When do birds of a feather flock together? k-means, proximity, and conic programming

Shuyang Ling

Courant Institute of Mathematical Sciences, NYU

May 14, 2018

Acknowledgement

Research in collaboration with:

- Prof. Xiaodong Li (Statistics, UC Davis)
- Prof. Thomas Strohmer, Yang Li (Mathematics, UC Davis)
- Prof. Ke Wei (School of Data Sciences, Fudan University, Shanghai)

k-means

Question: Given a set of N data points in $\mathbb{R}^m$, how to partition them into k clusters?

Criterion: minimize the k-means objective function:

$$\min_{\{\Gamma_l\}_{l=1}^k} \sum_{l=1}^{k} \underbrace{\sum_{i\in\Gamma_l} \|\boldsymbol{x}_i - \boldsymbol{c}_l\|^2}_{\text{within-cluster sum of squares}},$$

- $\{\Gamma_l\}$ is a partition of $\{1, \cdots, N\}$
- $\boldsymbol{c}_l$ is the sample mean of data points in Γ_l

k-means

Question: Given a set of N data points in $\mathbb{R}^m$, how to partition them into k clusters?

Criterion: minimize the k-means objective function:

$$\min_{\{\Gamma_l\}_{l=1}^k} \sum_{l=1}^{k} \underbrace{\sum_{i \in \Gamma_l} \|\boldsymbol{x}_i - \boldsymbol{c}_l\|^2}_{\text{within-cluster sum of squares}},$$

- $\{\Gamma_l\}$ is a partition of $\{1, \cdots, N\}$
- $\boldsymbol{c}_l$ is the sample mean of data points in Γ_l

Difficulty of k-means

Importance and Difficulties

- Widely used in vector quantization, unsupervised learning, Voronoi tessellation, etc.
- It is an NP-hard problem, even if $m = 2$. [Mahajan, etc 09]
- Heuristic method: Lloyd's algorithm [Lloyd 82] works well in practice. But convergence is not always guaranteed: it may take exponentially (in N) many steps to converge to stationary points (not even a local minimum).

Convex relaxation of k-means

Focus of talk

We are interested in the convex relaxation for k-means [Peng, Wei 07].

k-means

To minimize k-means objective, it suffices to optimize over all possible choices of partition $\{\Gamma_l\}$:

$$f(\{\Gamma_l\}) := \sum_{l=1}^{k} \sum_{i \in \Gamma_l} \|\boldsymbol{x}_i - \boldsymbol{c}_l\|^2$$

Convex relaxation of k-means

Focus of talk

We are interested in the convex relaxation for k-means [Peng, Wei 07].

k-means

To minimize k-means objective, it suffices to optimize over all possible choices of partition $\{\Gamma_l\}$:

$$f(\{\Gamma_l\}) := \sum_{l=1}^{k} \sum_{i \in \Gamma_l} \|\boldsymbol{x}_i - \boldsymbol{c}_l\|^2$$

Convex relaxation of k-means

Focus of talk

We are interested in the convex relaxation for k-means [Peng, Wei 07].

An equivalent form:

It suffices to minimize it over all choices of partition $\{\Gamma_l\}_{l=1}^k$:

$$f(\{\Gamma_l\}_{l=1}^k) := \sum_{l=1}^k \sum_{i\in\Gamma_l} \|\boldsymbol{x}_i - \boldsymbol{c}_l\|^2 = \sum_{l=1}^k \frac{1}{|\Gamma_l|} \sum_{i\in\Gamma_l, j\in\Gamma_l} \|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2$$

which is the sum of the squared pairwise deviations of points in the same cluster.

A bit more calculation

$f(\{\Gamma_l\}_{l=1}^k)$ is the inner product between two matrices

$$f(\{\Gamma_l\}) = \sum_{i=1}^N \sum_{j=1}^N \underbrace{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2}_{D_{ij}} \cdot \underbrace{\frac{1}{|\Gamma_l|} \mathbf{1}_{\{i\in\Gamma_l, j\in\Gamma_l\}}}_{X_{ij}} = \langle \boldsymbol{D}, \boldsymbol{X} \rangle$$

where $\boldsymbol{D} = (\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2)_{1\le i,j\le N}$ is the distance matrix and

$$\boldsymbol{X} = \left(\frac{1}{|\Gamma_l|} \cdot \mathbf{1}_{\{i\in\Gamma_l, j\in\Gamma_l\}} \right)_{1\le i,j\le N}$$

We simply call $\boldsymbol{X}$ the *partition matrix*.

What properties does $\boldsymbol{X}$ have for any given partition $\{\Gamma_l\}_{l=1}^k$?

A bit more calculation

$f(\{\Gamma_l\}_{l=1}^k)$ is the inner product between two matrices

$$f(\{\Gamma_l\}) = \sum_{i=1}^N \sum_{j=1}^N \underbrace{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2}_{D_{ij}} \cdot \underbrace{\frac{1}{|\Gamma_l|} \mathbf{1}_{\{i\in\Gamma_l, j\in\Gamma_l\}}}_{X_{ij}} = \langle \boldsymbol{D}, \boldsymbol{X} \rangle$$

where $\boldsymbol{D} = (\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2)_{1\leq i,j\leq N}$ is the distance matrix and

$$\boldsymbol{X} = \left(\frac{1}{|\Gamma_l|} \cdot \mathbf{1}_{\{i\in\Gamma_l, j\in\Gamma_l\}} \right)_{1\leq i,j\leq N}$$

We simply call $\boldsymbol{X}$ the *partition matrix*.

What properties does $\boldsymbol{X}$ have for any given partition $\{\Gamma_l\}_{l=1}^k$?

Relaxation

Up to certain permutation, the matrix $\boldsymbol{X}$ is a block-diagonal matrix:

$$\boldsymbol{X} = \begin{bmatrix} \frac{1}{|\Gamma_1|}\mathbf{1}_{|\Gamma_1|}\mathbf{1}_{|\Gamma_1|}^\top & \cdots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \frac{1}{|\Gamma_k|}\mathbf{1}_{|\Gamma_k|}\mathbf{1}_{|\Gamma_k|}^\top \end{bmatrix}$$

We want to find a larger and convex search space containing all $\boldsymbol{X}$ as a proper subset. What constraints does $\boldsymbol{X}$ satisfy?

Four constraints

- Nonnegativity: $\boldsymbol{X} \geq 0$.
- Positive semidefinite: $\boldsymbol{X} \succeq 0$.
- $\mathrm{Tr}(\boldsymbol{X}) = k$ (note that $\mathrm{rank}(\boldsymbol{X}) = k$ is nonconvex)
- Leading eigenvalues are 1 with multiplicities k: $\boldsymbol{X}\mathbf{1}_N = \mathbf{1}_N$.

Relaxation

Up to certain permutation, the matrix $\boldsymbol{X}$ is a block-diagonal matrix:

$$\boldsymbol{X} = \begin{bmatrix} \frac{1}{|\Gamma_1|}\mathbf{1}_{|\Gamma_1|}\mathbf{1}_{|\Gamma_1|}^\top & \cdots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \frac{1}{|\Gamma_k|}\mathbf{1}_{|\Gamma_k|}\mathbf{1}_{|\Gamma_k|}^\top \end{bmatrix}$$

We want to find a larger and convex search space containing all $\boldsymbol{X}$ as a proper subset. What constraints does $\boldsymbol{X}$ satisfy?

Four constraints

- Nonnegativity: $\boldsymbol{X} \geq 0$.
- Positive semidefinite: $\boldsymbol{X} \succeq 0$.
- $\mathrm{Tr}(\boldsymbol{X}) = k$ (note that $\mathrm{rank}(\boldsymbol{X}) = k$ is nonconvex)
- Leading eigenvalues are 1 with multiplicities k: $\boldsymbol{X}\mathbf{1}_N = \mathbf{1}_N$.

Relaxation

Up to certain permutation, the matrix $\boldsymbol{X}$ is a block-diagonal matrix:

$$\boldsymbol{X} = \begin{bmatrix} \frac{1}{|\Gamma_1|}\mathbf{1}_{|\Gamma_1|}\mathbf{1}_{|\Gamma_1|}^\top & \cdots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \frac{1}{|\Gamma_k|}\mathbf{1}_{|\Gamma_k|}\mathbf{1}_{|\Gamma_k|}^\top \end{bmatrix}$$

We want to find a larger and convex search space containing all $\boldsymbol{X}$ as a proper subset. What constraints does $\boldsymbol{X}$ satisfy?

Four constraints

- Nonnegativity: $\boldsymbol{X} \geq 0$.
- Positive semidefinite: $\boldsymbol{X} \succeq 0$.
- $\mathrm{Tr}(\boldsymbol{X}) = k$ (note that $\mathrm{rank}(\boldsymbol{X}) = k$ is nonconvex)
- Leading eigenvalues are 1 with multiplicities k: $\boldsymbol{X}\mathbf{1}_N = \mathbf{1}_N$.

Relaxation

Up to certain permutation, the matrix $\boldsymbol{X}$ is a block-diagonal matrix:

$$\boldsymbol{X} = \begin{bmatrix} \frac{1}{|\Gamma_1|}\mathbf{1}_{|\Gamma_1|}\mathbf{1}_{|\Gamma_1|}^\top & \cdots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \frac{1}{|\Gamma_k|}\mathbf{1}_{|\Gamma_k|}\mathbf{1}_{|\Gamma_k|}^\top \end{bmatrix}$$

We want to find a larger and convex search space containing all $\boldsymbol{X}$ as a proper subset. What constraints does $\boldsymbol{X}$ satisfy?

Four constraints

- Nonnegativity: $\boldsymbol{X} \geq 0$.
- Positive semidefinite: $\boldsymbol{X} \succeq 0$.
- $\mathrm{Tr}(\boldsymbol{X}) = k$ (note that $\mathrm{rank}(\boldsymbol{X}) = k$ is nonconvex)
- Leading eigenvalues are 1 with multiplicities k: $\boldsymbol{X}\mathbf{1}_N = \mathbf{1}_N$.

Relaxation

Up to certain permutation, the matrix $\boldsymbol{X}$ is a block-diagonal matrix:

$$\boldsymbol{X} = \begin{bmatrix} \frac{1}{|\Gamma_1|}\mathbf{1}_{|\Gamma_1|}\mathbf{1}_{|\Gamma_1|}^\top & \cdots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \frac{1}{|\Gamma_k|}\mathbf{1}_{|\Gamma_k|}\mathbf{1}_{|\Gamma_k|}^\top \end{bmatrix}$$

We want to find a larger and convex search space containing all $\boldsymbol{X}$ as a proper subset. What constraints does $\boldsymbol{X}$ satisfy?

Four constraints

- Nonnegativity: $\boldsymbol{X} \geq 0$.
- Positive semidefinite: $\boldsymbol{X} \succeq 0$.
- $\operatorname{Tr}(\boldsymbol{X}) = k$ (note that $\operatorname{rank}(\boldsymbol{X}) = k$ is nonconvex)
- Leading eigenvalues are 1 with multiplicities k: $\boldsymbol{X}\mathbf{1}_N = \mathbf{1}_N$.

Convex relaxation

Semidefinite programming relaxation [Peng, Wei, 07]

The convex relaxation of k-means is

$$\min \langle \boldsymbol{D}, \boldsymbol{Z} \rangle \quad s.t. \quad \boldsymbol{Z} \geq 0, \boldsymbol{Z} \succeq 0, \operatorname{Tr}(\boldsymbol{Z}) = k, \boldsymbol{Z}\mathbf{1}_N = \mathbf{1}_N.$$

Key question

Suppose we assume $\{\Gamma_l\}_{l=1}^k$ is the ground truth partition,

when does SDP relaxation recover $\boldsymbol{X} = \sum_{l=1}^k \frac{1}{|\Gamma_l|}\mathbf{1}_{\Gamma_l}\mathbf{1}_{\Gamma_l}^\top$?

Convex relaxation

Semidefinite programming relaxation [Peng, Wei, 07]

The convex relaxation of k-means is

$$\min \langle \boldsymbol{D}, \boldsymbol{Z} \rangle \quad s.t. \quad \boldsymbol{Z} \geq 0, \boldsymbol{Z} \succeq 0, \mathrm{Tr}(\boldsymbol{Z}) = k, \boldsymbol{Z}\mathbf{1}_N = \mathbf{1}_N.$$

Key question

Suppose we assume $\{\Gamma_l\}_{l=1}^k$ is the ground truth partition,

when does SDP relaxation recover $\boldsymbol{X} = \sum_{l=1}^k \frac{1}{|\Gamma_l|}\mathbf{1}_{\Gamma_l}\mathbf{1}_{\Gamma_l}^\top$?

A short literature review

Many excellent works for learning mixtures of distributions and SDP relaxation of k-means:

- SDP-relaxation of k-means: [Peng, Wei, 07], [Bandeira, Villar, Ward, etc, 17], [Mixon, Villar, etc, 15], etc.
- Spectral-projection based approaches: [Dasgupta, 99], [Vempala, Wang, 04], [Achlipotas, McSherry, 05], etc.

Almost all works have one thing in common: data are assumed to be sampled from a generative model, i.e., stochastic ball model, Gaussian mixture models, etc.

A short literature review

Many excellent works for learning mixtures of distributions and SDP relaxation of k-means:

- SDP-relaxation of k-means: [Peng, Wei, 07], [Bandeira, Villar, Ward, etc, 17], [Mixon, Villar, etc, 15], etc.
- Spectral-projection based approaches: [Dasgupta, 99], [Vempala, Wang, 04], [Achlipotas, McSherry, 05], etc.

Almost all works have one thing in common: data are assumed to be sampled from a generative model, i.e., stochastic ball model, Gaussian mixture models, etc.

A model-free framework?

Question: Can we establish a model-free framework to learn mixture of distributions?

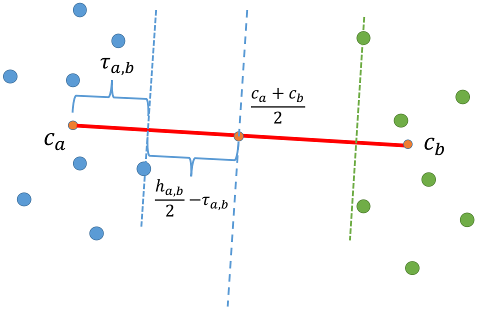
- Model-free: No assumption on data generative model.
- One model-free idea: different clusters are mutually *well-separated*.
- How large the separation is needed and in what sense?
- This is made possible by *proximity condition* [Kumar, Kannan, 10], [Awashi, Sheffet, 12].

A model-free framework?

Question: Can we establish a model-free framework to learn mixture of distributions?

- Model-free: No assumption on data generative model.
- One model-free idea: different clusters are mutually *well-separated*.
- How large the separation is needed and in what sense?
- This is made possible by *proximity condition* [Kumar, Kannan, 10], [Awashi, Sheffet, 12].

What is proximity condition?



- $h_{a,b}$: the distance between two centers
- $\tau_{a,b}$: the largest distance between data and their corresponding centers when projected on the line linking $\boldsymbol{c}_a$ with $\boldsymbol{c}_b$
- $d_{a,b} := \frac{h_{a,b}}{2} - \tau_{a,b}$ is the smallest distance between the middle point and projected data onto the line, which is a measure of separability

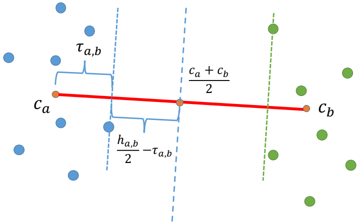
Proximity condition

Proximity condition

The partition $\Gamma = \sqcup_{l=1}^{k} \Gamma_l$ satisfies proximity condition if

$$d_{a,b} = \frac{h_{a,b}}{2} - \tau_{a,b} > \frac{1}{\sqrt{2}} \cdot \sqrt{k} \cdot \underbrace{\sqrt{\max \|\boldsymbol{\Sigma}_l\|}}_{\text{standard deviation}}$$

holds for any $a \neq b$ where $\boldsymbol{\Sigma}_l$ is the sample covariance matrix of data in Γ_l. Proximity condition quantifies how far each data point is away from the other clusters.



Main theorem

Theorem

Suppose the partition $\{\Gamma_l\}_{l=1}^k$ obeys the proximity condition, i.e.,

$$d_{a,b} \geq \frac{1}{\sqrt{2}} \cdot \underbrace{\sqrt{k}}_{tight?} \cdot \sqrt{\max \|\boldsymbol{\Sigma}_l\|}.$$

The minimizer of the SDP relaxation is unique and given by the ground truth partition $\boldsymbol{X}$.

- A purely deterministic and model-free condition.
- Conveniently apply to other data-generative models (shown in the next few slides).
- If all Γ_l are of the same size, the right hand side is replaced by $\sqrt{k} \cdot \sqrt{\max\{\|\boldsymbol{\Sigma}_a\|, \|\boldsymbol{\Sigma}_b\|\}}$ which only depends on the covariance matrix of group Γ_a and Γ_b.
- The dependence of Δ on $\sqrt{k}$ is not tight.

Main theorem

Theorem

Suppose the partition $\{\Gamma_l\}_{l=1}^k$ obeys the proximity condition, i.e.,

$$d_{a,b} \geq \frac{1}{\sqrt{2}} \cdot \underbrace{\sqrt{k}}_{tight?} \cdot \sqrt{\max \|\boldsymbol{\Sigma}_l\|}.$$

The minimizer of the SDP relaxation is unique and given by the ground truth partition $\boldsymbol{X}$.

- A purely deterministic and model-free condition.
- Conveniently apply to other data-generative models (shown in the next few slides).
- If all Γ_l are of the same size, the right hand side is replaced by $\sqrt{k} \cdot \sqrt{\max\{\|\boldsymbol{\Sigma}_a\|, \|\boldsymbol{\Sigma}_b\|\}}$ which only depends on the covariance matrix of group Γ_a and Γ_b.
- The dependence of Δ on $\sqrt{k}$ is not tight.

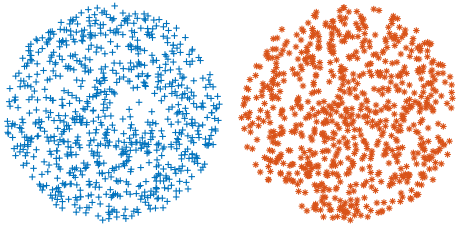
Data generative model - Stochastic ball model

Stochastic ball model

The data is generated from

$$\boldsymbol{x}_{a,i} = \boldsymbol{\mu}_a + \boldsymbol{r}_{a,i}, \quad 1 \leq i \leq n, \quad 1 \leq a \leq k$$

where $\boldsymbol{\mu}_a \in \mathbb{R}^m$ is the population center and $\boldsymbol{r}_{a,i}$ is uniform in $\mathcal{B}(\mathbb{R}^m)$.



Obviously, $\Delta = \min_{a \neq b} \|\boldsymbol{\mu}_a - \boldsymbol{\mu}_b\| > 2$ guarantees two balls are not overlapped and is necessary for exact recovery.

Data generative model - Stochastic ball model

- Our bound is slightly larger than 2 where the difference depends on the number of clusters k and dimension m.

Corollary

The proximity condition holds with high probability if

$$\Delta \geq 2 + \sqrt{2k \max \|\mathbf{\Sigma}_l\|} = 2 + \sqrt{\frac{2k}{m+2}} + o(1)$$

where Δ is the minimal separation $\Delta = \min_{a \neq b} \|\boldsymbol{\mu}_a - \boldsymbol{\mu}_b\|$ and m is the dimension.

State-of-the-art [Awashi, Bandeira, Villar, Ward, Mixon, etc, 2015, 2017]:

$$\Delta > \min\left\{2\sqrt{2}\left(1 + \frac{1}{\sqrt{m}}\right), 2 + \frac{k^2}{m}\right\}.$$

Data generative model - Stochastic ball model

- Our bound is slightly larger than 2 where the difference depends on the number of clusters k and dimension m.

Corollary

The proximity condition holds with high probability if

$$\Delta \geq 2 + \sqrt{2k \max \|\boldsymbol{\Sigma}_l\|} = 2 + \sqrt{\frac{2k}{m+2}} + o(1)$$

where Δ is the minimal separation $\Delta = \min_{a \neq b} \|\boldsymbol{\mu}_a - \boldsymbol{\mu}_b\|$ and m is the dimension.

State-of-the-art [Awashi, Bandeira, Villar, Ward, Mixon, etc, 2015, 2017]:

$$\Delta > \min \left\{ 2\sqrt{2}\left(1 + \frac{1}{\sqrt{m}}\right), 2 + \frac{k^2}{m} \right\}.$$

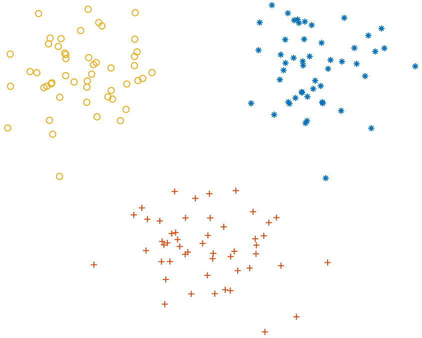
Data generative model - Gaussian mixture model

Gaussian mixture model

Consider

$$\boldsymbol{x}_{a,i} \sim \mathcal{N}(\boldsymbol{\mu}_a, \boldsymbol{\Sigma}_a), \quad 1 \leq i \leq n, 1 \leq a \leq k$$

where $\boldsymbol{\Sigma}_a$ is the covariance matrix.



Data generative model - Gaussian mixture model

Corollary

Assume $\boldsymbol{\Sigma}_a = \boldsymbol{I}_m$ for all $1 \leq a \leq k$, the proximity condition holds with high probability if

$$\Delta \geq 2\sqrt{k} + 4\sqrt{2}\log^{1/2}(kN^2) + o(1),$$

if $N \gg m^2 k^3 \log(k)$.

Gaussian mixture model: we achieve state-of-the-art result

$$\Delta \geq \mathcal{O}(\sqrt{k} + \log^{1/2}(kN))$$

for minimal separation by e.g. [Awasthi, Sheffet, 12] and [Mixon, Villar, Ward, 17], etc.

An impossibility theorem

Question: How tight is our bound?

The minimal separation Δ cannot be arbitrarily small, i.e., there is a lower bound for the separation for SDP to work. Here is one specific example:

Theorem

For stochastic ball model, the Peng-Wei relaxation fails to achieve exact recovery if N is large enough and

$$\Delta < 1 + \sqrt{1 + \frac{2}{m+2}} \approx 2 + \|\boldsymbol{\Sigma}\|$$

where $\|\boldsymbol{\Sigma}\| = \frac{1}{m+2}$.

Numerics: How does Δ depend on k?

Our bound: $\Delta \geq 2 + \sqrt{\frac{2k}{m+2}}$;

State-of-the-art bound: $\Delta \geq \min\left\{2\sqrt{2}\left(1 + \frac{1}{\sqrt{m}}\right), 2 + \frac{k^2}{m}\right\}$

The bound does not depend on k much.

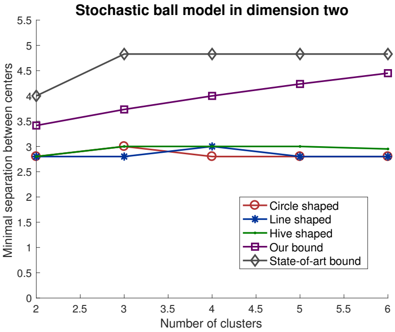


Figure: Numerical experiment on the stochastic ball model with dimension 2 and number of clusters k varies from 2 to 6.

Numerics: How does Δ depend on m?

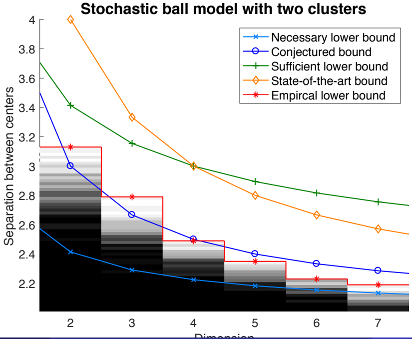
Here $k = 2$ and change m from 2 to 7.

Conjectured bound: $\Delta \geq 2 + \frac{2}{m+2}$

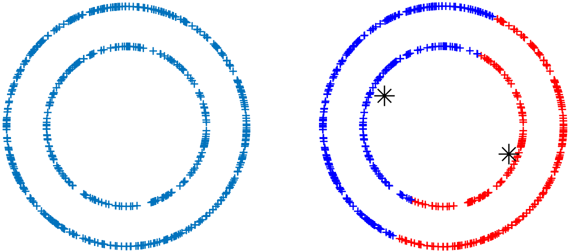
Necessary lower bound: $\Delta > 1 + \sqrt{1 + \frac{2}{m+2}}$

Sufficient lower bound: $\Delta > 2 + \frac{2}{\sqrt{m+2}}$

State-of-the-art: $\Delta > \min\left\{2\sqrt{2}\left(1 + \frac{1}{\sqrt{m}}\right), 2 + \frac{k^2}{m}\right\}$



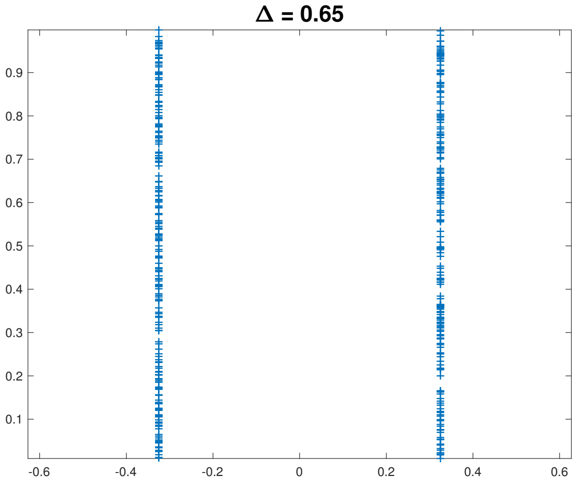
Example 1: data are on two circles with the same centers but different radius.



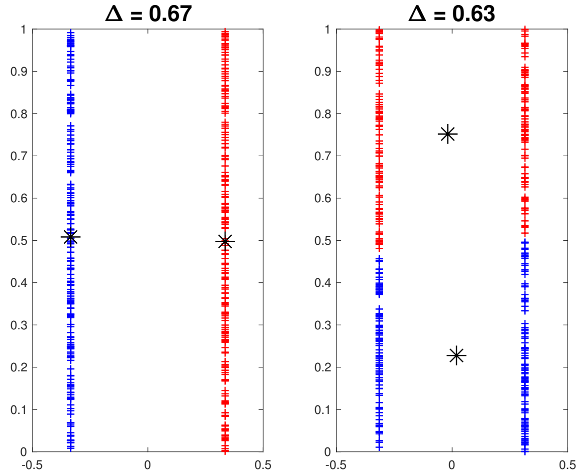
k-means does not work at all since it usually works for convex clusters.

Is k-means always a good choice? - toy example 2

Example 2: data are lying uniformly on two unit intervals with separation about $\Delta \approx 0.65$. Let's guess where the centers are?



Is k-means always a good choice? - toy example 2



Advertisement for an upcoming paper: kernel k-means?

- **Observation:** k-means does not work if the geometry of data is complicated.
- **Solution:** spectral clustering which consists of Laplacian eigenmap and k-means. However, many theoretic questions are not well understood.
- **Question:** Can we extend this convex relaxation framework to spectral clustering or kernel k-means?
- Yes, we will propose a convex relaxation of spectral clustering. It is also model-free and provably solves the previous two cases where ordinary k-means fails. The paper will be released soon!

Advertisement for an upcoming paper: kernel k-means?

- **Observation:** k-means does not work if the geometry of data is complicated.
- **Solution:** spectral clustering which consists of Laplacian eigenmap and k-means. However, many theoretic questions are not well understood.
- **Question:** Can we extend this convex relaxation framework to spectral clustering or kernel k-means?
- Yes, we will propose a convex relaxation of spectral clustering. It is also model-free and provably solves the previous two cases where ordinary k-means fails. The paper will be released soon!

Advertisement for an upcoming paper: kernel k-means?

- **Observation:** k-means does not work if the geometry of data is complicated.
- **Solution:** spectral clustering which consists of Laplacian eigenmap and k-means. However, many theoretic questions are not well understood.
- **Question:** Can we extend this convex relaxation framework to spectral clustering or kernel k-means?
- Yes, we will propose a convex relaxation of spectral clustering. It is also model-free and provably solves the previous two cases where ordinary k-means fails. The paper will be released soon!

Advertisement for an upcoming paper: kernel k-means?

- **Observation:** k-means does not work if the geometry of data is complicated.
- **Solution:** spectral clustering which consists of Laplacian eigenmap and k-means. However, many theoretic questions are not well understood.
- **Question:** Can we extend this convex relaxation framework to spectral clustering or kernel k-means?
- Yes, we will propose a convex relaxation of spectral clustering. It is also model-free and provably solves the previous two cases where ordinary k-means fails. The paper will be released soon!

Open problem and conclusions

Conclusions

- A model-free framework to certify the exactness of SDP relaxation applied to k-means.
- More details can be found *arXiv:1710.06008*.

Open problems

- For a mixture generated by the generalized stochastic ball model, is it possible to show

$$\Delta \geq 2 + \mathcal{O}\left(\frac{1}{m}\right),$$

suffices provided that the total number of points N is large enough.

- How to analyze misclassification rate via convex optimization approach?
- Understand the convergence of Lloyd's algorithm?

Open problem and conclusions

Conclusions

- A model-free framework to certify the exactness of SDP relaxation applied to k-means.
- More details can be found *arXiv:1710.06008*.

Open problems

- For a mixture generated by the generalized stochastic ball model, is it possible to show

$$\Delta \geq 2 + \mathcal{O}\left(\frac{1}{m}\right),$$

suffices provided that the total number of points N is large enough.

- How to analyze misclassification rate via convex optimization approach?
- Understand the convergence of Lloyd's algorithm?