Clustering bounds for hidden Partition Models

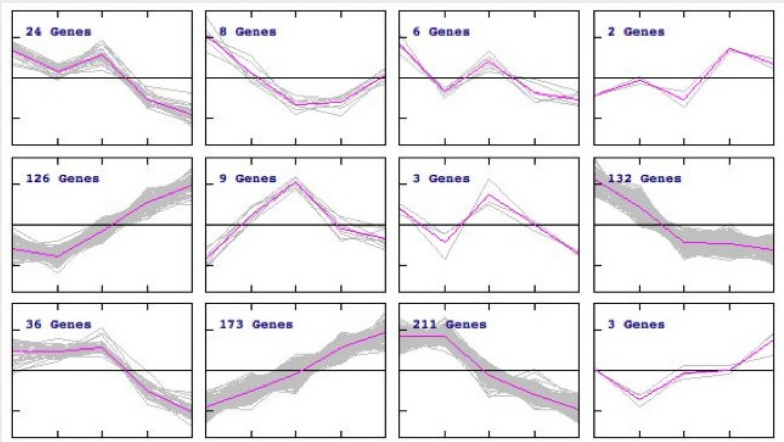
Nicolas Verzelen
INRAE, Montpellier

Mathematical Statistics in the Information Age

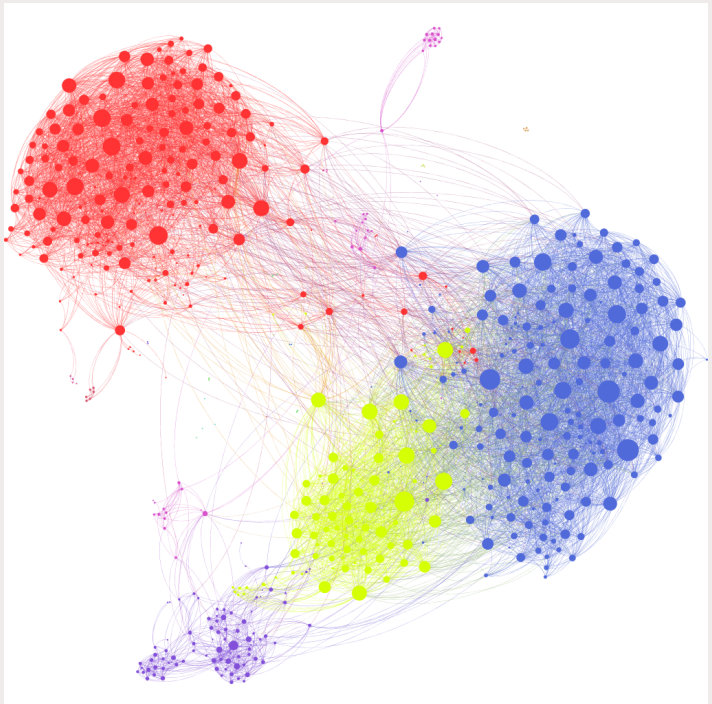
Clustering arises in various contexts

Clustering individuals w.r.t. features

Clustering features



Clustering graphs



Objectives

Topic of the talk

- Introduce several hidden partition problems.
- fresh view on some classical clustering algorithms.
- Some recovery error bounds for Gaussian Mixtures

Main Message

K-means and its relaxations (and corrections) are versatile and near optimal tools.

1 Hidden Partition problems

2 Omnibus Clustering Algorithms
- K-means criterion
- Lloyd's Algorithm
- Peng and Wei's SDP
- Spectral Clustering

3 Recovery bounds for (sub)Gaussian Mixtures
- Loss functions
- Minimax lower bound
- Recovery bounds for Peng and Wei' SDP

Mixture of Gaussian vectors [Pearson('1895)]

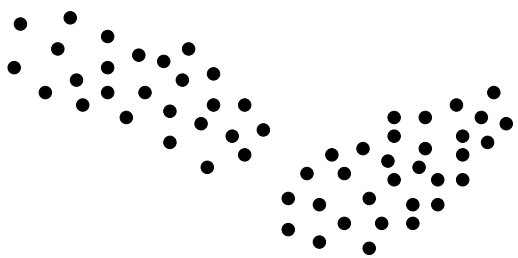
Partition

$G^* = \{G_1^*, \dots, G_K^*\}$ of $[n]$

Mixture of Gaussian variables (conditional)

$X_1, \dots, X_n \in \mathbb{R}^p$ are independent with $X_a \sim \mathcal{N}(\theta_k, \mathbf{\Sigma}_k)$ if $a \in G_k^*$

The observations are gathered in $\mathbf{X} = \begin{bmatrix} X_1 \\ \dots \\ X_n \end{bmatrix} \in \mathbb{R}^{n \times p}$



Mixture of Gaussian vectors [Pearson('1895)]

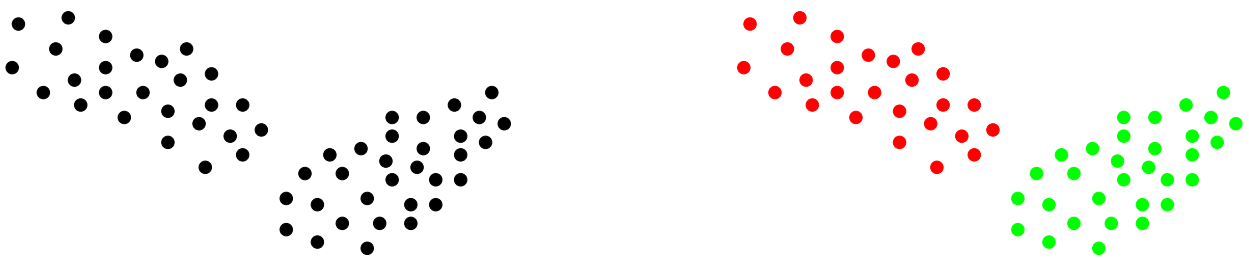
Partition

$G^* = \{G_1^*, \dots, G_K^*\}$ of $[n]$

Mixture of Gaussian variables (conditional)

$X_1, \dots, X_n \in \mathbb{R}^p$ are independent with $X_a \sim \mathcal{N}(\theta_k, \mathbf{\Sigma}_k)$ if $a \in G_k^*$

The observations are gathered in $\mathbf{X} = \begin{bmatrix} X_1 \\ \dots \\ X_n \end{bmatrix} \in \mathbb{R}^{n \times p}$



Objective : recovering G^* from $\mathbf{X}$ (θ and $\mathbf{\Sigma}$ are unknown but K is known)

Mixture of Gaussian vectors [Pearson('1895)]

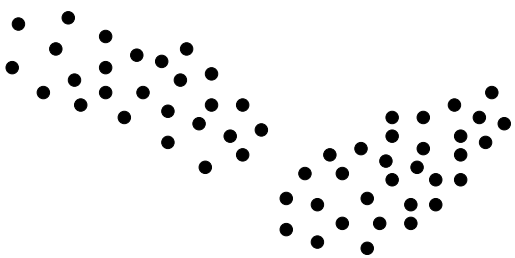
Partition

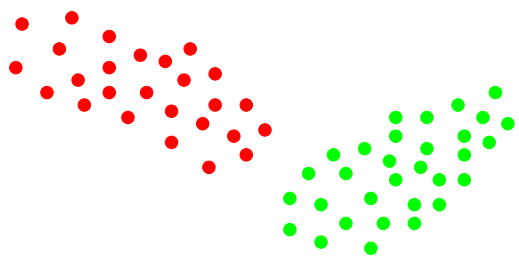
$G^* = \{G_1^*, \ldots, G_K^*\}$ of $[n]$

Mixture of Gaussian variables (conditional)

$X_1, \ldots, X_n \in \mathbb{R}^p$ are independent with $X_a \sim \mathcal{N}(\theta_k, \mathbf{\Sigma}_k)$ if $a \in G_k^*$

The observations are gathered in $\mathbf{X} = \begin{bmatrix} X_1 \\ \ldots \\ X_n \end{bmatrix} \in \mathbb{R}^{n \times p}$





Objective : recovering G^* from $\mathbf{X}$ (θ and $\mathbf{\Sigma}$ are unknown but K is known)

Clustering Problem $\neq$ Parameter Estimation

Mixture of Gaussians

Membership Matrix $\mathbf{A} \in \mathbb{R}^{n \times K}$ **defined by** $\mathbf{A}_{ak} = \mathbf{1}_{a \in G_k} = \mathbf{\Pi} \begin{bmatrix} 1 & 0 & 0 \\ & \dots & \\ 0 & 1 & 0 \\ & \dots & \\ 0 & 0 & 1 \end{bmatrix}$.

Mean component Matrix : $\mathbf{\Theta} = \begin{bmatrix} \theta_1 \\ \dots \\ \theta_K \end{bmatrix}$.

Population Version : $\mathbb{E}[\mathbf{X}] = \mathbf{A\Theta}$

$$\mathbf{X} = \mathbf{A\Theta} + \mathbf{E} = \text{“Signal”} + \text{“Noise”}$$

Stochastic Block Model (SBM)

Holland et al.('83), Abbe('17),. . . ,
$\mathbf{X}$ = adjacency matrix of an undirected graph $\in \{0,1\}^{n\times n}$.

Let $\mathbf{Q} \in [0,1]^{K\times K}_{sym}$

(conditional) SBM

The graph is generated by a SBM with partition G^* and matrix $\mathbf{Q}$ if $\mathbf{X}_{ab}$ with $a < b$ are independent and

$$\mathbb{P}[\mathbf{X}_{ab} = 1] = \mathbf{Q}_{jk} \quad \text{for any } a \in G^*_j \text{ and } b \in G^*_k \quad ,$$

Stochastic Block Model (SBM)

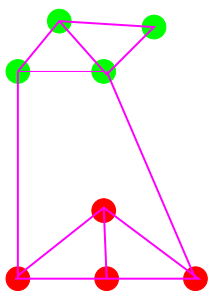
Holland et al.('83), Abbe('17),. . . ,
$\mathbf{X}$ = adjacency matrix of an undirected graph $\in \{0,1\}^{n\times n}$.

Let $\mathbf{Q} \in [0,1]^{K\times K}_{sym}$

(conditional) SBM

The graph is generated by a SBM with partition G^* and matrix $\mathbf{Q}$ if $\mathbf{X}_{ab}$ with $a < b$ are independent and

$$\mathbb{P}[\mathbf{X}_{ab} = 1] = \mathbf{Q}_{jk} \quad \text{for any } a \in G^*_j \text{ and } b \in G^*_k \quad ,$$



Stochastic Block Model (SBM)

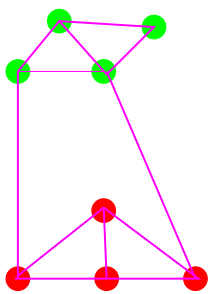
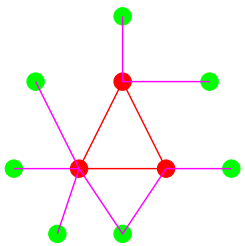
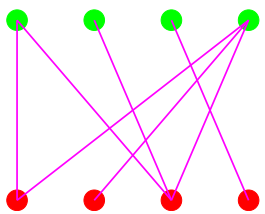
Holland et al.('83), Abbe('17),. . . ,
$\mathbf{X}$ = adjacency matrix of an undirected graph $\in \{0,1\}^{n\times n}$.

Let $\mathbf{Q} \in [0,1]^{K\times K}_{sym}$

(conditional) SBM

The graph is generated by a SBM with partition G^* and matrix $\mathbf{Q}$ if $\mathbf{X}_{ab}$ with $a < b$ are independent and

$$\mathbb{P}[\mathbf{X}_{ab} = 1] = \mathbf{Q}_{jk} \quad \text{for any } a \in G^*_j \text{ and } b \in G^*_k \quad ,$$

Stochastic Block Model (SBM)

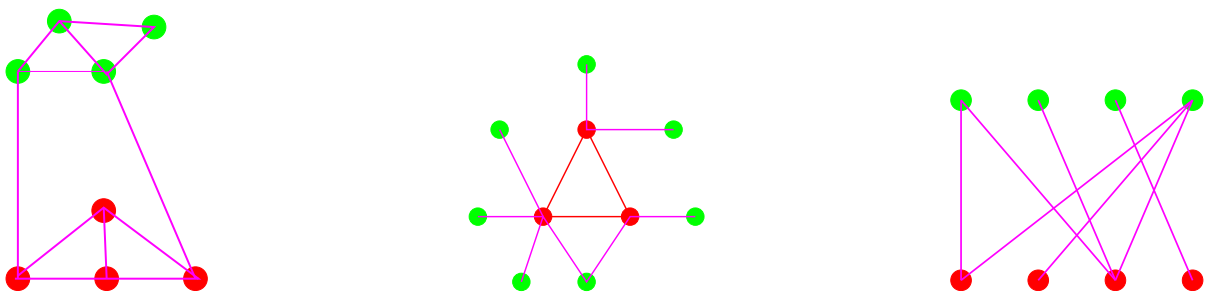
Holland et al.('83), Abbe('17),. . . ,
$\mathbf{X}$ = adjacency matrix of an undirected graph $\in \{0,1\}^{n\times n}$.

Let $\mathbf{Q} \in [0,1]^{K\times K}_{sym}$

(conditional) SBM

The graph is generated by a SBM with partition G^* and matrix $\mathbf{Q}$ if $\mathbf{X}_{ab}$ with $a < b$ are independent and

$$\mathbb{P}[\mathbf{X}_{ab} = 1] = \mathbf{Q}_{jk} \quad \text{for any } a \in G^*_j \text{ and } b \in G^*_k \quad ,$$



Objective : recovering G^* from $\mathbf{X}$ ($\mathbf{Q}$ is unknown.)

Population Version : $\mathbb{E}[\mathbf{X}] = \mathbf{AQA}^T - \mathrm{Diag}(\mathbf{AQA}^T) \approx \mathbf{A}[\mathbf{QA}^T]$

Outline

Model in mind

Algorithms do not depend on a particular model. To provide some intuition :

Mixture of Gaussian variables (conditional)

$X_1, \ldots, X_n \in \mathbb{R}^p$ are independent with $X_a \sim \mathcal{N}(\theta_k, \mathbf{\Sigma}_k)$ if $a \in G_k^*$

Signal + Noise decomposition

$$\mathbf{X} = \mathbf{A}\mathbf{\Theta} + \mathbf{E}$$

Outline

Maximum Likelihood Estimation

For the Gaussian Model, MLE

$$\hat{G}^{MLE} \in \arg\min_{G} \sum_{k=1}^{K} \min_{\mathbf{\Sigma}_k \in \mathcal{S}_p^+} \min_{\theta_k \in \mathbb{R}^p} \sum_{a \in G_k} \Big((X_i - \theta_k)^T \mathbf{\Sigma}_k^{-1} (X_i - \theta_k) + \log(\det(\mathbf{\Sigma}_k)) \Big)$$

Two Difficulties :

- **Computational** : (In principle) requires to scan over the space of partitions (size of order $K^n/K!$)
- **Statistical** : Estimation of $\mathbf{\Sigma}_k$ unstable for large p.

Maximum Likelihood Estimation

For the Gaussian Model, MLE

$$\hat{G}^{MLE} \in \arg\min_{G} \sum_{k=1}^{K} \min_{\mathbf{\Sigma}_k \in \mathcal{S}_p^+} \min_{\theta_k \in \mathbb{R}^p} \sum_{a \in G_k} \Big((X_i - \theta_k)^T \mathbf{\Sigma}_k^{-1} (X_i - \theta_k) + \log(\det(\mathbf{\Sigma}_k)) \Big)$$

Two Difficulties :

- **Computational** : (In principle) requires to scan over the space of partitions (size of order $K^n/K!$)
- **Statistical** : Estimation of $\mathbf{\Sigma}_k$ unstable for large p.

$\rightsquigarrow$ For the latter, assume in the criterion that $\mathbf{\Sigma}_k = \sigma^2 \mathbf{I}_p$

$$\hat{G} \in \arg\min_{G} \sum_{k=1}^{K} \min_{\theta_k \in \mathbb{R}^p} \sum_{a \in G_k} \|X_a - \theta_k\|_2^2$$

K-means criterion

$\widehat{G} \in \arg\min_G \mathsf{Crit}(\mathbf{X}, G)$ where

$$\mathsf{Crit}(\mathbf{X}, G) = \sum_{k=1}^{K} \sum_{a \in G_k} \|X_a - \overline{X}_{G_k}\|^2 = \frac{1}{2} \sum_{k=1}^{K} \frac{1}{|G_k|} \sum_{a,b \in G_k} \|X_a - X_b\|^2 \ ,$$

where $\overline{X}_{G_k} = \frac{1}{|G_k|} \sum_{a \in G_k} X_a$

K-means criterion

$\widehat{G} \in \arg\min_G \text{Crit}(\mathbf{X}, G)$ where

$$\text{Crit}(\mathbf{X}, G) = \sum_{k=1}^{K} \sum_{a \in G_k} \|X_a - \overline{X}_{G_k}\|^2 = \frac{1}{2} \sum_{k=1}^{K} \frac{1}{|G_k|} \sum_{a,b \in G_k} \|X_a - X_b\|^2 \ ,$$

where $\overline{X}_{G_k} = \frac{1}{|G_k|} \sum_{a \in G_k} X_a$

We did not address the computational issues

- There can be many local optima
- In worst-case solving K-means is NP-hard (Mahajan et al.('09))
- $(1+\epsilon)$-approximation is also NP-hard :
  Finding $\hat{G}$ s.t. $\text{Crit}(\mathbf{X}, \hat{G}) \leq (1+\epsilon) \min_G \text{Crit}(\mathbf{X}, G)$
- 8-approximation is possible in polynomial time (Kanungo et al.('04))

Encoding partition learning as a Matrix Problem.

Membership Matrix $\mathbf{A} \in \{0,1\}^{n \times K}$ defined by $\mathbf{A}_{ak} = \mathbf{1}_{a \in G_k}$ (or equivalently function $k : [n] \mapsto [K]$)
is NOT Identifiable. Why ?

Encoding partition learning as a Matrix Problem.

Membership Matrix $\mathbf{A} \in \{0,1\}^{n \times K}$ defined by $\mathbf{A}_{ak} = \mathbf{1}_{a \in G_k}$ (or equivalently function $k : [n] \mapsto [K]$)
is at best identifiable up to permutation

A more suitable object : The $n \times n$ **partnership** matrix
$\mathbf{B}^* = \mathbf{A}(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T$

$$\mathbf{B}^*_{ab} = \begin{cases} \frac{1}{|G^*_k|} & \text{if } a \text{ and } b \text{ belong to the same } G^*_k \\ 0 & \text{else} \end{cases}$$

Invariant with respect to the group labeling.

Encoding partition learning as a Matrix Problem.

Membership Matrix $\mathbf{A} \in \{0,1\}^{n \times K}$ defined by $\mathbf{A}_{ak} = \mathbf{1}_{a \in G_k}$ (or equivalently function $k : [n] \mapsto [K]$)
is at best identifiable up to permutation

A more suitable object : The $n \times n$ **partnership** matrix
$\mathbf{B}^* = \mathbf{A}(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T$

$$\mathbf{B}^*_{ab} = \begin{cases} \frac{1}{|G^*_k|} & \text{if } a \text{ and } b \text{ belong to the same } G^*_k \\ 0 & \text{else} \end{cases}$$

Invariant with respect to the group labeling.

Properties :
$\overline{\mathrm{tr}[\mathbf{B}^*] = K}$.

$$\mathbf{B}^* = \mathbf{\Pi}^T \begin{pmatrix} \frac{1}{|G^*_1|}\mathbf{J}_{|G^*_1|} & 0 & 0 \\ 0 & \frac{1}{|G^*_2|}\mathbf{J}_{|G^*_2|} & 0 \\ 0 & 0 & \frac{1}{|G^*_3|}\mathbf{J}_{|G^*_3|} \end{pmatrix} \mathbf{\Pi}$$

Rewriting K-means as a matrix estimation problem

$$
\begin{aligned}
\mathsf{Crit}(\mathbf{X}, G) &= \frac{1}{2} \sum_{k=1}^{K} \frac{1}{|G_k|} \sum_{a,b \in G_k} \|X_a - X_b\|^2 \\
&= -\sum_{k} \sum_{a,b \ \in G_k} \langle X_a, X_b \rangle \frac{1}{|G_k|} + \sum_{a=1}^{n} \|X_a\|_2^2 \\
&= -\langle \mathbf{X}\mathbf{X}^T, \mathbf{B} \rangle + \dots
\end{aligned}
$$

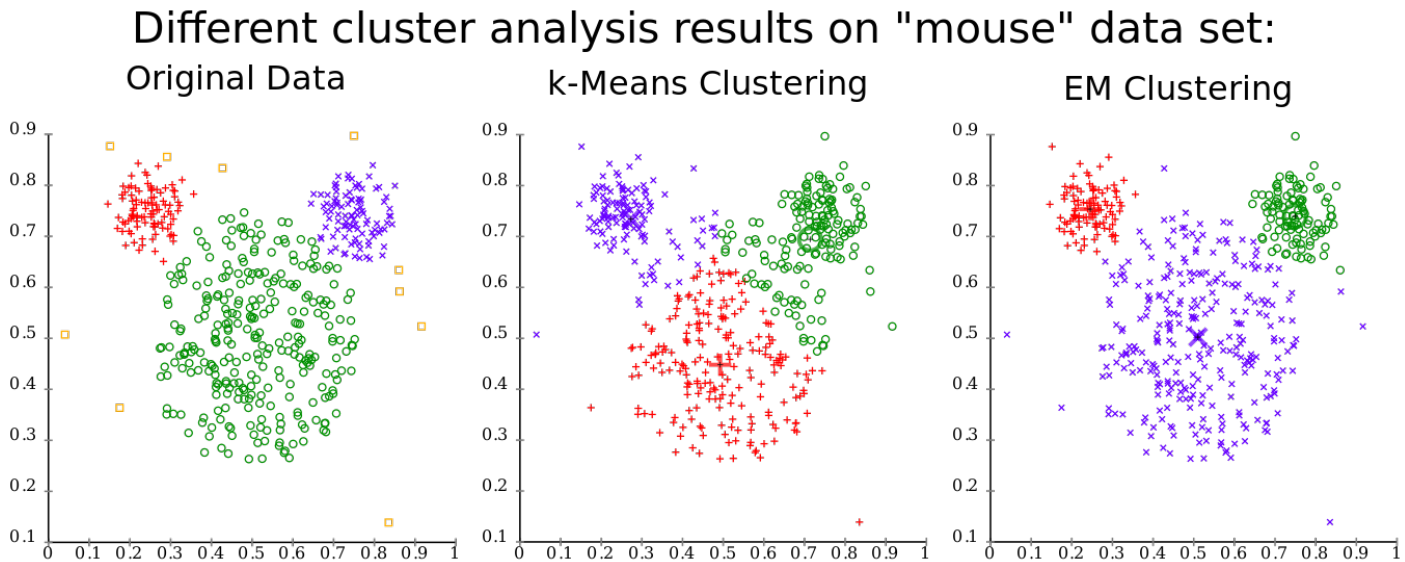
Rewriting K-means as a matrix estimation problem

$$\begin{aligned}
\text{Crit}(\mathbf{X}, G) &= \frac{1}{2}\sum_{k=1}^{K}\frac{1}{|G_k|}\sum_{a,b\in G_k}\|X_a - X_b\|^2 \\
&= -\sum_{k}\sum_{a,b\ \in G_k}\langle X_a, X_b\rangle\frac{1}{|G_k|} + \sum_{a=1}^{n}\|X_a\|_2^2 \\
&= -\langle \mathbf{X}\mathbf{X}^T, \mathbf{B}\rangle + \dots
\end{aligned}$$

K-means : linear minimization problem over the space of partnership matrices

$$\widehat{G} = \arg\min_{\mathbf{B}}\langle -\mathbf{X}\mathbf{X}^T, \mathbf{B}\rangle$$

One caveat of K-means criterion



https://en.wikipedia.org/wiki/K-means_clustering

Bias of K-means :

- Tends to split groups with large variance and favors small groups

Outline

Lloyd's algorithm Lloyd('82)

K-means as a least-square problem :

$$\hat{G} \in \arg\min_{G} \sum_{k=1}^{K} \min_{\theta_k \in \mathbb{R}^p} \sum_{a \in G_k} \|X_a - \theta_k\|_2^2$$

Lloyd's algorithm Lloyd('82)

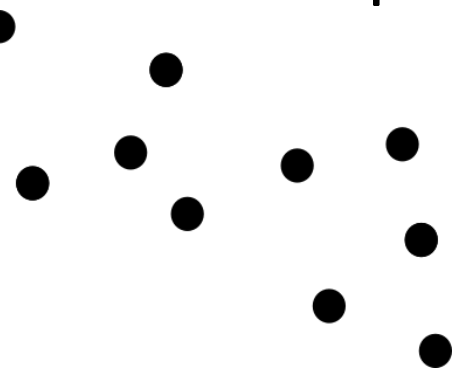
K-means as a least-square problem :

$$\hat{G} \in \arg\min_{G} \sum_{k=1}^{K} \min_{\theta_k \in \mathbb{R}^p} \sum_{a \in G_k} \|X_a - \theta_k\|_2^2$$

Alternate Minimization between estimation of the **centroids** and of the **partition**

Two steps :

1. Compute the centroids
2. Update the partition



Lloyd's algorithm Lloyd('82)

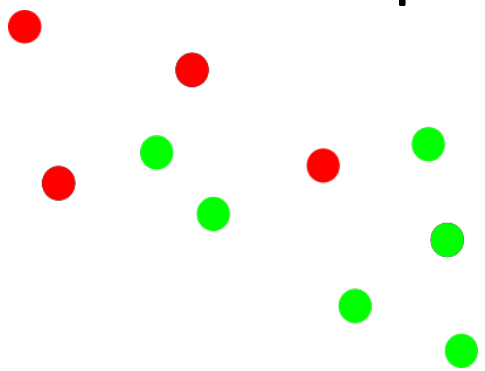
K-means as a least-square problem :

$$\hat{G} \in \arg\min_{G} \sum_{k=1}^{K} \min_{\theta_k \in \mathbb{R}^p} \sum_{a \in G_k} \|X_a - \theta_k\|_2^2$$

Alternate Minimization between estimation of the **centroids** and of the **partition**

Two steps :

1. Compute the centroids
2. Update the partition



Lloyd's algorithm Lloyd('82)

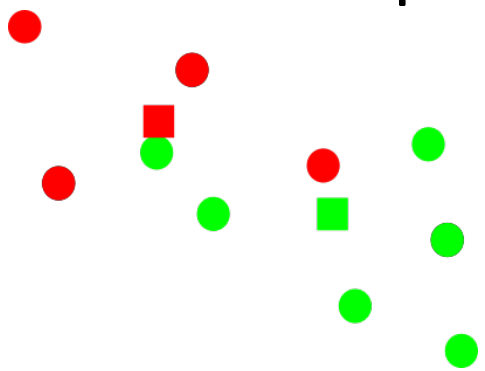
K-means as a least-square problem :

$$\hat{G} \in \arg\min_{G} \sum_{k=1}^{K} \min_{\theta_k \in \mathbb{R}^p} \sum_{a \in G_k} \|X_a - \theta_k\|_2^2$$

Alternate Minimization between estimation of the **centroids** and of the **partition**

Two steps :

1. Compute the centroids
2. Update the partition



Lloyd's algorithm Lloyd('82)

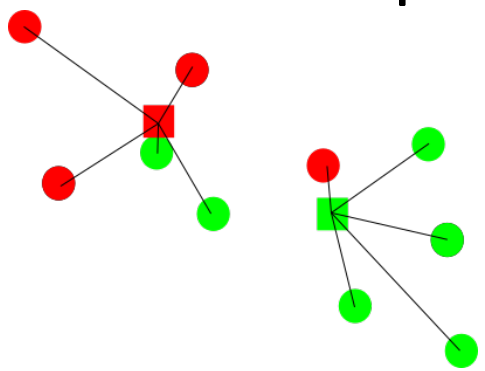
K-means as a least-square problem :

$$\hat{G} \in \arg\min_{G} \sum_{k=1}^{K} \min_{\theta_k \in \mathbb{R}^p} \sum_{a \in G_k} \|X_a - \theta_k\|_2^2$$

Alternate Minimization between estimation of the **centroids** and of the **partition**

Two steps :

1. Compute the centroids
2. Update the partition



Lloyd's algorithm Lloyd('82)

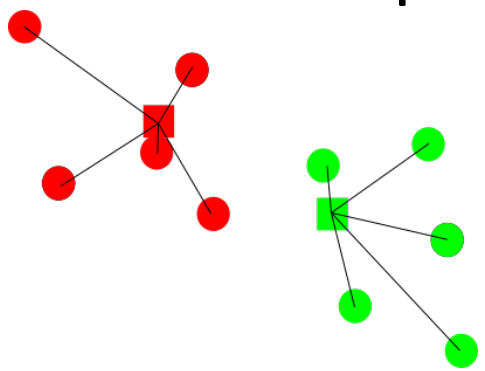
K-means as a least-square problem :

$$\hat{G} \in \arg\min_{G} \sum_{k=1}^{K} \min_{\theta_k \in \mathbb{R}^p} \sum_{a \in G_k} \|X_a - \theta_k\|_2^2$$

Alternate Minimization between estimation of the **centroids** and of the **partition**

Two steps :

1. Compute the centroids
2. Update the partition



Lloyd's algorithm Lloyd('82)

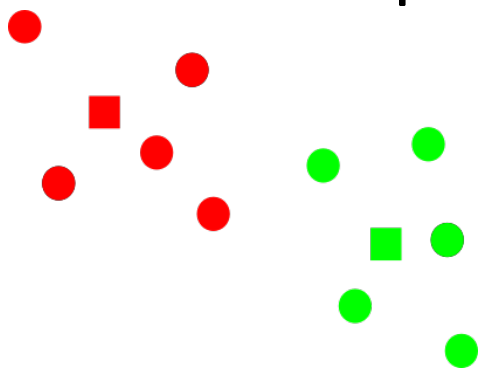
K-means as a least-square problem :

$$\hat{G} \in \arg\min_{G} \sum_{k=1}^{K} \min_{\theta_k \in \mathbb{R}^p} \sum_{a \in G_k} \|X_a - \theta_k\|_2^2$$

Alternate Minimization between estimation of the **centroids** and of the **partition**

Two steps :

1. Compute the centroids
2. Update the partition



Lloyd's algorithm Lloyd('82)

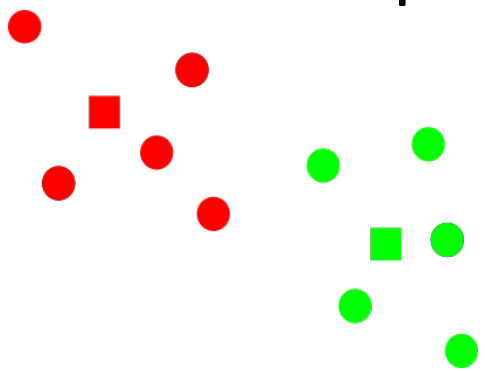
K-means as a least-square problem :

$$\hat{G} \in \arg\min_{G} \sum_{k=1}^{K} \min_{\theta_k \in \mathbb{R}^p} \sum_{a \in G_k} \|X_a - \theta_k\|_2^2$$

Alternate Minimization between estimation of the **centroids** and of the **partition**

Two steps :

1. Compute the centroids
2. Update the partition



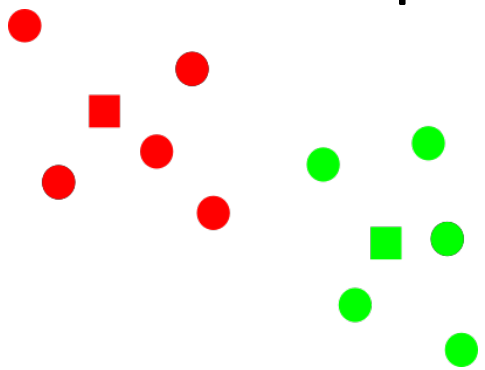
Caveats There can be many local optima. (depends on the initialization)

Lloyd's algorithm Lloyd('82)

K-means as a least-square problem :

$$\hat{G} \in \arg\min_{G} \sum_{k=1}^{K} \min_{\theta_k \in \mathbb{R}^p} \sum_{a \in G_k} \|X_a - \theta_k\|_2^2$$

Alternate Minimization between estimation of the **centroids** and of the **partition**



Two steps :

1. Compute the centroids
2. Update the partition

Caveats There can be many local optima. (depends on the initialization)

Many variants depend on the initialization :

Proposition (Arthur and Vassilvitskki('07))

K-means++ achieves a $\log(K)$ approximation of K-means criterion (in worst case).

Outline

Matrix Characterization of the Partnership matrix.

The $n \times n$ **Partnership** matrix

$$\mathbf{B}_{ab} = \begin{cases} \frac{1}{|G_k|} & \text{if } a \text{ and } b \text{ belong to the same } G_k \\ 0 & \text{else} \end{cases}$$

Lemma (Peng & Wei(07))

The K-means minimizer $\widehat{G}$ satisfies

$$\widehat{\mathbf{B}} \in \arg\min_{\mathbf{B} \in \mathcal{D}} \langle -\mathbf{X}\mathbf{X}^T, \mathbf{B} \rangle \ ,$$

$$\mathcal{D} := \left\{ \mathbf{B} \in \mathbb{R}^{n \times n} : \begin{array}{l} \bullet\ \mathbf{B} \succcurlyeq 0 \\ \bullet\ \sum_a \mathbf{B}_{ab} = 1, \forall b \\ \bullet\ \mathbf{B}_{ab} \geqslant 0,\ \forall a, b \\ \bullet\ \mathrm{Tr}(\mathbf{B}) = K \\ \bullet\ \mathbf{B}^2 = \mathbf{B} \end{array} \right\}$$

Proof : $\mathbf{B}$ is a bistochastic matrix with K eigenvalues equal to 1.

Matrix Characterization of the Partnership matrix.

The $n \times n$ **Partnership** matrix

$$\mathbf{B}_{ab} = \begin{cases} \frac{1}{|G_k|} & \text{if } a \text{ and } b \text{ belong to the same } G_k \\ 0 & \text{else} \end{cases}$$

Lemma (Peng & Wei(07))

The K-means minimizer $\widehat{G}$ satisfies

$$\widehat{\mathbf{B}} \in \arg\min_{\mathbf{B} \in \mathcal{D}} \langle -\mathbf{X}\mathbf{X}^T, \mathbf{B} \rangle \ ,$$

$$\mathcal{D} := \left\{ \mathbf{B} \in \mathbb{R}^{n \times n} : \begin{array}{l} \bullet\ \mathbf{B} \succcurlyeq 0 \\ \bullet\ \sum_a \mathbf{B}_{ab} = 1, \forall b \\ \bullet\ \mathbf{B}_{ab} \geqslant 0,\ \forall a, b \\ \bullet\ \mathrm{Tr}(\mathbf{B}) = K \\ \bullet\ \mathbf{B}^2 = \mathbf{B} \end{array} \right\}$$

Proof : $\mathbf{B}$ is a bistochastic matrix with K eigenvalues equal to 1.
Perron-Frobenius theorem $\rightsquigarrow$ support of $\mathbf{B}=$ adjacency matrix of a graph with K cc each block has rank 1 and is bistochastic.

Relaxed K-means

Idea : drop the $\mathbf{B}^2 = \mathbf{B}$ condition.

1. Estimate $\mathbf{B}^*$ using the semi-definite program (SDP)

$$\widehat{\mathbf{B}} = \underset{\mathbf{B} \in \mathcal{C}}{\arg\min} \langle -\mathbf{X}\mathbf{X}^T, \mathbf{B} \rangle$$

where

$$\mathcal{C} := \left\{ \mathbf{B} \in \mathbb{R}^{n \times n} : \begin{array}{l} \bullet\ \mathbf{B} \succcurlyeq 0 \\ \bullet\ \sum_a \mathbf{B}_{ab} = 1,\ \forall b \\ \bullet\ \mathbf{B}_{ab} \geqslant 0,\ \forall a, b \\ \bullet\ \mathrm{Tr}(\mathbf{B}) = K \end{array} \right\}$$

2. (Compute $\widehat{G}$ by applying any clustering algorithm on $\widehat{\mathbf{B}}$)

Remark :

- Convex optimization but many constraints :
  https://www.ams.jhu.edu/villar/research/ ($n \approx$ a few hundreds)
  Iguchi et al.('15), Mixon et al.('17)
- No information of the group sizes is needed.

Relaxed K-means

Idea : drop the $\mathbf{B}^2 = \mathbf{B}$ condition.

1. Estimate $\mathbf{B}^*$ using the semi-definite program (SDP)

$$\widehat{\mathbf{B}} = \underset{\mathbf{B} \in \mathcal{C}}{\arg\min} \langle -\mathbf{X}\mathbf{X}^T, \mathbf{B} \rangle$$

where

$$\mathcal{C} := \left\{ \mathbf{B} \in \mathbb{R}^{n \times n} : \begin{array}{l} \bullet\ \mathbf{B} \succcurlyeq 0 \\ \bullet\ \sum_a \mathbf{B}_{ab} = 1,\ \forall b \\ \bullet\ \mathbf{B}_{ab} \geqslant 0,\ \forall a, b \\ \bullet\ \mathrm{Tr}(\mathbf{B}) = K \end{array} \right\}$$

2. (Compute $\widehat{G}$ by applying any clustering algorithm on $\widehat{\mathbf{B}}$)

Remark :

- Convex optimization but many constraints :
  https://www.ams.jhu.edu/villar/research/ ($n \approx$ a few hundreds)
  Iguchi et al.('15), Mixon et al.('17)
- No information of the group sizes is needed.

Differs from the SDP relaxation of Max-Cut Problem Goeman and Williamson('95) (see e.g. Hajek et al.('16)) where the size of the communities has to be known...

Outline

Heuristic

$$\mathbf{E}[\mathbf{X}\mathbf{X}^T] = \mathbf{A}\mathbf{\Theta}\mathbf{\Theta}^T\mathbf{A}^T \left(+\mathbb{E}[\mathbf{E}\mathbf{E}^T]\right)$$

Lemma (e.g. Lei and Rinaldo('15))

Assume that $\mathbf{\Theta}\mathbf{\Theta}^T$ *has full rank.*

Let $\mathbf{A}\mathbf{\Theta}\mathbf{\Theta}^T\mathbf{A}^T = \sum_{k=1}^{K} d_k u_k u_k^T$ *and set* $\mathbf{U} = [u_1, \ldots, u_k] \in \mathbb{R}^{n\times K}$. *Then, there exist* $Z_1, \ldots, Z_K \in \mathbb{R}^K$, *such that*

$$U_{i:} = Z_k \quad \textit{for all } i \in G_k, \quad \textit{and} \quad \|Z_k - Z_\ell\|^2 = \frac{1}{|G_k|} + \frac{1}{|G_\ell|}.$$

$$\mathbf{U} = \begin{pmatrix} Z_{k(1)} \\ Z_{k(2)} \\ \cdots \\ Z_{k(n)} \end{pmatrix}$$

Spectral Clustering as relaxed K-means

Spectral Clustering

1. Compute the matrix $\widehat{\mathbf{U}}$ made of the K-leading eigenvectors of $\mathbf{X}^T\mathbf{X}$
2. Estimate $\widehat{G}$ by distance clustering on the rows of $\widehat{\mathbf{U}}$.

(e.g. Apply an approximate K-means algorithm to the rows of the matrix $\widehat{\mathbf{U}}$)

Spectral Clustering as relaxed K-means

Spectral Clustering

1. Compute the matrix $\widehat{\mathbf{U}}$ made of the K-leading eigenvectors of $\mathbf{X}^T\mathbf{X}$
2. Estimate $\widehat{G}$ by distance clustering on the rows of $\widehat{\mathbf{U}}$.

(e.g. Apply an approximate K-means algorithm to the rows of the matrix $\widehat{\mathbf{U}}$)

Lemma (Peng & Wei(07))

Spectral Clustering is equivalent to

1. *Estimate $\mathbf{B}^*$ using the semi-definite program (SDP)*

$$\overline{\mathbf{B}} = \underset{\mathbf{B}\in\overline{\mathcal{C}}}{\arg\min}\langle -\mathbf{X}\mathbf{X}^T, \mathbf{B}\rangle$$

$$\overline{\mathcal{C}} := \left\{ \mathbf{B}\in\mathbb{R}^{n\times n} : \begin{array}{l} \bullet\ 1 \succcurlyeq \mathbf{B} \succcurlyeq 0 \\ \bullet\ \mathrm{Tr}(\mathbf{B}) = K \end{array} \right\}$$

2. *(rounding) Compute $\widehat{G}$ by distance clustering on the rows of $\overline{\mathbf{B}}$*

$\Longrightarrow$ it amounts to dropping the constraints $\mathbf{B}1 = 1,\ \mathbf{B}_{ab} \geqslant 0$ in the former relaxation

Spectral Clustering as relaxed K-means

Spectral Clustering

1. Compute the matrix $\widehat{\mathbf{U}}$ made of the K-leading eigenvectors of $\mathbf{X}^T\mathbf{X}$
2. Estimate $\widehat{G}$ by distance clustering on the rows of $\widehat{\mathbf{U}}$.

(e.g. Apply an approximate K-means algorithm to the rows of the matrix $\widehat{\mathbf{U}}$)

Lemma (Peng & Wei(07))

Spectral Clustering is equivalent to

1. *Estimate $\mathbf{B}^*$ using the semi-definite program (SDP)*

$$\overline{\mathbf{B}} = \underset{\mathbf{B}\in\overline{\mathcal{C}}}{\arg\min}\langle -\mathbf{X}\mathbf{X}^T, \mathbf{B}\rangle$$

$$\overline{\mathcal{C}} := \left\{ \mathbf{B}\in\mathbb{R}^{n\times n} : \begin{array}{l} \bullet\ 1 \succcurlyeq \mathbf{B} \succcurlyeq 0 \\ \bullet\ \mathrm{Tr}(\mathbf{B}) = K \end{array} \right\}$$

2. *(rounding) Compute $\widehat{G}$ by distance clustering on the rows of $\overline{\mathbf{B}}$*

$\Longrightarrow$ it amounts to dropping the constraints $\mathbf{B}1 = 1,\ \mathbf{B}_{ab} \geqslant 0$ in the former relaxation

Proof : 1) $\overline{\mathbf{B}} = \widehat{\mathbf{U}}\widehat{\mathbf{U}}^T$

2) $(\widehat{\mathbf{U}}\widehat{\mathbf{U}}^T)_{a:}$ is some orthogonal transformation of $\widehat{\mathbf{U}}_{a:}$.

Outline

Partial recovery bounds

Proportion of misclustered points

$$err(\widehat{G}, G^*) = \min_{\pi \in \mathcal{S}_K} \frac{1}{2n} \sum_{k=1}^{K} \left| G_k^* \triangle \widehat{G}_{\pi(k)} \right|$$

Our goal

Prove that with high-probability, when s^2 is large

$$\text{prop. misclustered} = err(\widehat{G}, G^*) \leq e^{-cs^2}$$

where s^2 is an appropriate SNR.

Other related goals :

- **partial recovery** : Find the minimal s^2 such that $err(\widehat{G}, G^*)$ is smaller than random guess whp.
- **Almost full recovery** : Find the minimal s^2 such that $err(\widehat{G}, G^*) \to 0$
- **Perfect recovery** : Find the minimal s^2 such that $err(\widehat{G}, G^*) = 0$ whp.

Clustering Gaussian mixtures

Mixture of Gaussian variables

$X_1, \dots, X_n \in \mathbb{R}^p$ are independent with

- $X_a \sim \mathcal{N}(\theta_k, \mathbf{\Sigma})$ $a \in G_k^*$

many results discussed later readily extend to subGaussian random variables and to different covariances

$$\text{Set } \Delta^2 = \min_{j \neq k} \|\theta_k - \theta_j\|^2, \quad \sigma^2 = \|\mathbf{\Sigma}\|_{op} \quad \text{and} \quad R_{\mathbf{\Sigma}} = \frac{\|\mathbf{\Sigma}\|_F^2}{\|\mathbf{\Sigma}\|_{op}^2},$$

Specific case of Isovolumetric spherical Gaussians

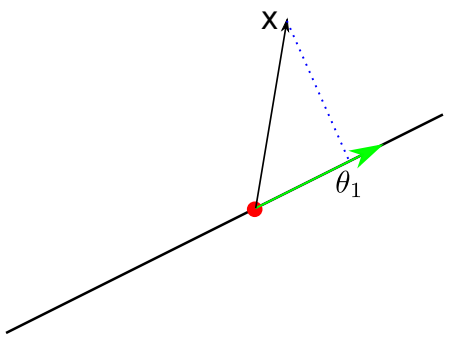
$\mathbf{\Sigma} = \sigma^2 \mathbf{I}_p$.

Outline

What should we expect ? What is the correct SNR ?

Toy example : $K = 2$, $|G_1^*| = |G_2^*| = n/2$, $\mathbf{\Sigma} = \sigma^2 \mathbf{I}_p$, $\theta_2 = -\theta_1$.

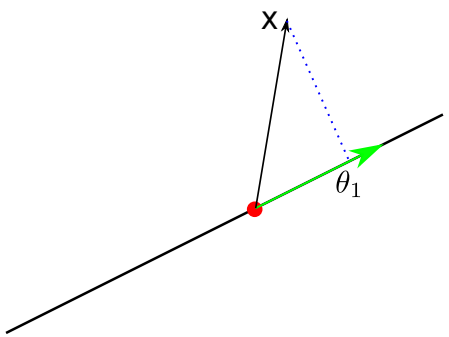
Simpler Problem 1 (known parameters) : θ_1 is known.



What should we expect ? What is the correct SNR ?

Toy example : $K = 2$, $|G_1^*| = |G_2^*| = n/2$, $\mathbf{\Sigma} = \sigma^2 \mathbf{I}_p$, $\theta_2 = -\theta_1$.

Simpler Problem 1 (known parameters) : θ_1 is known.



Bayes Classifier achieves :

$$\mathbb{E}[err(\widehat{G}, G^*)] = \mathbb{P}[\mathcal{N}(0, \sigma^2) > \|\theta_1\|] \leq \exp\left[-\frac{\Delta^2}{8\sigma^2}\right]$$

Benchmark 2 : Supervised case

Parameter : θ_1 is sampled uniformly on the sphere of radius $\Delta/2$.
Supervised observations : $\mathcal{L} = (X_a, Z_a)_{a=1,\dots,n}$ where $Z_a \in \{1, 2\}$ is the label of X_a and is sampled uniformly in $\{1, 2\}$

Objective : classify a new observation X.

Benchmark 2 : Supervised case

Parameter : θ_1 is sampled uniformly on the sphere of radius $\Delta/2$.
Supervised observations : $\mathcal{L} = (X_a, Z_a)_{a=1,\dots,n}$ where $Z_a \in \{1, 2\}$ is the label of X_a and is sampled uniformly in $\{1, 2\}$

Objective : classify a new observation X.

Optimal Classifier is achieved by LDA

$$\widehat{h}(x) = \begin{cases} 2 & \text{if } \mathbb{P}[Z = 2|X = x, \mathcal{L}] > \mathbb{P}[Z = 1|X = x, \mathcal{L}]\rangle \\ 1 & \text{if } \mathbb{P}[Z = 2|X = x, \mathcal{L}] \leq \mathbb{P}[Z = 1|X = x, \mathcal{L}]\rangle \end{cases}$$

$$\widehat{h}(x) = \frac{3}{2} + \frac{1}{2}\,\mathrm{sign}\left(\langle \frac{1}{n}\sum_{a=1}^{n}(2Z_a - 3)X_a, x\rangle\right) \ .$$

Benchmark 2 : Supervised case

Parameter : θ_1 is sampled uniformly on the sphere of radius $\Delta/2$.
Supervised observations : $\mathcal{L} = (X_a, Z_a)_{a=1,\dots,n}$ where $Z_a \in \{1,2\}$ is the label of X_a and is sampled uniformly in $\{1,2\}$

Objective : classify a new observation X.

Optimal Classifier is achieved by LDA

$$\widehat{h}(x) = \begin{cases} 2 & \text{if } \mathbb{P}[Z=2|X=x,\mathcal{L}] > \mathbb{P}[Z=1|X=x,\mathcal{L}]\rangle \\ 1 & \text{if } \mathbb{P}[Z=2|X=x,\mathcal{L}] \leq \mathbb{P}[Z=1|X=x,\mathcal{L}]\rangle \end{cases}$$

$$\widehat{h}(x) = \frac{3}{2} + \frac{1}{2}\,\mathrm{sign}\left(\langle \frac{1}{n}\sum_{a=1}^{n}(2Z_a - 3)X_a, x\rangle\right) \ .$$

$$\begin{aligned} \mathbb{P}[\widehat{h}(X) \neq Z] &= \mathbb{E}_\theta\,\mathbb{P}\left[\langle \theta_1 + \frac{\sigma}{\sqrt{n}}\epsilon, \theta_1 + \sigma\epsilon'\rangle < 0 | \theta\right] \\ &= \mathbb{P}\left[\frac{\Delta^2}{4\sigma^2} < \frac{\Delta}{2\sigma}(\frac{\epsilon_1}{\sqrt{n}} + \epsilon_1') - \frac{1}{\sqrt{n}}\langle\epsilon,\epsilon'\rangle\right] \\ &\sim_{\log} \begin{cases} \exp\left(-\frac{\Delta^2}{8\sigma^2}\right) & \text{if } \frac{\Delta^2}{\sigma^2} \gg [1 \vee \frac{p}{n}] \\ \exp\left(-\frac{n\Delta^4}{32p\sigma^2}\right) & \text{if } [1 \vee \sqrt{\frac{p}{n}}] \ll \frac{\Delta^2}{\sigma^2} \ll \frac{p}{n} \end{cases} \end{aligned}$$

Proper minimax lower bound

Toy Model : $K = 2$, $|G_1^*| = |G_2^*| = n/2$, $\mathbf{\Sigma} = \sigma^2 \mathbf{I}_p$

Theorem (Ndaoud('18))

$$\inf_{\widehat{G}} \sup_{\theta_1, \|\theta\| \geq \Delta/2; G^* \text{ balanced}} \mathbb{E}_{\theta, G^*} \left[\mathrm{err}(G^*; \widehat{G}) \right] \gtrsim e^{-c\left(\frac{\Delta^2}{\sigma^2} \wedge \frac{n\Delta^4}{p\sigma^4} \right)}$$

Signal To Noise Ratio

For more communities, define the **SNR**

$$s^2 = \frac{\Delta^2}{\sigma^2} \wedge \frac{m\Delta^4}{R_{\mathbf{\Sigma}}\sigma^4},$$

where

- m denotes the size of the smallest cluster.
- $R_{\mathbf{\Sigma}} = \frac{\|\mathbf{\Sigma}\|_F^2}{\|\mathbf{\Sigma}\|_{op}^2}$ is the effective rank of $\mathbf{\Sigma}$

Outline

Recall the relaxed K-means SDP

relaxed K-means

Solve the SDP

$$\widehat{B} \in \operatorname*{argmin}_{\mathbf{B} \in \mathcal{C}} \langle -\mathbf{X}^T\mathbf{X}, \mathbf{B} \rangle \ ,$$

with

$$\mathcal{C} := \left\{ \mathbf{B} \in \mathbb{R}^{n \times n} : \begin{array}{l} \bullet\ \mathbf{B} \succcurlyeq 0 \\ \bullet\ \sum_a \mathbf{B}_{ab} = 1, \forall b \\ \bullet\ \mathbf{B}_{ab} \geqslant 0,\ \forall a, b \\ \bullet\ \mathrm{Tr}(\mathbf{B}) = K \end{array} \right\}$$

Step 2 : Apply approximate K-medoid method Charikar et al.('02).

$$|\widehat{\mathbf{A}}\widehat{\mathbf{M}} - \widehat{\mathbf{B}}|_1 \leq \rho \min_{A,\ \mathrm{Rows}(M) \subset \mathrm{Rows}(\widehat{\mathbf{B}})} |\mathbf{AM} - \widehat{\mathbf{B}}|_1$$

Partial recovery bounds for relaxed K-means

$$s^2 = \frac{\Delta^2}{\sigma^2} \wedge \frac{m\Delta^4}{R_{\mathbf{\Sigma}}\sigma^4},$$

Theorem (Giraud and V. ('18))

If $s^2 \gtrsim n/m$, then $\mathbb{P}\left[err(\widehat{G}, G^*) > e^{-cs^2}\right] \lesssim \frac{1}{n^2}$.

Partial recovery bounds for relaxed K-means

$$s^2 = \frac{\Delta^2}{\sigma^2} \wedge \frac{m\Delta^4}{R_{\mathbf{\Sigma}}\sigma^4},$$

Theorem (Giraud and V. ('18))

If $s^2 \gtrsim n/m$, then $\mathbb{P}\left[err(\widehat{G}, G^*) > e^{-cs^2}\right] \lesssim \frac{1}{n^2}$.

$s^2 \gtrsim n/m = K$ is equivalent to $\Delta^2 \gtrsim \sigma^2 \frac{n}{m}\left(1 \vee \sqrt{\frac{R_{\mathbf{\Sigma}}}{n}}\right) = \sigma^2 K\left(1 \vee \sqrt{\frac{R_{\mathbf{\Sigma}}}{n}}\right)$.

Partial recovery bounds for relaxed K-means

$$s^2 = \frac{\Delta^2}{\sigma^2} \wedge \frac{m\Delta^4}{R_{\mathbf{\Sigma}}\sigma^4},$$

Theorem (Giraud and V. ('18))

If $s^2 \gtrsim n/m$, then $\mathbb{P}\left[err(\widehat{G}, G^*) > e^{-cs^2}\right] \lesssim \frac{1}{n^2}$.

$s^2 \gtrsim n/m = K$ is equivalent to $\Delta^2 \gtrsim \sigma^2 \frac{n}{m}\left(1 \vee \sqrt{\frac{R_{\mathbf{\Sigma}}}{n}}\right) = \sigma^2 K\left(1 \vee \sqrt{\frac{R_{\mathbf{\Sigma}}}{n}}\right)$.

Remarks :

1. Optimal convergence rate provided the SNR is large enough.
2. No restriction on the dimension

Partial recovery bounds for relaxed K-means

$$s^2 = \frac{\Delta^2}{\sigma^2} \wedge \frac{m\Delta^4}{R_{\mathbf{\Sigma}}\sigma^4},$$

Theorem (Giraud and V. ('18))

If $s^2 \gtrsim n/m$, then $\mathbb{P}\left[err(\widehat{G}, G^*) > e^{-cs^2}\right] \lesssim \frac{1}{n^2}$.

$s^2 \gtrsim n/m = K$ is equivalent to $\Delta^2 \gtrsim \sigma^2 \frac{n}{m}\left(1 \vee \sqrt{\frac{R_{\mathbf{\Sigma}}}{n}}\right) = \sigma^2 K\left(1 \vee \sqrt{\frac{R_{\mathbf{\Sigma}}}{n}}\right)$.

Remarks :

1. Optimal convergence rate provided the SNR is large enough.
2. No restriction on the dimension
3. Ideas come from Fei and Chen('17) for SBMs. See also Mixon et al.('16), Fei and Chen('18).

Partial recovery bounds for relaxed K-means

$$s^2 = \frac{\Delta^2}{\sigma^2} \wedge \frac{m\Delta^4}{R_{\mathbf{\Sigma}}\sigma^4},$$

Theorem (Giraud and V. ('18))

If $s^2 \gtrsim n/m$, then $\mathbb{P}\left[err(\widehat{G}, G^*) > e^{-cs^2}\right] \lesssim \frac{1}{n^2}$.

$s^2 \gtrsim n/m = K$ is equivalent to $\Delta^2 \gtrsim \sigma^2 \frac{n}{m}\left(1 \vee \sqrt{\frac{R_{\mathbf{\Sigma}}}{n}}\right) = \sigma^2 K\left(1 \vee \sqrt{\frac{R_{\mathbf{\Sigma}}}{n}}\right)$.

Remarks :

1. Optimal convergence rate provided the SNR is large enough.
2. No restriction on the dimension
3. Ideas come from Fei and Chen('17) for SBMs. See also Mixon et al.('16), Fei and Chen('18).
4. It does not recover the tight constant inside the exponential. This is possible in specific situations ; see Fei and Chen('19) for SBM.

Partial recovery bounds for relaxed K-means

$$s^2 = \frac{\Delta^2}{\sigma^2} \wedge \frac{m\Delta^4}{R_{\mathbf{\Sigma}}\sigma^4},$$

Theorem (Giraud and V. ('18))

If $s^2 \gtrsim n/m$, then $\mathbb{P}\left[err(\widehat{G}, G^*) > e^{-cs^2}\right] \lesssim \frac{1}{n^2}$.

$s^2 \gtrsim n/m = K$ is equivalent to $\Delta^2 \gtrsim \sigma^2 \frac{n}{m}\left(1 \vee \sqrt{\frac{R_{\mathbf{\Sigma}}}{n}}\right) = \sigma^2 K\left(1 \vee \sqrt{\frac{R_{\mathbf{\Sigma}}}{n}}\right)$.

Remarks :

1. Optimal convergence rate provided the SNR is large enough.
2. No restriction on the dimension
3. Ideas come from Fei and Chen('17) for SBMs. See also Mixon et al.('16), Fei and Chen('18).
4. It does not recover the tight constant inside the exponential. This is possible in specific situations ; see Fei and Chen('19) for SBM.
5. perfect recovery for $s^2 \gtrsim \log(n) \vee (n/m) \quad = \log(n) \vee K$

Large number K of clusters

Simplification : balanced partition + Isovolumetric spherical Gaussians.

- SNR condition $s^2 \geq K$ is needed for Peng and Wei's SDP or for spectral clustering.
- From Regev and Vijayaraghavan('17), the condition $s^2 \gtrsim \log(K)$ is needed for partial recovery to be information theoretically possible.

Large number K of clusters

Simplification : balanced partition + Isovolumetric spherical Gaussians.

- SNR condition $s^2 \geq K$ is needed for Peng and Wei's SDP or for spectral clustering.
- From Regev and Vijayaraghavan('17), the condition $s^2 \gtrsim \log(K)$ is needed for partial recovery to be information theoretically possible.

- In low-dimensional settings (e.g. $n \widetilde{\gg} p^3 K^2$), Vempala and Wang('04) achieve exact recovery if

$$s^2 \gtrsim \sqrt{K \log(n)} + \log(n) \ .$$

Large number K of clusters

Simplification : balanced partition + Isovolumetric spherical Gaussians.

- SNR condition $s^2 \geq K$ is needed for Peng and Wei's SDP or for spectral clustering.
- From Regev and Vijayaraghavan('17), the condition $s^2 \gtrsim \log(K)$ is needed for partial recovery to be information theoretically possible.
- In low-dimensional settings (e.g. $n \widetilde{\gg} p^3 K^2$), Vempala and Wang('04) achieve exact recovery if

$$s^2 \gtrsim \sqrt{K \log(n)} + \log(n) \ .$$

- In high-dimension $(p \geq n)$, it is **conjectured** that no polynomial-time estimator can beat the condition $s^2 \gtrsim K$ – Banks et al.('18)

General covariance matrices

- With a proper correction (see e.g. Bunea et al.('16)), general covariances $\mathbf{\Sigma}_1, \ldots, \mathbf{\Sigma}_K$ can be handled by K-means and its relaxations.

General covariance matrices

- With a proper correction (see e.g. Bunea et al.('16)), general covariances $\mathbf{\Sigma}_1, \ldots, \mathbf{\Sigma}_K$ can be handled by K-means and its relaxations.

- ... However, those methods do not build upon general covariances.

  Optimal convergence rate with respect to the **Mahalanobis** distance $(\mu_k - \mu_l)^T \mathbf{\Sigma}^{-1} (\mu_k - \mu_l)$

  $\rightsquigarrow$ For unknown $\mathbf{\Sigma}$, this is an active and challenging research direction Dabis et al.('21)

Summary

First Message

Various convex relaxations of K-means seem to work well...

Summary

First Message

Various convex relaxations of K-means seem to work well...

Second Message

... for a variety of models... (SBM, GMM,...)

Summary

First Message

Various convex relaxations of K-means seem to work well...

Second Message

... for a variety of models... (SBM, GMM,...)

Third Message

Large K asymptotic is still not completely understood.

Summary

First Message

Various convex relaxations of K-means seem to work well...

Second Message

... for a variety of models... (SBM, GMM,...)

Third Message

Large K asymptotic is still not completely understood.

Danke für Ihre Aufmerksamkeit !

References I


E. Abbe.
Community detection and stochastic block models : recent developments.
ArXiv e-prints, March 2017.


Pranjal Awasthi, Moses Charikar, Ravishankar Krishnaswamy, and Ali Kemal Sinop.
The Hardness of Approximation of Euclidean k-means.
arXiv preprint arXiv :1502.03316, 2015.


Emmanuel Abbe, Jianqing Fan, Kaizheng Wang, and Yiqiao Zhong.
Entrywise eigenvector analysis of random matrices with low expected rank.
arXiv preprint arXiv :1709.09565, 2017.


Matthew Brennan and Guy Bresler.
Average-case lower bounds for learning sparse mixtures, robust estimation and semirandom adversaries.
arXiv preprint arXiv :1908.06130, 2019.


Matthew Brennan and Guy Bresler.
Optimal Average-Case Reductions to Sparse PCA : From Weak Assumptions to Strong Hardness.
arXiv preprint arXiv :1902.07380, 2019.

References II


Matthew Brennan, Guy Bresler, and Wasim Huleihel.
Reducibility and computational lower bounds for problems with planted sparse structure.
arXiv preprint arXiv :1806.07508, 2018.


Matthew Brennan, Guy Bresler, and Wasim Huleihel.
Universality of computational lower bounds for submatrix detection.
arXiv preprint arXiv :1902.06916, 2019.


Florentina Bunea, Christophe Giraud, Martin Royer, and Nicolas Verzelen.
PECOK : a convex optimization approach to variable clustering.
arXiv preprint arXiv :1606.05100, 2016.


Jess Banks, Cristopher Moore, Roman Vershynin, Nicolas Verzelen, and Jiaming Xu.
Information-theoretic bounds and phase transitions in clustering, sparse PCA, and submatrix localization.
IEEE Transactions on Information Theory, 64(7) :4872–4894, 2018.


Moses Charikar, Sudipto Guha, Éva Tardos, and David B. Shmoys.
A Constant-Factor Approximation Algorithm for the k-Median Problem.
Journal of Computer and System Sciences, 65(1) :129–149, 2002.

References III

Peter Chin, Anup Rao, and Van Vu.
Stochastic Block Model and Community Detection in Sparse Graphs : A spectral algorithm with optimal rate of recovery.
In Peter Grünwald, Elad Hazan, and Satyen Kale, editors, Proceedings of The 28th Conference on Learning Theory, volume 40 of Proceedings of Machine Learning Research, pages 391–423, Paris, France, 03–06 Jul 2015. PMLR.

Yudong Chen and Jiaming Xu.
Statistical-computational tradeoffs in planted problems and submatrix localization with a growing number of clusters and submatrices.
Journal of Machine Learning Research, 17(27) :1–57, 2016.

Ilias Diakonikolas, Daniel M. Kane, and Alistair Stewart.
List-Decodable Robust Mean Estimation and Learning Mixtures of Spherical Gaussians.
CoRR, abs/1711.07211, 2017.

Yingjie Fei and Yudong Chen.
Hidden integrality of sdp relaxation for sub-gaussian mixture models.
arXiv preprint arXiv :1803.06510, 2018.

References IV


Yingjie Fei and Yudong Chen.
Achieving the Bayes Error Rate in Synchronization and Block Models by SDP, Robustly.
arXiv preprint arXiv :1904.09635, 2019.


Chao Gao, Zongming Ma, Anderson Y. Zhang, and Harrison H. Zhou.
Achieving Optimal Misclassification Proportion in Stochastic Block Models.
J. Mach. Learn. Res., 18(1) :1980–2024, January 2017.


Christophe Giraud and Nicolas Verzelen.
Partial recovery bounds for clustering with the relaxed K means.
arXiv preprint arXiv :1807.07547, 2018.


Michel X Goemans and David P Williamson.
Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming.
Journal of the ACM (JACM), 42(6) :1115–1145, 1995.


Samuel B. Hopkins and Jerry Li.
Mixture Models, Robustness, and Sum of Squares Proofs.
CoRR, abs/1711.07454, 2017.


Paul W Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt.
Stochastic blockmodels : First steps.
Social networks, 5(2) :109–137, 1983.

References V


B. Hajek, Y. Wu, and J. Xu.
Semidefinite Programs for Exact Recovery of a Hidden Community.
ArXiv e-prints, February 2016.


Takayuki Iguchi, Dustin G Mixon, Jesse Peterson, and Soledad Villar.
On the tightness of an SDP relaxation of k-means.
arXiv preprint arXiv :1505.04778, 2015.


Tapas Kanungo, David M Mount, Nathan S Netanyahu, Christine D Piatko, Ruth Silverman, and Angela Y Wu.
A local search approximation algorithm for k-means clustering.
Computational Geometry, 28(2-3) :89–112, 2004.


Pravesh K. Kothari and Jacob Steinhardt.
Better Agnostic Clustering Via Relaxed Tensor Norms.
CoRR, abs/1711.07465, 2017.


Stuart Lloyd.
Least squares quantization in PCM.
IEEE transactions on information theory, 28(2) :129–137, 1982.


Jing Lei and Alessandro Rinaldo.
Consistency of spectral clustering in stochastic block models.
Ann. Statist., 43(1) :215–237, 2015.

References VI


Y. Lu and H. H. Zhou.
Statistical and Computational Guarantees of Lloyd's Algorithm and its Variants.
ArXiv e-prints, December 2016.


Matthias Löffler, Anderson Y Zhang, and Harrison H Zhou.
Optimality of spectral clustering for gaussian mixture model.
arXiv preprint arXiv :1911.00538, 2019.


Meena Mahajan, Prajakta Nimbhorkar, and Kasturi Varadarajan.
The planar k-means problem is NP-hard.
In International Workshop on Algorithms and Computation, pages 274–285. Springer, 2009.


Ankur Moitra, William Perry, and Alexander S Wein.
How robust are reconstruction thresholds for community detection ?
In
Proceedings of the forty-eighth annual ACM symposium on Theory of Computing, pages 828–841. ACM, 2016.


Dustin G Mixon, Soledad Villar, and Rachel Ward.
Clustering subgaussian mixtures by semidefinite programming.
Information and Inference : A Journal of the IMA, 6(4) :389–415, 2017.

References VII


Mohamed Ndaoud.
Sharp optimal recovery in the Two Gaussian Mixture Model.
arXiv preprint arXiv :1812.08078, 2018.


Jiming Peng and Yu Wei.
Approximating K-means-type Clustering via Semidefinite Programming.
SIAM J. on Optimization, 18(1) :186–205, February 2007.


M. Royer.
Adaptive Clustering through Semidefinite Programming.
Advances in Neural Information Processing Systems (NIPS), 2017.


Oded Regev and Aravindan Vijayaraghavan.
On learning mixtures of well-separated gaussians.
In
2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS), pages 85–96. IEEE, 2017.


Nathan Srebro, Gregory Shakhnarovich, and Sam Roweis.
An investigation of computational and informational limits in gaussian mixture clustering.
In Proceedings of the 23rd international conference on Machine learning, pages 865–872. ACM, 2006.

References VIII

Yihong Wu and Harrison H Zhou.
Randomly initialized EM algorithm for two-component Gaussian mixture achieves near optimality in O(nî/2) iterations.
arXiv preprint arXiv :1908.10935, 2019.

Anru Zhang, T Tony Cai, and Yihong Wu.
Heteroskedastic PCA : Algorithm, optimality, and applications.
arXiv preprint arXiv :1810.08316, 2018.