Clustering with convex optimisation

Christophe Giraud

Université Paris-Sud
Université Paris Saclay

CIRM July 2017

Talks based on papers with/from

Flori (Cornell)

Nicolas (INRA Montpellier)

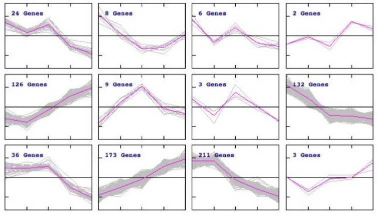
Martin (PhD, Paris Sud)

Ismael & Youssouf
(Master, Polytechnique)

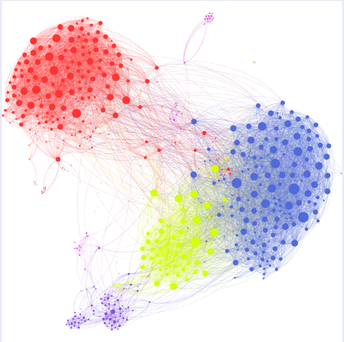
Clustering arises in various contexts

Clustering individuals w.r.t. features

Clustering features



Clustering graphs



Our objectives

Topic of the talk

- investigate "optimality" in clustering (in terms of exact recovery)
- probabilistic set-up: data generated by some (more or less) flexible models
- optimality in terms of rate-minimax "separation" between groups
- focus on polynomial time algorithms

Main message

A corrected convex relaxation of Kmeans achieves some rate-optimal performances in various settings.

Many classical algorithms (with caveats)

"Geometric" algorithms

- **Hierarchical clustering:** greedy, no global criterion
- **Kmeans:** multiple local minima, NP-hard, greedy approximations (Lloyd algorithm) very sensitive to initialization

"Model-based"-algorithms

- **Approximate MLE in mixture models** (with EM-like algorithms): multiples local minima, sensitive to initialization, issue of misspecification.

Spectral algorithms and SDP

Two popular alternatives

It has been shown that spectral clustering and some SDP have some (nearly)-optimal properties in some models (e.g. in assortative SBM, Gaussian mixture model)

In this talk

We will

- focus on a specific SDP derived from Kmeans, which achieves some optimal performances in a wide range of situations,
- connect this SDP to spectral clustering.

1- Relaxed Kmeans

Peng & Wei (07)

Kmeans criterion

Applying Kmeans on N data points $X_1, \ldots, X_N$ amounts to minimizes among all possible partitions $G = \{G_1, \ldots, G_K\}$ of $\{1, \ldots, N\}$

$$\begin{aligned}
\mathrm{Crit}(G) &= \sum_{k=1}^{K} \sum_{a \in G_k} \|X_a - \bar{X}_{G_k}\|^2 \\
&= \frac{1}{2} \sum_{k=1}^{K} \frac{1}{|G_k|} \sum_{a,b \in G_k} \|X_a - X_b\|^2 \\
&= -\sum_{k=1}^{K} \sum_{a,b \in G_k} \frac{1}{|G_k|} \langle X_a, X_b \rangle + \sum_{a=1}^{N} \|X_a\|^2 \\
&= -\langle B^G, X^T X \rangle + \|X\|_F^2
\end{aligned}$$

with $X = [X_1, \ldots, X_N]$ and

$$B^G_{ab} = 1/|G_k| \text{ if } a, b \text{ belong to the same group } G_k \text{ and } B^G_{ab} = 0 \text{ else.}$$

Kmeans criterion

Lemma (Peng & Wei (2007))

Solving Kmeans amounts to solve

$$\widehat{B}_{Kmeans} \in \underset{B \in \mathcal{D}}{\operatorname{argmin}} \langle -X^T X, B \rangle ,$$

with

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

Convexified Kmeans

Idea: drop the $B^2 = B$ constraint

Relaxed Kmeans (Peng & Wei (2007))

Solve the SDP

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

with

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

Remarks:

1. An additional clustering step is needed when $\widehat{B} \notin \mathcal{D}$.
2. Convex optimisation but with many constraints.

Spectral clustering

Drop the constraints $B_{ab} \geq 0$ and $\sum_a B_{ab} = 1$ but keep the (implicit) condition $I \succcurlyeq B$

Relaxed SDP

Solve the SDP

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

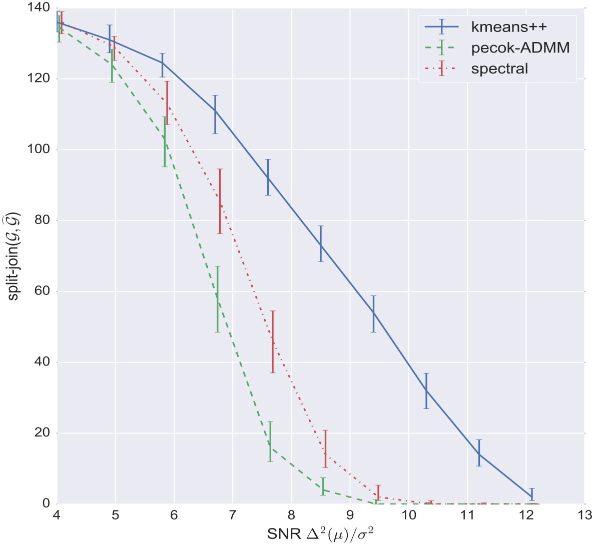
with

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

Relaxed SDP = Spectral clustering

The solution $\bar{B}$ is given by $\bar{B} = \bar{U}\bar{U}^T$ where $\bar{U}$ collects "the" K leading eigenvectors of $X^T X$.

Numerical comparison



2- Quantization versus clustering

Caveheat

A simple model

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

$$\mathbb{E}[X_a] = \nu_a \quad \text{and} \quad Tr(cov(X_a)) = \Gamma_a.$$

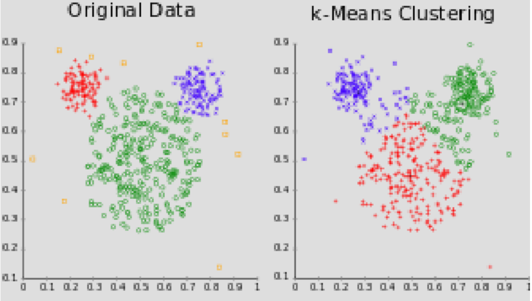
Mean value

For a partition G we have

$$\mathbb{E}[crit_{Kmeans}(G)] = \frac{1}{2}\sum_k \frac{1}{|G_k|}\sum_{a,b\in G_k}\|\nu_a-\nu_b\|^2 + \sum_a \Gamma_a - \sum_k \frac{1}{|G_k|}\sum_{a\in G_k}\Gamma_a$$

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

Example



Quantization rather than clustering

Estimation of Γ

Remark: If we knew the groups, we could estimate $\Gamma = diag(\Gamma_1, \dots, \Gamma_N)$ by

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

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

Definition

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

$\widehat{ne}_1(a) := \operatorname{argmin}_{b \in [n] \setminus \{a\}} U(a,b)$ and $\widehat{ne}_2(a) := \operatorname{argmin}_{b \in [n] \setminus \{a, \widehat{ne}_1(a)\}} U(a,b)$

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

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

Estimation of Γ

Remark: If we knew the groups, we could estimate $\Gamma = diag(\Gamma_1, \ldots, \Gamma_N)$ by

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

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

Definition

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

$\widehat{ne}_1(a) := \operatorname{argmin}_{b \in [n] \setminus \{a\}} U(a,b)$ and $\widehat{ne}_2(a) := \operatorname{argmin}_{b \in [n] \setminus \{a, \widehat{ne}_1(a)\}} U(a,b)$

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

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

Estimation of Γ

Remark: If we knew the groups, we could estimate $\Gamma = diag(\Gamma_1, \ldots, \Gamma_N)$ by

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

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

Definition

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

$\widehat{ne}_1(a) := \operatorname{argmin}_{b \in [n] \setminus \{a\}} U(a,b)$ and $\widehat{ne}_2(a) := \operatorname{argmin}_{b \in [n] \setminus \{a, \widehat{ne}_1(a)\}} U(a,b)$

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

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

Corrected convexified Kmeans

In the above simple model

$$\mathbb{E}[X^T X] = \text{a "block structured" matrix} + \Gamma.$$

Corrected convexified Kmeans (F. Bunea, C. G., M. Royer, N. Verzelen (2016))

Solve the SDP

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

with

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

Corrected convexified Kmeans

In the above simple model

$$\mathbb{E}[X^T X] = \text{a "block structured" matrix} + \Gamma.$$

Corrected convexified Kmeans (F. Bunea, C. G., M. Royer, N. Verzelen (2016))

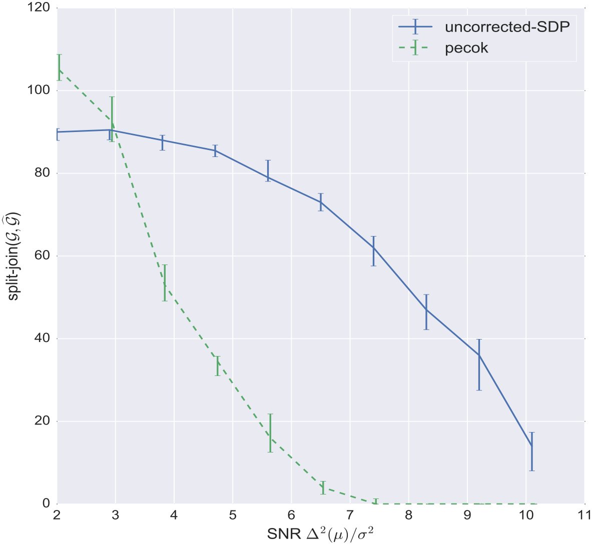
Solve the SDP

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

with

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

The correction can be useful



3- Some theory

Model 1: clustering "individuals"

Clustered independent subGaussian variables

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

- $\mathbb{E}[X_a] = \mu_k$ if $a \in G_k^*$
- $X_a \sim \mathrm{SubGauss}(\Sigma_a)$

For simplicity, we will focus here on the case where each group has the same size $|G_k^*| = n/K$

Exact recovery (M. Royer (2017))

Exact recovery with probability at least $1 - 1/n$ as soon as

$$\min_{j \neq k} \frac{\|\mu_j - \mu_k\|^2}{\max_a |\Sigma_a|_{op}} \gtrsim K \vee \log(n) + \sqrt{\frac{r^*(K \vee \log(n))}{n}} \quad \text{with} \quad r^* = \frac{\max_a Tr(\Sigma_a)}{\max_a |\Sigma_a|_{op}}.$$

Optimality

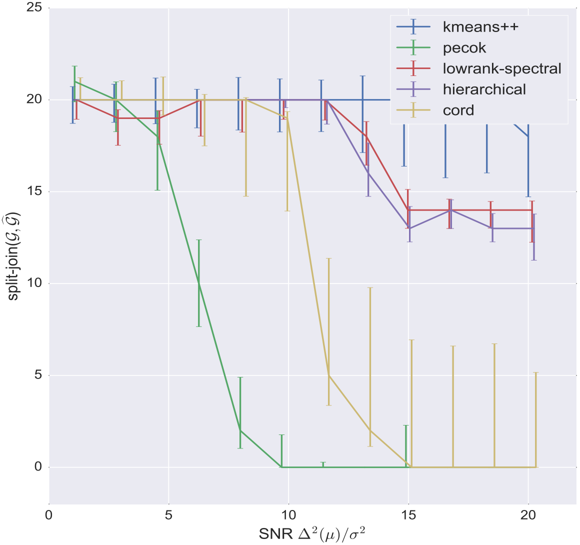
Some optimality

- Optimal rate when $\Sigma_a = \sigma^2 I_p$ and $K = O(\log(n))$.
- Computational gap for $K \gg \log(n)$? (as in SBM)

Remarks:

- The general case requires further investigations.
- The assumption of identical mean within groups can be relaxed.

Illustrations



Model 2: clustering "features"

We have n i.i.d. observations of a p-dimensional vector of features with $\mathcal{N}(0, \Sigma)$ distribution.

So the rows of the matrix $X = [X_{ia}]_{i=1,\ldots,n;\ a=1,\ldots,p}$ are independent, with $\mathcal{N}(0, \Sigma)$ distribution.

We want to cluster the features.

Block-structured covariance matrix

We assume the (unknown) block structure

- $\Sigma_{ab} = C_{kj}$ if $a \in G_k^*$, $b \in G_j^*$ and $a \neq b$
- $\Sigma_{aa} = C_{kk} + \Gamma_a$
- C is positive semi-definite ($\Longleftrightarrow$ a latent model)

For simplicity, we focus here on the case where each group of features has the same size $|G_k^*| = p/K$

Minimax-optimal recovery

Exact recovery (Bunea, G., Royer, Verzelen (2016))

Exact recovery with probability at least $1 - 1/p$ as soon as

$$\min_{j \neq k} \frac{C_{jj} + C_{kk} - 2C_{jk}}{|\Gamma|_\infty} \gtrsim \sqrt{\frac{\log(p) \vee K}{np/K}} + \frac{\log(p) \vee K}{n}.$$

- rate-minimax optimal for $K = O(\log(p))$,
- computational gap otherwise?
- can be extended to Subgaussian vectors,
- the same result can be achieved **when K is unknown**, with a slight variation of the SDP (drop the constraint $tr(B) = K$ and add $\widehat{\lambda} I$ to $\widehat{\Gamma}$).

Model 3: graph clustering

(conditional) SBM

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

<u>Remark:</u> the SDP is applied to $X^T X = X^2$ instead of X.

As before, we focus here on the case where each group of feature has the same size $|G_k^*| = N/K$

Exact recovery (Emin and Lemhadri (2017?))

Exact recovery with probability at least $1 - 1/N$ as soon as

$$\min_{j\neq k} \|Q_{\bullet j} - Q_{\bullet k}\|^2 \gtrsim |Q|_\infty \frac{K \vee \log(N)}{N/K} + \frac{\log(N)}{(N/K)^2}$$

4- Practice

Computational issues

Practical benefit?

- solving the SDP is very intensive when we cluster many "points" (many constraints)
- Intensive research for fast approximate solvers
- But does it make sense?
- May be: we can expect that approximate solvers are less greedy than Lloyd-like algorithms (under investigation...)

Many thanks to all the organizers for this great meeting!