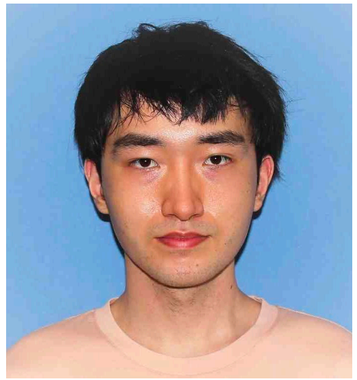
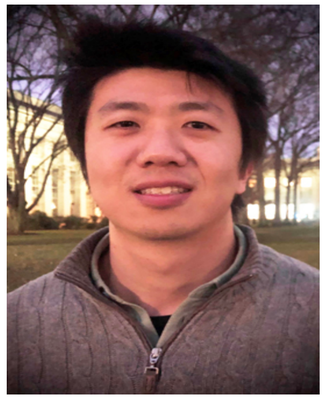
Statistical optimality and computational guarantee for K-means clustering

Xiaohui Chen
Department of Mathematics
University of Southern California
Email: xiaohuic@usc.edu
Website: https://the-xiaohuichen.github.io/

Yun Yang
(UIUC)



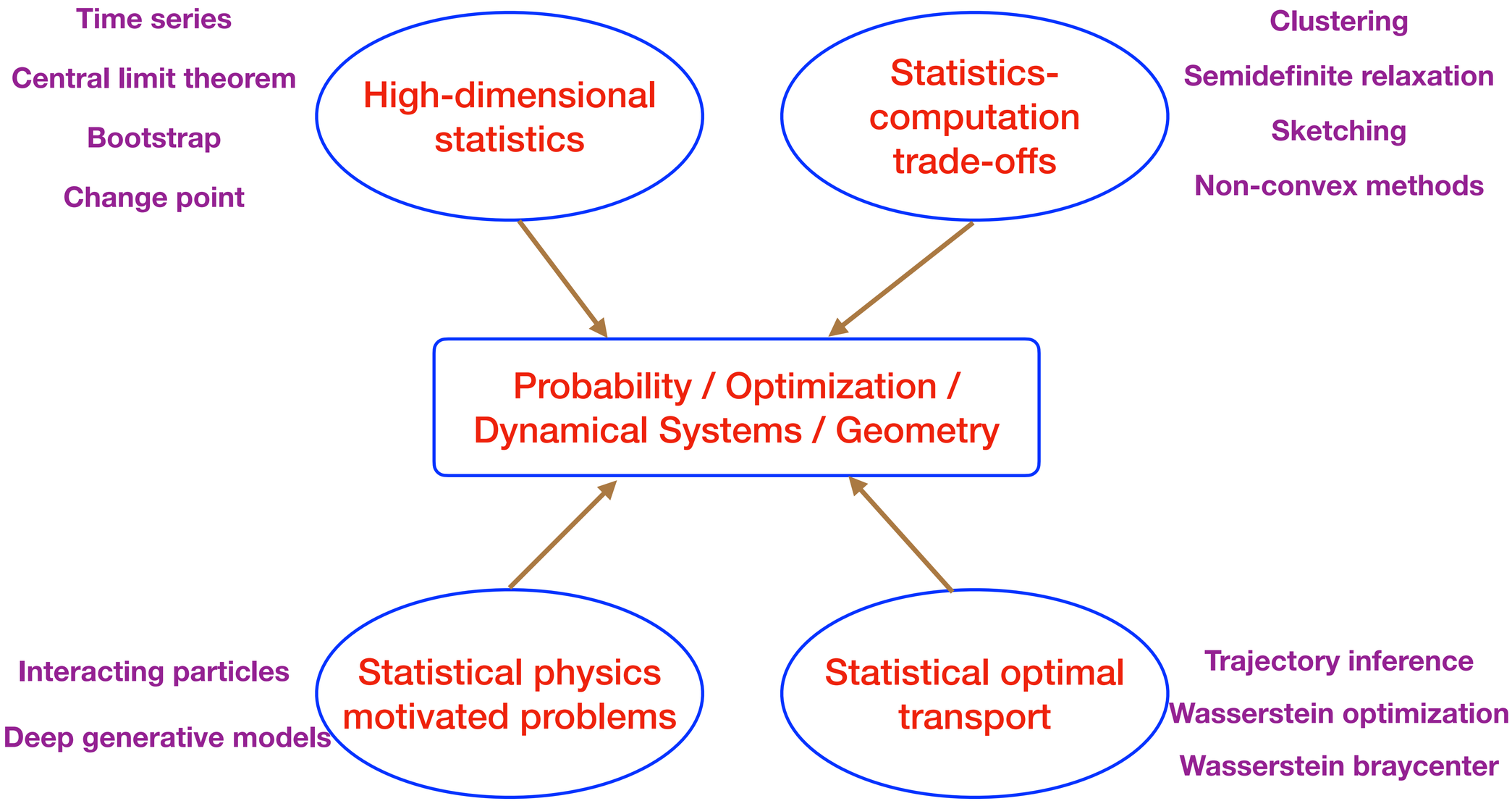
Yubo Zhuang
(UIUC)



Richard Y. Zhang
(UIUC)

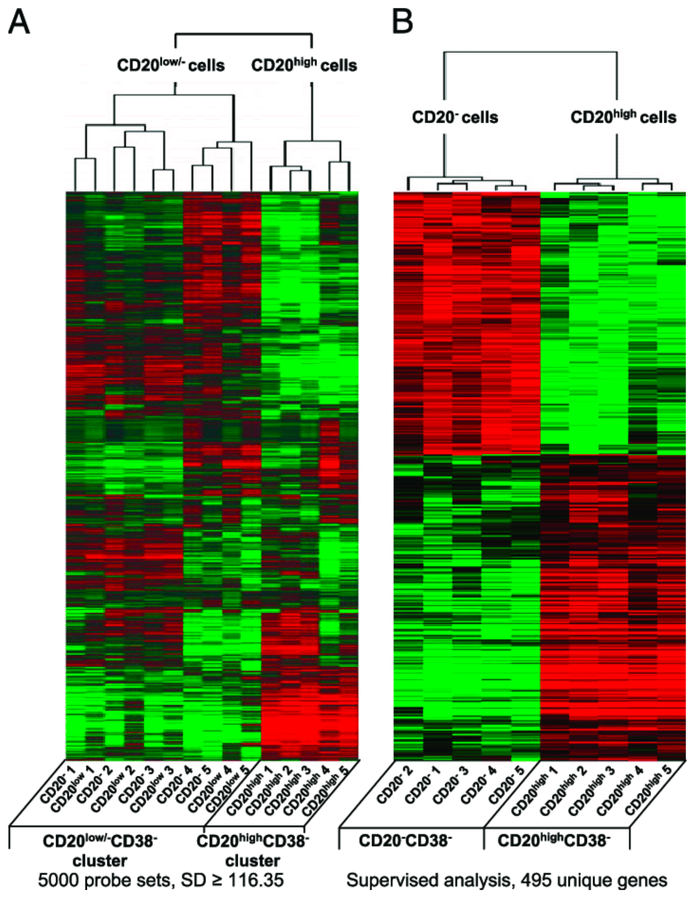
Reading group
10/5/2023

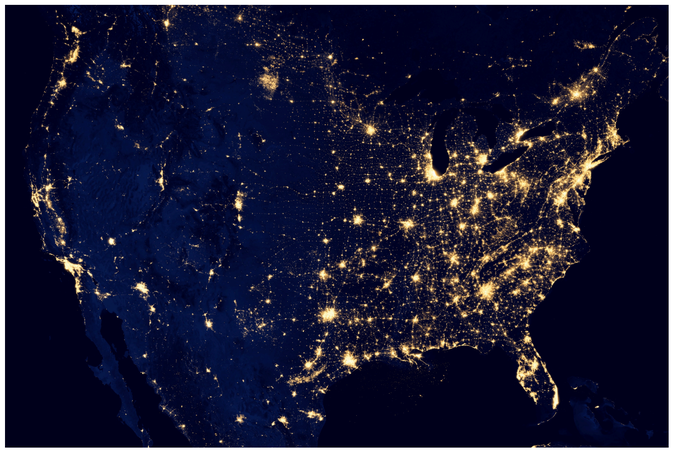
Overview of my research: statistics meets computation

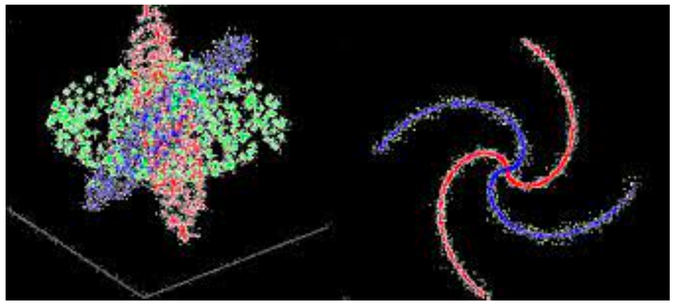


Clustering analysis

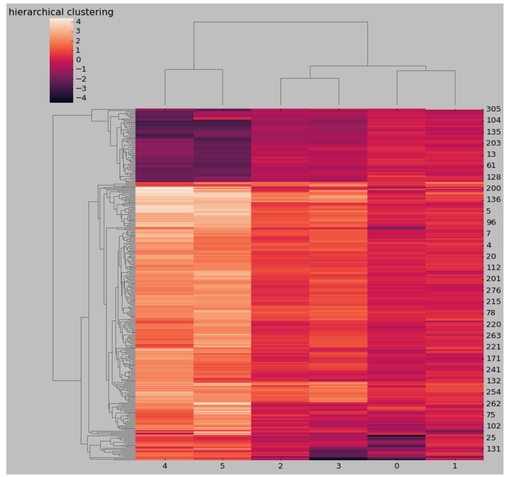
- Divide data points $\{x_1, \ldots, x_n\}$ into K disjoint groups based on data similarity.



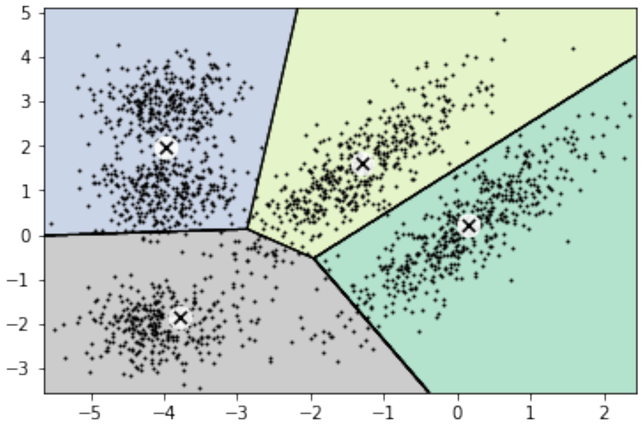




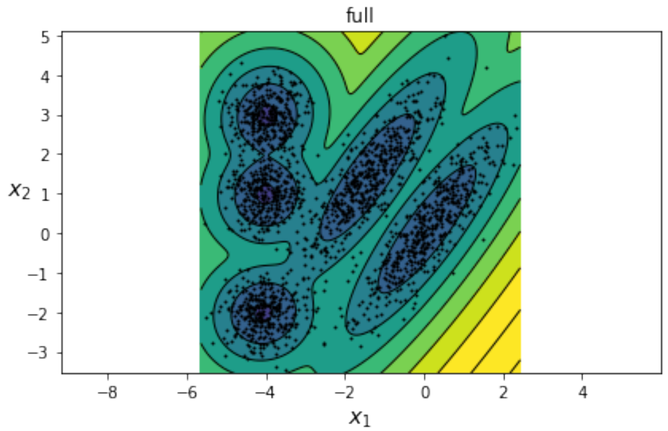
Clustering methods



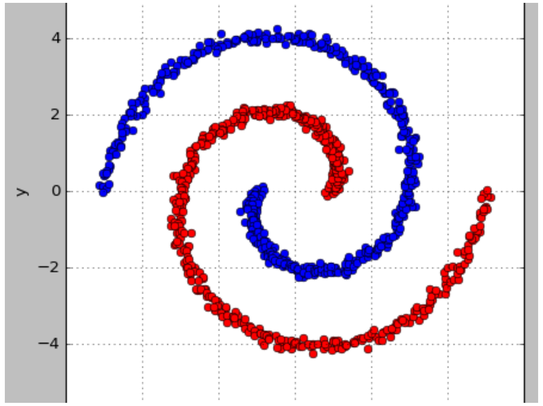
Linkage-based clustering
(Hierarchical clustering)



Centroid-based clustering
(K-means)



Model-based clustering
(Gaussian mixture model)



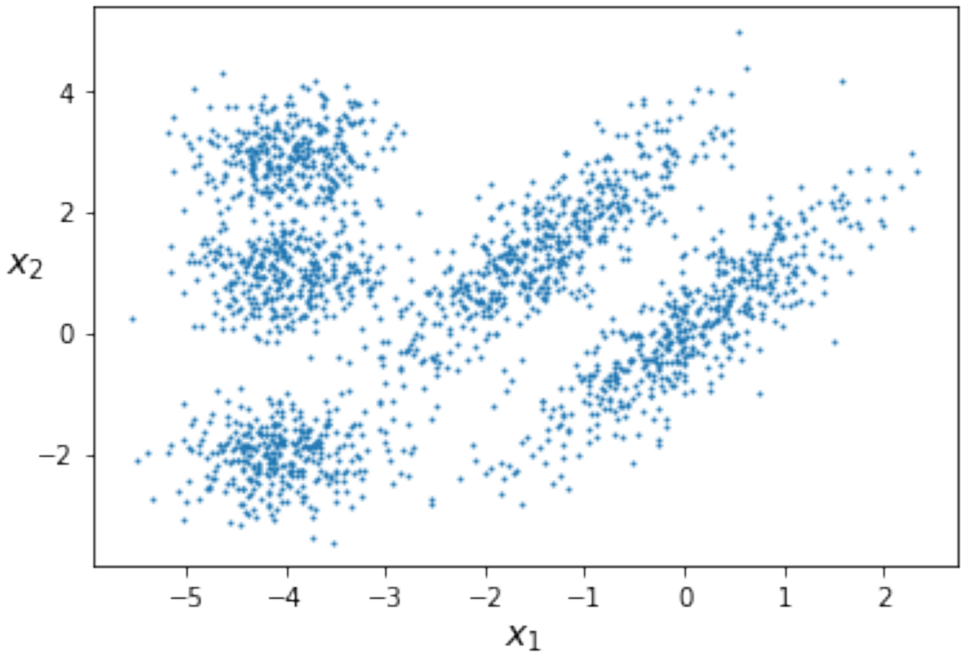
Graph-based clustering
(Spectral clustering)

- **Challenges**: no labels for $\{x_1, \ldots, x_n\}$.
- Clustering is **hard**: computationally and statistically.
  - Exact solution involves **combinatorial optimization**.
  - Practical algorithms often are **heuristic approximations**.

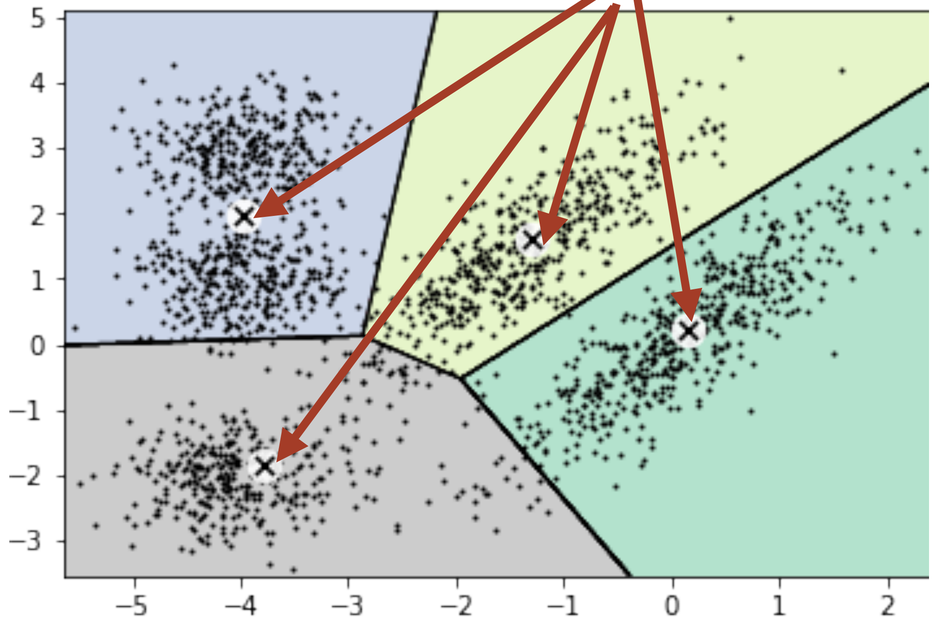
- **Data types**:
  - Heterogeneous features
  - Geometric and non-Euclidean
  - Variable dimensions: finite and infinite

This talk: develop **computationally tractable (or even scalable)** algorithms with **strong theoretical guarantees** for recovering the true clustering structure.

K-means clustering



Centroids (arithmetic means)



Voronoi tessellation

Centroid-based formulation

$$(\beta_1^*, \ldots, \beta_K^*) = \arg \min_{\beta_1, \ldots, \beta_K \in \mathbb{R}^p} \sum_{i=1}^{n} \min_{k \in [K]} \|x_i - \beta_k\|_2^2$$

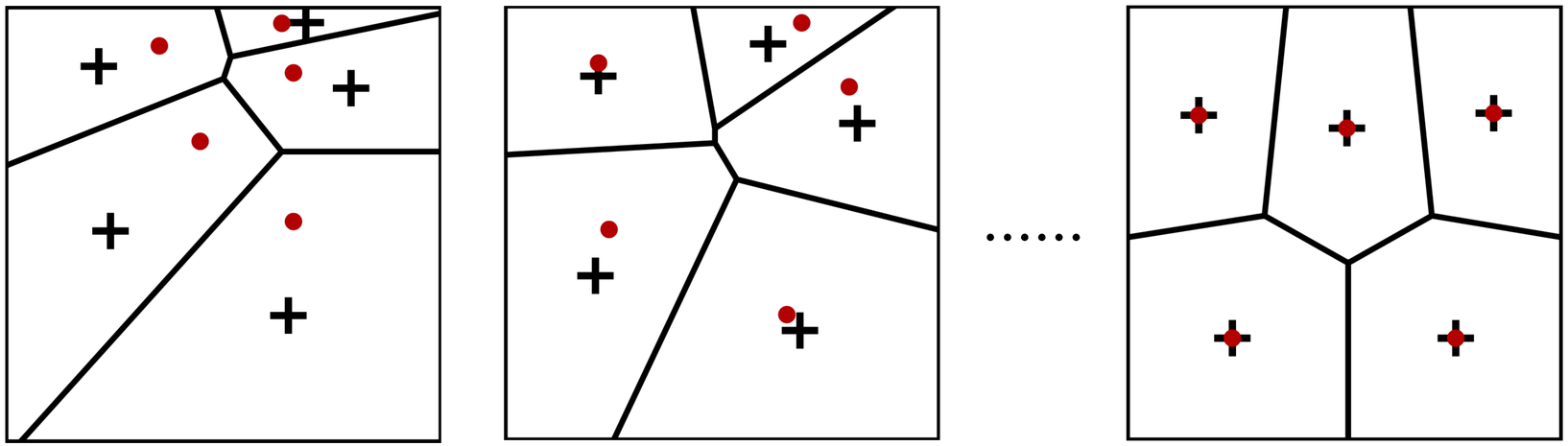
Data label

$$z_i^* = \arg \min_{k \in [K]} \|x_i - \beta_k^*\|_2^2$$

Lloyd's algorithm

- Solving the exact K-means clustering is worst-case **NP-hard**: combinatorial learning.
- Heuristic **approximation** algorithms to obtain local optimum: sensitive to initialization.
- **Lloyd algorithm**: iteratively perform the following steps until convergence:

1. Compute the Voronoi diagram of the K sites: $G_k^{(t)} = \left\{ i \in [n] : \|x_i - \beta_k^{(t)}\|_2 \leq \|x_i - \beta_j^{(t)}\|_2, \ \forall j \in [K] \right\}$.
2. Compute the centroid of each Voronoi cell: $\beta_k^{(t+1)} = |G_k^{(t)}|^{-1} \sum_{i \in G_k^{(t)}} x_i$ and then move each K site to the corresponding Voronoi cell centroid.



Iteration 1 Iteration 2 Iteration 15

K-means as combinatorial optimization

- **Partition-based formulation** of K-means clustering: minimize the **total intra-cluster squared Euclidean distances** in $\mathbb{R}^p$:

$$\min_{G_1,\ldots,G_K} \sum_{k=1}^{K} \frac{1}{|G_k|} \sum_{i \in G_k} \sum_{j \in G_k} \|x_i - x_j\|_2^2 \quad \text{subject to} \quad \bigsqcup_{k=1}^{K} G_k = [n]$$

over all possible partitions of $[n]$, where $|G_k|$ is the cardinality of G_k.

Generalized parallelogram law

$$\sum_{i=1}^{n} \sum_{j=1}^{n} \|x_i - x_j\|_2^2 = 2n \sum_{i=1}^{n} \|x_i - \bar{x}_n\|_2^2$$

- Dropping the sum of squared norms $\sum_{i=1}^{n} \|x_i\|_2^2$, K-means is the same as maximizing the **total within-cluster covariances**:

$$\max_{G_1,\ldots,G_K} \sum_{k=1}^{K} \frac{1}{|G_k|} \sum_{i \in G_k} \sum_{j \in G_k} x_i^T x_j \quad \text{subject to} \quad \bigsqcup_{k=1}^{K} G_k = [n]\,.$$

- Here, $a_{ij} = x_i^T x_j$ can be viewed as a **similarity measure** (i.e., **affinity**) of two data points x_i and x_j, specified by the Euclidean space inner product $\langle x_i, x_j \rangle_{\mathbb{R}^p} = x_i^T x_j$.

Convex relaxation

- **Convexification**: approximation the K-means by convex relaxation.
- Recall: K-means objective

$$\max_{G_1,\ldots,G_K} \sum_{k=1}^{K} \frac{1}{|G_k|} \sum_{i \in G_k} \sum_{j \in G_k} x_i^T x_j$$

$$\text{subject to } \bigsqcup_{k=1}^{K} G_k = [n] .$$

- Partition $G_1, \ldots, G_K \iff$ assignment matrix $H_{n \times K}$

	1	2	3	G_k
x_1	1	0	0	1
x_2	1	0	0	1
x_3	0	1	0	2
x_4	1	0	0	1
x_5	0	1	0	2
x_6	0	0	1	3
x_7	0	0	1	3
x_8	0	0	1	3
x_9	0	1	0	2
x_{10}	0	1	0	2

Semi-definite programming (SDP) relaxation

- **Reparametrization** as a **0-1 integer program** (NP-hard): $A = XX^T$ (similarity / affinity / Gram matrix) and $B = \mathrm{diag}(|G_1|^{-1}, \ldots, |G_K|^{-1})$:

$$\max\{\langle A, HBH^T\rangle : H \in \{0,1\}^{n\times K}, H1_K = 1_n\}.$$

- **Change of variable**: $Z = HBH^T$. What properties should Z be preserved?

$$Z^T = Z,\ Z \succeq 0,\ Z \geq 0,\ \mathrm{tr}(Z) = K,\ Z1_n = 1_n.$$

- **SDP relaxed K-means** [Peng, Wei (2007)]: convex optimization

$$\max_{Z\in\mathbb{R}^{n\times n}} \langle A, Z\rangle$$

subject to $Z^T = Z,\ Z \succeq 0,\ Z \geq 0,\ \mathrm{tr}(Z) = K,\ Z1_n = 1_n$

Statistical optimality

- Balanced Gaussian mixture model (GMM) with K components of equal size $|G_1^*| = \cdots = |G_K^*|$:

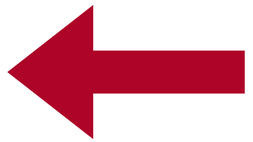
$$x_i = \mu_k + \varepsilon_i, \quad \varepsilon_1, \ldots, \varepsilon_n \sim N(0,\, \sigma^2 I_p) \text{ are i.i.d.}$$

- **Minimal separation**: $\Theta_{\min}^2 = \min_{1 \le k \ne l \le K} \|\mu_k - \mu_l\|_2^2$.

- **Theorem** [C, Yang (2021) *IEEE Trans. Inf. Theory*] Suppose $K \le \dfrac{\log n}{\log\log n}$. Exact recovery information-theoretic limit:

  1. If $\Theta_{\min}^2 \ge (1+\alpha)\overline{\Theta}^2$, where

$$\overline{\Theta}^2 = 4\sigma^2 \left(1 + \sqrt{1 + \frac{Kp}{n \log n}}\right) \log n,$$

  Sharp threshold

  then the SDP solution $\hat{Z} = Z^*$ (thus $\hat{G}_1 = G_1^*, \ldots, \hat{G}_K = G_K^*$) with probability tending to one.

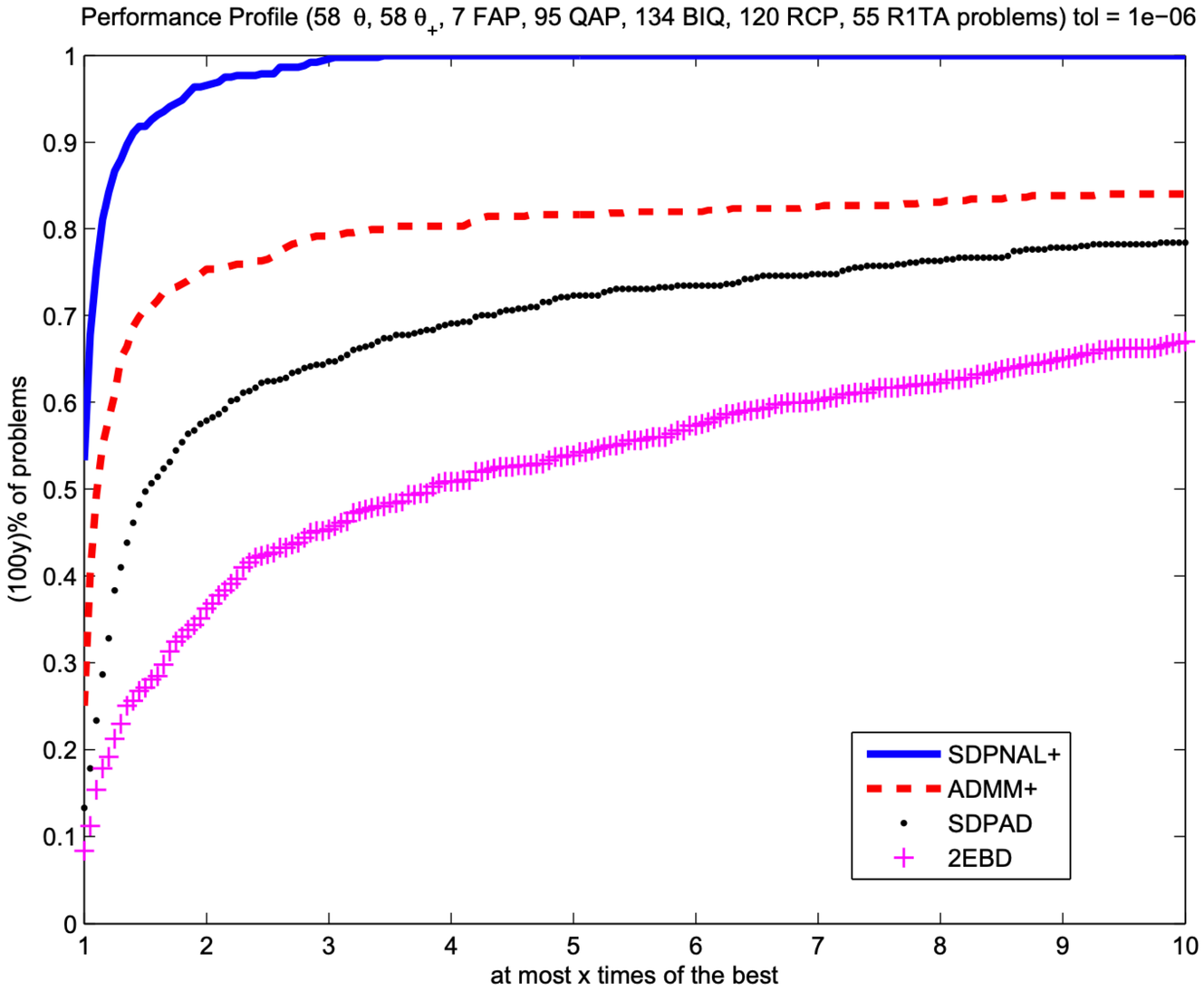
  2. If $\Theta_{\min}^2 \le (1-\alpha)\overline{\Theta}^2$, then probability of exact recovery of any estimator vanishes to zero.

Computational complexity

- Interior point algorithm: $O(n^{3.5})$ runtime complexity.
- **SDPNAL+** [Su, Toh, Yuan, Zhao (2020)]: a large-scale SDP solver based on augmented Lagrangian method implemented in MATLAB.

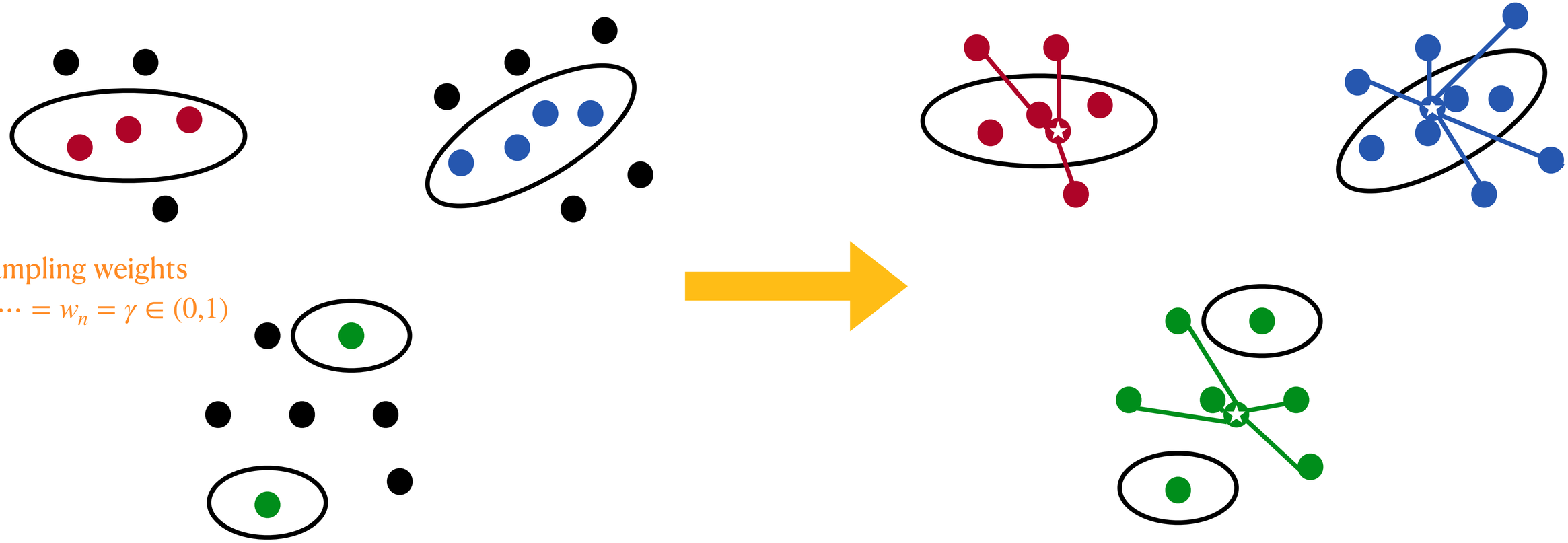
Accuracy (effective zero) $\varepsilon = 10^{-6}$

Yang, Sun, Toh (2015). SDPNAL+: A Majorized Semismooth Newton-CG Augmented Lagrangian Method for Semidefinite Programming with Nonnegative Constraints. *Mathematical Programming Computation*. 7, 331-366.



Sketch-and-Lift

- **Idea**: sub-sampling m data points for SDP with $m \ll n$.
- **Sketch-and-Lift** [Zhuang, C, Yang (2022) *AISTATS*]: $O(n + m^{3.5})$.



Statistical guarantee

- **Theorem** (**Exact recovery**). Suppose $K \leq \dfrac{\log(\gamma n)}{\log\log(\gamma n)}$ and $p \leq (\gamma n/K)^2$. If $\Theta^2_{\min} \geq (1+\alpha)\overline{\Theta}^2_\gamma$, where

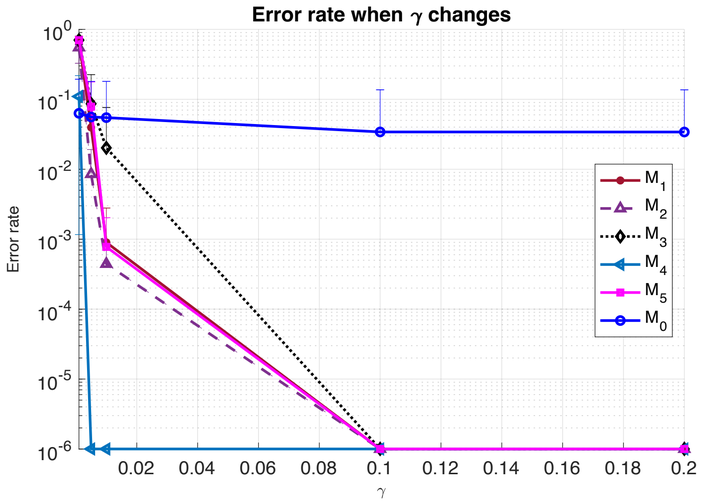
$$\overline{\Theta}^2_\gamma = 4\sigma^2 \left(1 + \sqrt{1 + \frac{Kp}{\gamma n \log n}}\right) \log n,$$

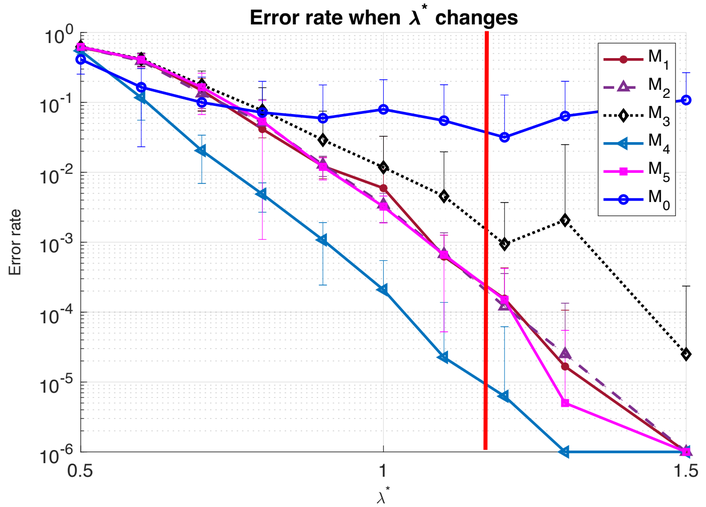
then the sketch-and-lift output $(\hat{G}_1, \ldots, \hat{G}_K)$ equals to $(G_1^*, \ldots, G_K^*)$ with probability at least $1 - (\log(\gamma n))^{-C}$.

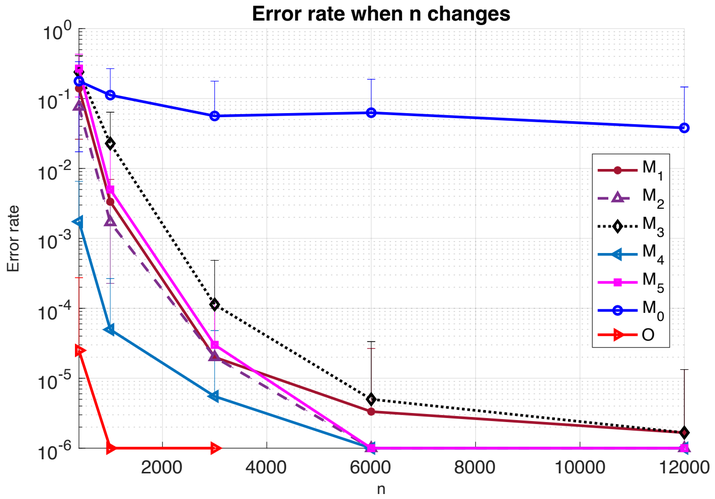
- **Remarks**

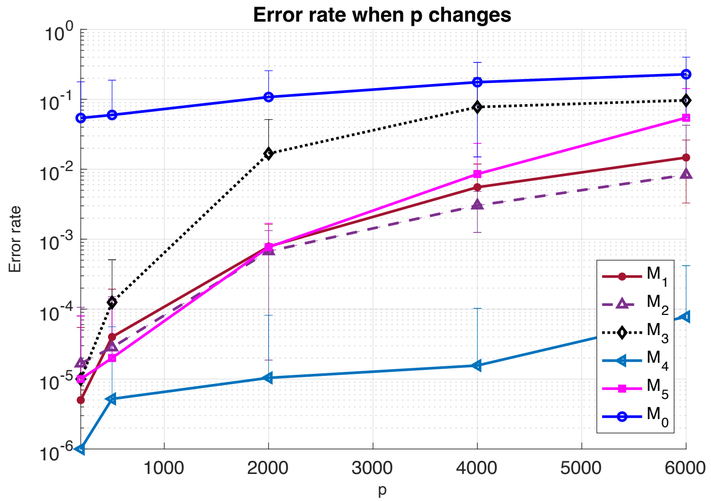
1. Full sampling case $\gamma = 1$ recovers the sharp threshold of K-means SDP.
2. Sub-sampling ratio $\gamma \gg K/n$ (i.e., $m \gg K$ solve a sub-sampled SDP on a subset with at least a constant number of data points) ensures exact recovery.

Numeric examples







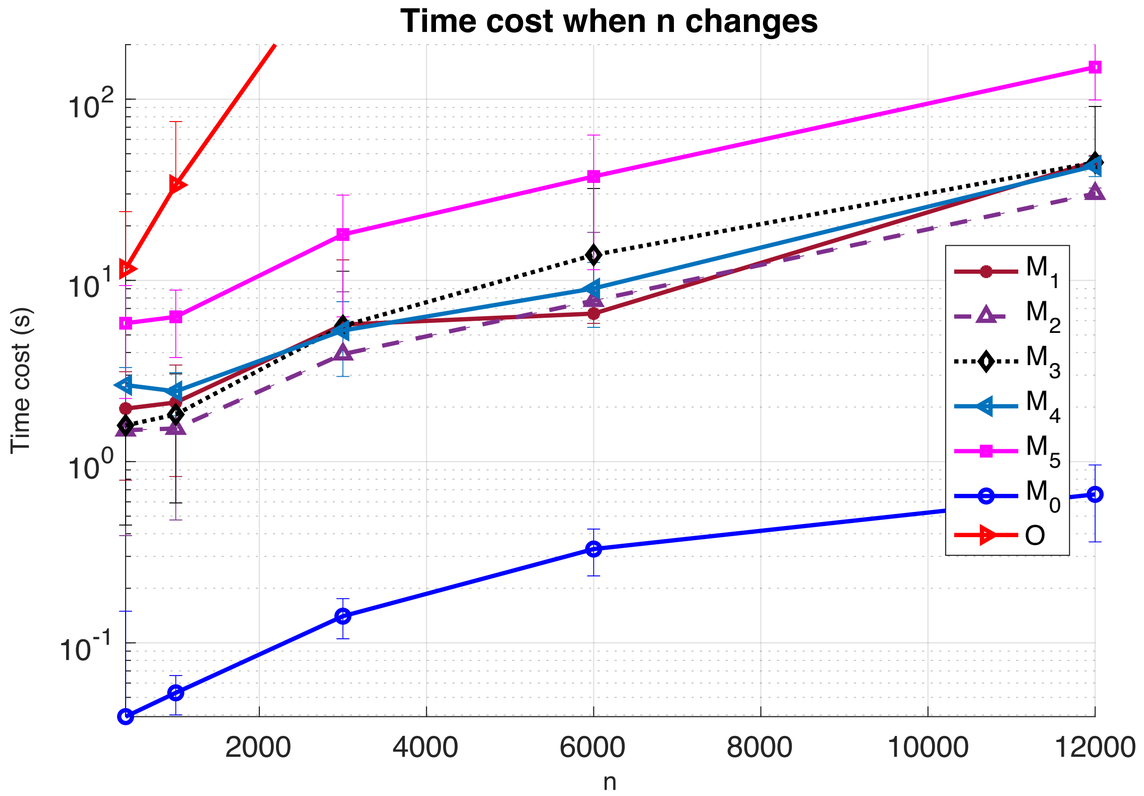


- M_0: MATLAB build-in K-means
- M_1: sketch-and-lift (SL)
- M_2: bias-corrected SL
- M_3: weighted SL
- M_4: multi-epoch SL
- M_5: multi-round SL
- O: full sample SDP

Separation signal

$$\Theta^2_{\min} = (\lambda^* \overline{\Theta})^2$$

Runtime comparison



Log-scale runtime with error bars (100 repetitions) v.s. sample size n.

- Setup:
  $p = 1000$, $\gamma = 0.1$, $\lambda^* = 1.2$.
- SL methods have the same **linear $O(n)$ runtime complexity** as the fast K-means++ (with difference only in the leading constants.)
- Full sample K-means SDP has super-linear complexity.

Low-rank SDP

- **Idea**: true membership matrix $Z^*_{n\times n}$ is a block diagonal matrix with rank K.
- **Restriction of low-rank SDP**: for $r \geq K$,

$$\min_{U\in\mathbb{R}^{n\times r}} \left\{ \langle A, UU^T\rangle : \|U\|_F^2 = K, UU^T 1_n = 1_n, U \geq 0 \right\}.$$

- **Burer–Monteiro** (BM) is a non-convex approach for solving an SDP problem when its solution is expected to be low rank.
- Manifold-like set: $\Omega := \{U \in \mathbb{R}^{n\times r} : \|U\|_F^2 = K,\ U \geq 0\}$ and its projection operator

$$\Pi_\Omega(V) := \arg\min_{U\in\Omega} \|U - V\|_F = \frac{\sqrt{K}\cdot (V)_+}{\|(V)_+\|_F}.$$

- Standard equality-constrained BM on manifold

$$\min_{U\in\Omega} \left\{ \langle A, UU^T\rangle : UU^T 1_n = 1_n \right\}.$$

Primal-dual gradient descent-ascent algorithm

- Augmented Lagrangian method:

$$\min_{U\in\Omega}\left\{\langle L\cdot \mathrm{Id}_n + A, UU^T\rangle + \frac{\beta}{2}\|UU^T 1_n - 1_n\|_2^2 : UU^T 1_n = 1_n\right\}.$$

- Augmented Lagrangian function:

$$\min_{U\in\Omega}\mathscr{L}_\beta(U,y) := \langle L\cdot \mathrm{Id}_n + A, UU^T\rangle + \langle y, UU^T 1_n - 1_n\rangle + \frac{\beta}{2}\|UU^T 1_n - 1_n\|_2^2.$$

- Iterating between primal and dual updates:

$$U_{\text{new}} = \arg\min_{U\in\Omega}\mathscr{L}_\beta(U,y),$$

$$y_{\text{new}} = y + \beta(U_{\text{new}}U_{\text{new}}^T 1_n - 1_n).$$

- Primal descent via projected GD steps: $U^{t+1} = \Pi_\Omega(U^t - \alpha\nabla_U\mathscr{L}_\beta(U^t,y))$.

Linear convergence rate

- **Theorem** [Zhuang, C, Yang, Zhang (2023)] (case $r = K$) Under the GMM with equal cluster size, $p = O(n \log n)$ and $\Theta^2_{\min} = \Omega(\log n)$, for appropriate choice of L, β, α, and initialization U^0 such that the initialization error $\Delta^0 = U^0 - U^*$ satisfies

$$\|\Delta^0_{S^c}\|_\infty = O\left(\sqrt{\frac{K}{n}}\right) \quad \text{and} \quad \|\Delta^0\|_F = O\left(\frac{\min\{1, K^{-2.5}\Theta^2_{\min}/\log n\}}{K^6}\right),$$

where S is the support of the block diagonal matrix class $\mathscr{F}$, then we have for any $t \geq O(K^3)$,

$$U^t_{n\times K} \in \mathscr{F} \quad \text{and} \quad \|U^{t+1} - U^*\|_F \leq \gamma \|U^t - U^*\|_F,$$

where $U^{t+1} = \Pi_\Omega(U^t - \alpha \nabla_U \mathscr{L}_\beta(U^t, y^*))$ and $\gamma = 1 - O(K^{-6})$.

- **Time complexity**: $O(nrK^6)$.
- **Optimal solution** $U^* = \text{blkdiag}(n_1^{-1/2}1_{n_1}, \ldots, n_K^{-1/2}1_{n_K})$ satisfies $\|U^*\|_\infty = \sqrt{K/n}$ and $\|U^*\|_F = \sqrt{K}$. This implies that the initialization requires a ***constant*** (in n) relative error!

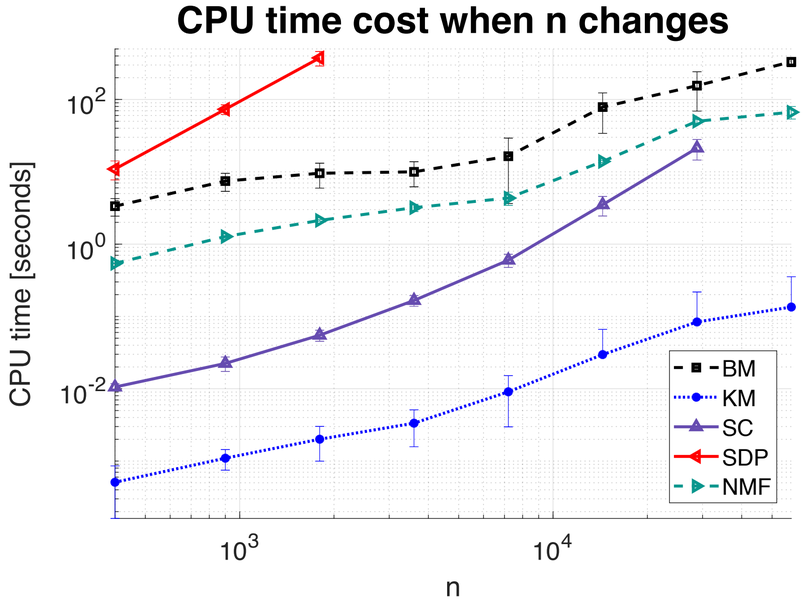
Numeric experiments

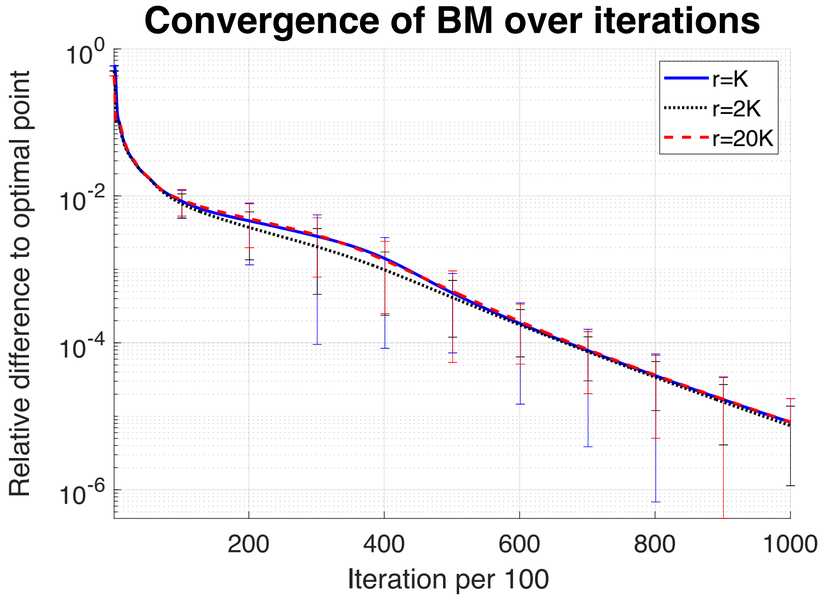
- Clustering via non-negative matrix factorization (NMF)

$$\min_{U \in \mathbb{R}^{n \times r}} \left\{ \|A + UU^T\|_F^2 : U \geq 0 \right\}.$$

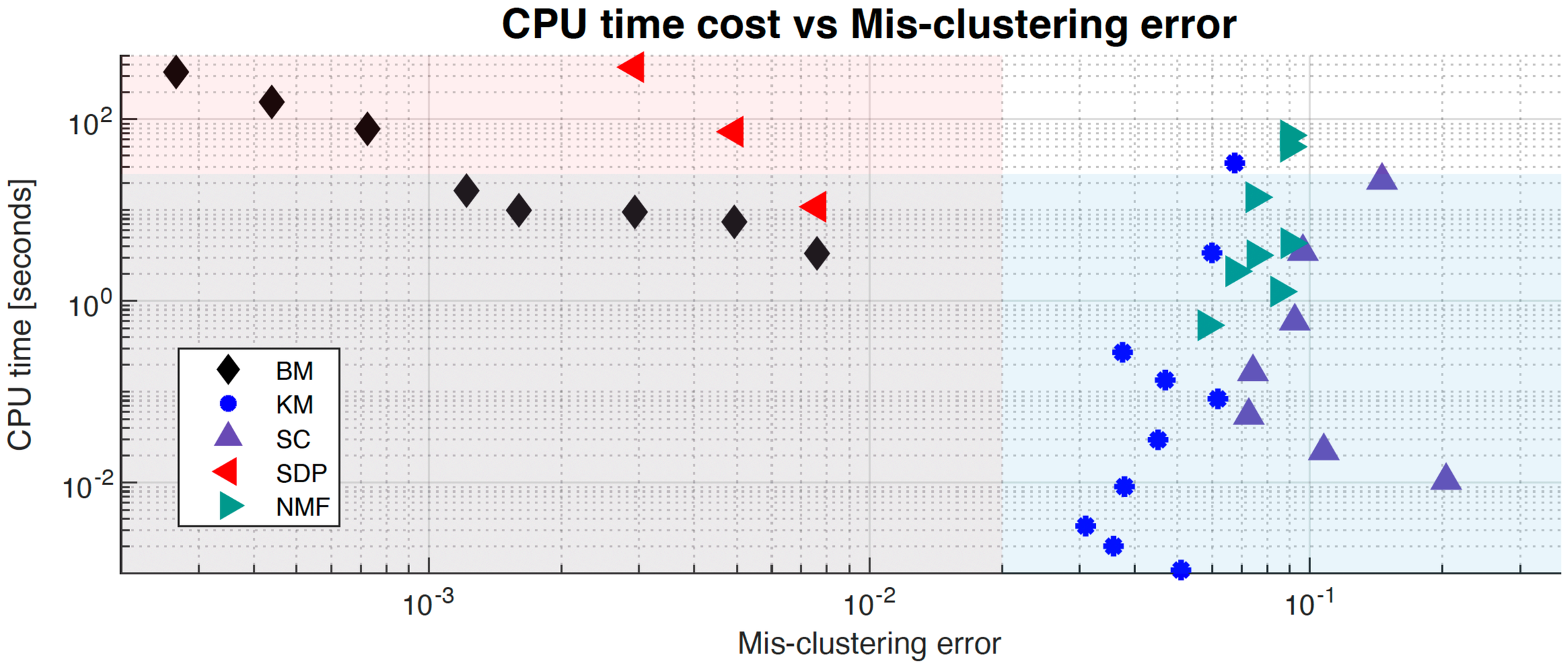
- Simulation setup: $K = 4$, GMM with $\Theta_{\min} = 0.8 \times \overline{\Theta}$, $p = 20$, and $n \in [400, 57{,}600]$.
- BM rank parameter: $r = 2K$.

$$\Theta_{\min} = 1.2 \times \overline{\Theta}$$





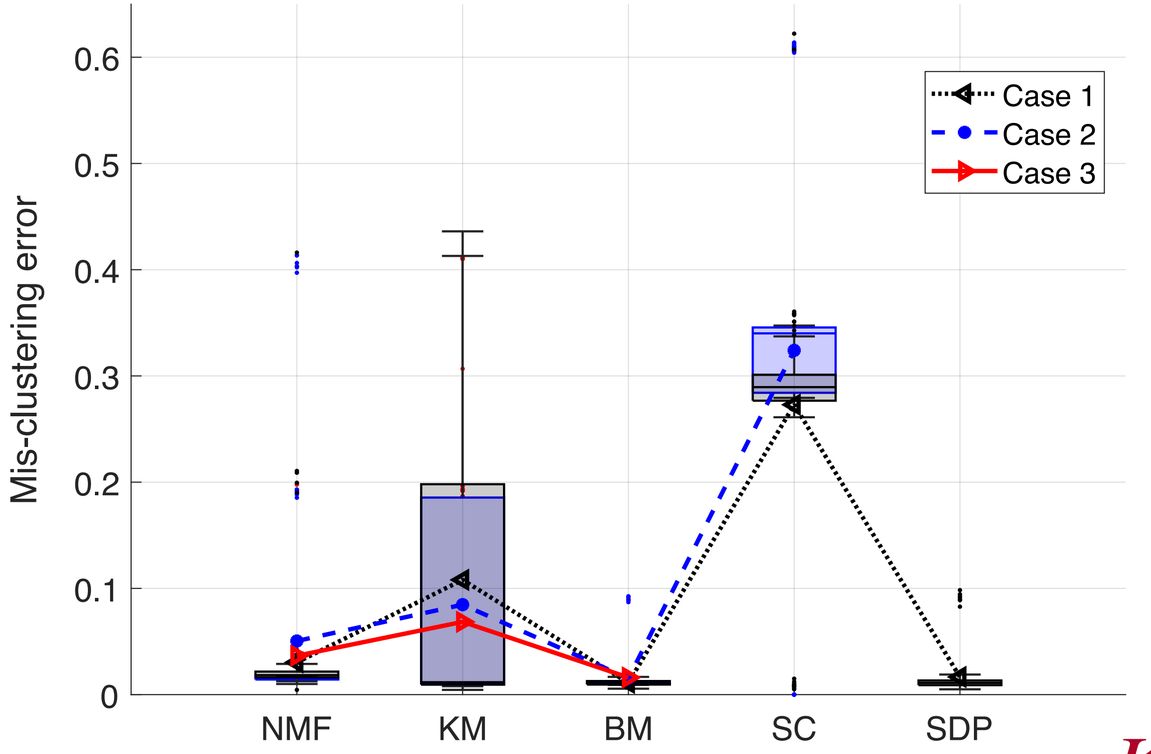
Numeric experiments



Real data examples

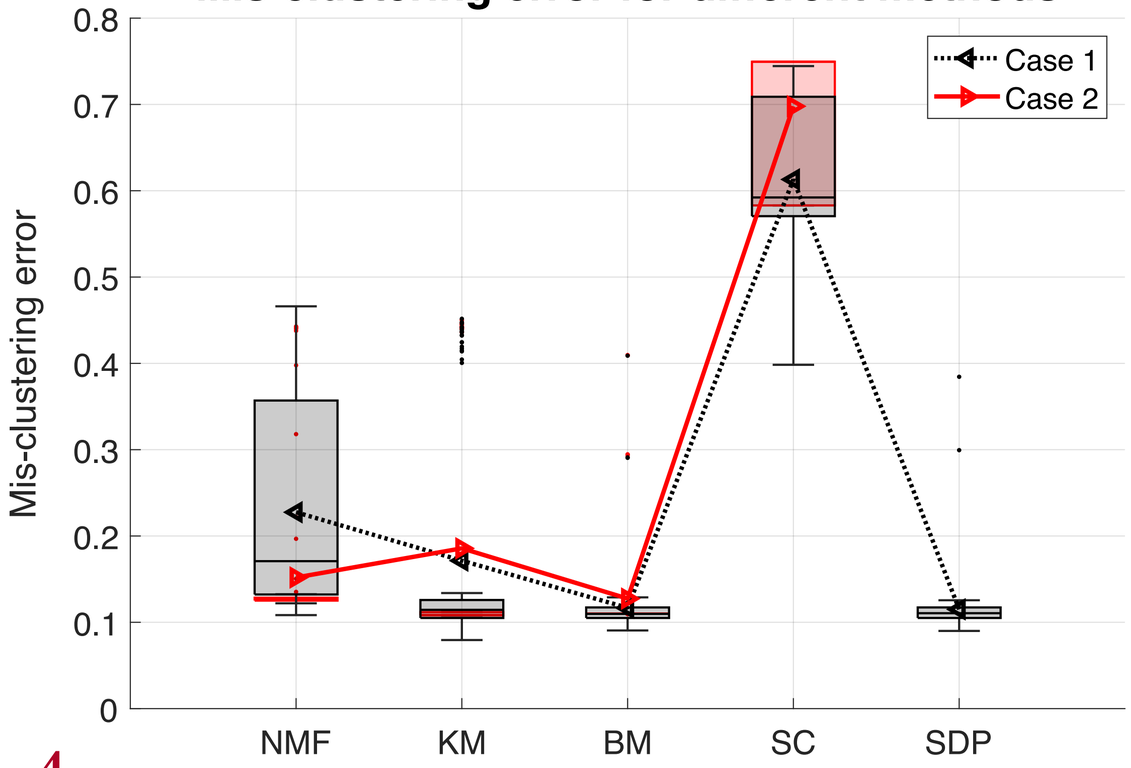
- Mass cytometry (CyTOF) dataset
- 32 protein markers
- Sample size $n = 1{,}800,\ 20{,}000,\ 46{,}258$.

Mis-clustering error for different methods



- CIFAR-10 dataset (colored images of size $32 \times 32 \times 3$)
- Inception v3 model and PCA to $p = 50$
- Sample size $n = 1{,}800,\ 4{,}000$.

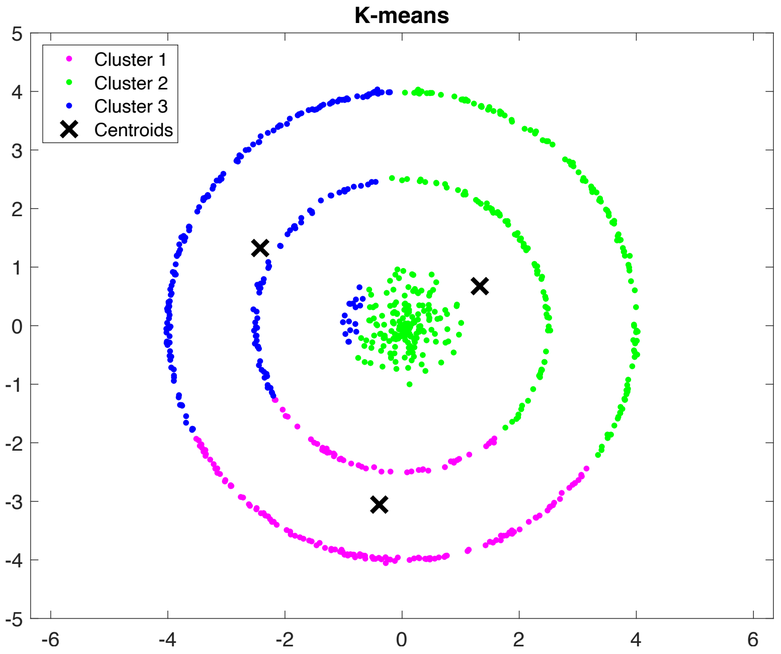
Mis-clustering error for different methods

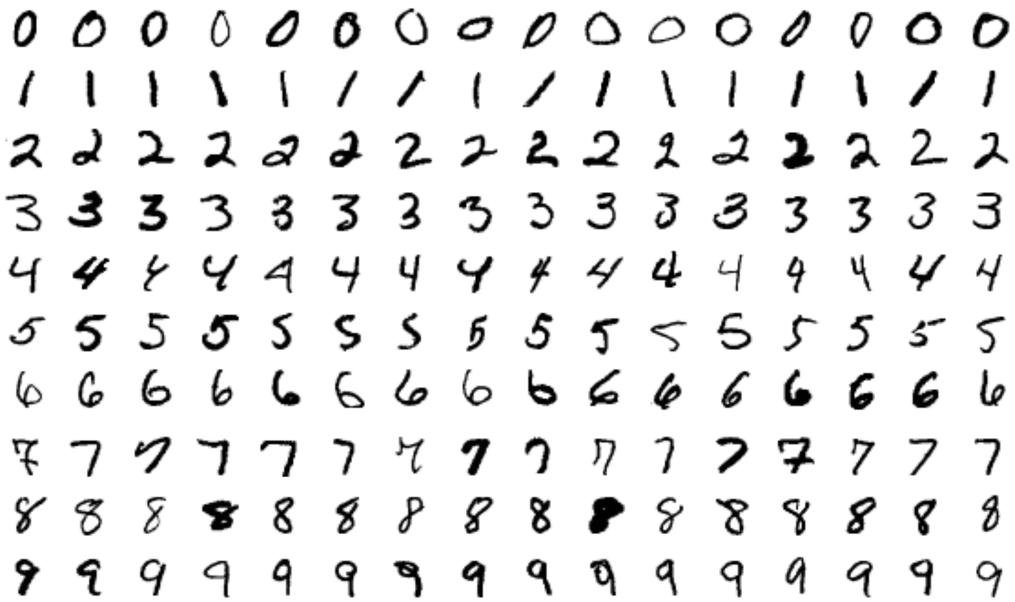


$K = 4$

Beyond Euclidean data

- Data with geometric features:
  - **manifold-valued data** [C, Yang (2021) *Appl. Comput. Harmon. Anal.*]: non-Euclidean data with low-dimensional structures.
  - **measure-valued data** [Zhuang, C, Yang (2022) *NeurIPS*]: images after normalization to probability distributions.





MNIST dataset

Clustering probability measures

- Input data: $\mu_1, \ldots, \mu_n$ probability measures.
- Example: 2D images, volumetric data (3D shape).
- **Vectorization**: rich geometric information lost.
- Natural distance: **optimal transport metric**

$$W_2^2(\mu, \nu) := \min_{\gamma \in \Gamma(\mu,\nu)} \left\{ \int_{\mathbb{R}^p \times \mathbb{R}^p} \|x - y\|_2^2 \, d\gamma(x, y) \right\}.$$

- **Challenges of Wasserstein space**:
  - Infinite-dimensional
  - Curved space: non-negative Alexandrov curvature

Wasserstein barycenter

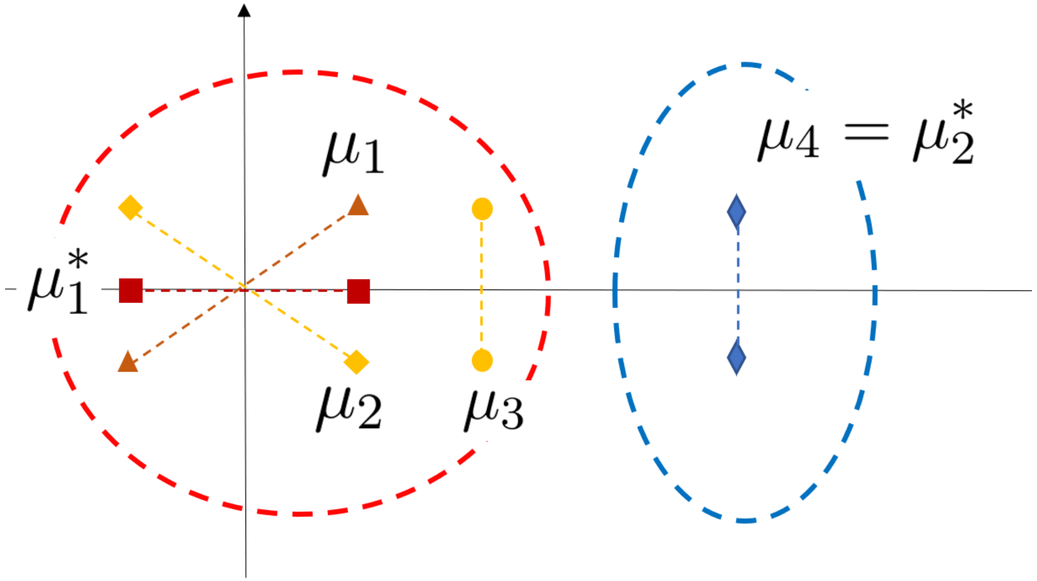
$$\bar{\nu} = \arg\min_{\nu} \sum_{i=1}^{n} \lambda_i W_2^2(\mu_i, \nu)$$

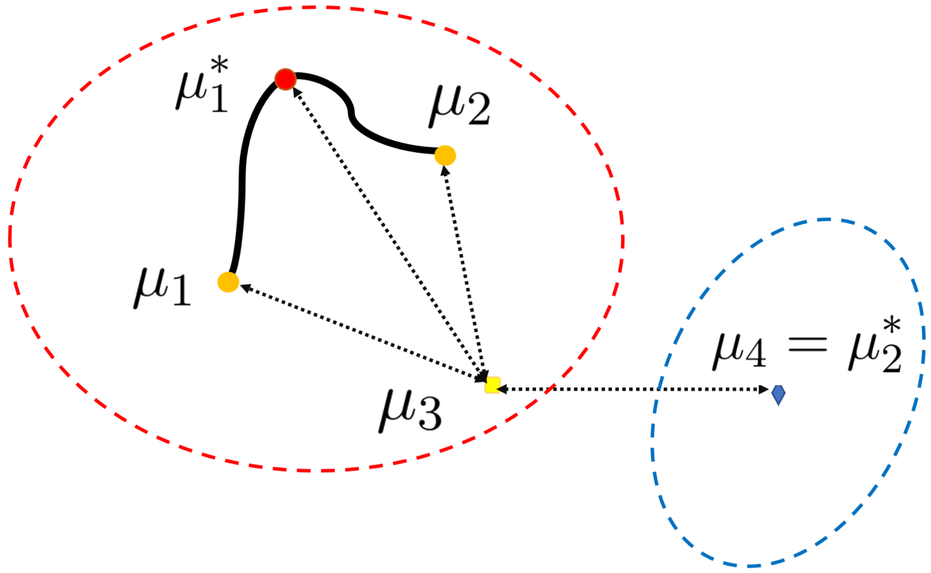
- Euclidean space: **vector space** and **flat** (i.e., zero-curvature).

Adversarial configuration

- Failure of centroid-based Wasserstein K-means:

$$W_2(\mu_3, \mu_1^*) > W_2(\mu_3, \mu_2^*) > \max\left\{W_2(\mu_3, \mu_1), W_2(\mu_3, \mu_2)\right\}$$

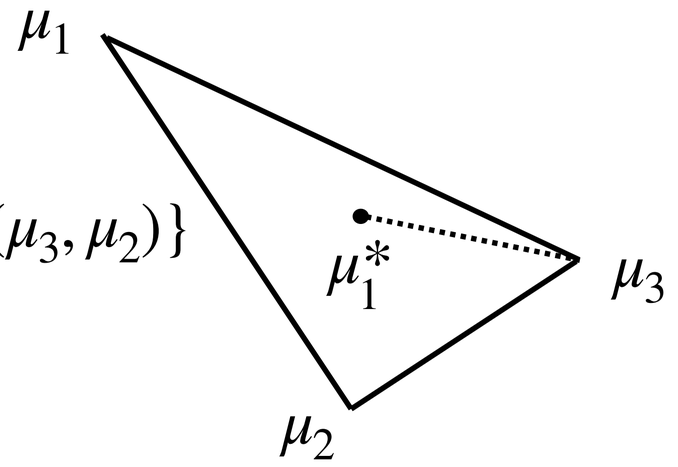




Violation of parallelogram law!

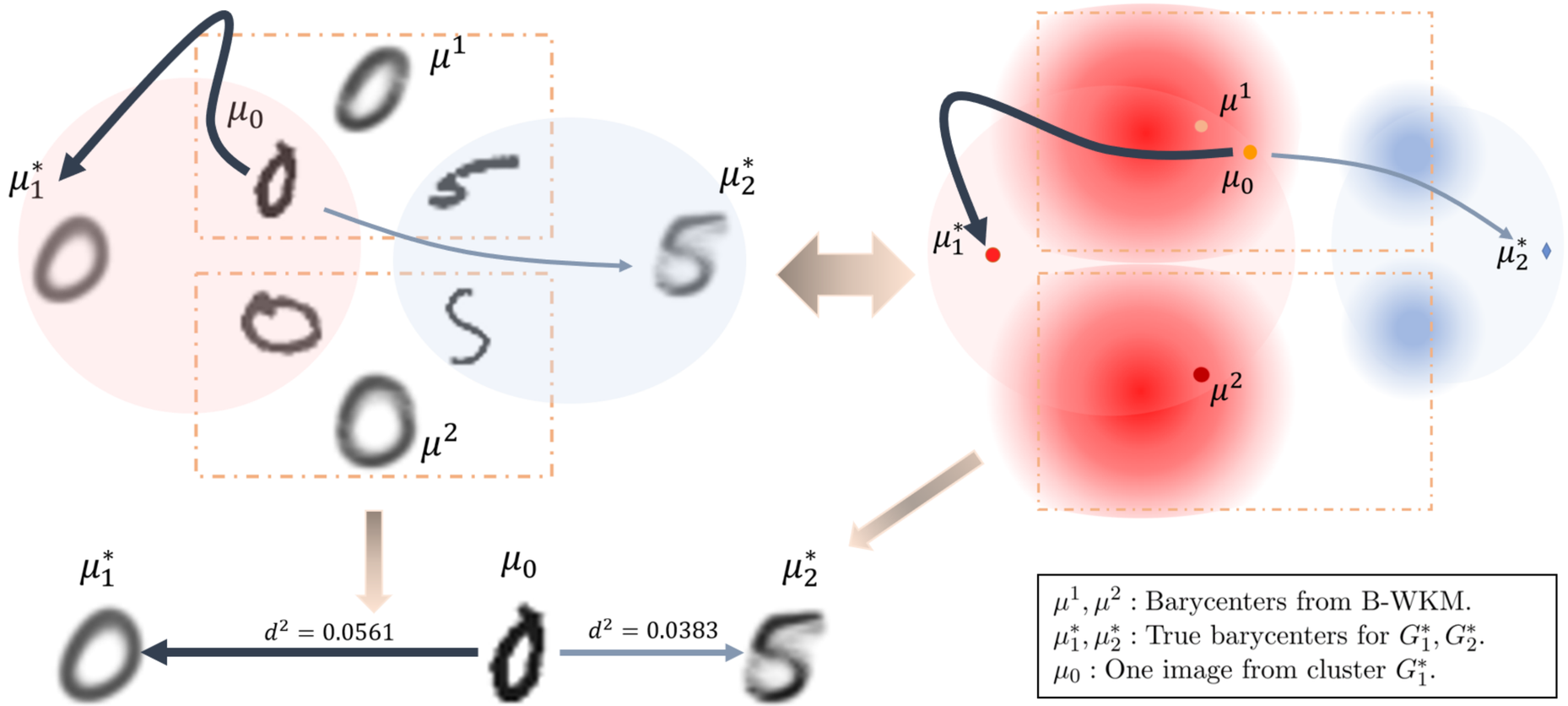
In Euclidean space,

$$d(\mu_3, \mu_1^*) \leq \max\{d(\mu_3, \mu_1), d(\mu_3, \mu_2)\}$$



Avoid centroid-based formulation in Wasserstein space!

MNIST data



- Mis-clustered example for the barycenter-based Wasserstein K-means (B-WKM) on a randomly sampled subset from MNIST (200 digit ‘0’ and 100 digit ‘5’).

Wasserstein K-means

- **Wasserstein K-means** [Zhuang, C, Yang (2022) *NeurIPS*]: partition-based formulation

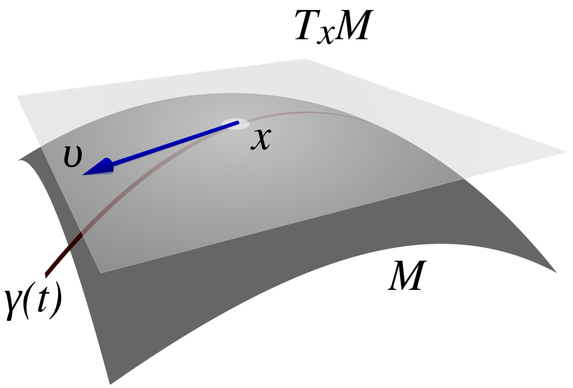
$$\min_{G_1,\ldots,G_K} \left\{ \sum_{k=1}^{K} \frac{1}{|G_k|} \sum_{i\in G_k} \sum_{j\in G_k} W_2^2(\mu_i,\mu_j) : \bigsqcup_{k=1}^{K} G_k = [n] \right\}.$$

- **Algorithms**:
  - Iterative greedy algorithm: mimic the Voronoi diagram.
  - SDP relaxation:

$$\min_{Z\in\mathbb{R}^{n\times n}} \left\{ \langle A, Z\rangle : Z^\top = Z, Z \succeq 0,\ \mathrm{Tr}(Z) = K, Z1_n = 1_n, Z \geq 0 \right\},$$

    where $a_{ij} = W_2^2(\mu_i,\mu_j)$.

Statistical guarantee



- Gaussian measures $\mu_i = N(0, V_i)$ from K groups $G_1^*, \ldots, G_K^*$ of size $n_1, \ldots, n_K$.
- Clustering structure on covariance matrices: if $i \in G_k^*$, then

Geodesic from $V^{(k)}$ ➡ $V_i = (I + tX_i)V^{(k)}(I + tX_i)$ with $X_1, \ldots, X_n \overset{i.i.d.}{\sim} SymN(0,1)$,

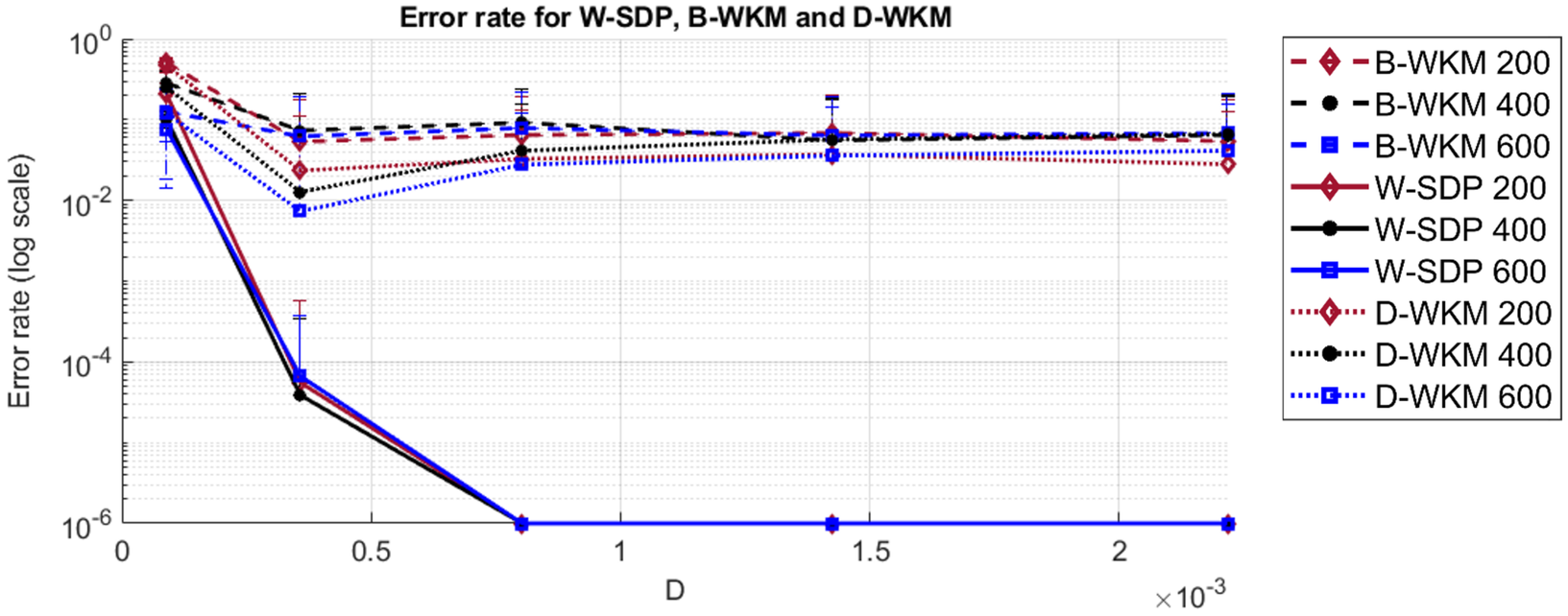
where $V^{(K)}$ is the center of k-th cluster, $SymN(0,1)$ denotes symmetric random matrix with i.i.d. standard Gaussian entries, t is a small perturbation parameter.

- **Separation**: **Bures distance**

$$W_2^2(N(0,V), N(0,U)) = \mathrm{Tr}\left[V + U - 2\left(V^{1/2}UV^{1/2}\right)^{1/2}\right] =: d^2(V,U).$$

- **Theorem**. If $\Theta^2 := \min_{k \neq l} d^2(V^{(k)}, V^{(l)}) \geq \overline{\Theta}^2 := t^2 \mathscr{V}\, p^2 \log n$ with $\mathscr{V} = \max_k \|V^{(k)}\|_{\mathrm{op}}$ and $\underline{n} := \min\{n_1, \ldots, n_K\} \geq \log^2 n$, then the Wasserstein K-means SDP achieves exact recovery with probability tending to 1.

Numeric examples



Gaussian simulation

W-SDP: Wasserstein K-means SDP
D-WKM: distance-based Wasserstein K-means (greedy)
B-WKM: barycenter-based Wasserstein K-means
KM: K-means

Table 2: Error rate (SD) for clustering three benchmark datasets: MNIST, Fashion-MNIST and USPS handwriting digits. MNIST_1 (MNIST_2) refers to the results of Case 1 (Case 2) for MNIST dataset.

	W-SDP	D-WKM	B-WKM	KM
MNIST_1	0.235 (0.045)	0.156 (0.057)	0.310 (0.069)	0.295 (0.066)
MNIST_2	0.279 (0.050)	0.185 (0.097)	0.324 (0.032)	0.362 (0.033)
Fashion-MNIST	0.082 (0.020)	0.056 (0.014)	0.141 (0.059)	0.138 (0.099)
USPS handwriting	0.206 (0.020)	0.159 (0.061)	0.240 (0.045)	0.284 (0.025)

References

- **Xiaohui Chen**, Yun Yang. (2021) Cutoff for Exact Recovery of Gaussian Mixture Models. *IEEE Transactions on Information Theory.*
- **Xiaohui Chen**, Yun Yang. (2021) Diffusion K-means clustering on manifolds: Provable exact recovery via semidefinite relaxations. *Appl. Comput. Harmon. Anal.*
- Yubo Zhuang, **Xiaohui Chen**, Yun Yang. (2022) Sketch-and-lift: scalable subsampled semidefinite program for K-means clustering. *AISTATS.*
- Yubo Zhuang, **Xiaohui Chen**, Yun Yang. (2022) Wasserstein K-means for clustering probability distributions. *NeurIPS.*
- Yubo Zhuang, **Xiaohui Chen**, Yun Yang, Richard Y. Zhang. (2023) Statistically Optimal K-means Clustering via Nonnegative Low-rank Semidefinite Programming. (arXiv:2305.18436)

THANK YOU!

- Research support:
  - NSF CAREER Award
  - Simons Fellowship in Mathematics
  - Arnold O. Beckman Award (UIUC)