K-means: Relaxation and Correction

Nicolas Verzelen

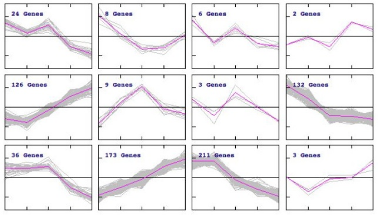
Joint works with C. Giraud, F. Bunea and M. Royer.

HDPA - July 2nd

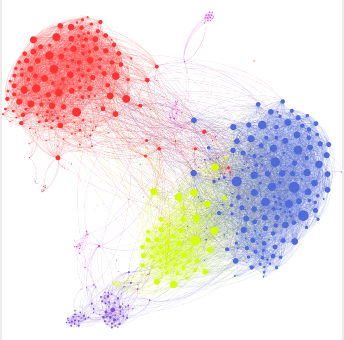
Clustering arises in various contexts

Clustering individuals w.r.t. features

Clustering features



Clustering graphs



Objectives

Topic of the talk

- K-means (relaxed or not) must and can be debiased
- we derive some non-asymptotic partial recovery bounds for a relaxed K-means
- some optimality in terms of exponential exponent

Main message

A corrected convex relaxation of K-means achieves some rate-optimal performances in various settings including (conditional) mixture of sub-Gaussian and (conditional) Stochastic Block Model.

Only tuning Parameter is K

1 Two clustering Models

2 K-means and relaxed K-means

3 Corrected K-means

4 Partial Recovery bounds
- subGaussian Mixtures
- Stochastic Block Models

Mixture of subGaussian variables Pearson('1895)

Partition

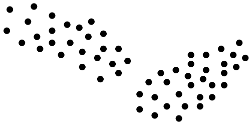
Partition $G^* = \{G_1^*, \dots, G_K^*\}$ of $\{1, \dots, n\}$

Mixture of subGaussian variables (conditional)

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

- $\mathbb{E}[X_a] = \mu_k$ if $a \in G_k^*$
- $\mathbf{\Sigma}_a^{-1/2} X_a$ is SubGauss$(L^2 I_p)$ where $\mathbf{\Sigma}_a = \mathrm{Cov}\,(X_a)$ and $L \geq 1$.

The observations are gathered in $\mathbf{X} = [X_1, \dots, X_n] \in \mathbb{R}^{p \times n}$



Mixture of subGaussian variables Pearson('1895)

Partition

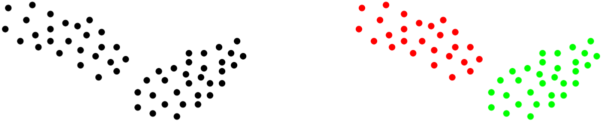
Partition $G^* = \{G_1^*, \ldots, G_K^*\}$ of $\{1, \ldots, n\}$

Mixture of subGaussian variables (conditional)

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

- $\mathbb{E}[X_a] = \mu_k$ if $a \in G_k^*$
- $\mathbf{\Sigma}_a^{-1/2} X_a$ is SubGauss$(L^2 I_p)$ where $\mathbf{\Sigma}_a = \mathrm{Cov}\,(X_a)$ and $L \geq 1$.

The observations are gathered in $\mathbf{X} = [X_1, \ldots, X_n] \in \mathbb{R}^{p \times n}$



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

Stochastic Block Model (SBM)

Holland et al(83), Abbe('17),
Let $\mathbf{X}$ = 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),
Let $\mathbf{X}$ = 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.)

How do we encode partition learning ?

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 ?

How do we encode partition learning ?

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.

K-means criterion

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

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

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

In practice, iterative minimization based on Lloyd's algorithm LLoyd('82) .

K-means criterion

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

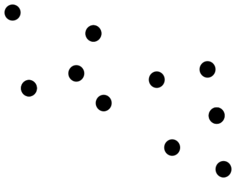
$$\mathsf{Crit}(\mathbf{X}, G) = \sum_{k=1}^{K} \sum_{a \in G_k} \|X_a - \overline{X}_{G_k}\|^2 \ ,$$

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

In practice, iterative minimization based on Lloyd's algorithm LLoyd('82) .

Two steps :

1. Compute the centroids
2. Update the partition



K-means criterion

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

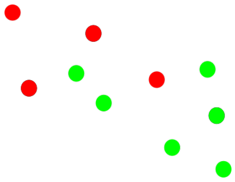
$$\mathsf{Crit}(\mathbf{X}, G) = \sum_{k=1}^{K} \sum_{a \in G_k} \|X_a - \overline{X}_{G_k}\|^2 \ ,$$

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

In practice, iterative minimization based on Lloyd's algorithm LLoyd('82) .

Two steps :

1. Compute the centroids
2. Update the partition



K-means criterion

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

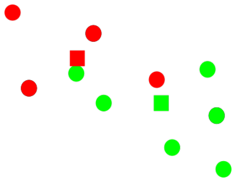
$$\mathsf{Crit}(\mathbf{X}, G) = \sum_{k=1}^{K} \sum_{a \in G_k} \|X_a - \overline{X}_{G_k}\|^2 \ ,$$

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

In practice, iterative minimization based on Lloyd's algorithm LLoyd('82) .

Two steps :

1. Compute the centroids
2. Update the partition



K-means criterion

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

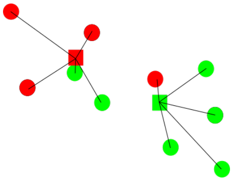
$$\mathsf{Crit}(\mathbf{X}, G) = \sum_{k=1}^{K} \sum_{a \in G_k} \|X_a - \overline{X}_{G_k}\|^2 \ ,$$

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

In practice, iterative minimization based on Lloyd's algorithm LLoyd('82) .

Two steps :

1. Compute the centroids
2. Update the partition



K-means criterion

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

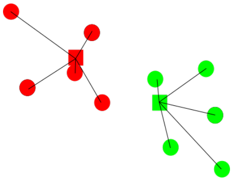
$$\mathsf{Crit}(\mathbf{X}, G) = \sum_{k=1}^{K} \sum_{a \in G_k} \|X_a - \overline{X}_{G_k}\|^2 \ ,$$

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

In practice, iterative minimization based on Lloyd's algorithm LLoyd('82) .

Two steps :

1. Compute the centroids
2. Update the partition



K-means criterion

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

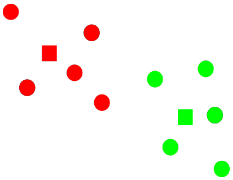
$$\mathsf{Crit}(\mathbf{X}, G) = \sum_{k=1}^{K} \sum_{a \in G_k} \|X_a - \overline{X}_{G_k}\|^2 \ ,$$

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

In practice, iterative minimization based on Lloyd's algorithm LLoyd('82) .

Two steps :

1. Compute the centroids
2. Update the partition



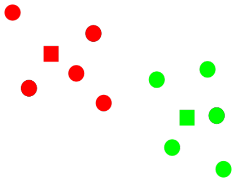
K-means criterion

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

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

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

In practice, iterative minimization based on Lloyd's algorithm LLoyd('82) .



Two steps :

1. Compute the centroids
2. Update the partition

Two caveats :

- There can be many local optima.
- In worst-case solving K-means is NP-hard (Mahajan et al.('09))

Rewriting K-means

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

Rewriting K-means

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

Lemma (Peng & Wei(07))

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

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

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

Proof : Perron-Frobenius

Relaxed K-means

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

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

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

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://cims.nyu.edu/~villar/mnist.html
- No information of the group sizes are needed.

A second relaxation : Spectral Clustering

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 $\widehat{\mathbf{U}}$)

A second relaxation : Spectral Clustering

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)*

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

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

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

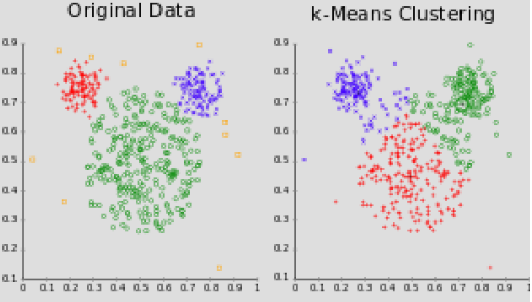
$\Longrightarrow$ it amounts to dropping the constraints $\mathbf{B}1 = 1,\ \mathbf{B}_{ab} \geqslant 0$ in the former relaxation

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

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

Wikipedia knows

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



Quantization rather than clustering
https://en.wikipedia.org/wiki/K-means_clustering

Caveat

A simple model

Assume that the "points" X_a are independent random variables with

$$\mathbb{E}[X_a] = \mu_a \quad \text{and} \quad \mathrm{Tr}(\mathrm{Cov}\,(X_a)) = \Gamma_a.$$

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

Expected value at G

For a partition G we have

$$\mathbb{E}[\mathsf{Crit}_K(G)] = \frac{1}{2} \sum_k \frac{1}{|G_k|} \sum_{a,b \in G_k} \|\mu_a - \mu_b\|^2 + \sum_a \Gamma_a - \sum_k \frac{1}{|G_k|} \sum_{a \in G_k} \Gamma_a$$

$\longrightarrow$ tends to split "wide" clusters : a correction is needed !

Caveat (alternative view)

Recall our Minimization Problem : $\langle -\mathbf{X}^T\mathbf{X}, \mathbf{B}\rangle$

sGaussian Mixtures are of the form : $X_a = \mathbb{E}[X_a] + E_a = Information + Noise$,

$$\mathbb{E}[\mathbf{X}^T\mathbf{X}] = \mathbb{E}[\mathbf{X}]^T\,\mathbb{E}[\mathbf{X}] + \mathbf{\Gamma}\ , \qquad \text{where } \mathbf{\Gamma}_{aa} = \mathrm{Tr}[\mathrm{Cov}\,(E_a)]$$

Population K-means vs Ideal K-means

$$\begin{aligned} \mathbf{B}^{pop} &= \arg\min_{\mathbf{B}\in\mathcal{D}} \langle -\,\mathbb{E}[\mathbf{X}]^T\,\mathbb{E}[\mathbf{X}] - \mathbf{\Gamma}, \mathbf{B}\rangle \\ \mathbf{B}^{id} &= \arg\min_{\mathbf{B}\in\mathcal{D}} \langle -\,\mathbb{E}[\mathbf{X}]^T\,\mathbb{E}[\mathbf{X}], \mathbf{B}\rangle \end{aligned}$$

- Since $\mathrm{Tr}[\mathbf{B}] = K$, we have $\mathbf{B}^{pop} = \mathbf{B}^{id}$ when $\mathbf{\Gamma} = \gamma\mathbf{I}$.
- For heterogeneous $\mathbf{\Gamma}$, $\mathbf{B}^{pop}_{aa}$ tends to take large values for large $\mathbf{\Gamma}_{aa}$ (it splits wide clusters).

Estimation of Γ

Remark : If we knew the groups, we could estimate $\mathbf{\Gamma} = \mathrm{diag}(\Gamma_1, \dots, \Gamma_n)$ by

$$\widehat{\mathbf{\Gamma}}_{aa} = \langle X_a - X_{ne_1(a)}, X_a - X_{ne_2(a)} \rangle$$

with $ne_1(a)$ and $ne_2(a)$ two "neighbors" of a.

Estimation of $\mathbf{\Gamma}$

Remark : If we knew the groups, we could estimate $\mathbf{\Gamma} = \mathrm{diag}(\Gamma_1, \ldots, \Gamma_n)$ by

$$\widehat{\mathbf{\Gamma}}_{aa} = \langle X_a - X_{ne_1(a)}, X_a - X_{ne_2(a)} \rangle$$

with $ne_1(a)$ and $ne_2(a)$ two "neighbors" of a.

Definition

Then, the estimator $\widehat{\mathbf{\Gamma}}$ is the diagonal matrix defined by

$$\widehat{\mathbf{\Gamma}}_{aa} = \langle X_a - X_{\widehat{ne}_1(a)}, X_a - X_{\widehat{ne}_2(a)} \rangle$$

Estimation of $\mathbf{\Gamma}$

Remark : If we knew the groups, we could estimate $\mathbf{\Gamma} = \mathrm{diag}(\Gamma_1, \ldots, \Gamma_n)$ by

$$\widehat{\mathbf{\Gamma}}_{aa} = \langle X_a - X_{ne_1(a)}, X_a - X_{ne_2(a)} \rangle$$

with $ne_1(a)$ and $ne_2(a)$ two "neighbors" of a.

Definition

Set $U(a,b) := \max_{c,d \in [n] \setminus \{a,b\}} \left| \left\langle X_a - X_b, \frac{X_c - X_d}{\|X_c - X_d\|} \right\rangle \right|$ and

$\widehat{ne}_1(a) := \underset{b \in [n] \setminus \{a\}}{\arg\min}\ U(a,b)$ and $\widehat{ne}_2(a) := \underset{b \in [n] \setminus \{a, \widehat{ne}_1(a)\}}{\arg\min}\ U(a,b)$

Then, the estimator $\widehat{\mathbf{\Gamma}}$ is the diagonal matrix defined by

$$\widehat{\mathbf{\Gamma}}_{aa} = \langle X_a - X_{\widehat{ne}_1(a)}, X_a - X_{\widehat{ne}_2(a)} \rangle$$

Corrected relaxed K-means

Corrected relaxed K-means (Bunea et al.('16))

Solve the SDP

$$\widehat{B} \in \underset{\mathbf{B}\in\mathcal{C}}{\text{argmin}} \langle \quad -\mathbf{X}^T\mathbf{X}, \mathbf{B}\rangle \ ,$$

with

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

Similarly, one may define a corrected spectral clustering.

Corrected relaxed K-means

Corrected relaxed K-means (Bunea et al.('16))

Solve the SDP

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

with

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

Similarly, one may define a corrected spectral clustering.

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.
- **Perfect recovery** : Find the minimal s^2 such that $err(\widehat{G}, G^*) = 0$ whp.

Clustering subGaussian mixtures

Mixture of subGaussian variables

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

- $\mathbb{E}[X_a] = \mu_k$ if $a \in G_k^*$
- $\mathbf{\Sigma}_a^{-1/2} X_a$ is SubGauss$(L^2 \mathbf{I}_p)$ where $\mathbf{\Sigma}_a = \mathrm{cov}(X_a)$

We set

$$\Delta^2 = \min_{j \neq k} \|\mu_k - \mu_j\|^2, \quad \sigma^2 = L^2 \max_k |\mathbf{\Sigma}_k|_{op} \quad \text{and} \quad R_{\mathbf{\Sigma}} = \max_k \frac{|\mathbf{\Sigma}_k|_F^2}{|\mathbf{\Sigma}_k|_{op}^2},$$

and 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.

Toy Examples

Simplification : $K = 2$, $|G_1^*| = |G_{-1}^*| = n/2$, $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_{-1} = \sigma^2 \mathbf{I}_p$, $\mu_{-1} = -\mu_1$.

Simplified Model 1 : μ_1 is known. Bayes Classifier achieves :

$$\mathbb{E}[err(\widehat{G}, G^*)] = 2\,\mathbb{P}[\mathcal{N}(0,\sigma^2) > \|\mu_1\|] \leq 2\exp\left[-\frac{\Delta^2}{8\sigma^2}\right]$$

Simplified Model 2 : μ_1 is sampled uniformly on the sphere of radius $\Delta/2$. Labels $Z_a \in \{-1, 1\}$ for $a = 1, \dots, n$ are known.

Objective : classify a new observation X.

Optimal Classifier : $\widehat{h}(x) = \operatorname{sign}\left(\langle \frac{1}{n}\sum_{a=1}^n Z_a X_a, x\rangle\right)$:

- achieves the rate $e^{-c\Delta^2/\sigma^2}$ if $\frac{\Delta^2}{\sigma^2} \gtrsim 1 \vee \frac{p}{n}$.
- achieves the rate $e^{-cn\Delta^4/(p\sigma^4)}$ if $1 \vee \sqrt{\frac{p}{n}} \lesssim \frac{\Delta^2}{\sigma^2} \lesssim 1 \vee \frac{p}{n}$.

See Ndaoud('18) for proper lower bounds.

Partial recovery bounds

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

Theorem (Giraud and V. ('18))

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

Partial recovery bounds

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

Theorem (Giraud and V. ('18))

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

$s^2 \gtrsim n/m$ 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. s^2 reduces to Δ^2/σ^2 when $\Delta^2/\sigma^2 \geq R_{\mathbf{\Sigma}}/m$
   Fei and Chen ('18) , See Lu and Zhou ('16), Ndaoud('18) for sharp constants

Partial recovery bounds

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

Theorem (Giraud and V. ('18))

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

$s^2 \gtrsim n/m$ 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. s^2 reduces to Δ^2/σ^2 when $\Delta^2/\sigma^2 \geq R_{\mathbf{\Sigma}}/m$
   Fei and Chen ('18) , See Lu and Zhou ('16), Ndaoud('18) for sharp constants
2. perfect recovery for $s^2 \gtrsim \log(n) \vee (n/m) \quad = \log(n) \vee K$
   Dependency in K is suboptimal.
   Vempala and Wang('04) $\rightsquigarrow s^2 \gtrsim \log(n) \vee \sqrt{K\log(n)}$ when $n \gg p^3$.
3. Do not cover the case where the proportion of error is $\geq e^{-c''K}$.

Mild price for $\mathbf{\Gamma}$ estimation : $\frac{\|\mathbf{\Sigma}_k\|_{op}\mathrm{Tr}(\mathbf{\Sigma}_k)}{\|\mathbf{\Sigma}_k\|_F^2} \lesssim \frac{n}{\log(n)}$

Benefit of Corrected K-means

Mild price for $\mathbf{\Gamma}$ estimation : $\frac{\|\mathbf{\Sigma}_k\|_{op}\mathrm{Tr}(\mathbf{\Sigma}_k)}{\|\mathbf{\Sigma}_k\|_F^2} \lesssim \frac{n}{\log(n)}$

Without correction, additional assumption is required :

$$\Delta^2 \gtrsim \frac{\max \mathbf{\Gamma}_a - \min \mathbf{\Gamma}_a}{m}$$

For a balanced Partition, it amounts to

$$\Delta^2 \gtrsim \sigma^2 K \left(1 \vee \sqrt{\frac{R_{\mathbf{\Sigma}}}{n}} \vee \frac{\max_k tr[\mathbf{\Sigma}_k] - \min_k tr[\mathbf{\Sigma}_k]}{n}\right] \ .$$

Proof Ideas

Simple Versions : All $\|\mu_i - \mu_j\|_2$ are equal

Step 1 : $|\widehat{\mathbf{B}} - \mathbf{B}^*|_1$ small implies that $err(\widehat{G}, G)$ is small.

New Objective : Show that $\langle \mathbf{X}^T\mathbf{X} - \widehat{\mathbf{\Gamma}}, \mathbf{B}^* - \mathbf{B} \rangle > 0$ as long as $|\mathbf{B}^* - \mathbf{B}|_1$ is not small

Proof Ideas

Simple Versions : All $\|\mu_i - \mu_j\|_2$ are equal

Step 1 : $|\widehat{\mathbf{B}} - \mathbf{B}^*|_1$ small implies that $err(\widehat{G}, G)$ is small.

New Objective : Show that $\langle \mathbf{X}^T\mathbf{X} - \widehat{\mathbf{\Gamma}}, \mathbf{B}^* - \mathbf{B} \rangle > 0$ as long as $|\mathbf{B}^* - \mathbf{B}|_1$ is not small

$$\begin{aligned} \langle \mathbf{X}^T\mathbf{X} - \widehat{\mathbf{\Gamma}}, \mathbf{B}^* - \mathbf{B} \rangle \quad &= \quad \langle \mathbf{A}\mu\mu^T\mathbf{A}^T, \mathbf{B}^* - \mathbf{B} \rangle + \langle \mathbf{E}^T\mathbf{E} - \mathbf{\Gamma}, \mathbf{B}^* - \mathbf{B} \rangle \\ &\quad + \langle \mathbf{\Gamma} - \widehat{\mathbf{\Gamma}}, \mathbf{B}^* - \mathbf{B} \rangle + \langle \mathbf{A}\mu\mathbf{E}^T + \mathbf{E}\mu\mathbf{A}^T, \mathbf{B}^* - \mathbf{B} \rangle \end{aligned}$$

We focus on the two first terms

Signal Term : $\langle \mathbf{A}\mu\mu^T\mathbf{A}^T, \mathbf{B}^* - \mathbf{B} \rangle = \frac{1}{4}\Delta^2 |\mathbf{B}^* - \mathbf{B}^*\mathbf{B}|_1$

Control of the quadratic term : $\langle \mathbf{E}^T\mathbf{E} - \Gamma, \mathrm{B}^* - \mathrm{B} \rangle$

$\mathbf{B}^*$ is projection operator that averages over element of the same group.
$\rightsquigarrow$ Decomposition of $\mathbf{E}^T\mathbf{E} - \mathbf{\Gamma}$ by applying $\mathbf{B}^*$ or $(\mathbf{I} - \mathbf{B}^*)$.

Control of the quadratic term : $\langle \mathrm{E}^T\mathrm{E} - \Gamma, \mathrm{B}^* - \mathrm{B} \rangle$

$\mathbf{B}^*$ is projection operator that averages over element of the same group.
$\rightsquigarrow$ Decomposition of $\mathbf{E}^T\mathbf{E} - \mathbf{\Gamma}$ by applying $\mathbf{B}^*$ or $(\mathbf{I} - \mathbf{B}^*)$.

Step 3 : Control of the Projection Along $\mathrm{Im}(\mathbf{B}^*)$

$$\begin{aligned} \langle (\mathbf{I} - \mathbf{B}^*)(\mathbf{E}^T\mathbf{E} - \mathbf{\Gamma})(\mathbf{I} - \mathbf{B}^*), \mathbf{B}^* - \mathbf{B} \rangle \quad &\leq \quad \|\mathbf{E}^T\mathbf{E} - \mathbf{\Gamma}\|_{op} \|(\mathbf{I} - \mathbf{B}^*)[\mathbf{B}^* - \mathbf{B}](\mathbf{I} - \mathbf{B}^*)\|_* \\ &= \|\mathbf{E}^T\mathbf{E} - \mathbf{\Gamma}\|_{op} \frac{1}{2m} |\mathbf{B}^* - \mathbf{B}^*\mathbf{B}|_1 \end{aligned}$$

$\rightsquigarrow$ *Concentration of $\mathbf{E}^T\mathbf{E}$ in operator norm*

Control of the quadratic term : $\langle \mathbf{E}^T\mathbf{E} - \mathbf{\Gamma}, \mathbf{B}^* - \mathbf{B} \rangle$

$\mathbf{B}^*$ is projection operator that averages over element of the same group.
$\rightsquigarrow$ Decomposition of $\mathbf{E}^T\mathbf{E} - \mathbf{\Gamma}$ by applying $\mathbf{B}^*$ or $(\mathbf{I} - \mathbf{B}^*)$.

Step 3 : Control of the Projection Along $\mathrm{Im}(\mathbf{B}^*)$

$$\begin{aligned} \langle (\mathbf{I} - \mathbf{B}^*)(\mathbf{E}^T\mathbf{E} - \mathbf{\Gamma})(\mathbf{I} - \mathbf{B}^*), \mathbf{B}^* - \mathbf{B} \rangle \quad &\leq \quad \|\mathbf{E}^T\mathbf{E} - \mathbf{\Gamma}\|_{op} \|(\mathbf{I} - \mathbf{B}^*)[\mathbf{B}^* - \mathbf{B}](\mathbf{I} - \mathbf{B}^*)\|_* \\ &= \|\mathbf{E}^T\mathbf{E} - \mathbf{\Gamma}\|_{op} \frac{1}{2m} |\mathbf{B}^* - \mathbf{B}^*\mathbf{B}|_1 \end{aligned}$$

$\rightsquigarrow$ *Concentration of $\mathbf{E}^T\mathbf{E}$ in operator norm*

Step 4 : Control of $\langle \mathbf{B}^*(\mathbf{E}^T\mathbf{E} - \mathbf{\Gamma}), \mathbf{B}^* - \mathbf{B} \rangle$.

A First try : $\langle \mathbf{A}, \mathbf{B} \rangle \leq |\mathbf{A}|_\infty |\mathbf{B}|_1$ does not lead to exponential bounds.

A Second try (Fei and Chen('17)) : $\langle \mathbf{A}, \mathbf{B} \rangle \leq \sum_{i=1}^{|\mathbf{B}|_1} \mathbf{A}_{(i)}$, where $\mathbf{A}_{(1)} \geq \mathbf{A}_{(2)} \geq \ldots$
Control of the order statistics $\mathbf{B}^*(\mathbf{E}^T\mathbf{E} - \mathbf{\Gamma})$ by Hanson-Wright inequality + Union bound

Model 2 : graph clustering

(conditional) SBM

Assume that the graph is generated by a SBM with $\mathbf{Q}_{jk}$ = probability of connection between nodes of groups j and k.
Let $\mathbf{X}$ = adjacency matrix of the graph $\in \{0,1\}^{n\times n}$.

$$\underline{\text{For } a \in G_k^* :}\;\; X_a = [\mathbf{QA}]_{k:} - \mathbf{Q}_{kk} e_a + E_a, \quad \text{where} \;\; E_a = X_a - \mathbb{E}[X_a]$$

$$\Delta^2 = \min_{j\neq k} \|[\mathbf{QA}]_{k:} - [\mathbf{QA}]_{j:}\|^2 \geq m \times \min_{j\neq k} \|\mathbf{Q}_{k:} - \mathbf{Q}_{j:}\|^2 \;\; \left(\;\geq 2m\, \lambda_{\min}(\mathbf{Q})^2 \right)$$

Partial recovery for SBM

$$\Delta^2 = \min_{j \neq k} \|[\mathbf{QA}]_{k:} - [\mathbf{QA}]_{j:}\|^2 \geq m \times \min_{j \neq k} \|\mathbf{Q}_{k:} - \mathbf{Q}_{j:}\|^2 \quad \left(\; \geq 2m\, \lambda_{\min}(\mathbf{Q})^2 \right)$$

$$L \geq \|\mathbf{Q}\|_{op} \vee 1/m$$

Theorem (Giraud and V.('18))

We set $s^2 = \Delta^2 / L$. If $s^2 \gtrsim n/m$ we have $\mathbb{P}[\mathrm{err}(G, \widehat{G}) > e^{-cs^2}] \lesssim 1/n^2$

Partial recovery for SBM

$$\Delta^2 = \min_{j \neq k} \|[\mathbf{QA}]_{k:} - [\mathbf{QA}]_{j:}\|^2 \geq m \times \min_{j \neq k} \|\mathbf{Q}_{k:} - \mathbf{Q}_{j:}\|^2 \quad \left(\ \geq 2m\, \lambda_{\min}(\mathbf{Q})^2 \right)$$

$$L \geq \|\mathbf{Q}\|_{op} \vee 1/m$$

Theorem (Giraud and V.('18))

We set $s^2 = \Delta^2/L$. If $s^2 \gtrsim n/m$ we have $\mathbb{P}[\mathrm{err}(G, \widehat{G}) > e^{-cs^2}] \lesssim 1/n^2$ if we have enforced $\|\mathbf{B}\|_{op} \leq \frac{K^3}{n} e^{4nL}$.

Comments

Assortative case : $\mathbf{Q} = (p-q)\mathbf{I} + q11^T$ and $m = n/K$

1. $s^2 = 2m(p-q)^2/p$ for $p \geq K/n$.
   tight constants in Gao et al.('17), Yun and Proutière('14)
2. perfect recovery for
$$\frac{(p-q)^2}{p} \gtrsim \frac{K^2 \vee K\log(n)}{n}$$
   Matches Best known polynomial time algorithm condition Chen and Xu('16)

Comments (General $\mathbf{Q}$)

Exponential decay :
Abbe and Sandon('15) consider the scaling $\mathbf{Q} = \mathbf{Q}_0 \log(n)/n$ for a fixed K. Results not completely comparable.

Perfect recovery : if $\|\mathbf{Q}\|_{op} = O(\min_{j,k} \mathbf{Q}_{jk})$, we recover (up to constant) the optimal condition of Abbe and Sandon('15)

Comments (General $\mathbf{Q}$)

Exponential decay :
Abbe and Sandon('15) consider the scaling $\mathbf{Q} = \mathbf{Q}_0 \log(n)/n$ for a fixed K. Results not completely comparable.

Perfect recovery : if $\|\mathbf{Q}\|_{op} = O(\min_{j,k} \mathbf{Q}_{jk})$, we recover (up to constant) the optimal condition of Abbe and Sandon('15)

Other SDP for SBM : Relaxed K-means differs from Chen & Xu('16), Hajek et al.('16), Guédon & Vershynin('16), Perry & Wein ('16)...

$$\widetilde{\mathbf{B}} = \arg\max_{\mathbf{B} \in \mathcal{C}'} \langle \mathbf{X}, \mathbf{B} \rangle$$

for assortative graphs (diag($\mathbf{Q}$)> nondiag($\mathbf{Q}$))

Proof

Same arguments, but :

- spectral control requires trimming arguments in the proof
- control of quadratic terms quite messy due to the symmetry of $\mathbf{X}$ (peeling, conditionning, ...)

Summary

Main message

A corrected convex relaxation of K-means achieves some rate-optimal performances in various settings including (conditional) mixture of sub-Gaussian and (conditional) Stochastic Block Model.

Only tuning Parameter is K

F. Bunea, C. Giraud, M. Royer, N. V. **PECOK : a convex optimization approach to variable clustering**. *Annals of Statistics*. `ArXiv:1606.05100`

C. Giraud and N.V. **Partial recovery bounds for clustering with the relaxed K-means**. *Mathematical Statistics and Learning* `ArXiv:1807.07547`

Summary

Main message

A corrected convex relaxation of K-means achieves some rate-optimal performances in various settings including (conditional) mixture of sub-Gaussian and (conditional) Stochastic Block Model.

Only tuning Parameter is K

F. Bunea, C. Giraud, M. Royer, N. V. **PECOK : a convex optimization approach to variable clustering**. *Annals of Statistics*. `ArXiv:1606.05100`

C. Giraud and N.V. **Partial recovery bounds for clustering with the relaxed K-means**. *Mathematical Statistics and Learning* `ArXiv:1807.07547`

Merci pour votre attention !

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 and Colin Sandon.
Community Detection in General Stochastic Block Models : Fundamental Limits and Efficient Algorithms for Recovery.
In Proceedings of the 2015 IEEE 56th Annual Symposium on Foundations of Computer Scie
FOCS '15, pages 670–688, Washington, DC, USA, 2015. IEEE Computer Society.

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.

References II

Y. Fei and Y. Chen.
Exponential error rates of SDP for block models : Beyond Grothendieck's inequality.
ArXiv e-prints, 2017.

Y. Fei and Y. Chen.
Hidden Integrality of SDP Relaxation for Sub-Gaussian Mixture Models.
ArXiv e-prints, March 2018.

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.

Olivier Guédon and Roman Vershynin.
Community detection in sparse networks via Grothendieck's inequality.
arXiv preprint arXiv :1411.4686, 2014.

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

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

S. Lloyd.
Least Squares Quantization in PCM.
IEEE Trans. Inf. Theor., 28(2) :129–137, September 1982.

Mohamed Ndaoud.
Sharp optimal recovery in the Two Component Gaussian Mixture Model.
arXiv e-prints, page arXiv :1812.08078, Dec 2018.

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

A. Perry and A. S. Wein.
A semidefinite program for unbalanced multisection in the stochastic block model.
ArXiv e-prints, July 2015.

Santosh Vempala and Grant Wang.
A spectral algorithm for learning mixture models.
Journal of Computer and System Sciences, 68(4) :841–860, 2004.
Special Issue on FOCS 2002.

References IV

Se-Young Yun and Alexandre Proutière.
Accurate Community Detection in the Stochastic Block Model via Spectral Algorithms.
CoRR, abs/1412.7335, 2014.