Mean-shift Algorithms for Manifold Denoising, Matrix Completion and Clustering

Weiran Wang

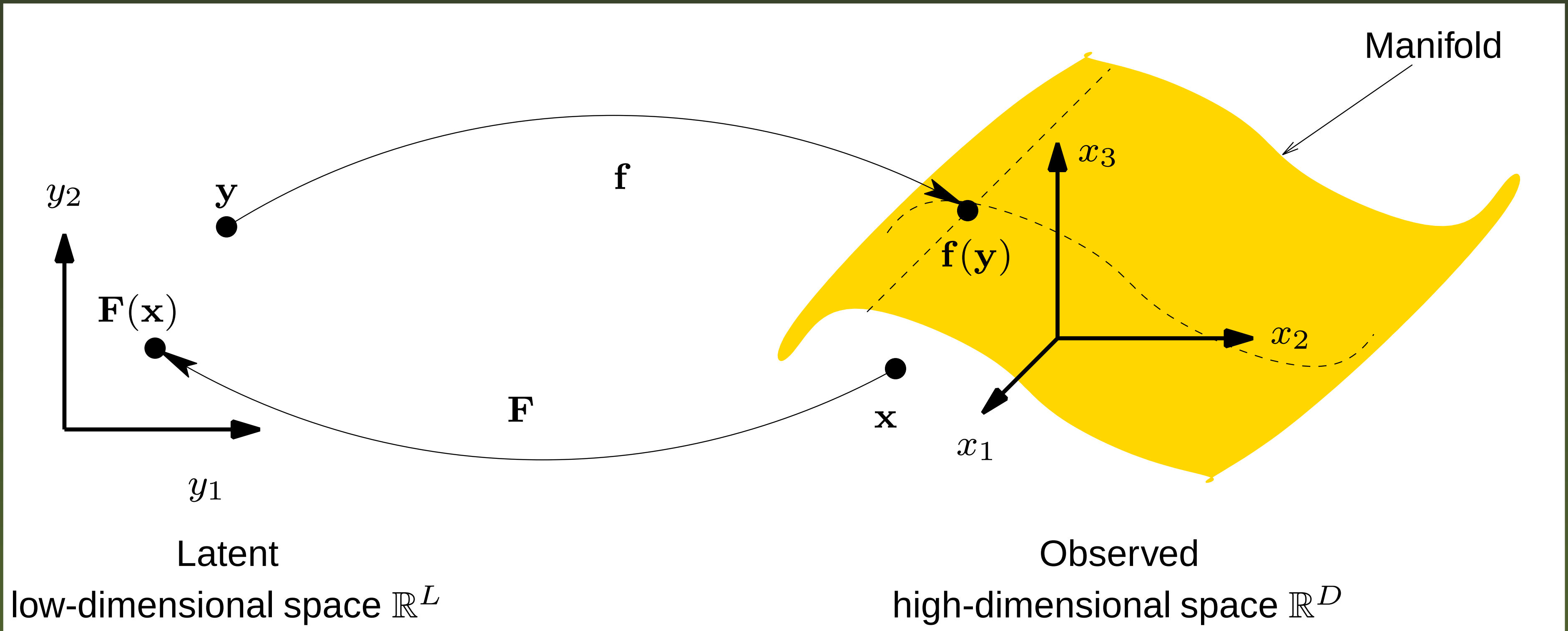
wwang5@ucmerced.edu

EECS Department, UC Merced

Manifold Learning

High dimensional dataset with manifold structure.

- Variations within the dataset can be modeled by a few latent variables.
- Small variation in latent space leads to small variation in data space.
- Local neighborhood of each data point can be approximated by a tangent space.



An example: MNIST

Small variations in translation, rotation, scaling and different writing styles change the image appearance slightly, and do not change the identity.

Mean-shift update

Given a set of data points $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_N] \subset \mathbb{R}^D$.

- Maximizes kernel density estimate (mode finding)

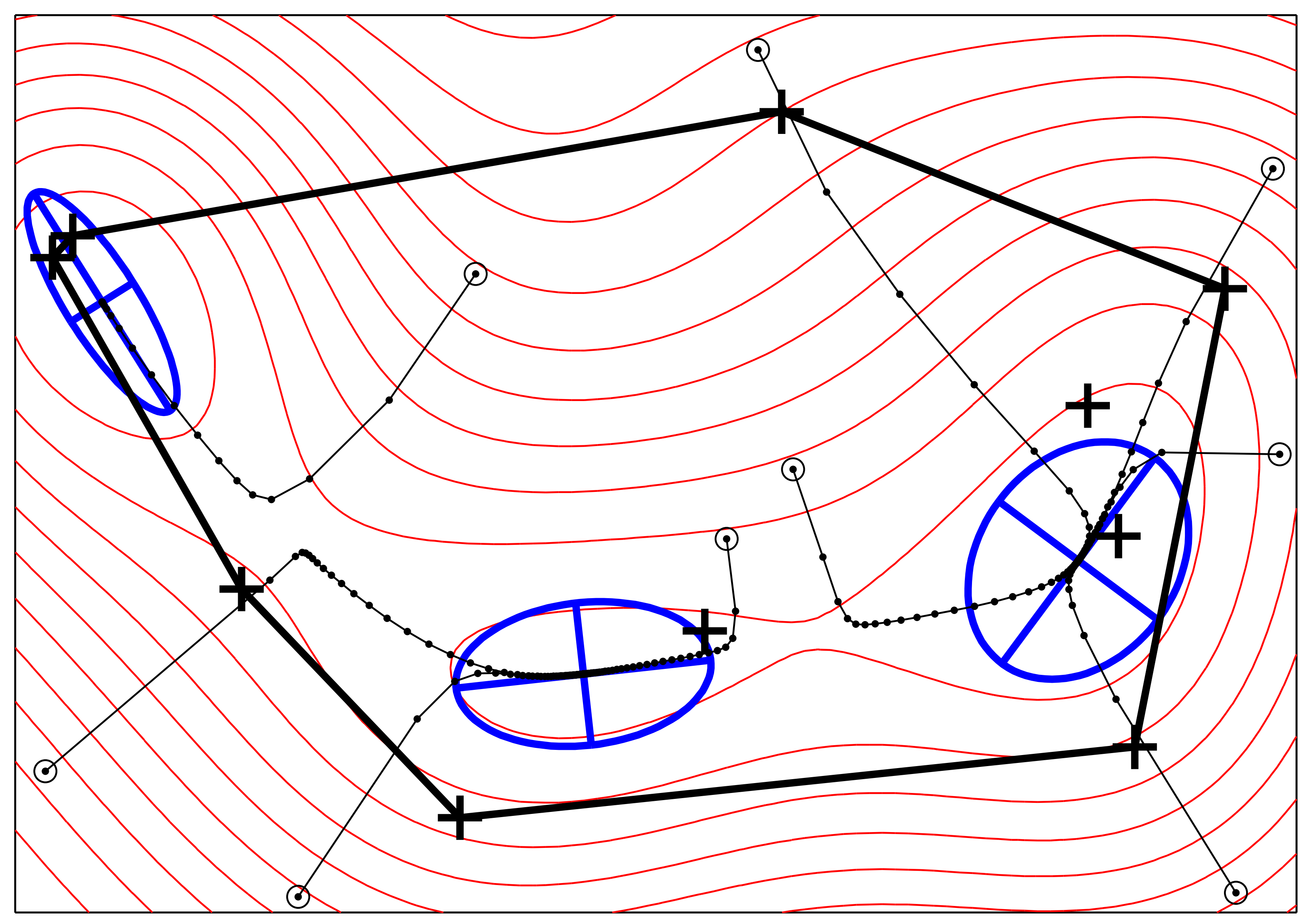
$$p(\mathbf{x}) = \frac{1}{N} \sum_{n=1}^{N} G\left(\left\| \frac{\mathbf{x} - \mathbf{x}_n}{\sigma} \right\|^2\right), \qquad G(t) = e^{-t/2}.$$

- Applies the mean-shift update (fixed point iteration) iteratively

$$p(n|\mathbf{x}) = \frac{G\left(\left\| \frac{\mathbf{x}-\mathbf{x}_n}{\sigma} \right\|^2\right)}{\sum_{n'=1}^{N} G\left(\left\| \frac{\mathbf{x}-\mathbf{x}_{n}}{\sigma} \right\|^2\right)}, \qquad \mathbf{x} \leftarrow \mathbf{f}(\mathbf{x}) = \textstyle\sum_{n=1}^{N} p(n|\mathbf{x})\mathbf{x}_n$$

- Gradient ascent. Linear convergence rate.

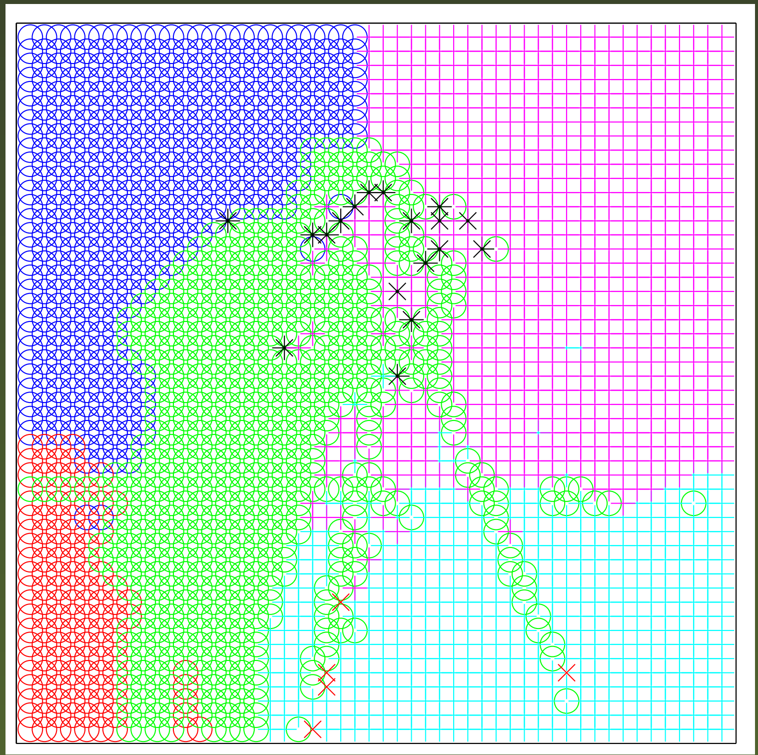
Mean-shift update



Paths followed by GMS for various starting points.

Mean-shift clustering

- Gaussian Mean-shift (GMS): points that converge to the same mode/centroid define a cluster. Number of clusters depends on σ.
- Gaussian Burring Mean-shift (GBMS): update dataset after each mean-shift step, has much faster (cubic) convergence rate and strong (isotropic) denoising effect.



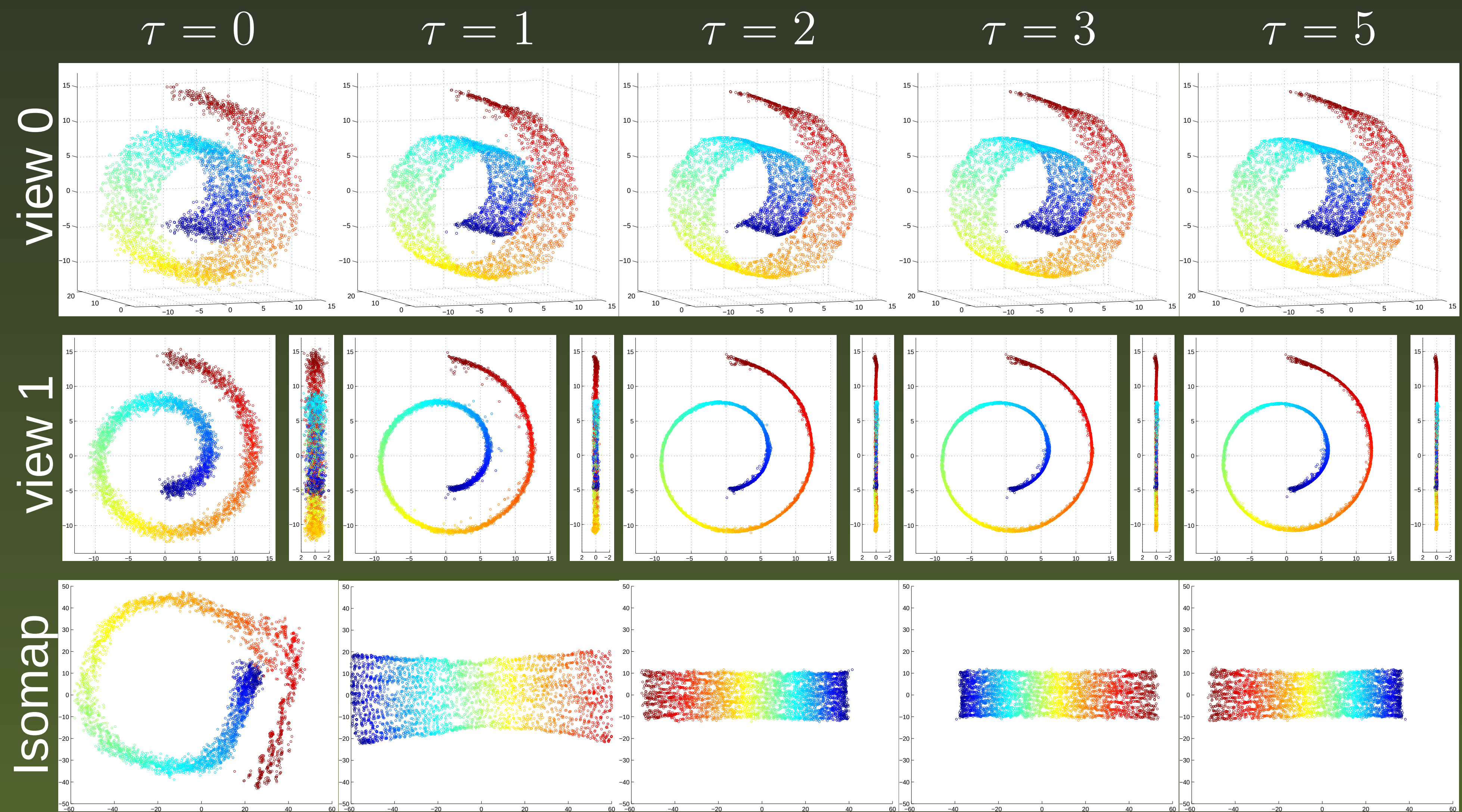
☞

Outline

- Manifold Blurring Mean-shift (MBMS) algorithm for manifold denoising
- MBMS for matrix completion
- K-modes algorithm for clustering
- Laplacian K-modes algorithm for clustering

Motivation

We develop an algorithm that denoises the dataset, and acts as a preprocessing step for unsupervised/supervised learning.



Manifold Blurring Mean-Shift

- Predictor averaging step: local clustering with GBMS, moves data point to the kernel average of its neighbors

$$\mathbf{x}_n \leftarrow \sum_{m \in \mathcal{N}_n} \frac{G\left(\|(\mathbf{x}_n - \mathbf{x}_m)/\sigma\|^2\right)}{\sum_{m' \in \mathcal{N}_n} G\left(\|(\mathbf{x}_n - \mathbf{x}_{m'})/\sigma\|^2\right)} \mathbf{x}_m$$

- Corrector projective step: estimate local tangent space with PCA, gives the best linear L-dimensional manifold in terms of reconstruction error (orthogonal projection on the manifold)

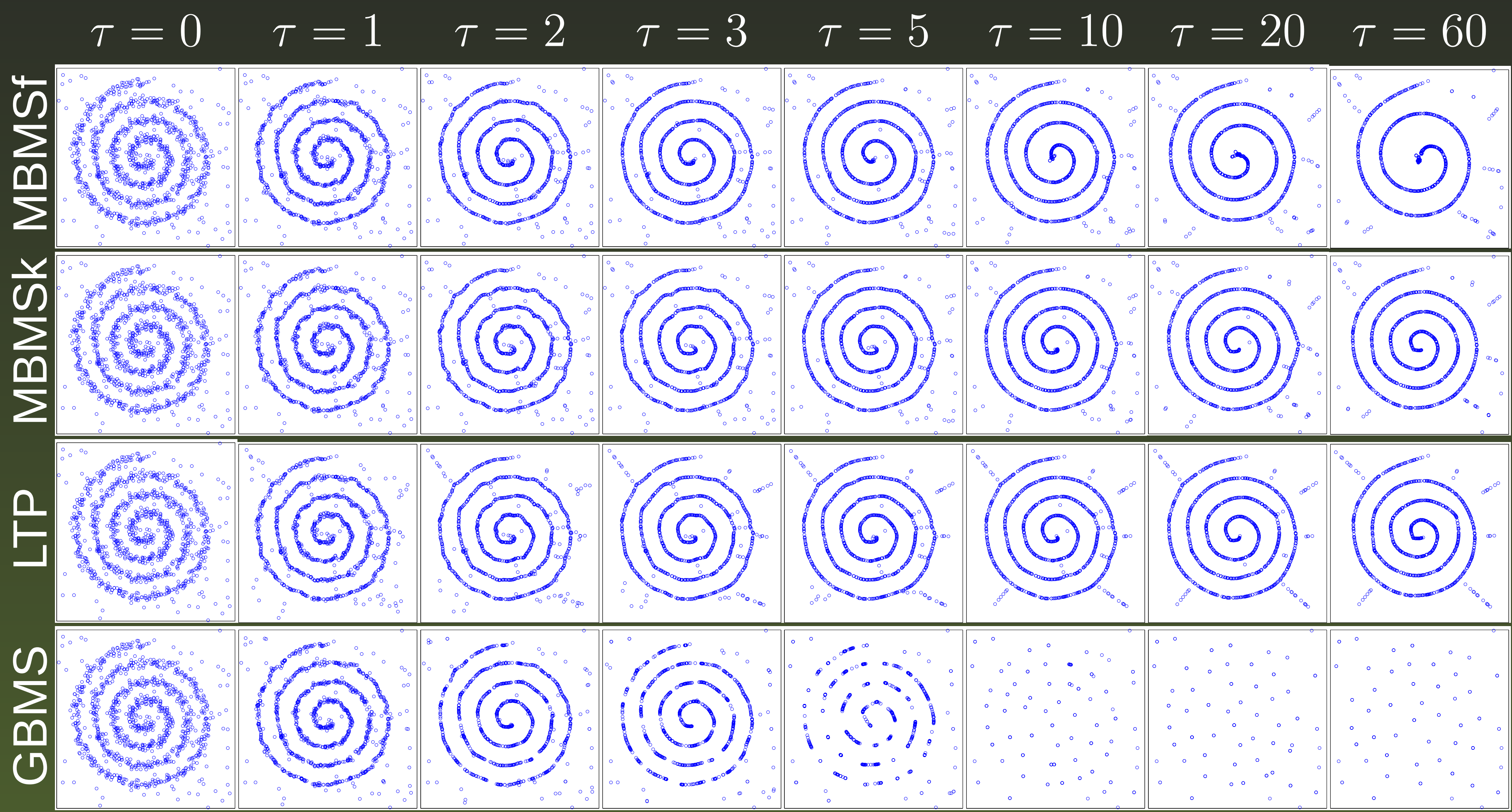
$$\min_{\boldsymbol{\mu}, \mathbf{U}} \sum_{m \in \mathcal{N}'_n} \left\| \mathbf{x}_m - (\mathbf{U}\mathbf{U}^T(\mathbf{x}_m - \boldsymbol{\mu}) + \boldsymbol{\mu}) \right\|^2$$

- User parameters: σ, K, L.

Practicalities

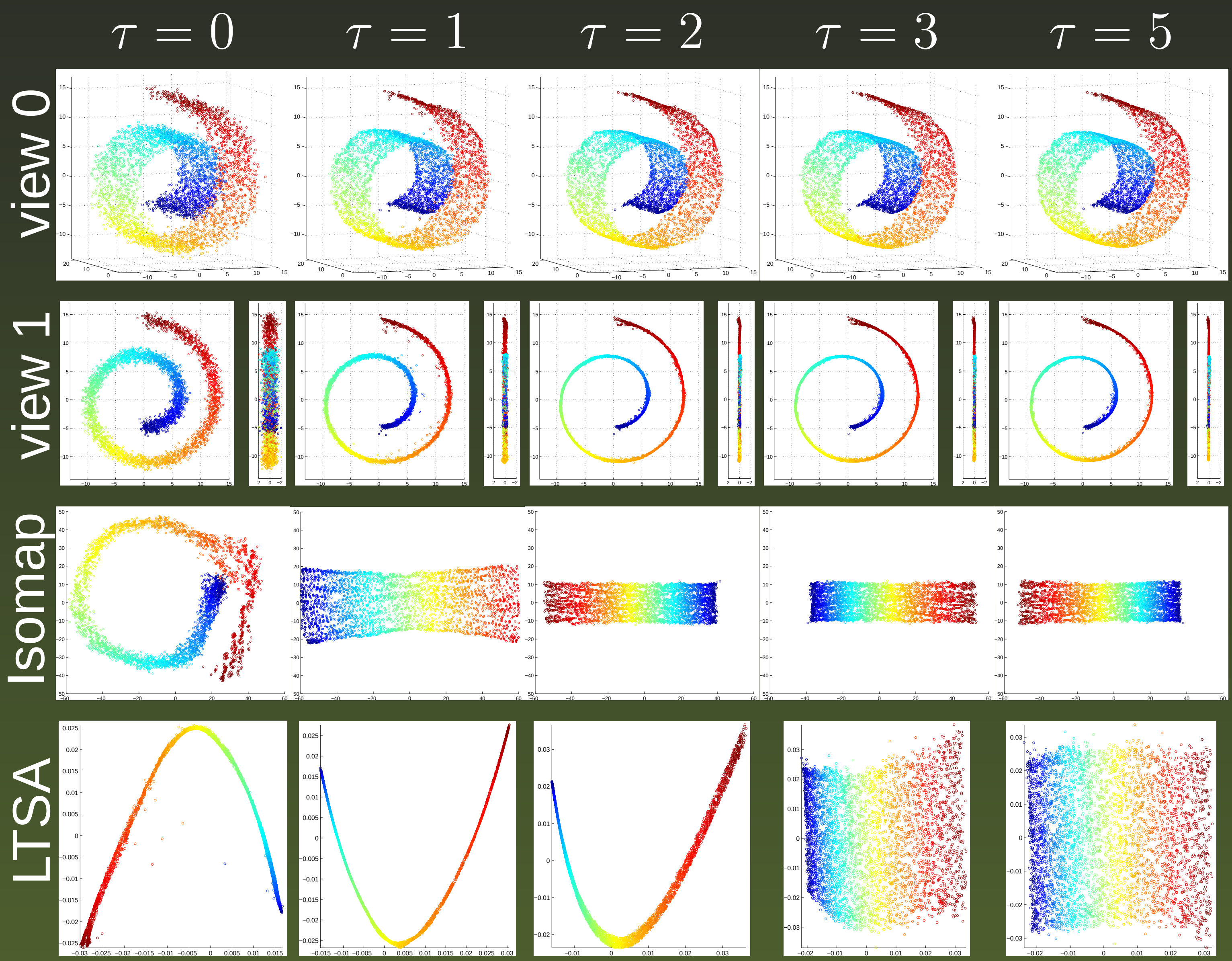
- Variations of MBMS:
  - MBMSf/MBMSk: use full/knn graph in predictor step.
  - Local Tangent Projection (LTP): MBMSk with $\sigma = \infty$.
  - GBMS: $L = 0$, no corrector step.
- User parameters can be determined by cross-validation for supervised problem.
- Stopping criteria: orthogonal variance $\lambda_\perp$ (sum of the trailing $D - L$ eigenvalues of $\mathbf{x}_n$'s local covariance) is small.

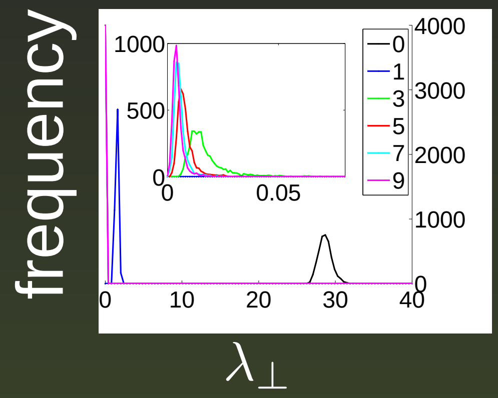
Experiment: noisy spiral

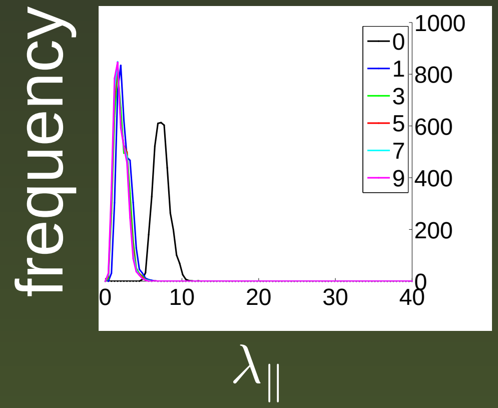


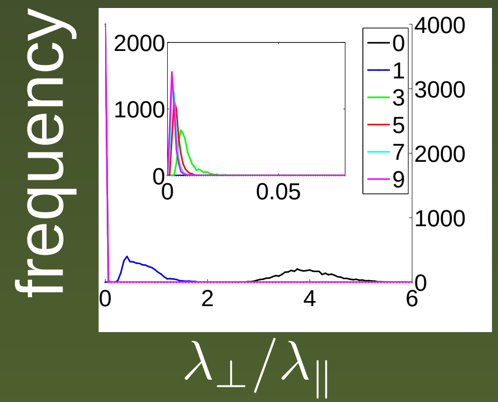
Denoising a noisy spiral with outliers over iterations. ☞

Experiment: preprocessing for spectral methods









Dimensionality reduction with Isomap and LTSA for iterations of MBMSk. ☞

Experiment: preprocessing for classifying MNIST

We denoise images of each digit separately using MBMSk.

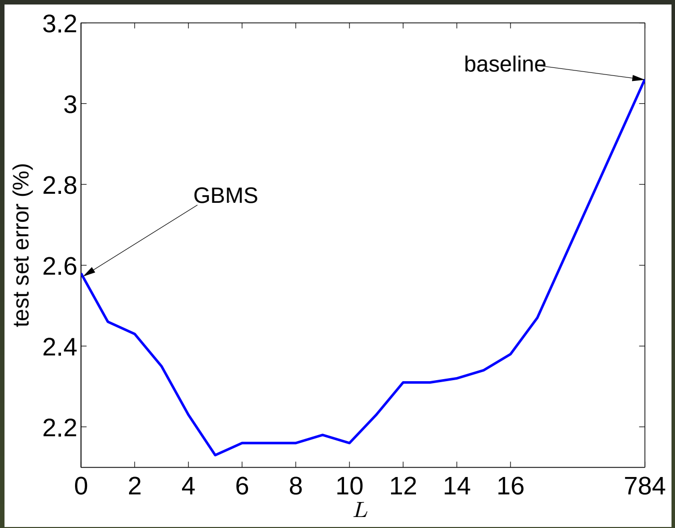
Sample pairs of (original,denoised) images from the training set.

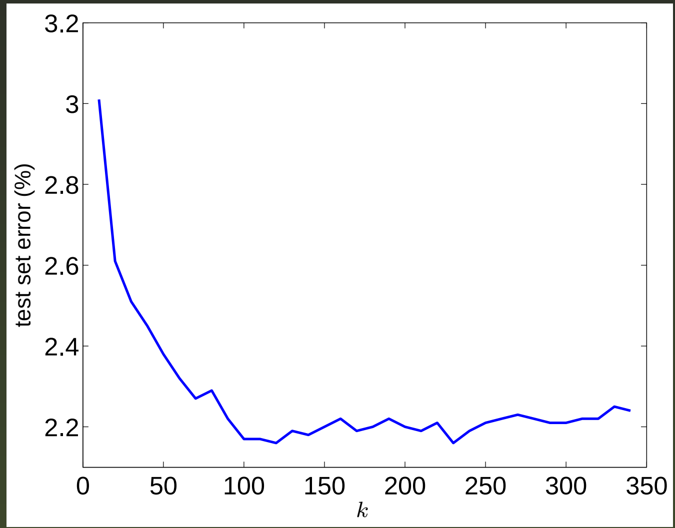
Experiment: preprocessing for classifying MNIST

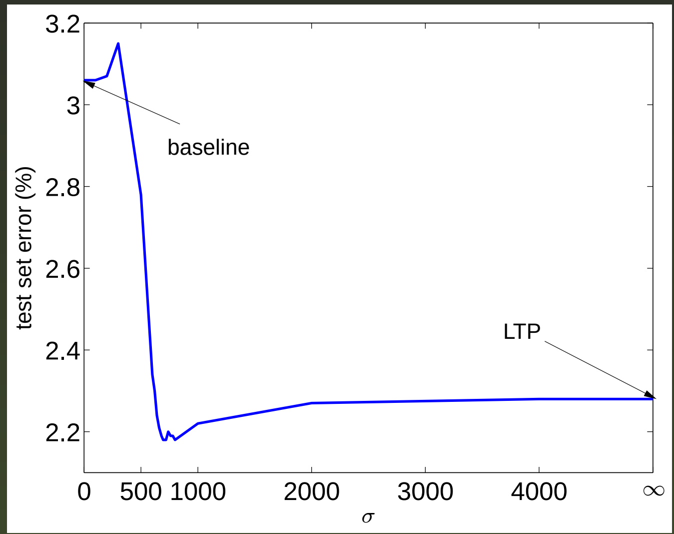
Classify test set using denoised training set and Nearest Neighbor.

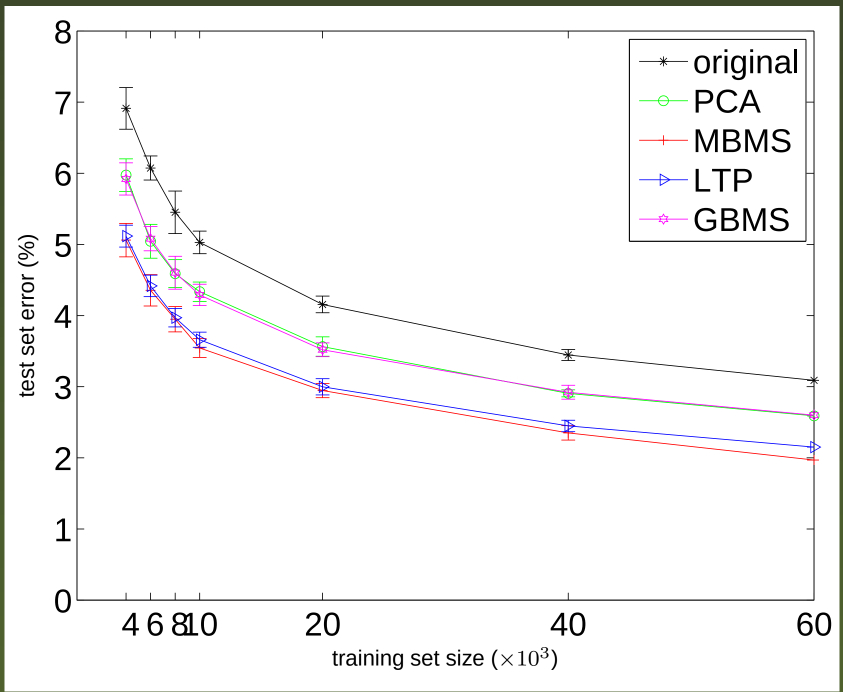
Some misclassified images. Each triplet is (test, original-nearest-neighbor, denoised-nearest-neighbor) and the corresponding label is above each image, with errors highlighted.

Experiment: preprocessing for classifying MNIST









Top 3 plots: 5–fold cross-validation error (%) curves with a nearest-neighbor classifier on training set using MBMSk.

Bottom left plot: denoising and classification of the MNIST test set, by training on the entire training set and smaller subsets.

Conclusion

- Very effective at denoising in a handful of iterations.
- Nonparametric and deterministic.
- Causing very small shrinkage or distortion.
- Able to handle large noise and extreme outliers.

Outline

- Manifold Blurring Mean-shift (MBMS) algorithm for manifold denoising
- MBMS for matrix completion
- K-modes algorithm for clustering
- Laplacian K-modes algorithm for clustering

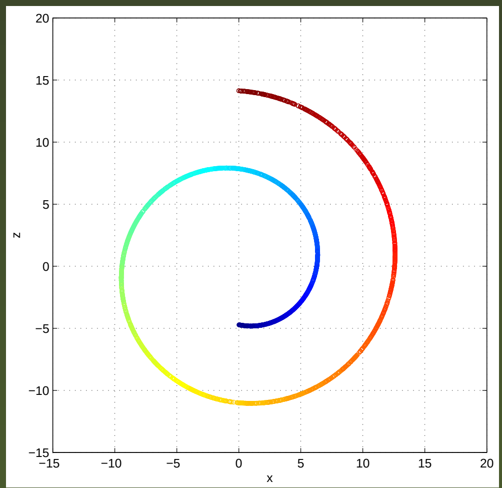
Problem Setting

- Given a set of data points $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_N] \subset \mathbb{R}^D$, where each point may contain missing entries.
  - $\mathbf{X}^{\mathcal{M}}$ and $\mathbf{X}^{\mathcal{P}}$ indicate the selection of missing or present entries $\mathbf{X}$, where $\mathcal{P} \subset \mathcal{U}$, $\mathcal{M} = \mathcal{U} - \mathcal{P}$ and $\mathcal{U} = \{(d, n) : \ d = 1, \ldots, D, \ n = 1, \ldots, N\}$.
  - Indices $\mathcal{P}$ and values $\overline{\mathbf{X}}^{\mathcal{P}}$ of the present entries are the data of the problem.
- An ill-posed problem. Very important in industrial applications.

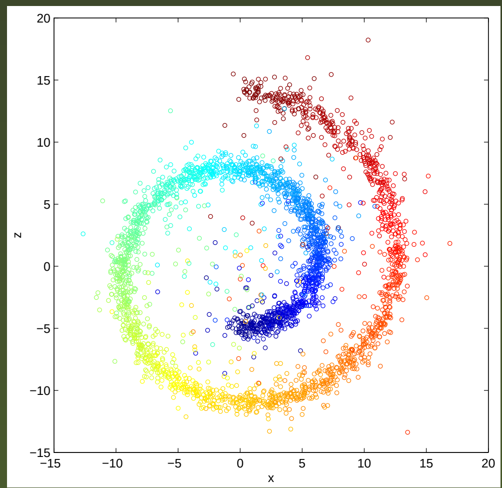
5	1	?	2	3	?	?
?	2	?	4	1	?	1
?	?	?	5	?	3	2
4	?	?	1	2	?	4
2	3	5	?	?	?	?
?	4	2	?	5	1	3
?	?	3	?	1	2	2

Motivation

- Popular approaches for matrix completion
  - Low-rank: $\min_{\mathbf{X}} \|\mathbf{X}\|_*$ s.t. $\mathbf{X}_{\mathcal{P}} = \overline{\mathbf{X}}_{\mathcal{P}}$
  - Matrix factorization (probabilistic and nonlinear extensions): $\min_{\mathbf{L},\mathbf{R}} \sum_{(i,j)\in\mathcal{P}} (\mathbf{X}_{ij} - \mathbf{L}_i \mathbf{R}_j^T)^2 + \lambda(\|\mathbf{L}\|_{\text{Fro}}^2 + \|\mathbf{R}\|_{\text{Fro}}^2)$.
- Globally low-rank assumption is too restrictive for nonlinear manifold. We use locally low-rank assumption instead.



data



SVP

MBMS for matrix completion

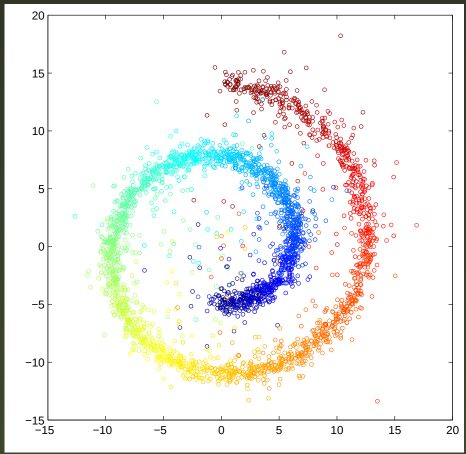
- GBMS maximizes the following objective function by taking parallel steps of the mean-shift form for each point:

$$E(\mathbf{X}) = \frac{1}{N} \sum_{n,m=1}^{N} G\left(\left\| \frac{\mathbf{x}_n - \mathbf{x}_m}{\sigma} \right\|^2 \right)$$

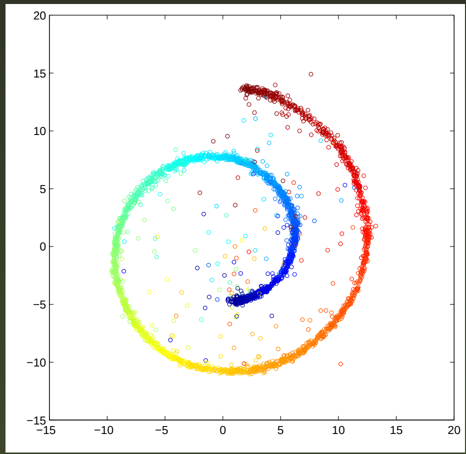
- Apply GBMS to matrix completion by adding the constraints given by the present values $\mathbf{X}_{\mathcal{P}} = \overline{\mathbf{X}}_{\mathcal{P}}$.
- We iteratively carry out a GBMS denoising step on $\mathbf{X}$ and refill $\mathbf{X}_{\mathcal{P}}$ to the present values; equivalent to a gradient projection algorithm.
- MBMS can be applied instead to prevent shrinkage.
- Hyperparameters and number of iterations can be cross-validated on held out present entries.

Synthetic example

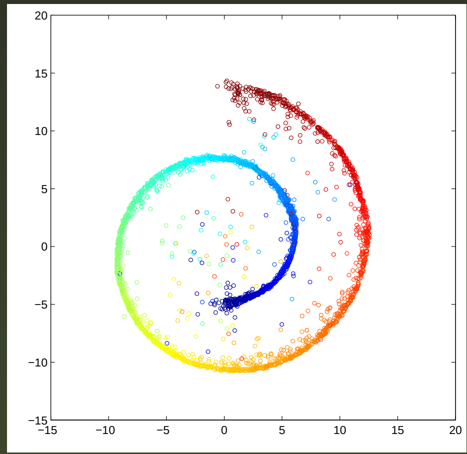
SVP
$\tau = 0$



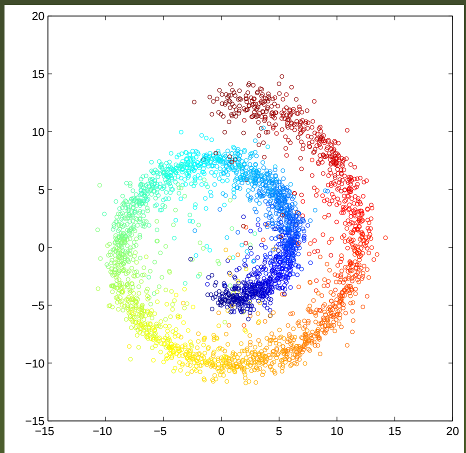
SVP + GBMS
$\tau = 1$



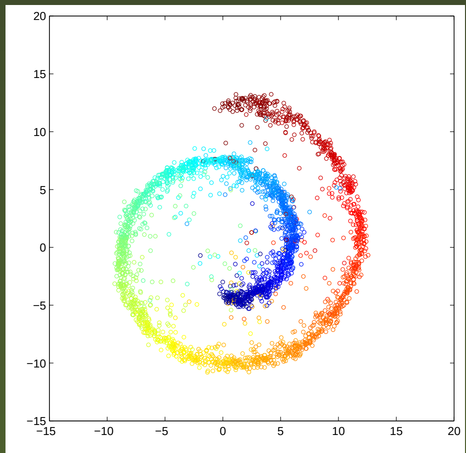
SVP + MBMS
$\tau = 2$



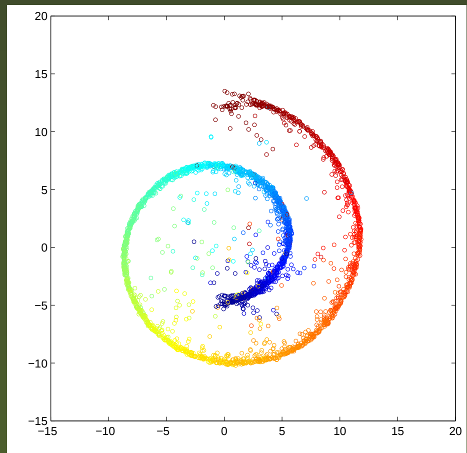
Gaussian
$\tau = 0$



Gaussian + GBMS
$\tau = 1$



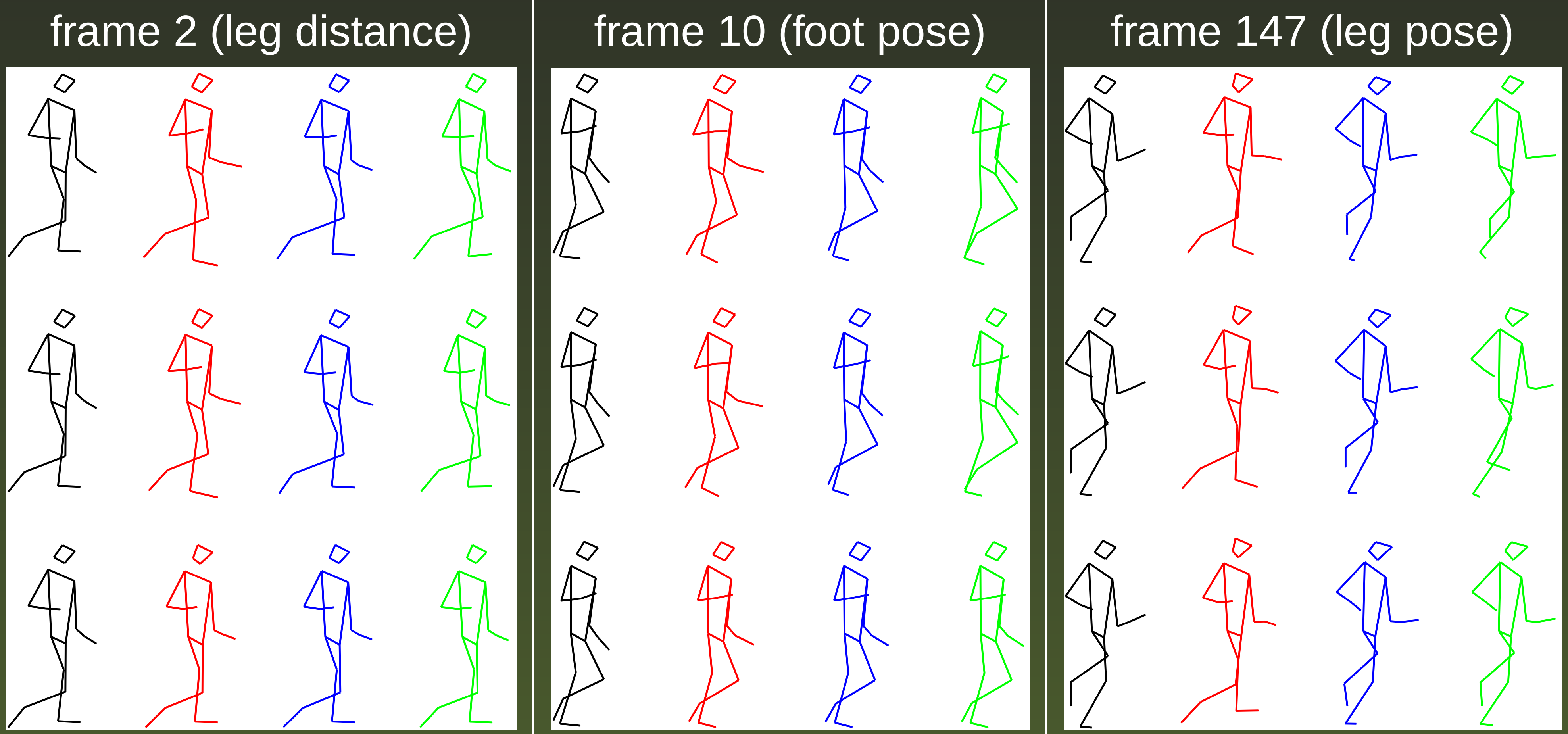
Gaussian + MBMS
$\tau = 25$



Denoising effect of different algorithms on 100D swissroll.

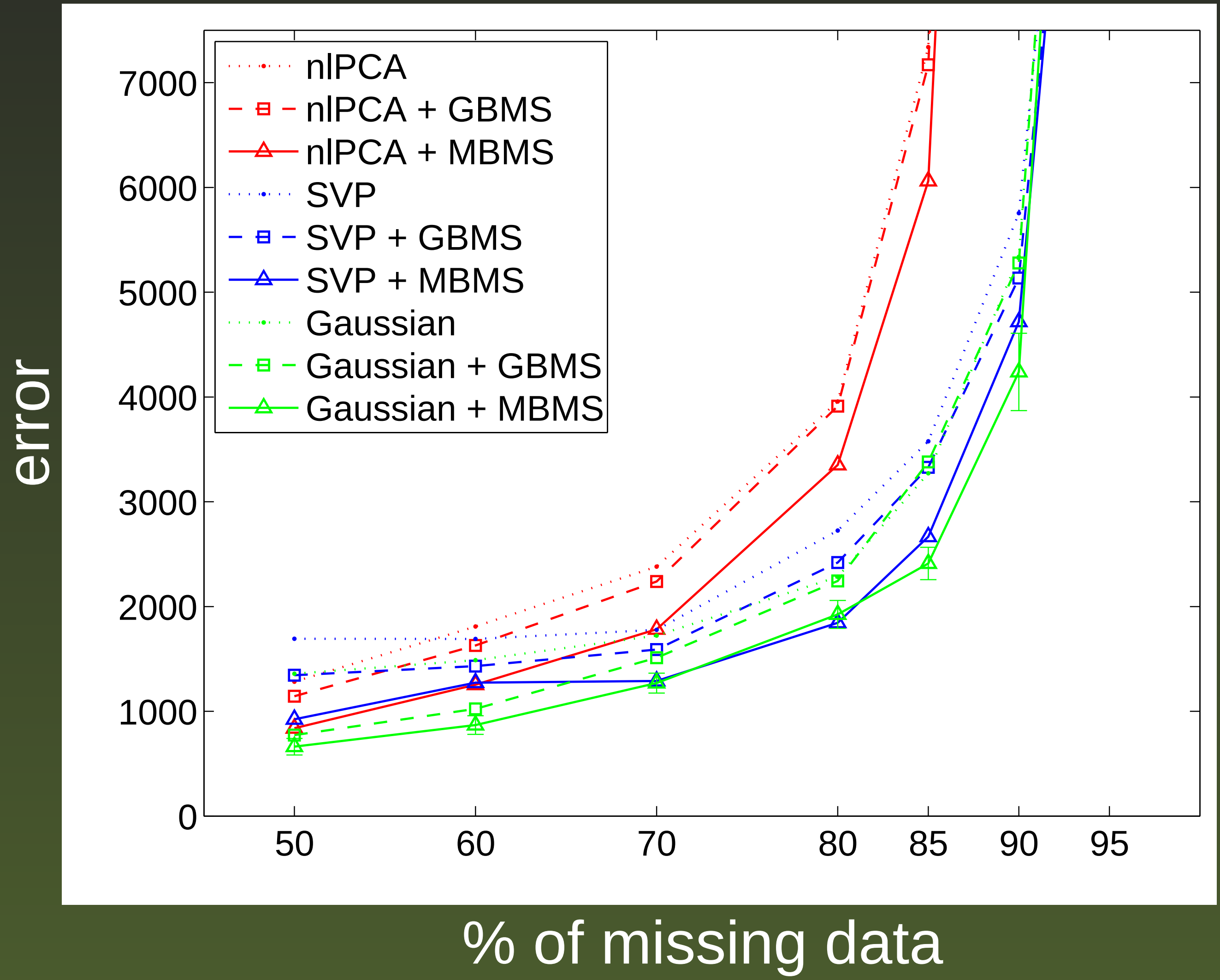
Experiment: Mocap

Running sequence with 148 samples of 150D sensor readings. ☞



Sample reconstructions when 85% percent data is missing. *Row 1*: initialization. *Row 2*: init+GBMS. *Row 3*: init+MBMS. Color indicates different initialization: original data, nlPCA, SVP, Gaussian.

Experiment: Mocap



Results on Mocap dataset. Mean of errors (RSSE) of 5 runs obtained by different algorithms for varying percentage of missing values.

Experiment: MNIST digit 7

$6\,265$ greyscale images of size 28×28, 50% entries missing.

Methods	RSSE	mean	stdev
nlPCA	7.77	26.1	42.6
SVP	6.99	21.8	39.3
+ GBMS (400,140,0,1)	6.54	18.8	37.7
+ MBMS (500,140,9,5)	6.03	17.0	34.9

Reconstruction errors of different algorithms at their optimal parameters.

Experiment: MNIST digit 7

Selected reconstructions of MNIST block-occluded digits ‘7’.

Conclusion

- We propose new denoising paradigm for matrix completion, which generalizes the commonly used assumption of low rank.
- MBMS-based algorithm bridges the gap between pure denoising (GBMS) and local low rank.
- Denoising works due to the fundamental fact that a missing value can be predicted by averaging nearby present values, a common approach in recommender systems.

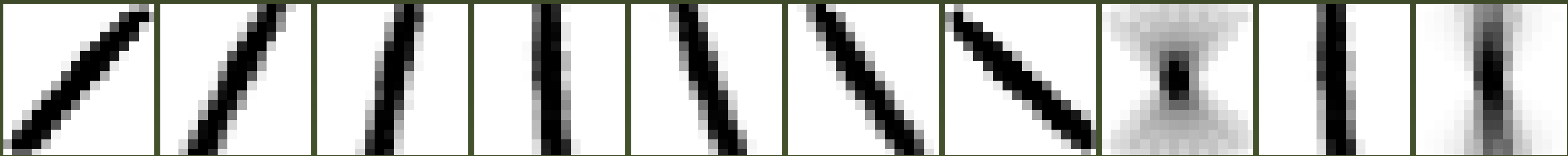
Outline

- Manifold Blurring Mean-shift (MBMS) algorithm for manifold denoising
- MBMS for matrix completion
- K-modes algorithm for clustering
- Laplacian K-modes algorithm for clustering

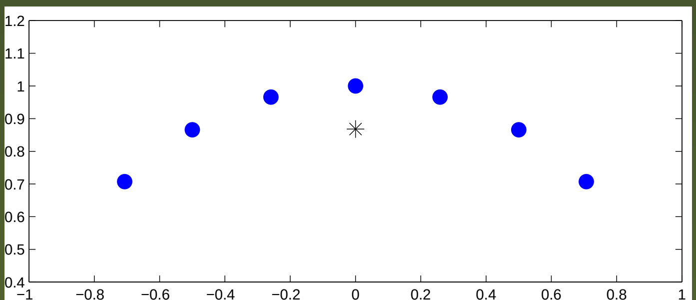
Motivation

- Given a dataset $\mathbf{x}_1, \dots, \mathbf{x}_N \in \mathbb{R}^D$, centroids-based clustering
  - partition data points into groups,
  - estimate a representative $\mathbf{c}_k \in \mathbb{R}^D$ of each cluster k.
- Popular algorithms of this type: K-means, mean-shift, K-medoids.
- No K-modes algorithm exists. Mode $\Rightarrow$ high density $\Rightarrow$ representativeness.

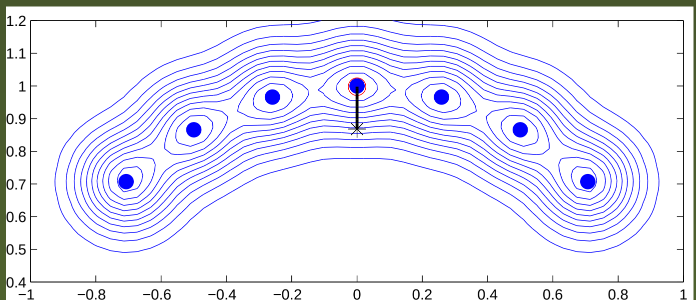
data K-means K-modes GMS



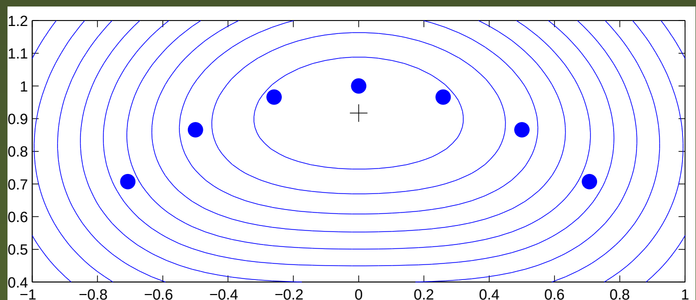
K-means



K-modes ($\sigma = 0.1$)



GMS ($\sigma = 0.45$)



K-means algorithm

Optimizes over assignment $\mathbf{Z}$ and centroids $\mathbf{C}$

$$\min_{\mathbf{Z},\mathbf{C}} \quad \sum_{k=1}^{K} \sum_{n=1}^{N} z_{nk} \left\| \mathbf{x}_n - \mathbf{c}_k \right\|^2$$

$$\text{s.t.} \quad z_{nk} \in \{0,1\}, \sum_{k=1}^{K} z_{nk} = 1, \text{ for } n = 1, \ldots, N.$$

- Efficient algorithm alternates $\mathbf{Z}$-step (computes assignment) and $\mathbf{C}$-step (computes mean).
- Can only produce convex clusters (Voronoi tessellation).
- Cluster mean may not be valid pattern.
- Sensitive to noise and outliers.

K-modes: objective function

$$\max_{\mathbf{Z},\mathbf{C}} \quad \sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk} G\left(\left\| \frac{\mathbf{x}_n - \mathbf{c}_k}{\sigma} \right\|^2 \right)$$

$$\text{s.t.} \quad z_{nk} \in \{0,1\}, \quad \sum_{k=1}^{K} z_{nk} = 1, \ \text{for } n = 1,\dots,N,$$

- Sum of KDE but separately for each cluster.
- Combines the notions of assignment and density estimation.
- Two limit cases: "K-medoids" when $\sigma \to 0$, K-means when $\sigma \to \infty$.
- Alternating optimization with guaranteed convergence
  - $\mathbf{Z}$-step: decouples over points, same assignment rule as K-means.
  - $\mathbf{C}$-step: decouples over clusters, mode-finding within each cluster.

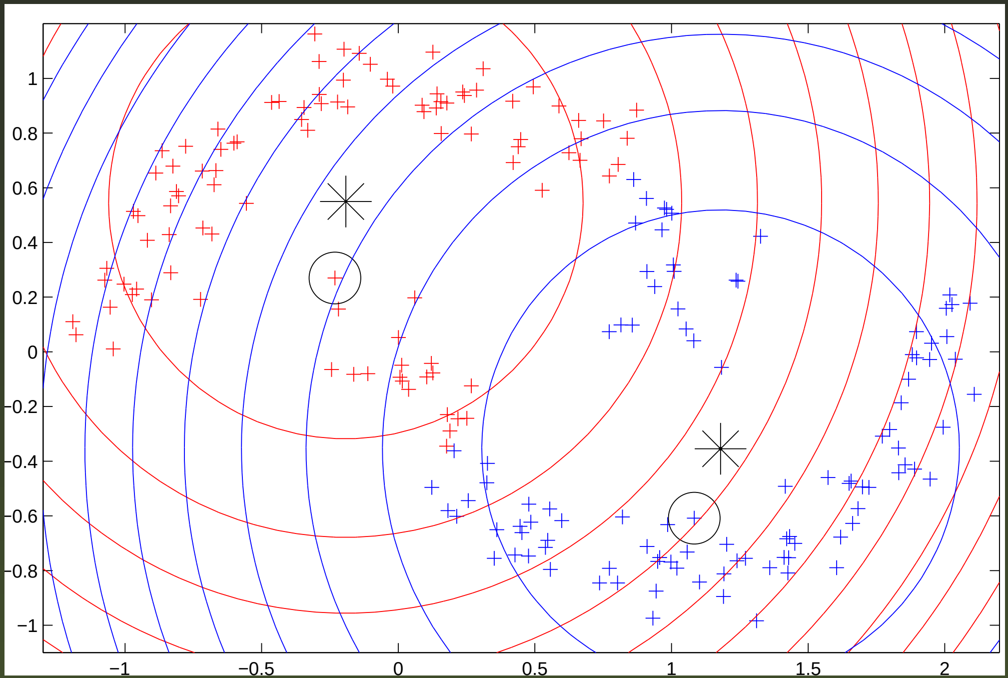
K-modes: homotopy algorithm

Start with $\sigma = \infty$ (K-means), gradually decrease σ while running J iterations of the fixed-σ K-modes algorithm for each value of σ, until reach a target value σ^*.

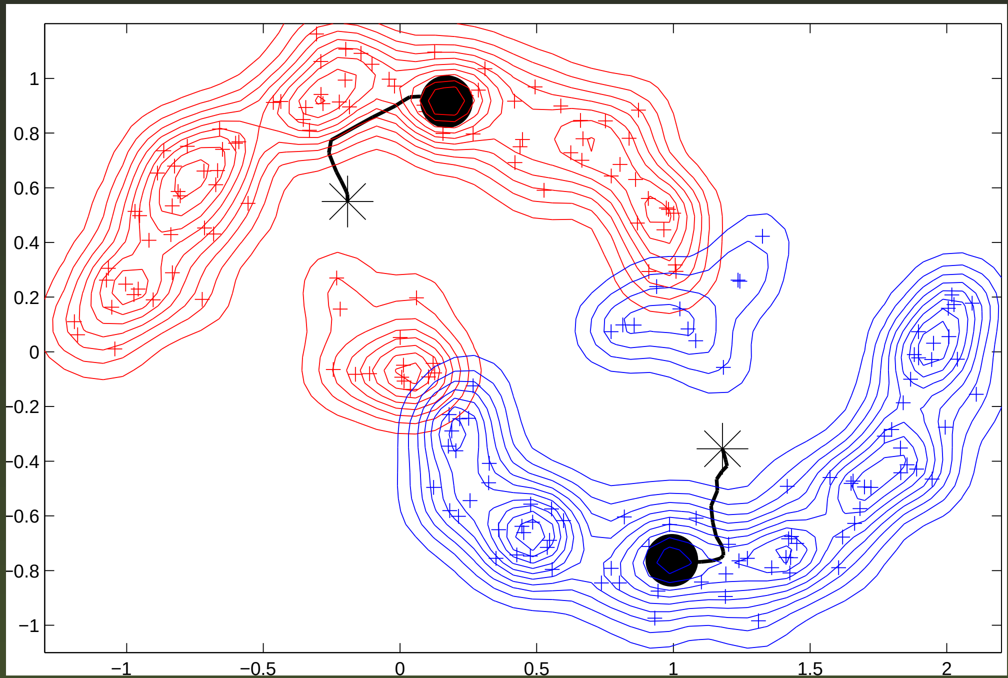
- A deterministic algorithm given local optimum found by K-means.
- Follows an optimum path $(\mathbf{Z}(\sigma), \mathbf{C}(\sigma))$ for $\sigma \in [\sigma^*, \infty)$.
- Homotopy techniques tends to find better optima than starting directly at the target value σ^*.
- Representative, valid centroids are obtained for a wide range of intermediate σ values.

K-modes: homotopy algorithm

$\sigma = \infty$



$\sigma = 0.1$

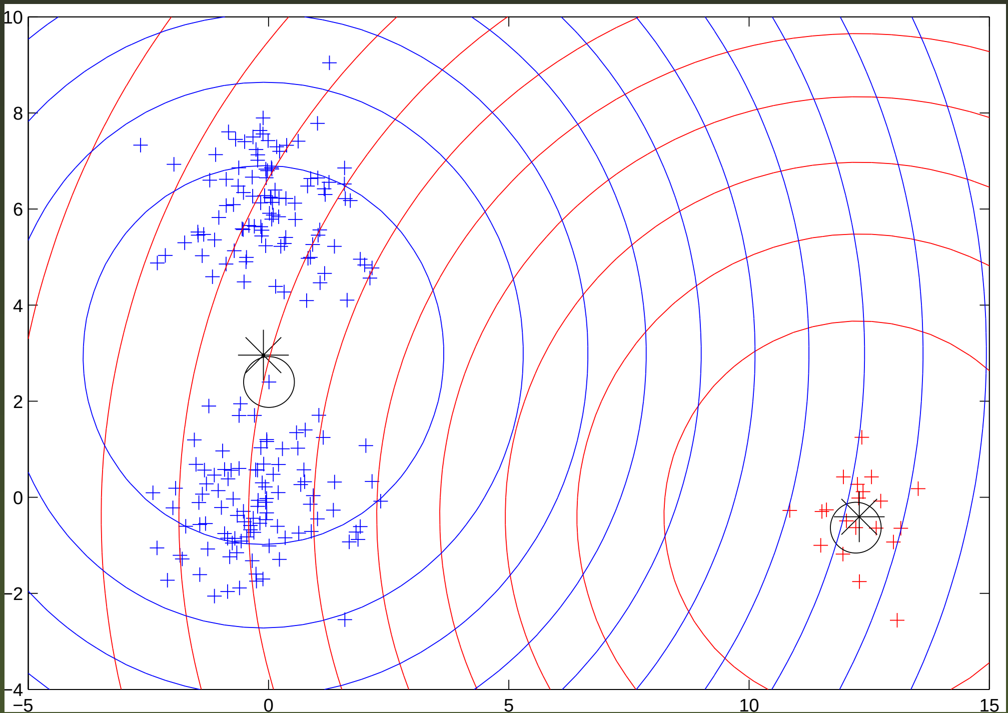


☞ $K = 2$. No value of σ results in two modes that separate the (nonconvex) moons.

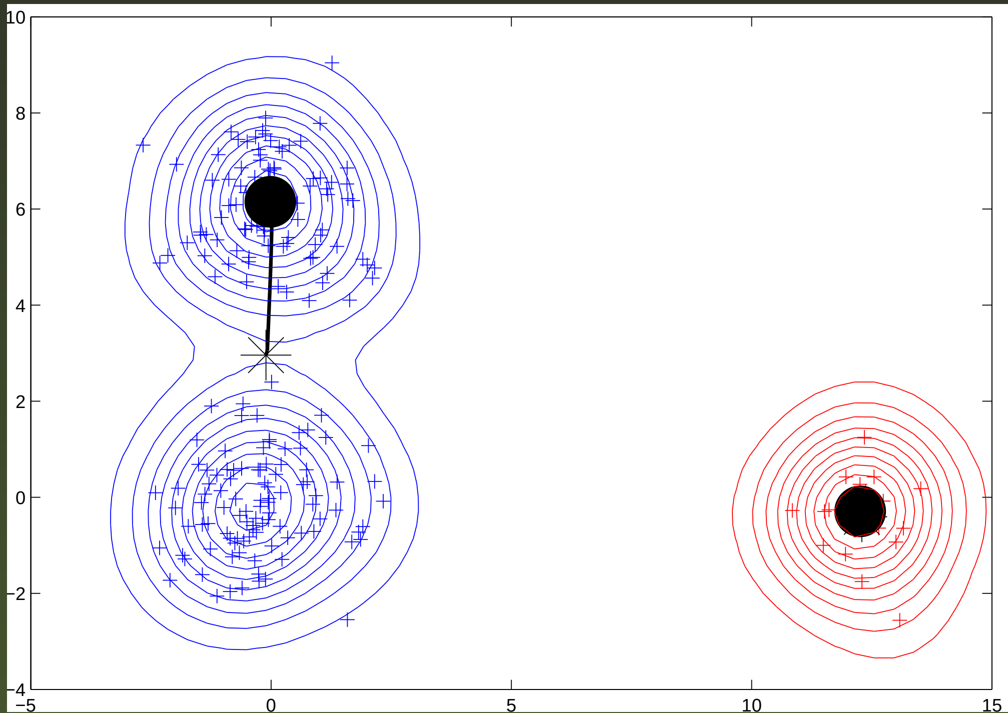
Experiment: misspecification of K

3 natural clusters, but use $K = 2$. ☞

$\sigma = \infty$

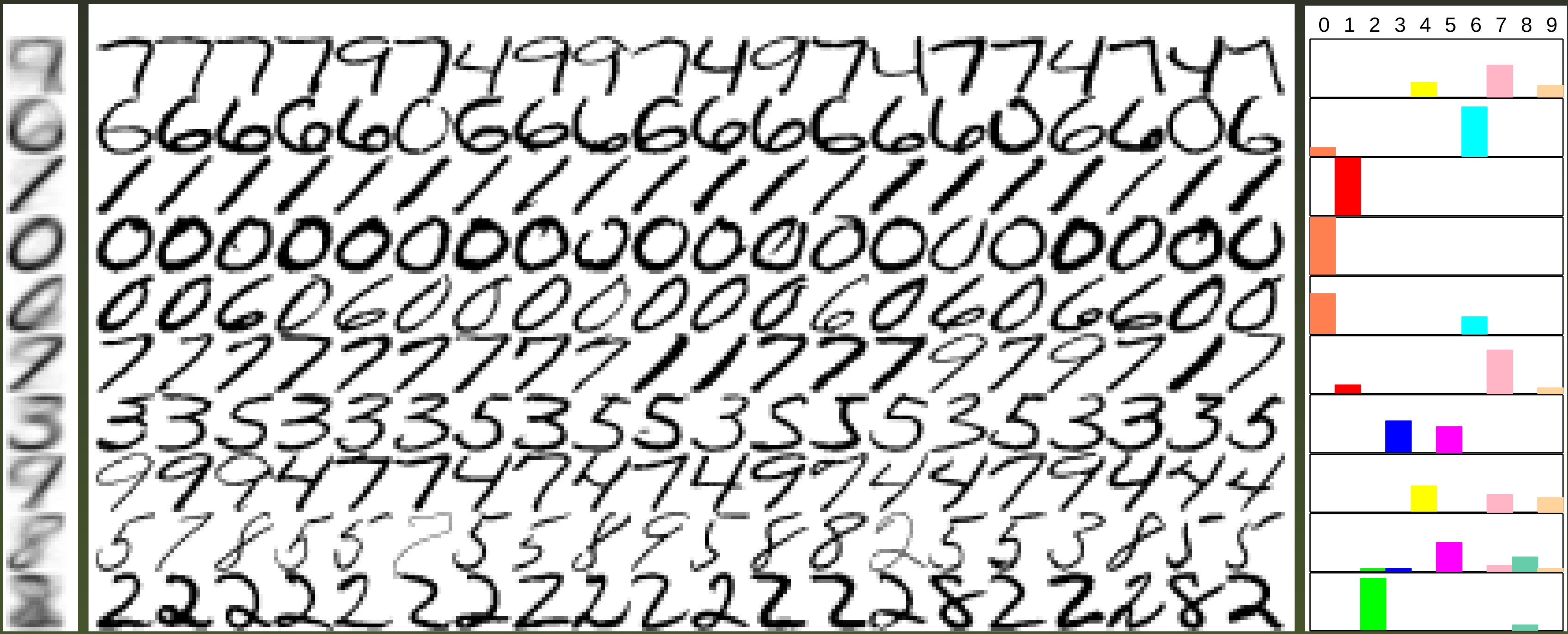


$\sigma = 1$



Experiment: handwritten digit images

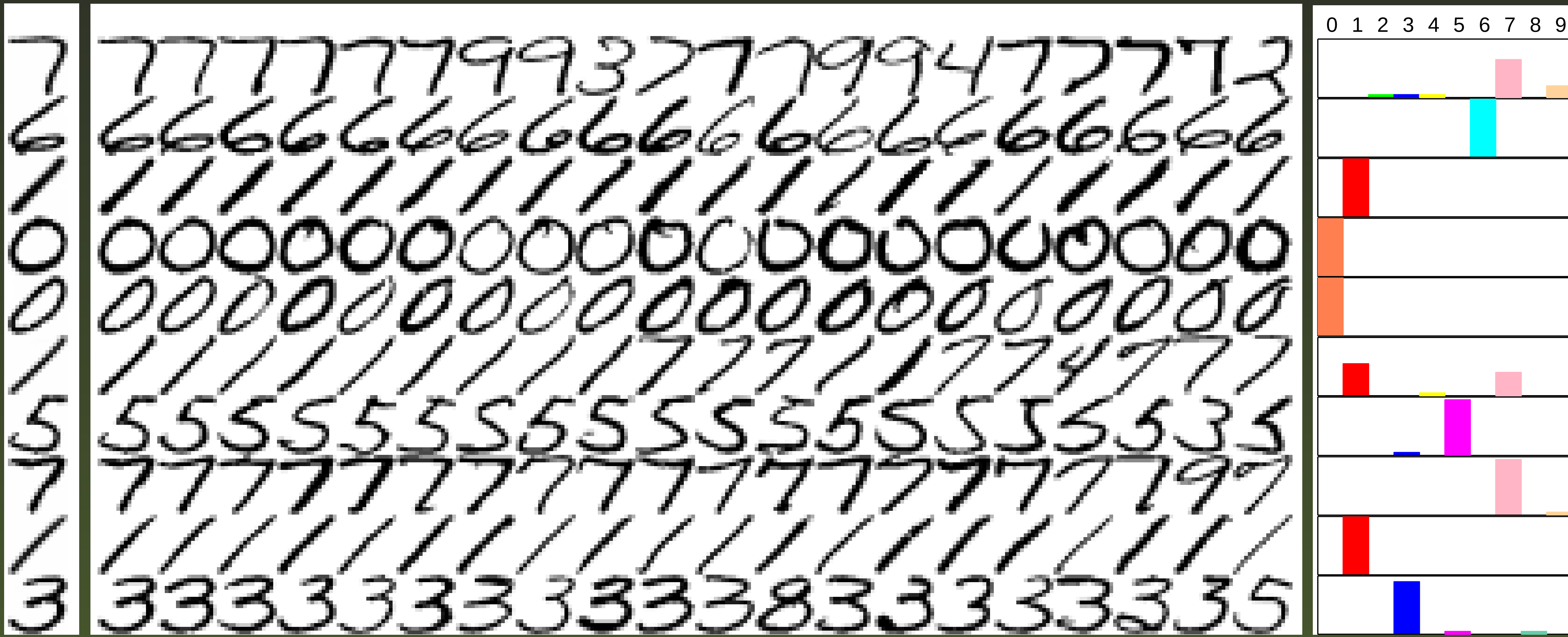
K-means result ($K = 10$, $\sigma = \infty$)



- Centroids are average of different classes.
- Neighborhoods are not homogeneous/pure.

Experiment: handwritten digit images

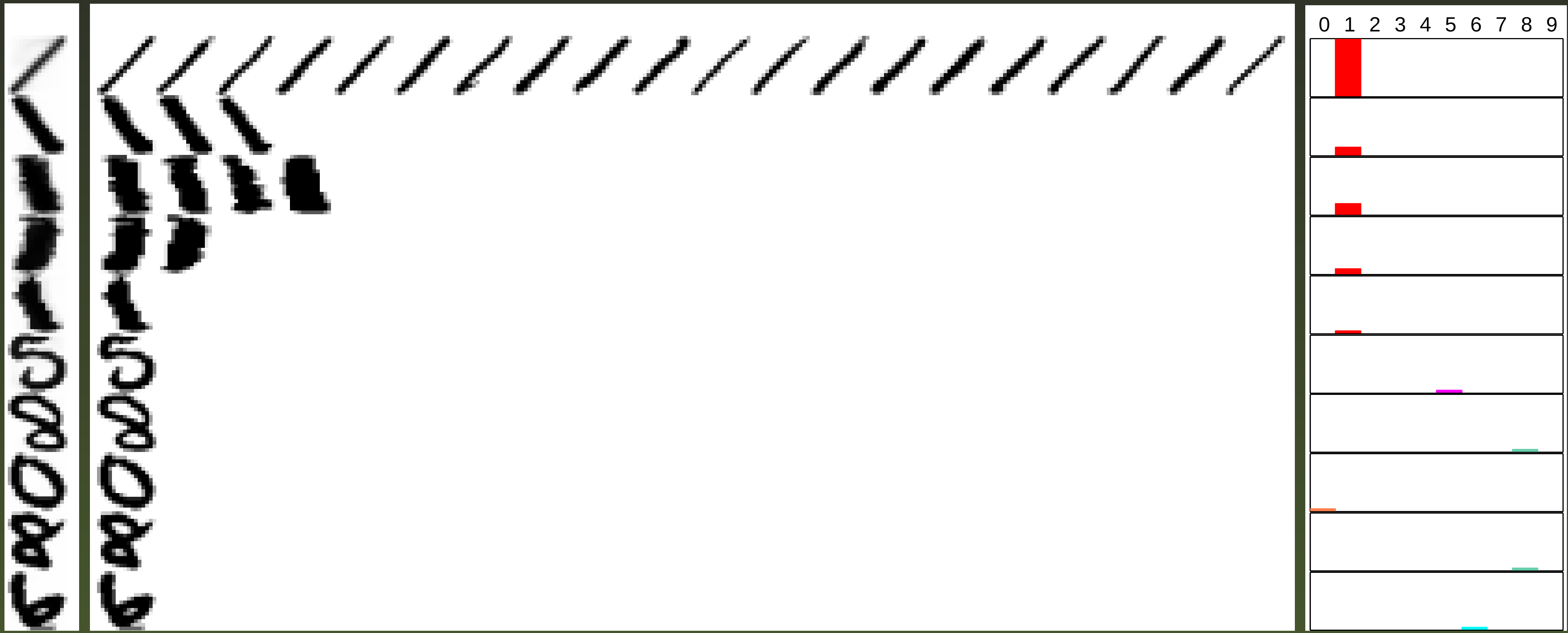
K-modes result ($K = 10$, $\sigma = 1$) ☞



- Centroids are very representative.
- Neighborhoods are homogeneous/pure.

Experiment: handwritten digit images

Mean-shift result ($\sigma = 1.8369$)



- In high dimensions, many modes have very few associated points.

Summary

- K-modes is more robust than K-means and GMS to outliers and parameter misspecification.
- K-modes will return exactly K modes (one per cluster) no matter the value of σ, and whether the dataset KDE has more or fewer than K modes.
- Centroids are representative, valid patterns.

Outline

– Manifold Blurring Mean-shift (MBMS) algorithm for manifold denoising

– MBMS for matrix completion

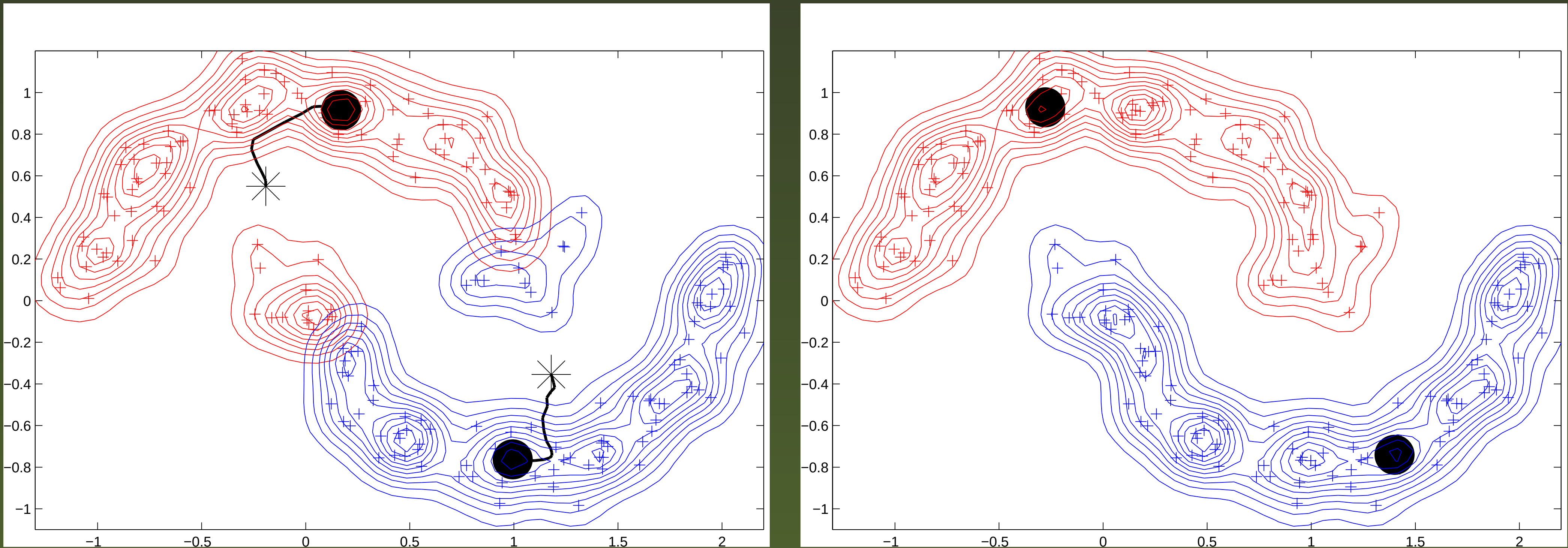
– K-modes algorithm for clustering

– Laplacian K-modes algorithm for clustering

Motivation

- Limitation of K-modes assignment: can only find convex clusters.
- In addition to representative centroids and density estimate, we want more flexible assignment.

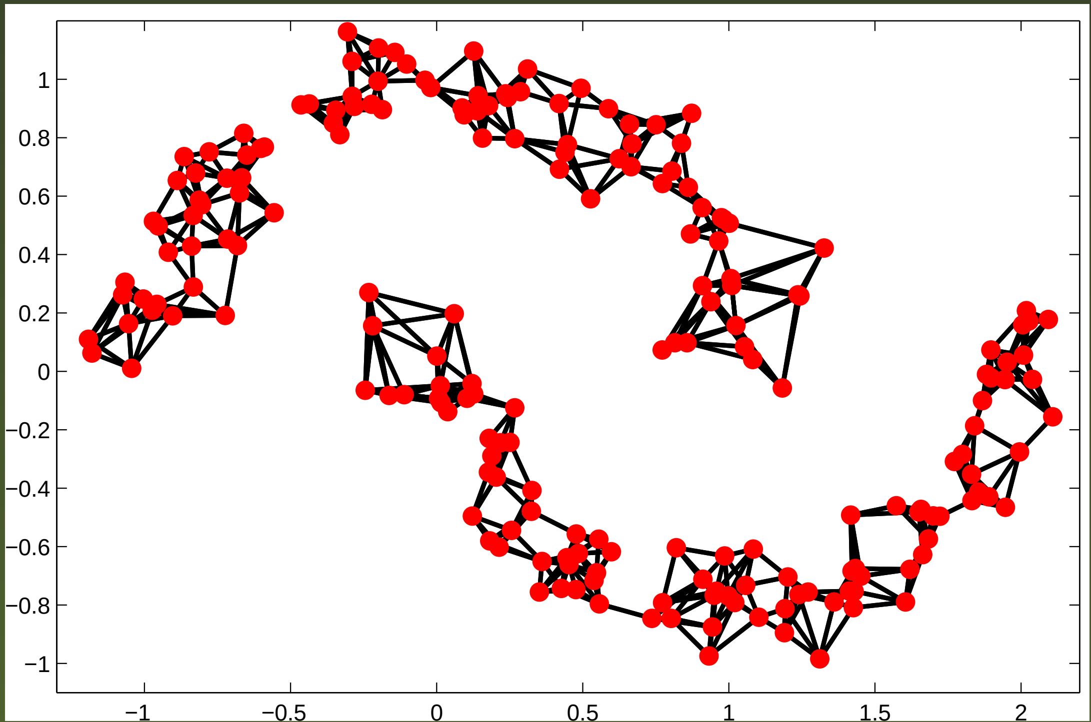
K-modes ($\sigma = 0.1$) Laplacian K-modes ($\sigma = 0.1$)



Laplacian smoothing

Key to separate clusters with manifold structure: nearby data points should have similar assignment.

1. Relax the assignment to be continuous, but constrain them to probabilities.
2. Build a graph on the dataset, let w_{mn} be the weight between $\mathbf{x}_m$ and $\mathbf{x}_n$.
3. Add Laplacian smoothing term $\frac{\lambda}{2}\sum_{m=1}^{N}\sum_{n=1}^{N} w_{mn}\left\|\mathbf{z}_m - \mathbf{z}_n\right\|^2$.



Laplacian K-modes: objective function

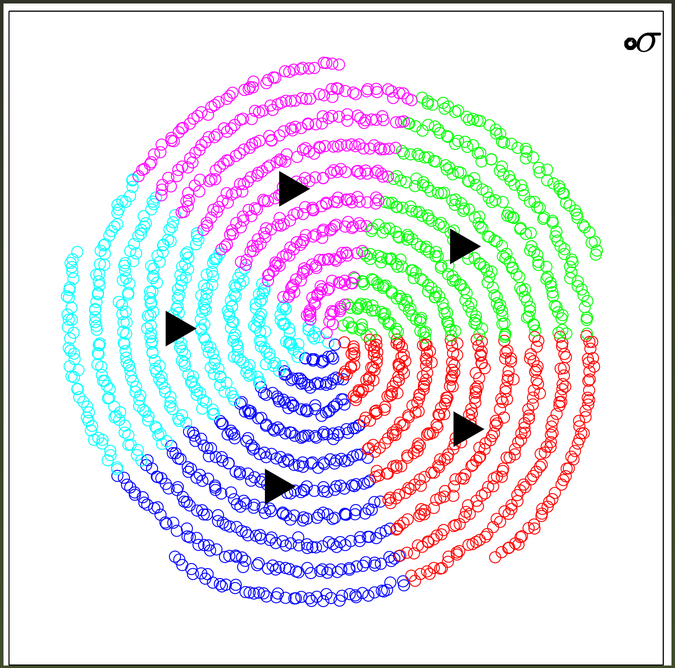
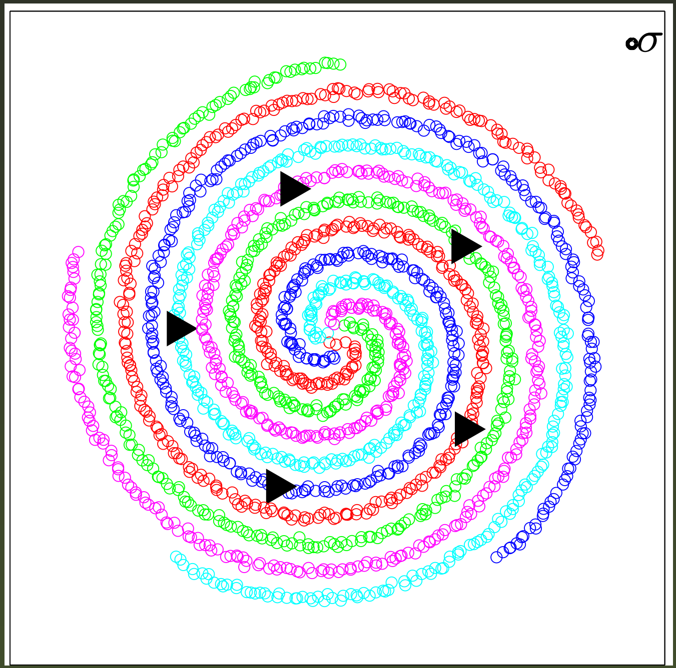
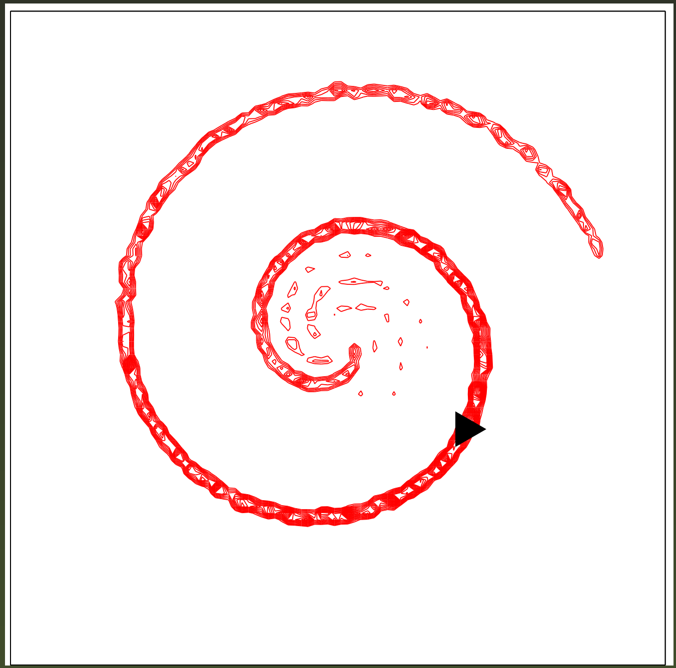
$$\min_{\mathbf{Z},\mathbf{C}} \quad \frac{\lambda}{2}\sum_{m=1}^{N}\sum_{n=1}^{N} w_{mn}\left\|\mathbf{z}_m-\mathbf{z}_n\right\|^2-\sum_{n=1}^{N}\sum_{k=1}^{K} z_{nk}G\left(\left\|\frac{\mathbf{x}_n-\mathbf{c}_k}{\sigma}\right\|^2\right)$$

$$\text{s.t.} \quad \sum_k z_{nk}=1,\ \text{for } n=1,\dots,N,$$

$$z_{nk}\geq 0,\ \text{for } n=1,\dots,N,\ k=1,\dots,K.$$

- Obtain hard assignment by choosing largest assignment probability.
- Alternating optimization
  - $\mathbf{C}$-step: decouples over clusters, mode-finding within each cluster.
  - $\mathbf{Z}$-step: convex quadratic program, solved with gradient projection.
- Homotopy in (σ,λ) can be done similarly as in K-modes.

Effect of Laplacian smoothing

K-modes	Laplacian K-modes	KDE
		

- K=5. K-modes assignment rule can never separate the spirals.

Out-of-sample problem

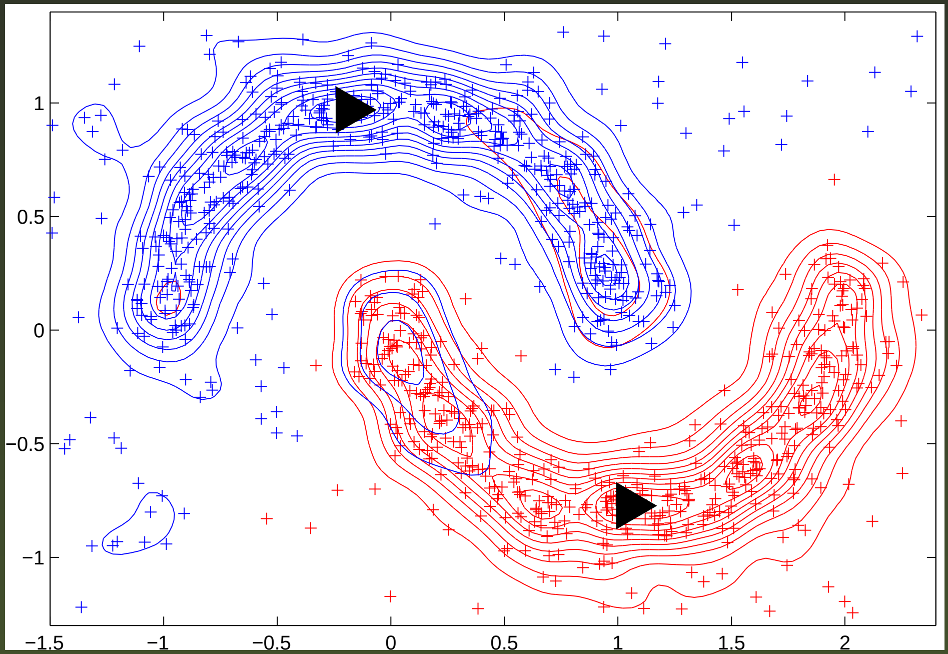
- Optimize assignment $\mathbf{z}$ of new point $\mathbf{x}$ given $\mathbf{Z}$ and $\mathbf{C}$ from training.
- The out-of-sample problem is equivalently

$$\min_{\mathbf{z}} \quad \frac{1}{2}\left\|\mathbf{z} - \bar{\mathbf{z}} - \gamma\mathbf{q}\right\|^2, \qquad \text{s.t.} \quad \mathbf{z}^\top \mathbf{1}_K = 1, \quad \mathbf{z} \geq 0,$$

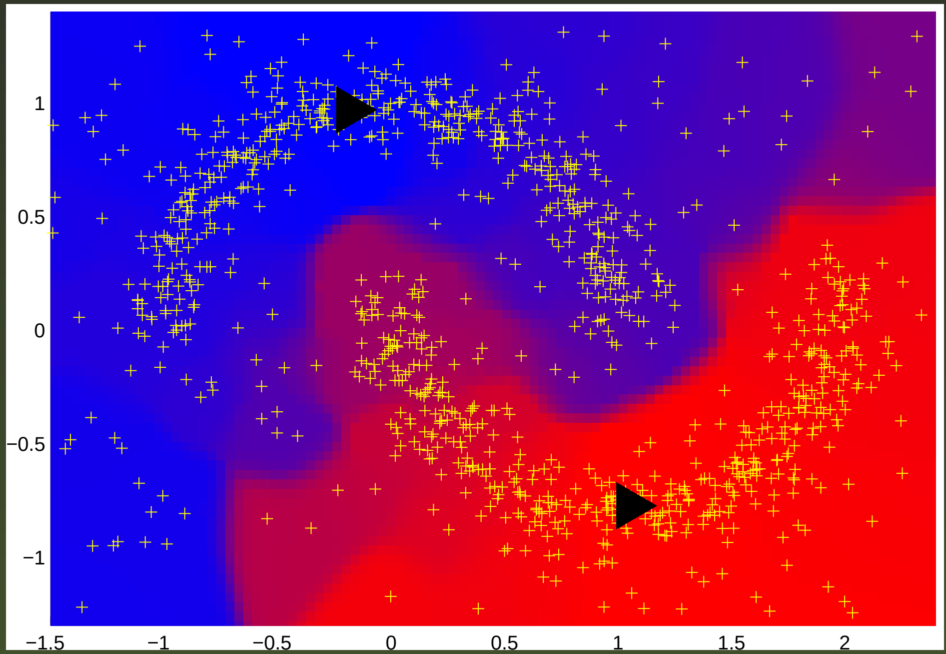
  where $\bar{\mathbf{z}}$ is the weighted mean of training assignments, $\mathbf{q}$ is soft distance to centroids.
- Projection of $\bar{\mathbf{z}} + \gamma\mathbf{q}$ onto the probability simplex.
- It is a mixture of two assignment rules and a nonlinear mapping.

Out-of-sample problem

Laplacian K-modes



Out-of-sample



☞ $K = 2$. Homotopy in σ for Laplacian K-modes.

Clustering analysis

Statistics of datasets.

dataset	size (N)	dimensionality (D)	# of classes (K)
MNIST (digit image)	2000	768	10
COIL20 (object image)	1440	1024	20
TDT2 (document)	9394	36771	30

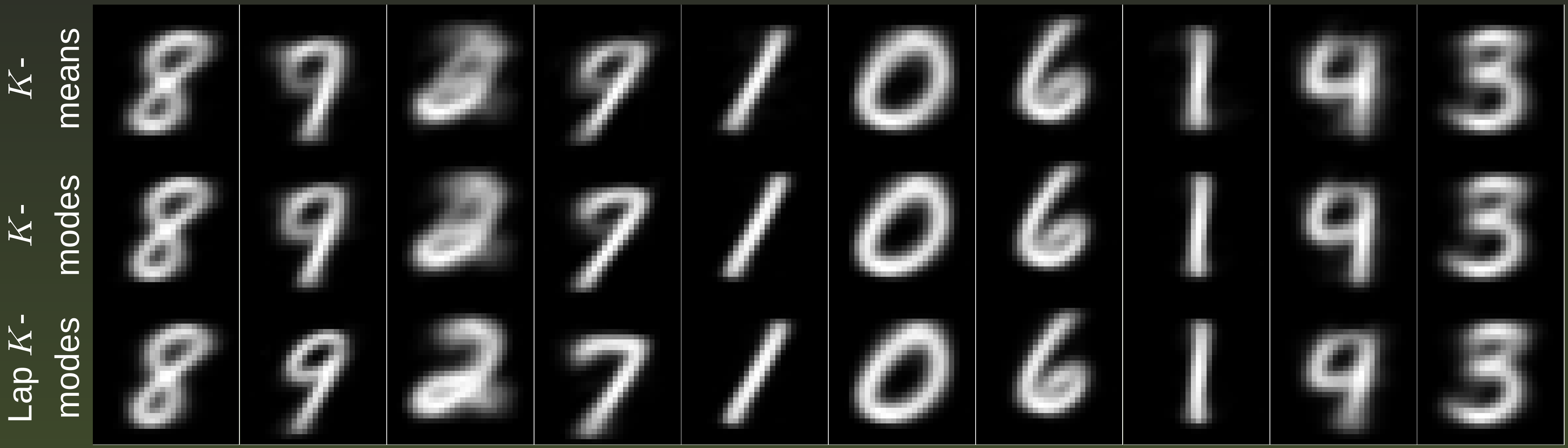
Clustering accuracy (%).

dataset	K-means	K-modes	GMS	NCut	GNMF	DCD	Lap. K-modes
MNIST	58.2	59.2	15.9	65.5	66.2	69.4	70.5
COIL–20	66.5	67.2	27.2	79.0	75.3	71.5	81.0 (81.5)
TDT2	68.9	70.0	N/A	88.4	88.6	55.1	91.4

Normalized Mutual Information (%).

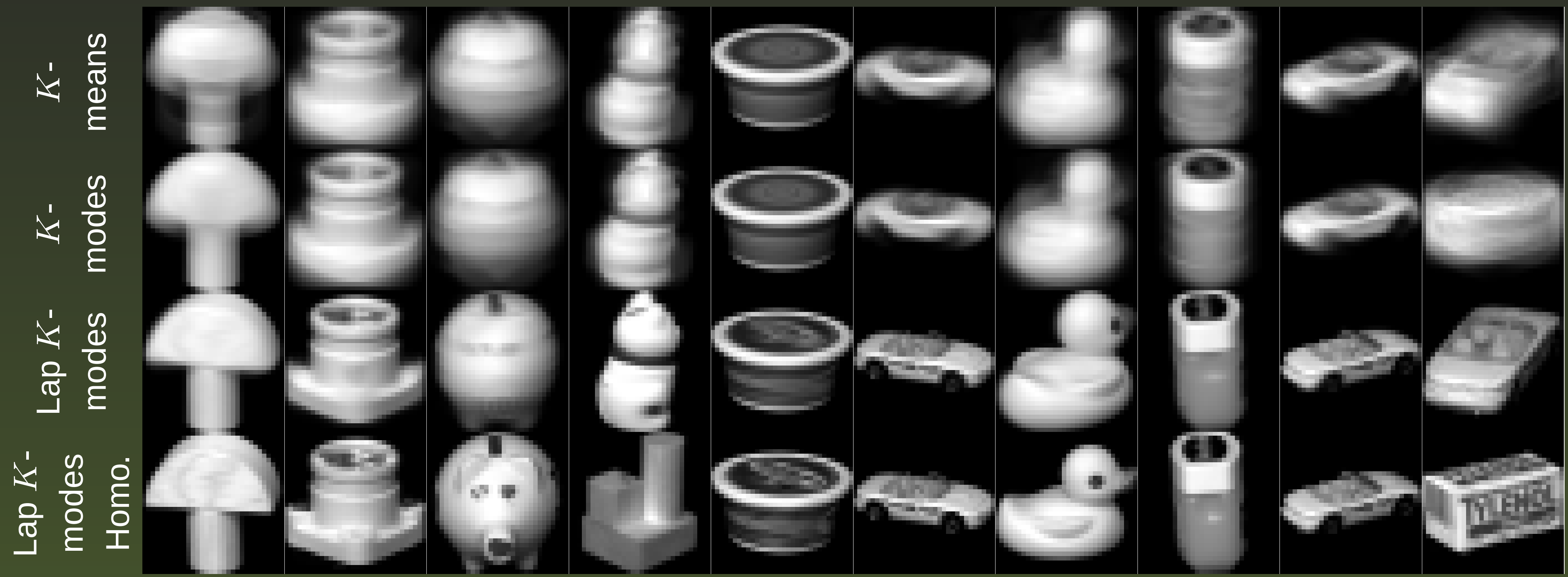
dataset	K-means	K-modes	GMS	NCut	GNMF	DCD	Lap. K-modes
MNIST	53.3	53.6	6.51	66.9	64.9	65.6	68.8
COIL–20	75.3	75.9	38.9	88.0	87.5	77.6	87.3 (88.0)
TDT2	75.3	75.8	N/A	83.7	83.7	68.6	88.8

Clustering analysis



Centroids found by different algorithms on MNIST.

Clustering analysis



Centroids found by different algorithms on COIL–20.

Summary

Comparison of different clustering algorithms.

	K-means	K-medoids	Mean-shift	Spectral clustering	K-modes	Laplacian K-modes
Centroids	likely invalid	“valid”	“valid”	N/A	valid	valid
Nonconvex clusters	no	depends	yes	yes	no	yes
Density	no	no	yes	no	yes	yes
Assignment	hard	hard	hard	hard	hard	soft
Cost/iteration	KND	KN^2D	N^2D	$N^2 \sim N^3$	KND	KND

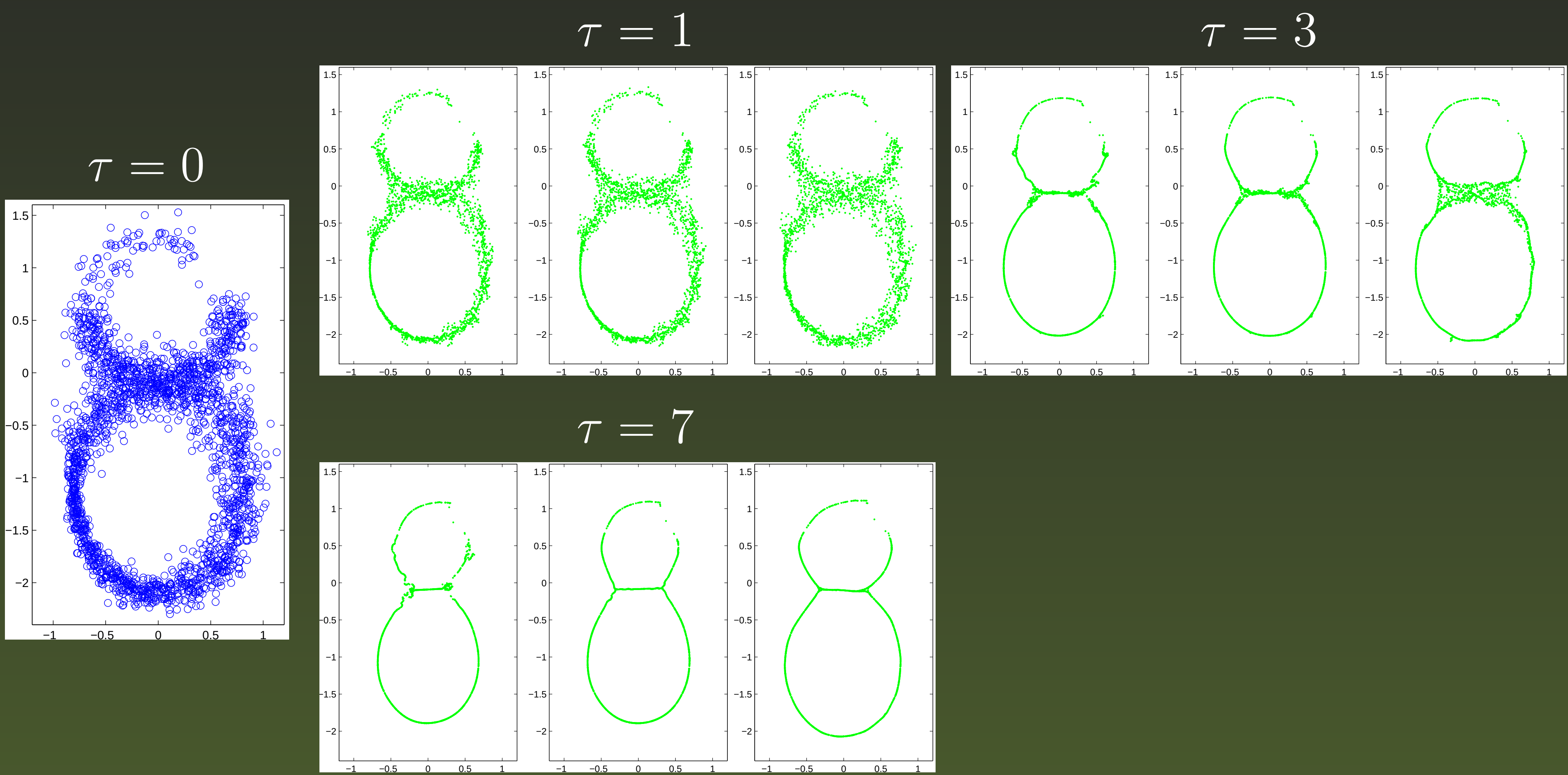
Conclusion

- We develop mean-shift algorithms to analyze dataset with low degrees of freedom.
- Future directions:
  - Theoretical analysis
  - Speedup training and testing
  - Incorporating more domain knowledge

Papers

- Miguel A. Carreira-Perpinan and Weiran Wang. *A simple assignment model with Laplacian smoothing*. Unpublished manuscript.
- Weiran Wang and Miguel A. Carreira-Perpinan. *The role of dimensionality reduction in classification*. Unpublished manuscript.
- Weiran Wang and Miguel A. Carreira-Perpinan. *The Laplacian K-modes algorithm for clustering*. Unpublished manuscript.
- Miguel A. Carreira-Perpinan and Weiran Wang. *The K-modes algorithm for clustering*. Unpublished manuscript, Apr. 23, 2013, arXiv:1304.6478 [cs.LG].
- Miguel A. Carreira-Perpinan and Weiran Wang. *Distributed optimization of deeply nested systems*. Unpublished manuscript, Dec. 24, 2012, arXiv:1212.5921 [cs.LG].
- Weiran Wang and Miguel A. Carreira-Perpinan. *Nonlinear low-dimensional regression using auxiliary coordinates*. AISTATS 2012.
- Weiran Wang, Miguel A. Carreira-Perpinan and Zhengdong Lu. *A denoising view of matrix completion*. NIPS 2011.
- Weiran Wang and Miguel A. Carreira-Perpinan. *Manifold blurring mean shift algorithms for manifold denoising*. CVPR 2010.

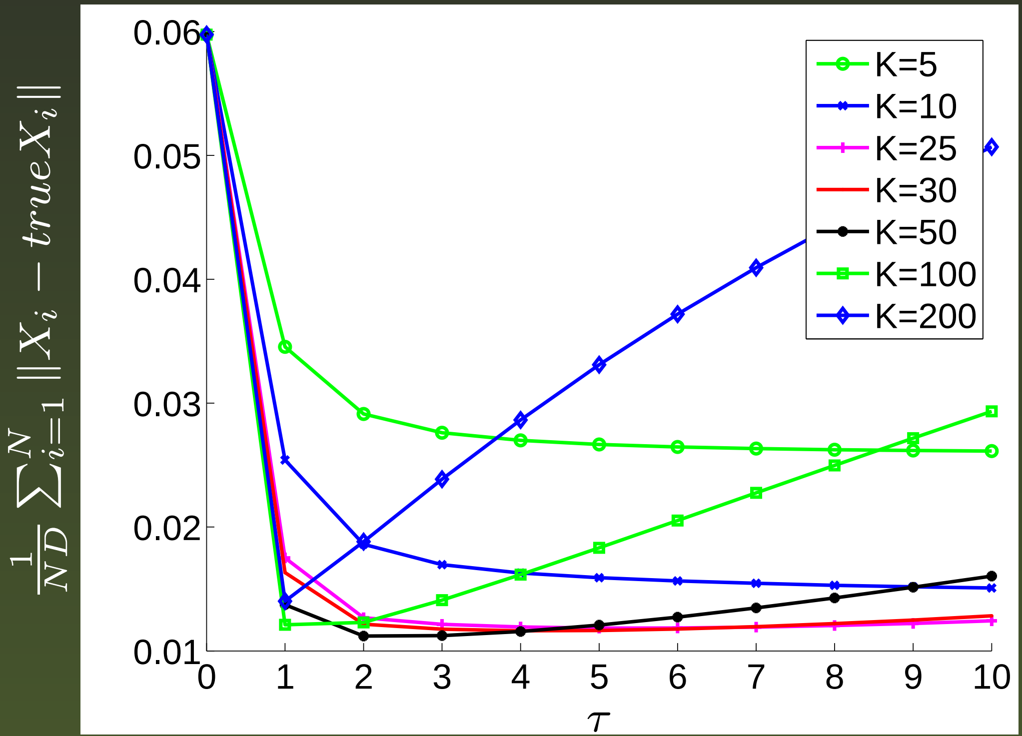
MBMS Experiment: complex shape

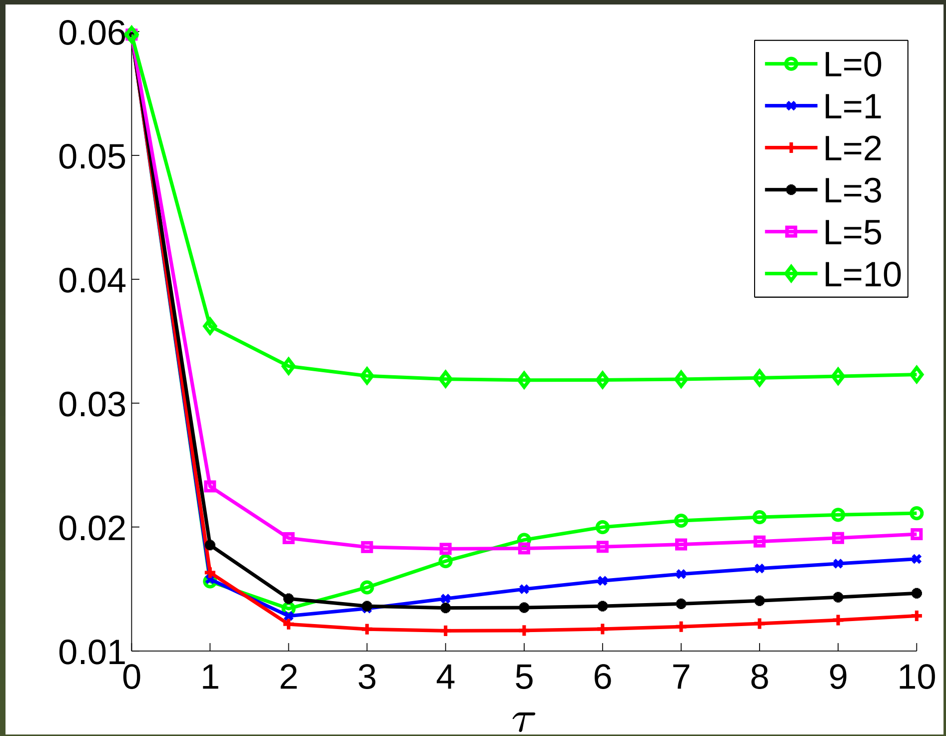


Denoising a complex shape with nonuniform density and noise with MBMSf using different affinity (left: normal, middle: diffusion maps, right: entropic affinity).

MBMS Experiment: Robustness to parameters choice

For swissroll dataset, there is a wide range for each parameter in which MBMS works well.

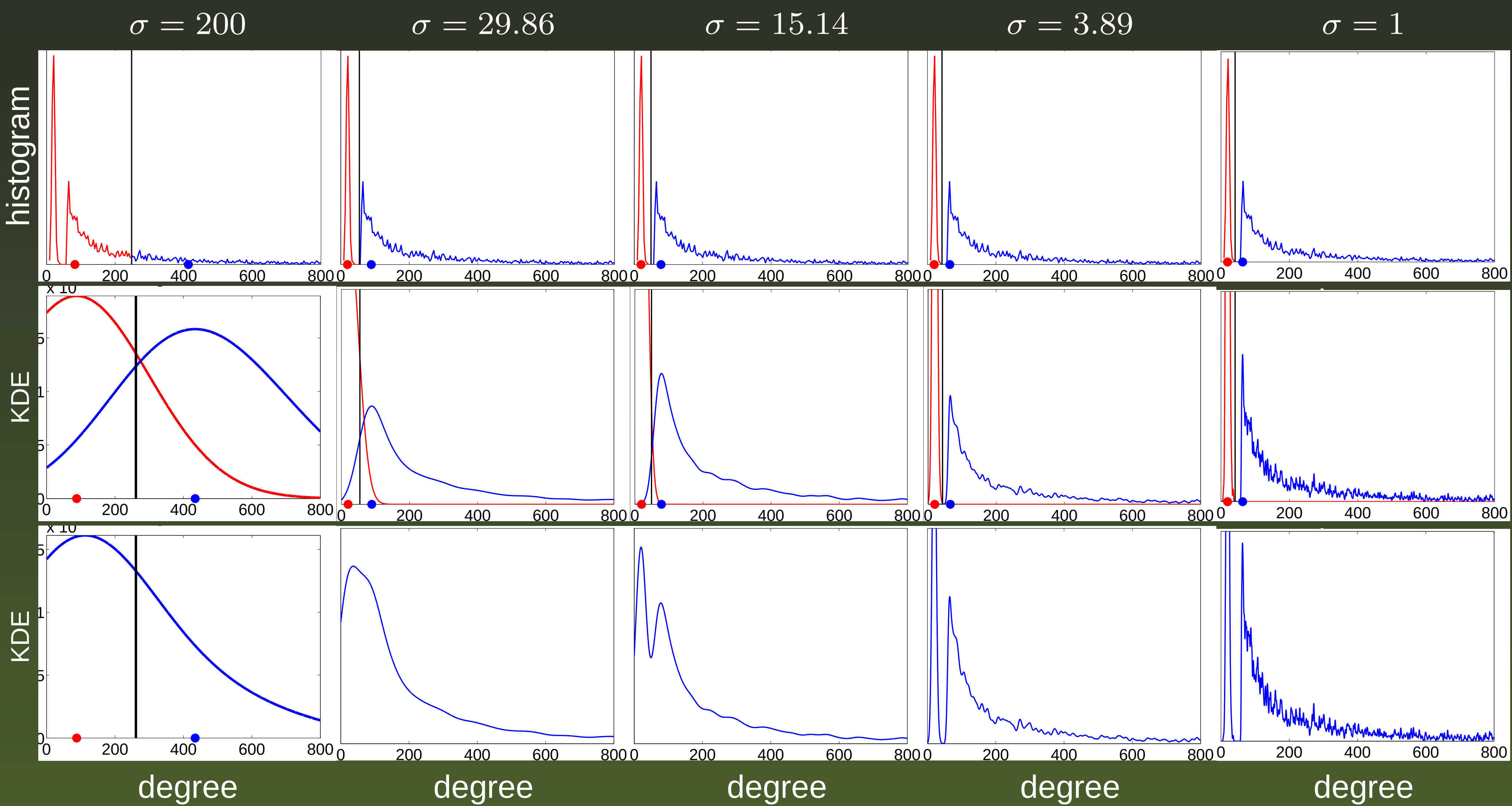




$\sigma = \infty, L = 2$ $\qquad$ $\sigma = \infty, K = 30$

Behavior of LTP for different parameters K and L. Error decreases for all parameter choices.

K-modes Experiment: heavy tailed distribution



☞ $K = 2$. Separating mixture of a Gaussian component and a power-law component.

Laplacian K-modes: alternating optimization

- **C**-step: decouples over different cluster. For cluster k, solve $\max_{\mathbf{c}_k} \sum_{\{n: z_{nk}>0\}} z_{nk} G\left(\left\|\frac{\mathbf{x}_n - \mathbf{c}_k}{\sigma}\right\|^2\right)$ with mean-shift updates.
- **Z**-step: nolonger decouples over different points.

$$\begin{aligned} \min_{\mathbf{Z}} \quad & \lambda \operatorname{tr}\left(\mathbf{Z}^\top \mathbf{L} \mathbf{Z}\right) - \operatorname{tr}\left(\mathbf{B}^\top \mathbf{Z}\right) \\ \text{s.t.} \quad & \mathbf{Z}\mathbf{1}_K = \mathbf{1}_N, \\ & \mathbf{Z} \geq 0, \end{aligned}$$

where $\mathrm{B}_{nk} = G\left(\left\|\frac{\mathbf{x}_n - \mathbf{c}_k}{\sigma}\right\|^2\right)$, $\mathbf{L}$ is the graph Laplacian.

- Quadratic program of NK variables.
- Interior point method is too slow for large dataset.
- We use first order method instead.

Laplacian K-modes: Z-step

- The ISTA/FISTA framework (gradient proximal method):
  - Solves $\min_{\mathbf{x}} f(\mathbf{x}) = g(\mathbf{x}) + h(\mathbf{x})$. g is convex and has Lipschitz continuous gradient (with constant L). h is convex and not necessarily differentiable.
  - $\mathbf{x}_{n+1} = \arg\min_{\mathbf{y}} \frac{L}{2} \left\| \mathbf{y} - (\mathbf{x}_n - \frac{1}{L}\nabla g(\mathbf{x}_n)) \right\|^2 + h(\mathbf{y})$.
  - Convergence: $f(\mathbf{x}_T) - f(\mathbf{x}^*) \approx \mathcal{O}(\frac{1}{T})$ for constant stepsize $\frac{1}{L}$.
  - Nesterov's acceleration scheme improves the rate to $\mathcal{O}(\frac{1}{T^2})$.
- Apply to our Z-step:
  - g is the quadratic objective function, with $L = 2\lambda\sigma_1(\mathbf{L})$.
  - h is the indicator function of probability simplex, therefore the proximal step is computing Euclidean projection.

Accelerated gradient projection for $\mathbf{Z}$-step

Input: Initial $\mathbf{Z}_0 \in \mathbf{R}^{N\times K}$, $s = \frac{1}{2\lambda\sigma_1(\mathbf{L})}$.

1: Set $\mathbf{Y}_1 = \mathbf{Z}_0,\ t_1 = 1,\ k = 1$.
2: **repeat**
3: Compute gradient at $\mathbf{Y}_k$: $\mathbf{G}_k = 2\lambda\mathbf{L}\mathbf{Y}_k - \mathbf{B}$,
4: $\mathbf{Z}_k =$ SimplexProj$(\mathbf{Y}_k - s\mathbf{G}_k)$ where SimplexProj$()$ projects each row of the argument onto the probability simplex,
5: $t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2}$,
6: $\mathbf{Y}_{k+1} = \mathbf{Z}_k + (\frac{t_k-1}{t_{k+1}})(\mathbf{Z}_k - \mathbf{Z}_{k-1})$,
7: $k = k+1$,
8: **until** convergence.

Output: $\mathbf{Z}_k$ is the solution of $\mathbf{Z}$ given $\mathbf{C}$.

Projection onto the probability simplex

Input: A vector $\mathbf{v} \in \mathbf{R}^K$

1: Sort $\mathbf{v}$ into $\mathbf{u}$: $u_1 \geq u_2 \geq \cdots \geq u_K$

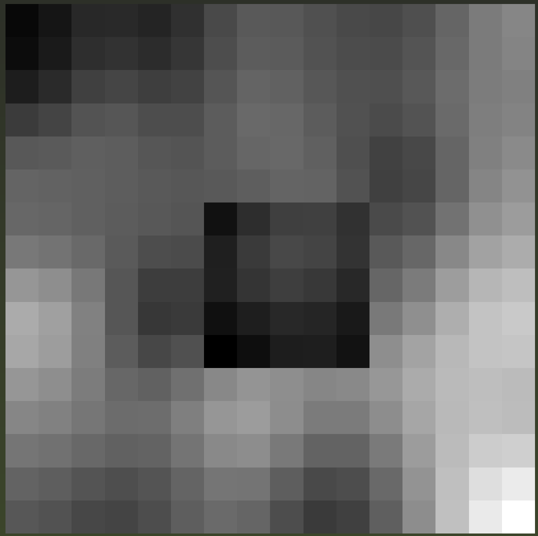
2: Find $\rho = \max\{1 \leq j \leq K : u_j - \frac{1}{j}(\sum_{r=1}^{j} u_r - 1) > 0\}$

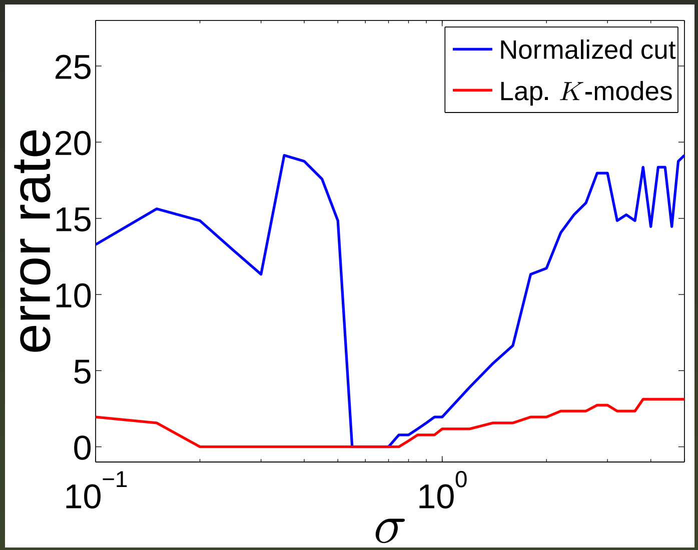
3: Define $\theta = \frac{1}{\rho}(\sum_{r=1}^{\rho} u_r - 1)$

Output: $\mathbf{w}$ s.t. $w_i = \max\{v_i - \theta, 0\}$

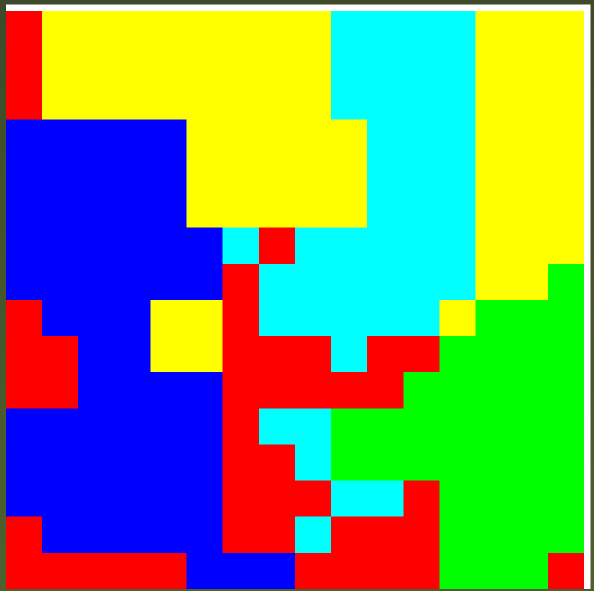
Computational complexity: $\mathcal{O}(K \log K)$.

Laplacian K-modes: occluder segmentation

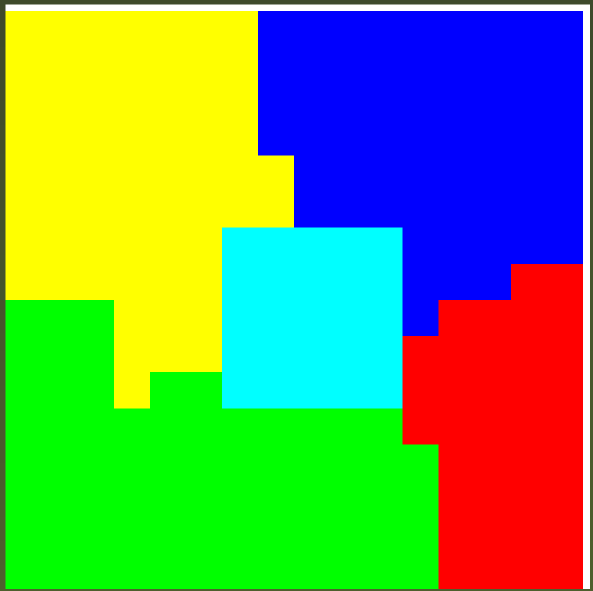




Normalized cut ($\sigma = 0.2$)



Laplacian K-modes ($\sigma = 0.2$)



☞ $K = 5$. Each pixel is connected with nearby eight pixels with edge weighted using heat kernel.