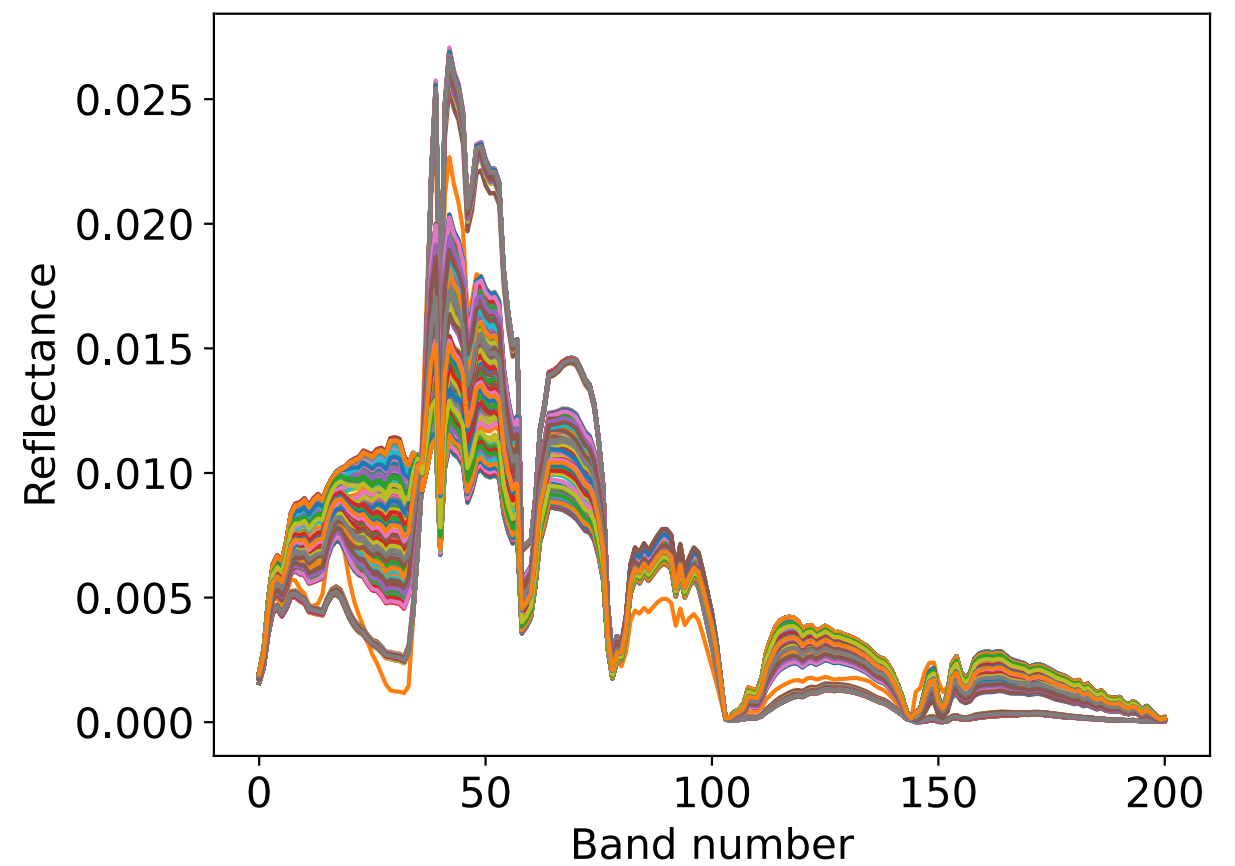
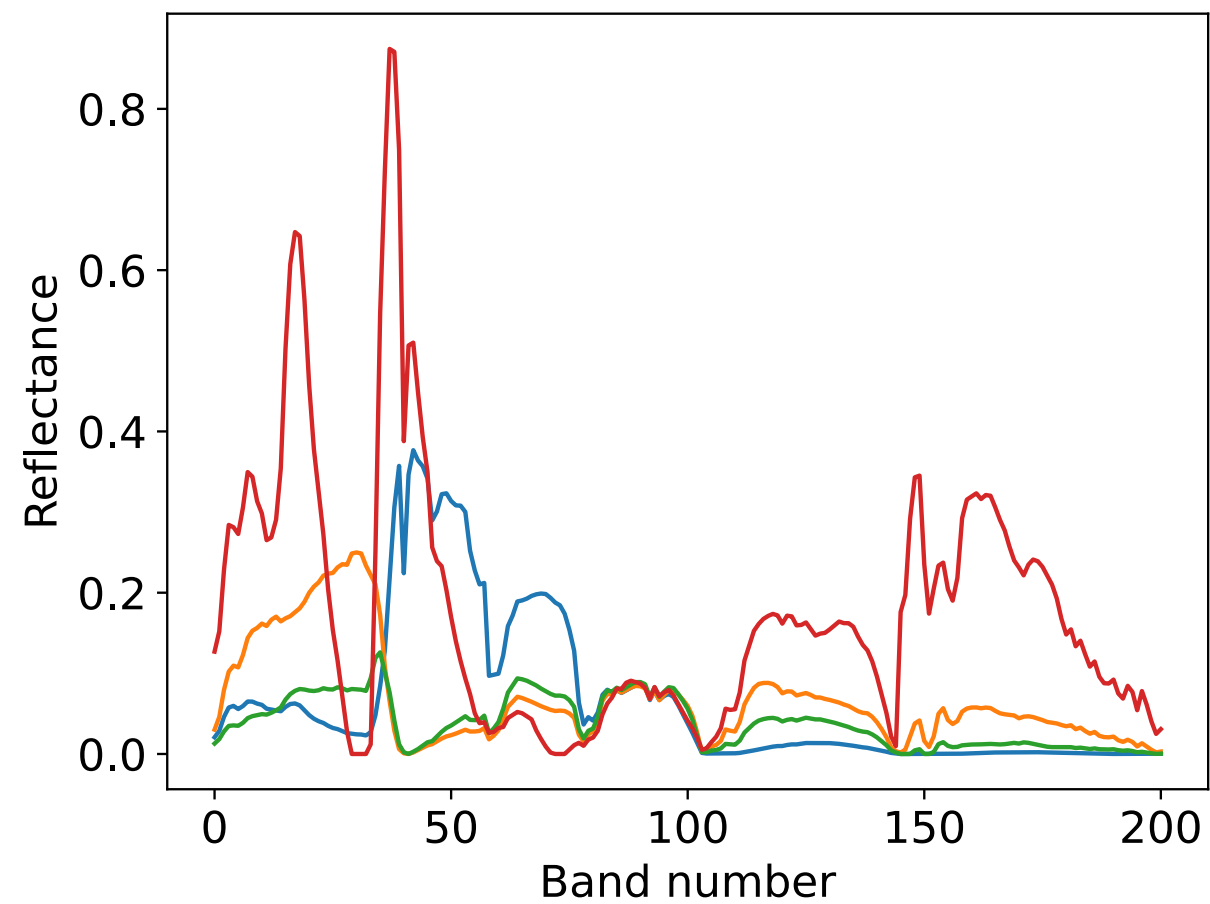
Salinas A Spectra



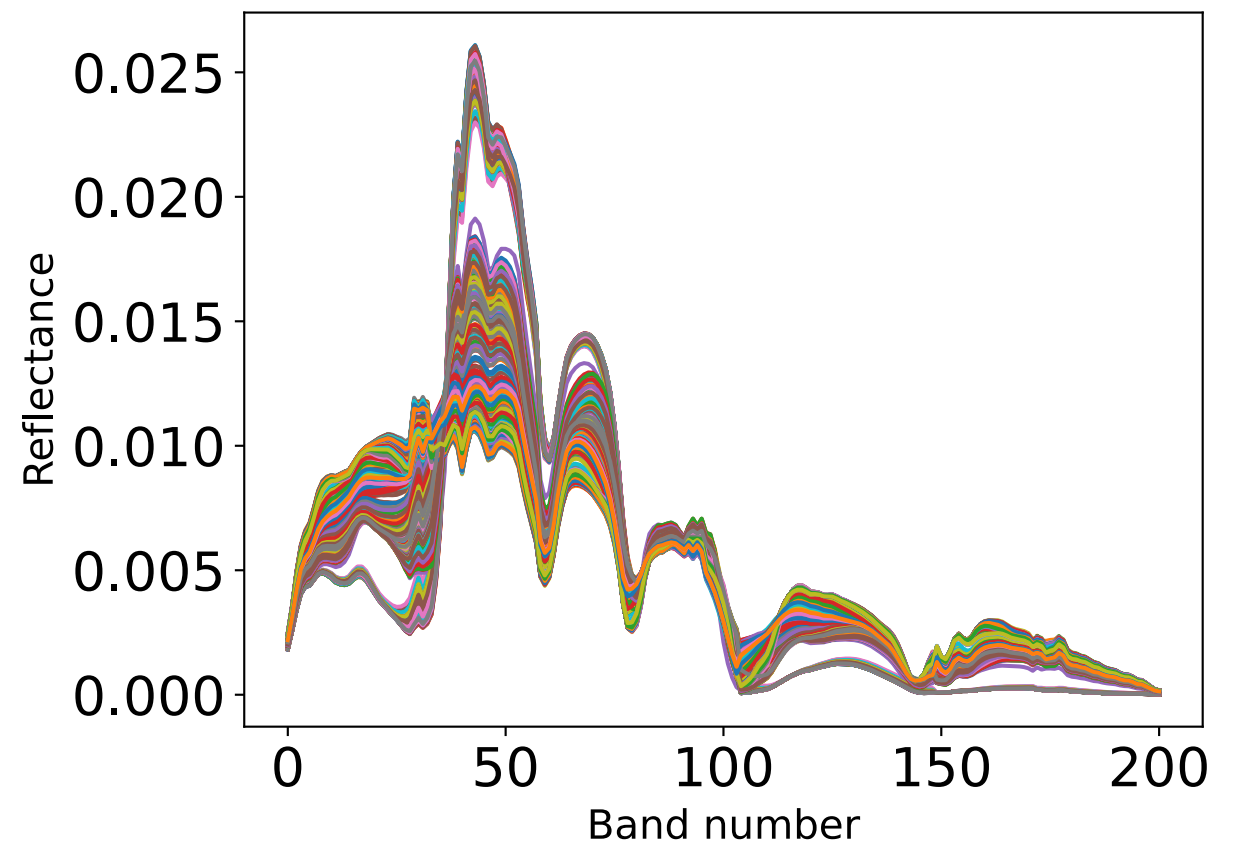
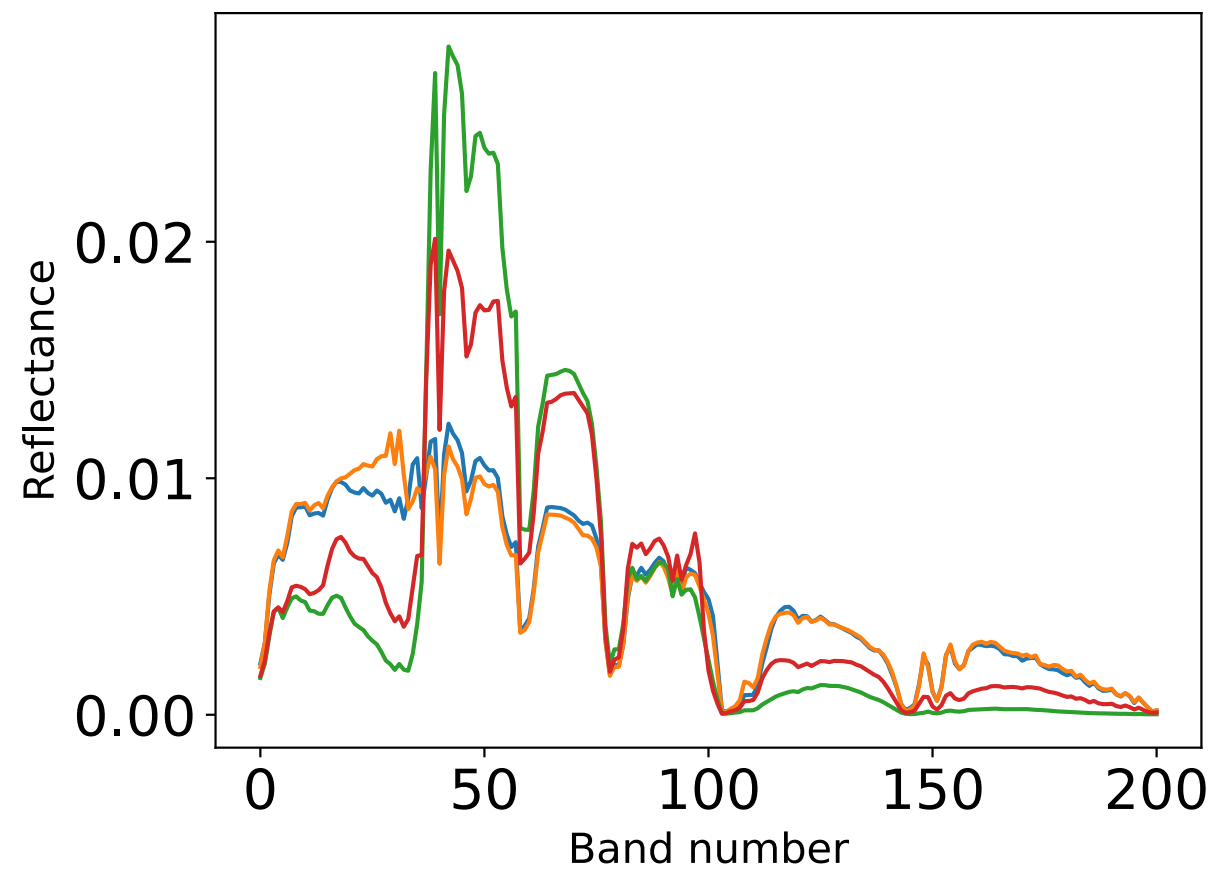
PCA on Salinas A



NMF on Salinas A



GeoWDL on Salinas A



BCM Representational Capacity

- How can we choose references measures $\{\mu_i\}_{i=1}^{\infty}$ so that

$$\text{BCM}(\{\mu_i\}_{i=1}^{\infty}) = \mathcal{P}_2(\mathbb{R}^d)?$$

- Mixtures of scale-translation families are closed under barycentric combinations.
- So, natural choices in the linear perspective (e.g., mixtures of Gaussians) fail in the barycentric contexts.

The Linear Barycentric Coding Model

- A useful alternative to the BCM is the *linear barycentric coding model (LBCM)*.
- For measures μ_0 and $\{\mu_i\}_{i=1}^m$, let

$$\text{LBCM}(\mu_0; \{\mu_i\}_{i=1}^m) = \{(\sum_{i=1}^m \lambda_i T_{\mu_0 \rightarrow \mu_i}) \# \mu_0 \mid \lambda \in \Delta^m\}$$

- BCM and LBCM coincide for compatible measures, in particular in 1 dimension for absolutely continuous measures.

Universal Approximation in the LBCM

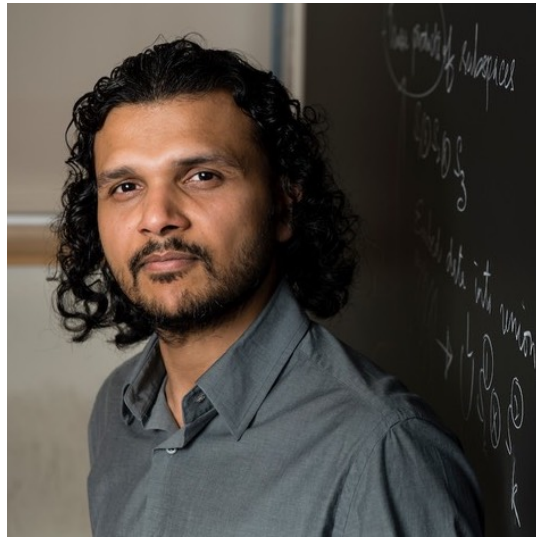
Theorem. (Werenski et al.) The set

$$\text{LBCM}(Unif([0, 1]); \{\alpha\delta_0 + (1 - \alpha)\delta_1 \mid a \in [0, 1]\})$$

is dense in $\mathcal{P}([0, 1])$.

- This set of reference measures is not dense under linear mixing.
- Generalizing to higher dimensions is challenging and naive extensions fail.

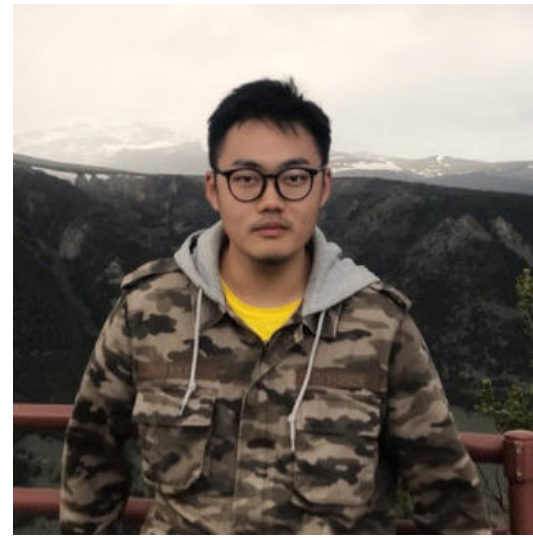
Collaborators



S. Aeron (Tufts ECE)



S. Fullenbaum (Tufts Math -> Fidelity)



R. Jiang (Tufts CS -> LinkedIn)



B. Mallery (Tufts Math)



S. Bin Masud (Tufts ECE -> FL Blue)



M. Mueller (Tufts Math -> ERSI)



A. Tasissa (Tufts Math)



M. Werenski (Tufts CS -> Us Govt.)

Papers

Fullenbaum, Mueller, Tasissa, Murphy

“Nonlinear Unmixing of Hyperspectral Images via Regularized Wasserstein Dictionary Learning”

IGARSS, 2024

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

Mallery, Murphy, Aeron

“Synthesis and Analysis of Data as Probability Measures with Entropy-Regularized Optimal Transport”

AISTATS, 2025

Masud, Werenski, Murphy, Aeron

"Multivariate Soft Rank Via Entropic Optimal Transport: Sample Efficiency and Generative Modeling"

Journal of Machine Learning Research, 2023

Mueller, Aeron, Murphy, Tasissa

“Geometric Sparse Coding in Wasserstein Space”

TAG in ML Workshop at ICML, 2023

Werenski, Jiang, Tasissa, Aeron, Murphy

“Measure Estimation in the Barycentric Coding Model”

ICML, 2022

Werenski, Mallery, Aeron, Murphy

“Linearized Wasserstein Barycenters: Synthesis, Analysis, Representational Capacity, and Applications”

AISTATS, 2025

Code and Contact Information

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

Contact: jm.murphy@tufts.edu

Thanks for Your Attention!

