Statistical and Computational challenges in clustering

Nicolas Verzelen

INRAE, Mistea

NetBio - September 15th

Acknowledgements

Alexandra Carpentier
(Potsdam)

Bertrand Even
(Paris-Saclay)

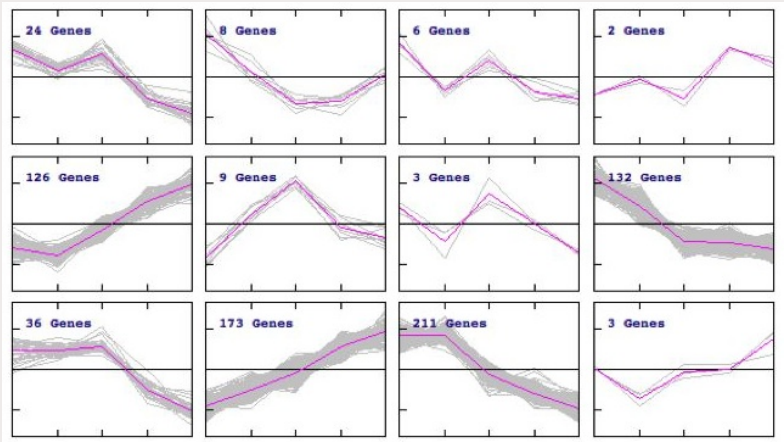
Simone Giancola
(Paris-Saclay)

Christophe Giraud
(Paris-Saclay)

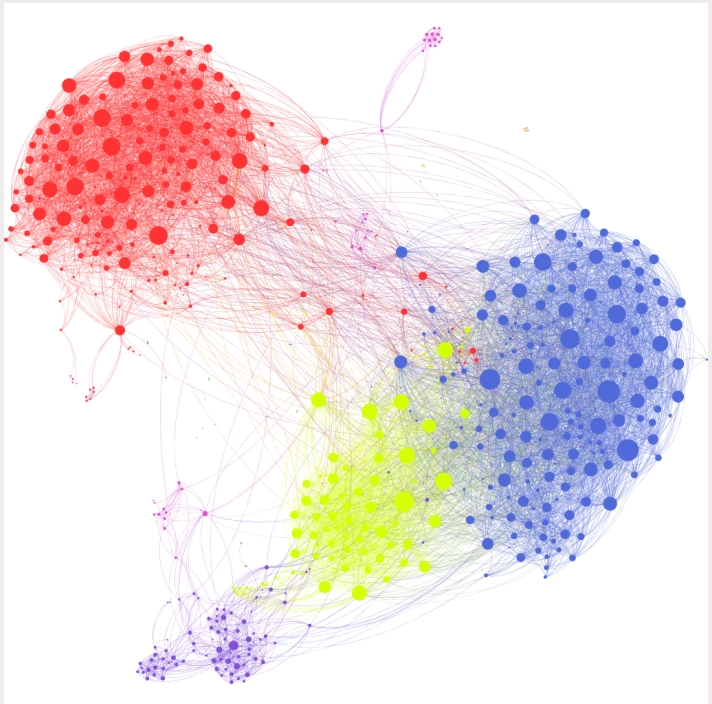
Clustering arises in various contexts

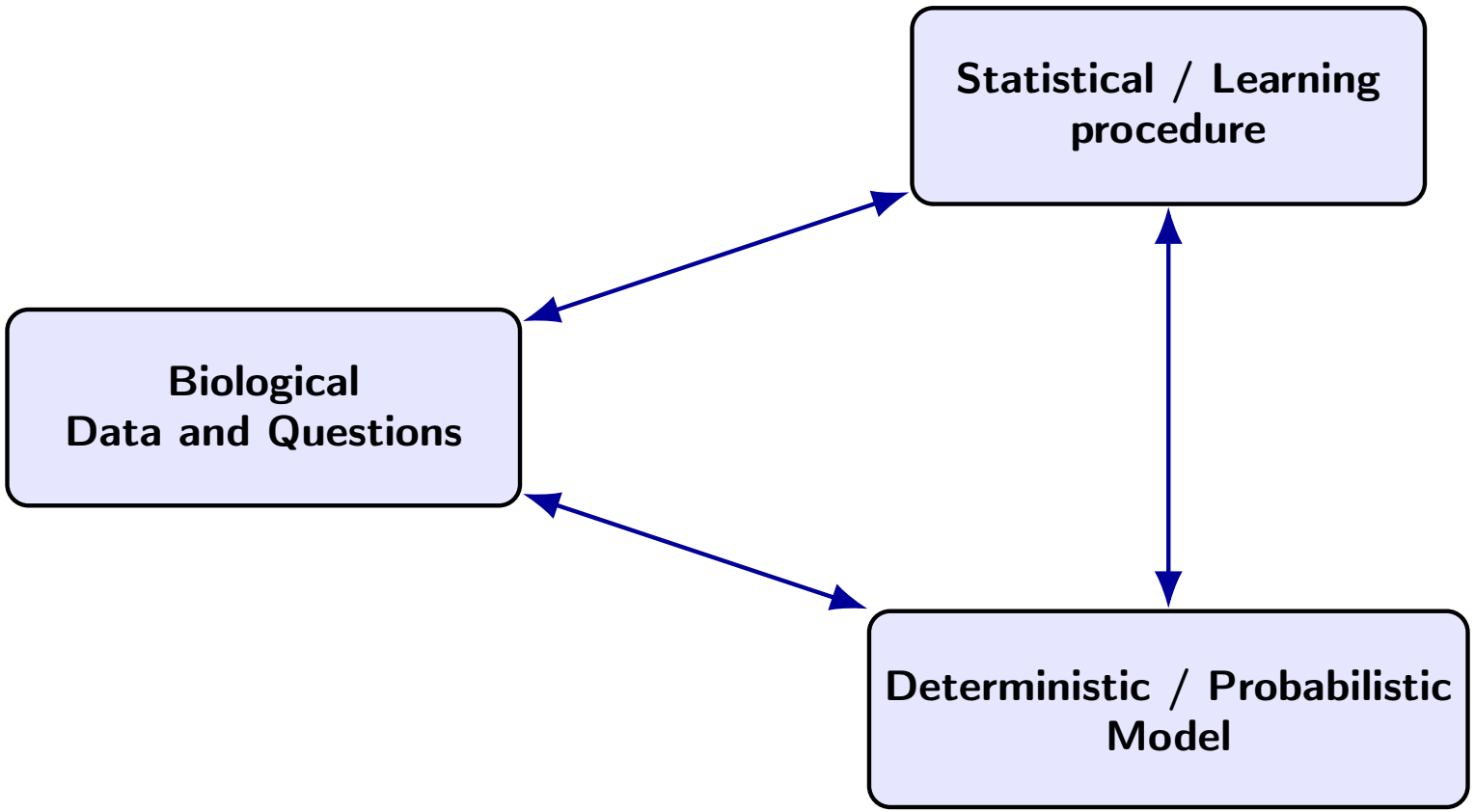
Clustering individuals w.r.t. features

Clustering features



Clustering graphs




Statistical / Learning procedure
Biological Data and Questions
Deterministic / Probabilistic Model

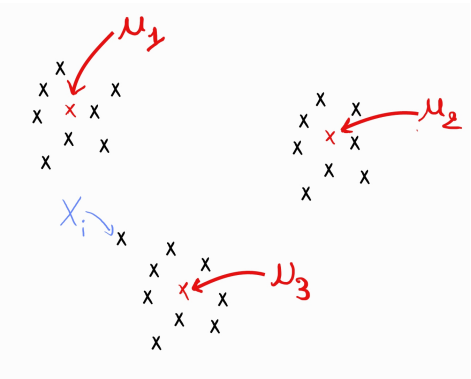
Model-Based Point Clustering

Gaussian Mixture Modeling (Pearson '1894)

Model

Observe $X_1, \dots, X_n \in \mathbb{R}^d$ independent
$\exists\ G^* = (G_1^*, \dots, G_K^*)$ partition of $[n]$ into K groups

$$X_i \sim \mathcal{N}(\mu_k, \sigma^2 I_d), \qquad \forall i \in G_k^*$$

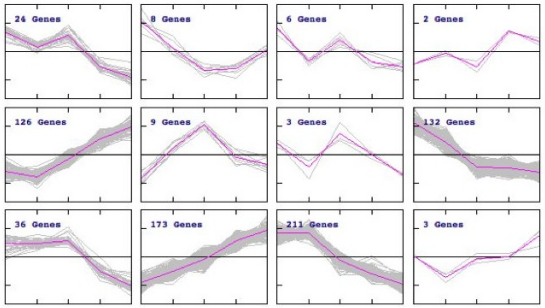


Goal

Recover G^* from $\mathbf{X} = [X_1, X_2, \dots, X_n]^T \in \mathbb{R}^{n \times d}$

- Approximately wrt $err(\hat{G}, G^*) = \frac{1}{2n} \min_{\pi \in \mathcal{S}_K} \sum_{k=1}^{K} |G_k^* \Delta \hat{G}_{\pi(k)}|$
  $=$ "proportion of missclassified nodes"

Model-Based Variable Clustering (Bunea et al.'20)

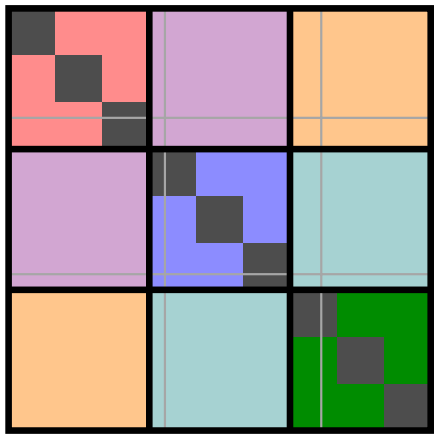


Model

Observe $X_1, \dots, X_n \sim \mathcal{N}(0, \Sigma) \in \mathbb{R}^d$ independent
$\exists\ G^* = (G_1^*, \dots, G_K^*)$ partition of $[d]$ into K groups

$$\Sigma_{i,j} = \Sigma_{i',j} \text{ if } i \sim_{G^*} i'$$

$\Sigma =$



By homogeneity, take $\sigma = 1$.

Goal

Recover G^* from $\mathbf{X} = [X_1, X_2, \dots, X_n]^T \in \mathbb{R}^{n \times d}$

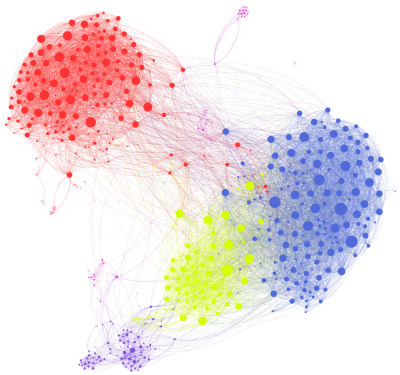
- Approximately wrt $err(\hat{G}, G^*) = \frac{1}{2d} \min_{\pi \in \mathcal{S}_K} \sum_{k=1}^{K} |G_k^* \Delta \hat{G}_{\pi(k)}|$
  $=$ "proportion of missclassified nodes"

Model-Based Graph Clustering

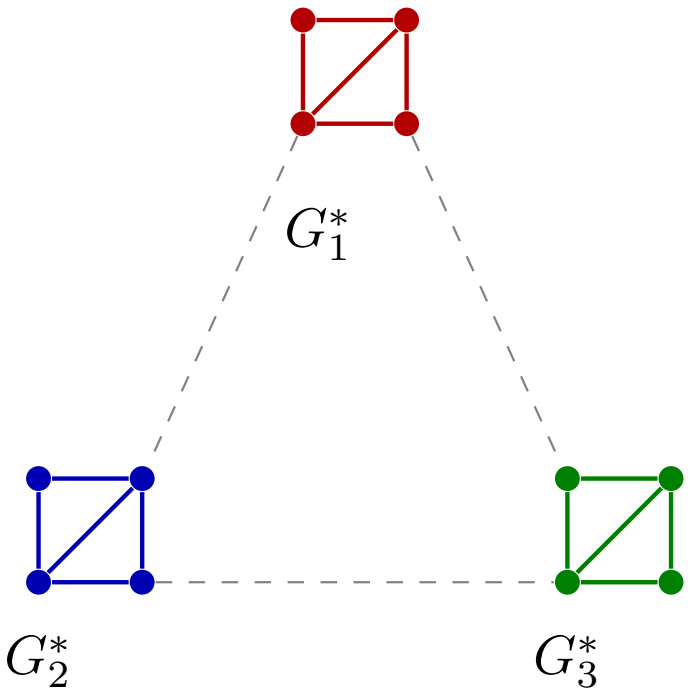
Stochastic Block Model (Holland, et al. '1983)

Observe a graph $\mathcal{G} = ([n], E)$ on n nodes.
$\exists\ G^* = (G_1^*, \dots, G_K^*)$ partition of $[n]$ into K communities and
$B = (B_{kl}) \in [0,1]^{K \times K}$ symmetric such that, independently for all $i < j$,

$$\{i,j\} \in E \quad \text{with probability} \quad B_{k,l}, \quad \text{for } i \in G_k^*,\ j \in G_l^*.$$



- Adjacency matrix
  $X_{ij} = \mathbb{1}\{\{i,j\} \in E\}$
- Within community k : connection probability B_{kk} (dense)
- Between communities $k \neq l$: probability B_{kl} (sparse)



Main Questions in this talk

- Can we have confidence on our statistical procedure ?
- Is the problem actually feasible ?
- Can we craft a better procedure ?
- Are all the previous responses sensible to the model choice ?

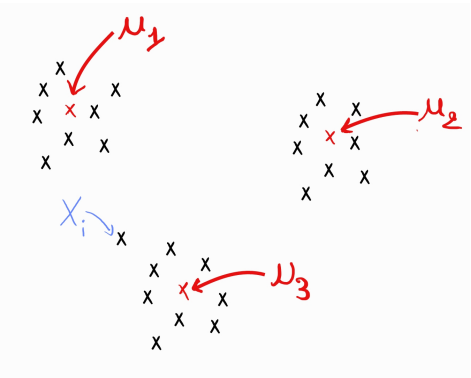
Focus on Gaussian Mixture Clustering

Gaussian Mixture Modeling (Pearson '1894)

Model

Observe $X_1, \dots, X_n \in \mathbb{R}^d$ independent
$\exists\ G^* = (G_1^*, \dots, G_K^*)$ partition of $[n]$ into K groups

$$X_i \sim \mathcal{N}(\mu_k, \sigma^2 I_d), \qquad \forall i \in G_k^*$$



Goal

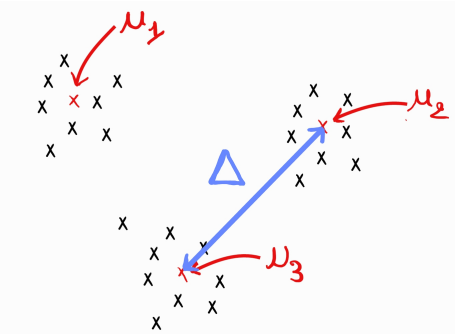
Recover G^* from $\mathbf{X} = [X_1, X_2, \dots, X_n]^T \in \mathbb{R}^{n \times d}$

- Approximately wrt $err(\hat{G}, G^*) = \frac{1}{2n} \min_{\pi \in \mathcal{S}_K} \sum_{k=1}^{K} |G_k^* \Delta \hat{G}_{\pi(k)}|$
  $=$ "proportion of missclassified nodes"

Relevant quantities ?

Separation : $\Delta^2 := \min_{k \neq l} \|\mu_k - \mu_l\|^2$

Assumption : $|G_k^*| \asymp \frac{n}{K}$



Specific Questions ?

- Conditions on (Δ, n, d, K) to recover exactly (resp. approximately) G^* ?
- What about classical methods ?
- Which conditions under computational constraints ?

Minimax risk/Information bound

Old paradigm in Statistics ('70's)

Collection of Mixture with minimal separation

Fix separation threshold $\Delta_0 > 0$.

$$\boldsymbol{\Theta}(\Delta_0) := \{(G^*, \mu) : \Delta \geq \Delta_0\}$$

Minimax risk

$$\inf_{\hat{G}} \sup_{(G^*,\mu)\in\boldsymbol{\Theta}(\Delta_0)} \mathbb{E}[\mathrm{err}(\hat{G}, G^*)]$$

Goal : Find Δ_I^* and Procedure $\hat{G}$ such that :

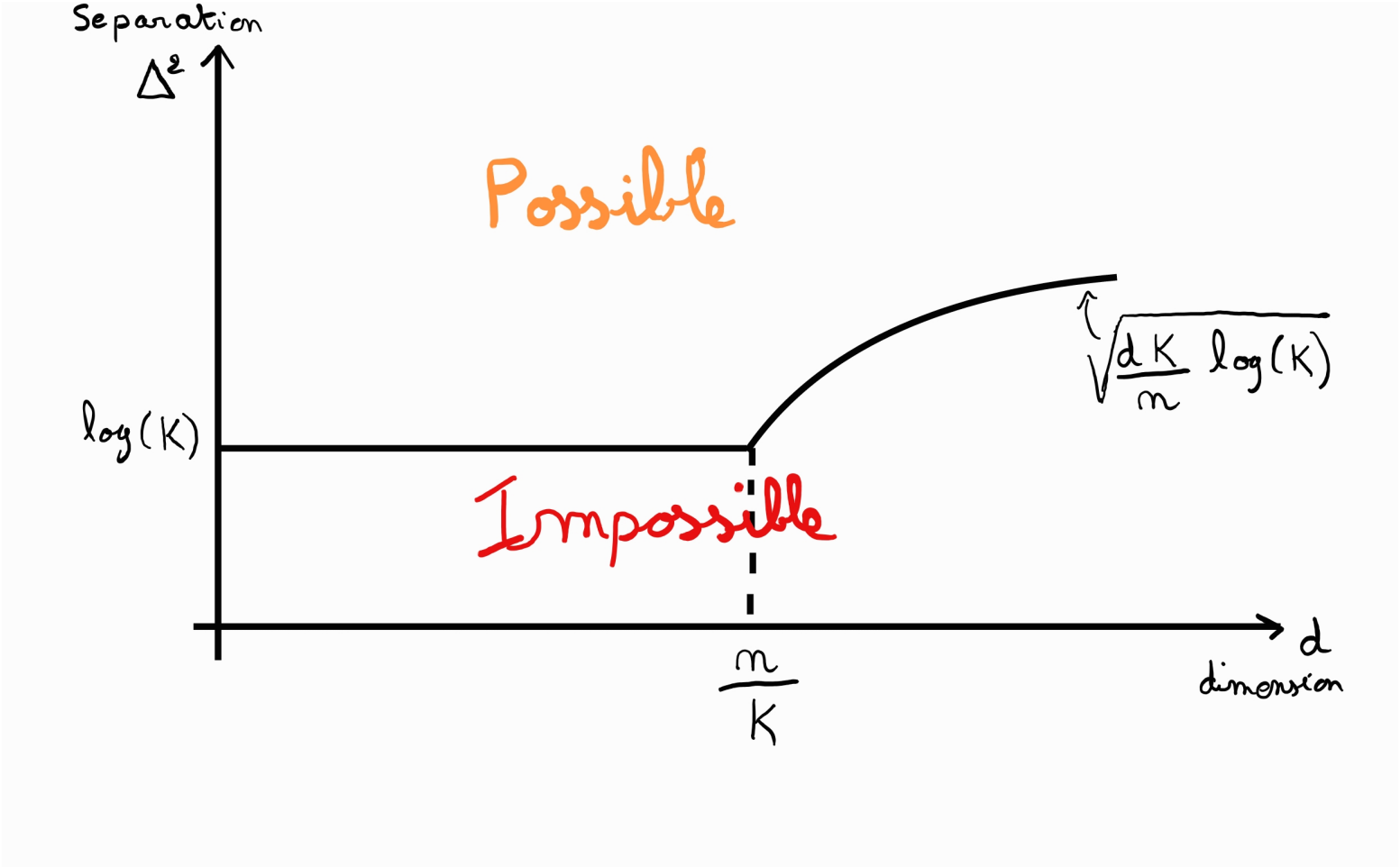
- If $\Delta_0 \ll \Delta_I^*$ clustering is impossible (minimax risk is large)
- If $\Delta_0 \gg \Delta_I^*$ clustering is possible (minimax risk is small)

Information barrier

Theorem (Regev and Vijayaraghavan'17, Kwon and Caramanis '20, EGV'24)

Partial recovery minimax impossible if $\Delta^2 \lesssim \Delta_I^2 := \log(K) \vee \sqrt{\frac{dK}{n}\log(K)}$

possible with exact K*-means if* $\Delta^2 \gtrsim \Delta_I^2$



Exact K-means

Exact K-means

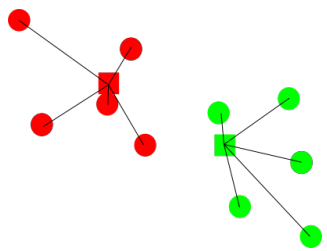
$\widehat{G}_K$ is any global minimizer of

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

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

Remark :

- NP-hard in worst case (Mahajan et al. ('09))
- Possibly many local optima
- In practice, use of local optimization (Lloyds'algorithm (1982))



$\rightsquigarrow$ **Impossible to compute this on large scale problems !**

Some efficient clustering algorithms

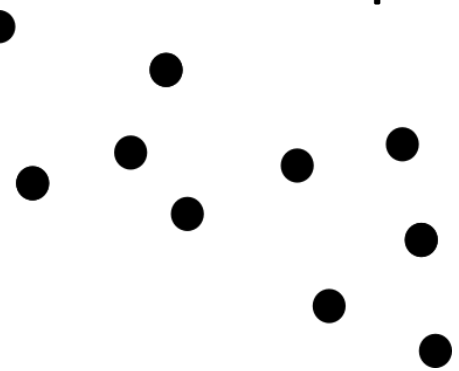
- Lloyd/ EM algorithm
- Spectral Clustering (aka PCA + Lloyd)
- Hierarchical Clustering Algorithms

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

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

Two steps :

1. Compute the centroids
2. Update the partition



Some efficient clustering algorithms

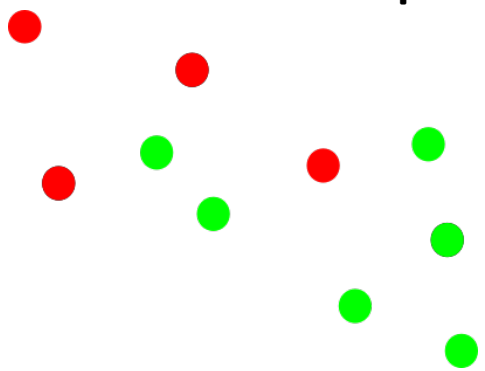
- Lloyd/ EM algorithm
- Spectral Clustering (aka PCA + Lloyd)
- Hierarchical Clustering Algorithms

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

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

Two steps :

1. Compute the centroids
2. Update the partition



Some efficient clustering algorithms

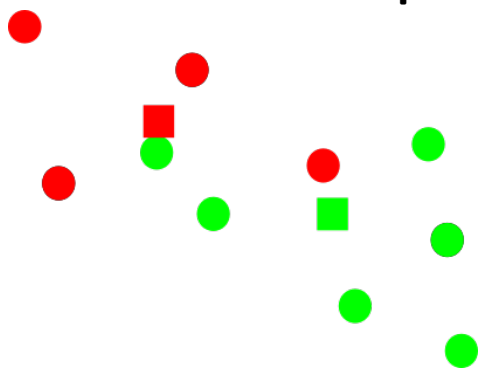
- Lloyd/ EM algorithm
- Spectral Clustering (aka PCA + Lloyd)
- Hierarchical Clustering Algorithms

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

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

Two steps :

1. Compute the centroids
2. Update the partition



Some efficient clustering algorithms

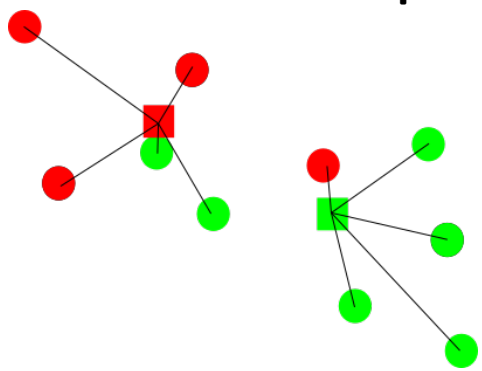
- Lloyd/ EM algorithm
- Spectral Clustering (aka PCA + Lloyd)
- Hierarchical Clustering Algorithms

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

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

Two steps :

1. Compute the centroids
2. Update the partition



Some efficient clustering algorithms

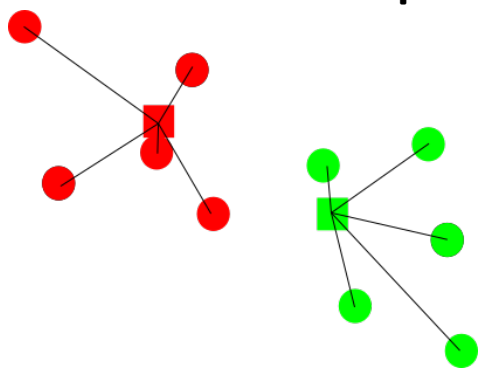
- Lloyd/ EM algorithm
- Spectral Clustering (aka PCA + Lloyd)
- Hierarchical Clustering Algorithms

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

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

Two steps :

1. Compute the centroids
2. Update the partition



Some efficient clustering algorithms

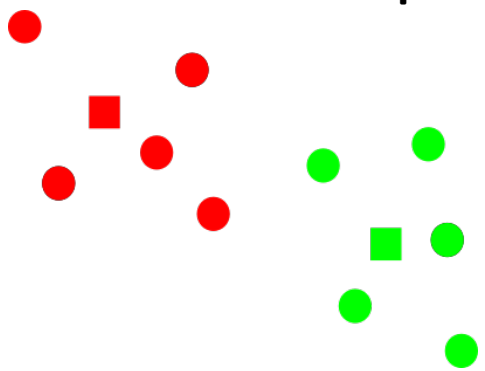
- Lloyd/ EM algorithm
- Spectral Clustering (aka PCA + Lloyd)
- Hierarchical Clustering Algorithms

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

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

Two steps :

1. Compute the centroids
2. Update the partition

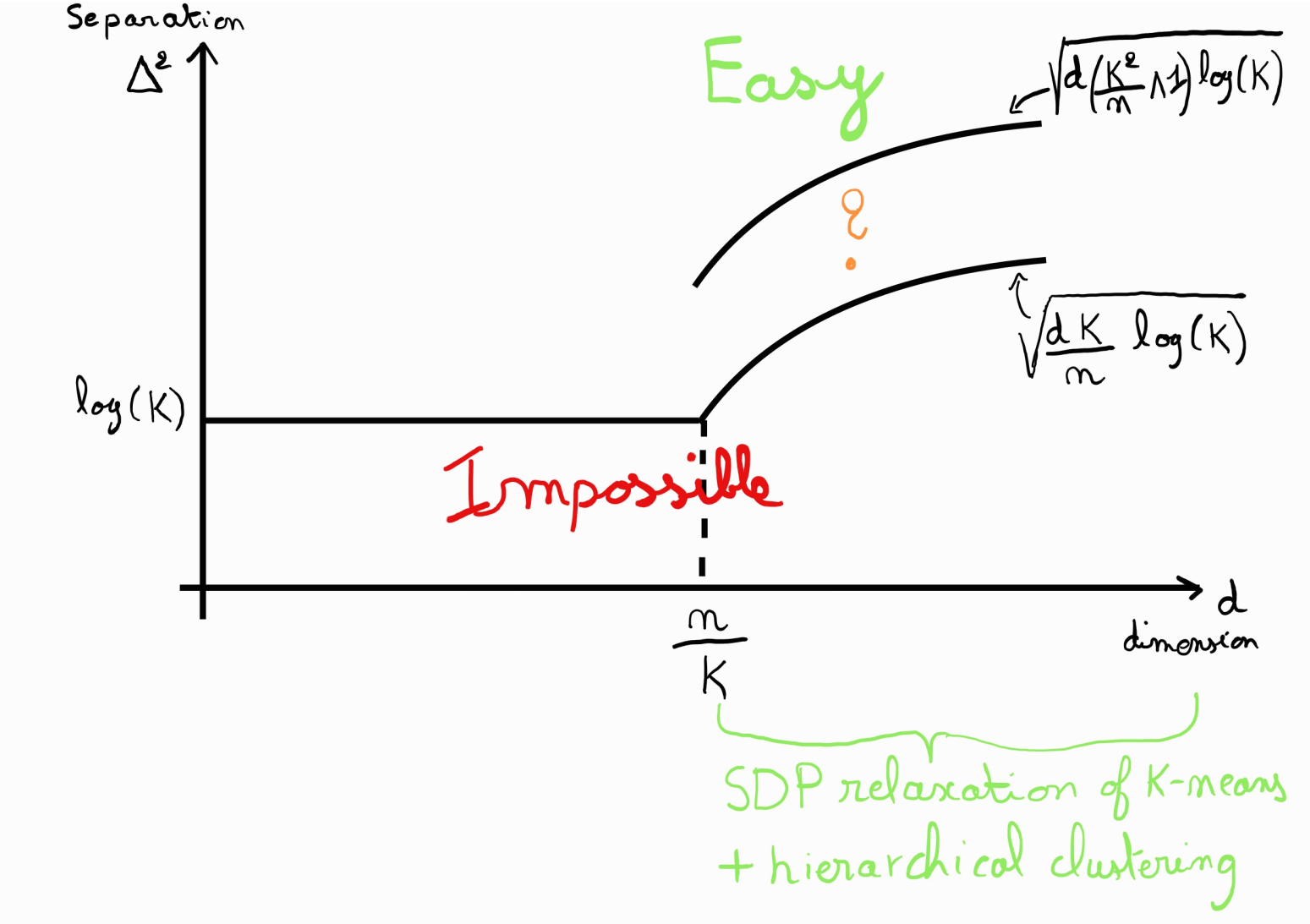


Caveats There can be many local optima. (depends on the initialization)

Many variants depend on the initialization : K-means++ Arthur and Vassilvitskki('07)

Analysis of Spectral and/or Hierarchical Clustering

High-dimensional regime $n \leq dK \vee K^2$.



Statistical Computational Trade-off (SCTO)

All Estimators are not equally relevant !

Computational Efficiency

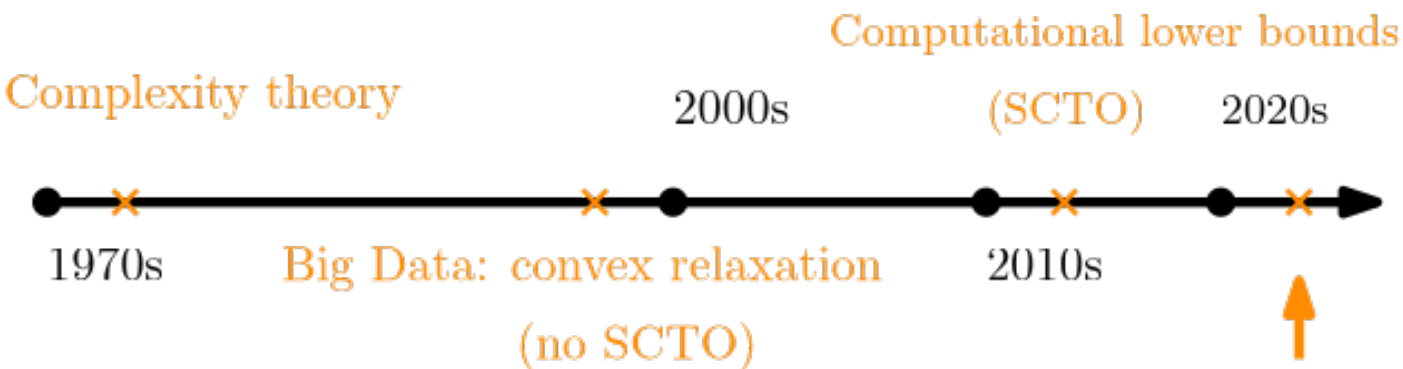
$$\inf_{\hat{G} \text{ efficiently computation.}} \sup_{(G^*,\mu)\in\Theta(\Delta_0)} \mathbb{E}[\mathrm{err}(\hat{G}, G^*)]$$

$\rightsquigarrow$ Boundary Δ^*_{comp}

Computational-Statistical gaps

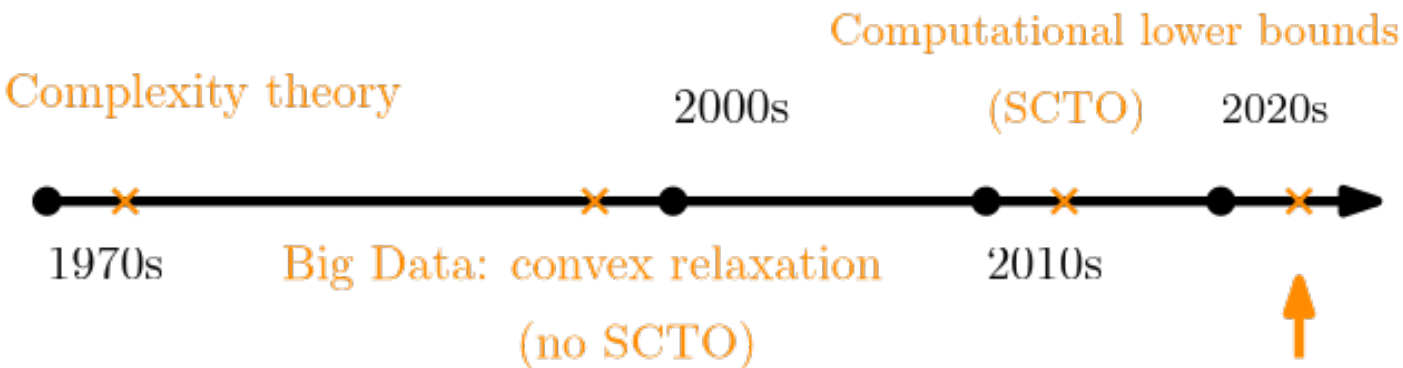
If $\Delta^*_I \asymp \Delta^*_{\mathrm{comp}}$: no SCTO
If $\Delta^*_I \ll \Delta^*_{\mathrm{comp}}$: SCTO

Complexity Theory in Statistics and Machine Learning



Algorithms

- Convex relaxations (Tibshirani '96)
- Spectral methods (Weiss et al.'99)
- Iterative methods (Kilmer et al.'01)
- Hierarchical methods (Murtah et al.'12)
- Tensor-based methods (Liu et al. '23)



Complexity Theory in Statistics and Machine Learning

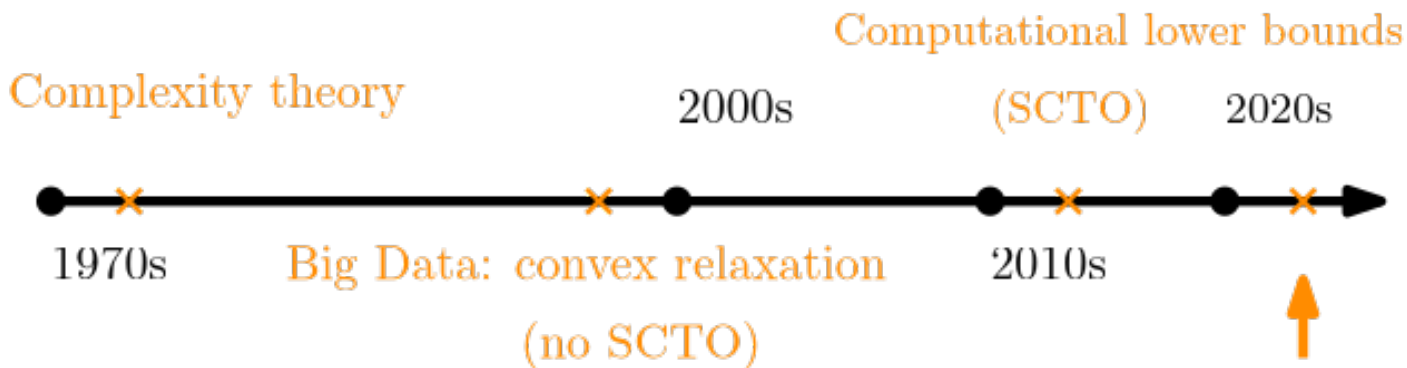
Algorithms

- Convex relaxations (Tibshirani '96)
- Spectral methods (Weiss et al.'99)
- Iterative methods (Kilmer et al.'01)
- Hierarchical methods (Murtah et al.'12)
- Tensor-based methods (Liu et al. '23)

Computational lower bounds

- Reduction (Berthet and Rigollet '13)
- SOS hierarchies (Barak et al. '19)
- SQ algorithms (Feldman et al. '17)
- Local algorithms (MCMC) landscape analysis (Bandeira et al. '22)
- Low-degree polynomials (Hopkins '18)

Connections between these notions (Brennan et al. '21, Bandeira et al. '22)



Low-degree Polynomials for detection (Hopkins'18, Wein'25)

Conjecture/Informal

For 'noisy' problems, if methods based on polynomials on the data, then all polynomial-time methods will fail.

Low-degree Polynomials for detection (Hopkins'18, Wein'25)

Conjecture/Informal

For 'noisy' problems, if methods based on polynomials on the data, then all polynomial-time methods will fail.

Wide class of efficient algorithms approximated by low-degree polynomials :

- spectral methods (power method)

$$\lambda_1(M) \approx \left[\sum_{i=1}^{d} \lambda_i^k(M)\right]^{1/k} = [tr(M^k)]^{1/k}$$

- approximate message passing (AMP)

Low-degree Polynomials for detection (Hopkins'18, Wein'25)

Conjecture/Informal

For 'noisy' problems, if methods based on polynomials on the data, then all polynomial-time methods will fail.

Wide class of efficient algorithms approximated by low-degree polynomials :

- spectral methods (power method)

$$\lambda_1(M) \approx \left[\sum_{i=1}^{d} \lambda_i^k(M)\right]^{1/k} = [tr(M^k)]^{1/k}$$

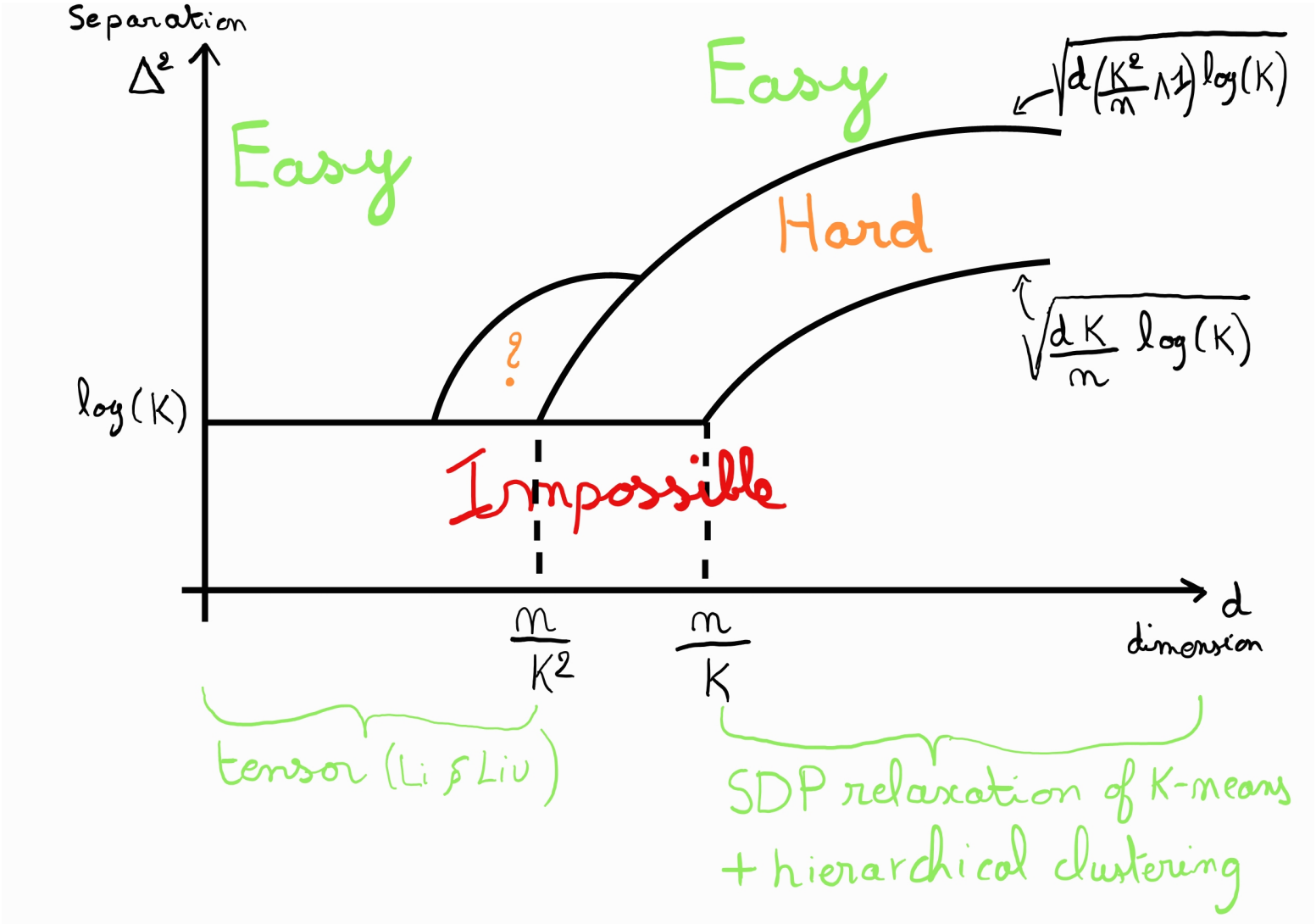
- approximate message passing (AMP)

Other reasons :

- In some settings (e.g. Gaussian additive model), close connections with SQ, SOS, MCMC type lower bounds (Kunisky et al.'18, Bandeira et al. '22)
- Recovers conjectured CSTO for many problems (planted cliques, SBM,...)
- Generic approach for detection (Kunisky et al.'18) and for estimation (Schramm and Wein '22).

Theorem (Informal ; EGV'25+)

Clustering is low-degree polynomial hard if $\Delta^2 \overset{\log^\beta}{\gtrsim} \sqrt{d\left(\frac{k^2}{n} \wedge 1\right)} + 1$



Summary

- CSTO for clustering with many groups
- **Spectral Clustering** and Hierarchical Clustering Algorithms seemingly optimal.
- Similar conclusion holds for SBM Chin et al.'25