Introduction au domaine de recherche: Clustering d'une mixture de Gaussiennes isotropiques.

Bertrand Even

Le Clustering est un problème classique en apprentissage supervisé. Etant donné un ensemble de données observées, il s'agit d'exhiber une structure cachée. Plus précisément, il faut trouver une partition de ces données qui soit pertinente. Dans ce manuscrit, nous allons voir ce problème d'un point de vue inférence statistique sur le modèle simple d'une mixture de Gaussiennes isotropiques.

Le problème de Clustering possède des applications sur des données réelles, notamment en astronomie [1], en biologie [2] et en Computer Vision [2].

Dans ce manuscrit, nous allons;

- Définir le modèle statistique considéré,
- Donner un aperçu de quelques algorithmes étudiés dans la littérature,
- Etudier les limites informationnelles de ce problème,
- Evoquer l'existence d'une barrière computationnelle.

1 Modèle Statistique

Nous allons considérer le modèle simple d'une mixture conditionnelle de Gaussiennes isotropiques. Nous supposons que les données $X_1, \dots, X_n \in \mathbb{R}^p$ sont générés de la manière suivante. Il existe une partition inconnue $G^* = G_1^*, \dots, G_K^*$ de $[1, n]$, des vecteurs inconnus $\mu_1, \dots, \mu_K \in \mathbb{R}^p$, tel que, si $i \in G_k^*$, $X_i \sim \mathcal{N}(\mu_k, I_p)$.

Le but est de retrouver la partition G^* à l'aide de l'observation de $X_1, \dots, X_n$, le nombre de clusters K étant lui aussi connu.

Dans la suite du manuscrit, on suppose que la partition G^* est presque équilibré; on suppose qu'il existe des constantes $\alpha, \beta > 0$ tel que pour tout $k \in [1, K]$, $\alpha \frac{n}{K} < |G_k^*| < \beta \frac{n}{K}$. Les constantes numériques ne pourront dépendre que de α et β.

1.1 Quantification du signal

La quantité qui définit la difficulté du problème est la séparation

$$\Delta := \min_{k \neq l} \|\mu_k - \mu_l\| \ .$$

En effet, plus cette séparation Δ est grande, plus il est aisé de retrouvé la partition. Lorsque cette séparation est nulle, le modèle est non identifiable et il est alors impossible de retrouver la partition G^* avec forte probabilité. A l'inverse, lorsque $\Delta \to \infty$, il est trivial de retrouver G^* seulement en regroupant les points les plus proches entre eux.

On introduit le signal-to-noise ratio s^2 définit par

$$s^2 = \Delta^2 \wedge \frac{\Delta^4 n}{pK} \ .$$

En petite dimension, s correspond à la séparation Δ. En grande dimension, le terme qui va compter sera $\frac{\Delta^4 n}{pK}$ et provient de la difficulté d'estimer les centres des clusters en grande dimension.

1.2 Estimation exacte ou estimation partielle

On peut s'interresser à la fois à l'estimation exacte ou à l'estimation partielle de la partition G^*. Pour l'estimation exacte, il faut trouver les régimes dans lesquels un estimateur donné $\hat{G}$ vérifie $\hat{G} = G^*$ avec grande probabilité.

Pour l'estimation partielle, on souhaite borner le nombre d'erreurs faites par un estimateur $\hat{G}$. Pour cela, on définit l'erreur de misclassifiation correspondant à la proportion de points mal-classifiés

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

ou $\mathcal{S}_K$ désigne l'ensemble des permutations de $[1, K]$ et $G_k^* \Delta \hat{G}_{\pi(k)}$ désigne la différence symmétrique entre les deux ensembles.

1.3 Notation

Pour $i \in [1, n]$, on décompose $X_i = \mathbb{E}(X_i) + E_i = \mu_k + E_i$, si $i \in G_k^*$. Alors, on dispose de $(E_i)_{i \in [1,n]}$ vecteurs indépendants, de loi $\mathcal{N}(0, I_p)$. On pose:

- $X \in \mathbb{R}^{n \times p}$ dont la i-ème ligne est le vecteur X_i,
- $A \in \mathbb{R}^{n \times K}$ la matrice d'appartenance définie par $A_{ik} = \mathbf{1}_{i \in G_k^*}$,
- $E \in \mathbb{R}^{n \times p}$ la matrice de bruit dont le ieme vecteur est E_i,
- $\mu \in \mathbb{R}^{K \times p}$ dont la k-ième ligne est μ_k.

On obtient la décomposion de type signal+bruit

$$X = A\mu + E.$$

Dans tout le manuscrit, on note $\|\cdot\|_q$ la norme L_q d'un vecteur ou d'une matrice. On écrit $\|\cdot\|_{op}$ et $\|\cdot\|_*$ respectivement pour la norme opérateur et la norme nucléaire d'une matrice.

On note $\mathcal{S}_K$ le groupe des permutations de $[1, K]$.

1.4 Une inégalité de concentration bien utile dans la suite

Pour les différentes méthodes de clustering, il sera nécessaire de controler les différentes déviations de certaines fonctions du bruit. Ainsi, il est important de s'armer des bonnes inégalités de concentration.

L'inégalité suivante permet de controler les déviations de formes quadratiques. Il s'agit du lemme d'Hanson-Wright. Pour plus de détails, se réferer notamment au livre [3].

Lemma 1. (Hanson-Wright inequality) *Soit $\varepsilon \sim \mathcal{N}(0, I_d)$ pour $d > 0$. Soit S une matrice réelle symétrique $p \times p$. Alors, pour tout $x > 0$,*

$$\mathbb{P}\left[\varepsilon^T S\varepsilon - \mathrm{Tr}(S) > \sqrt{8\|S\|_F^2 x} \vee (8\|S\|_{op} x)\right] \leq e^{-x}.$$

2 Quelques méthodes de clustering.

2.1 Hierarchical Clustering

Une première méthode est de fusionner les points entre eux petit à petit. On commence avec la partition trivial contenant que des singletons $G^0 = \{\{1\}, \ldots, \{n\}\}$. Ensuite, tant que $G^{(t)}$ contient plus que K groupes, on construit la partition $G^{(t+1)}$ en fusionnant les deux groupes contenant les points les plus proches. L'algorithme s'arrête quand le nombre de groupe dans la partition est K, ce qui arrive pour $t = n - K$.

On définit la fonction de lien entre deux ensembles $A, B \subset [1, n]$ comme $l(A, B) = \min_{(i,j)\in A\times B} \|X_i - X_j\|^2$. On obtient alors l'Algorithme 1.

Algorithm 1: Hierarchical Clustering

Data: $X_1, \ldots, X_n$
$t \leftarrow 0$;
$G^{(0)} \leftarrow \{\{1\}, \ldots, \{n\}\}$;
while $t < n - K$ **do**
  Trouve $\hat{a}, \hat{b}$ minimisant $l\left(G_{\hat{a}}^{(t)}, G_{\hat{b}}^{(t)}\right)$;
  Construit $G^{(t+1)}$ en fusionnant les groupes $G_{\hat{a}}^{(t)}$ et $G_{\hat{b}}^{(t)}$, les autres groupes restant inchangés;
  $t \leftarrow t + 1$;
end
Result: La partition $G^{(n-K)}$.

La proposition suivante donne des conditions pour retrouver exactement la partition G^* avec grande probabilité en utilisant l'Algorithme 1.

Proposition 1. *Il existe des constantes numeriques c_1 et c_2 telles que, si $\Delta^2 \geq c_1\left(\log(n) + \sqrt{p\log(n)}\right)$, l'Algorithme 1 retrouve exactement la partition G^* avec probabilité au moins $1 - \frac{c_2}{n^2}$.*

Proof of Proposition 1. Pour $i \in G_k^*$, on définit $k_i^* = k$. Soit $i \neq j \in [1, n]$. Alors:

$$\|X_i - X_j\|^2 = \|E_i - E_j\|^2 + 2\langle E_i - E_j, \mu_{k_i^*} - \mu_{k_j^*}\rangle + \|\mu_{k_i^*} - \mu_{k_j^*}\|^2 \ .$$

Pour prouver la proposition 1, on prouve que, avec forte probabilité, la quantité au dessus est plus petite quand $k_i^* = k_j^*$ que quand $k_i^* \neq k_j^*$. Pour $i \neq j \in [1, n]$, en utilisant Lemma 1, on dispose d'une constante numérique $c > 0$ telle que, pour tout $x > 0$,

$$\mathbb{P}[|\|E_i - E_j\|^2 - 2p| > c\,(\sqrt{px} + x)] \leq 2e^{-x} \ .$$

Pour $e^{-x} = \frac{1}{n^4}$ et en utilisant une borne d'union pour les coupes $i \neq j$, on obtient

$$\mathbb{P}\left[\forall i \neq j \in [1, n],\ \ |\|E_i - E_j\|^2 - 2p| > 4c\left(\sqrt{p\log(n)} + \log(n)\right)\right] \leq \frac{1}{n^2} \ .$$

Controllons maintenant le terme croisé $\langle E_i - E_j, \mu_{k_i^*} - \mu_{k_j^*} \rangle$ uniformément sur tous les couples $i \neq j \in [1,n]$. Pour de tels $i \neq j \in [1,n]$, $\langle E_i - E_j, \mu_{k_i^*} - \mu_{k_j^*} \rangle \sim \sqrt{2}\|\mu_{k_i^*} - \mu_{k_j^*}\| \mathcal{N}(0, I_p)$. Ainsi, il existe une constante numérique $c' > 0$, telle qu'avec probailité au moins $1 - e^{-x^2}$, on ait $\langle E_i - E_j, \mu_{k_i^*} - \mu_{k_j^*} \rangle \geq -c' x \|\mu_{k_i^*} - \mu_{k_j^*}\|$. Posons $e^{-x^2} = \frac{1}{n^4}$ et faisons une borne d'union sur les couples $i \neq j$, on obtient alors

$$\mathbb{P}\left[\forall i \neq j,\ \langle E_i - E_j, \mu_{k_i^*} - \mu_{k_j^*} \rangle \geq -2c'\sqrt{\log(n)}\|\mu_{k_i^*} - \mu_{k_j^*}\|\right] \geq 1 - \frac{1}{n^2}\ .$$

Alors, avec probabilité au moins $1 - \frac{2}{n^2}$, uniformément sur tous les couples $i \neq j \in [1,n]$, on a les deux inégalités

$$\begin{aligned} \langle E_i - E_j, \mu_{k_i^*} - \mu_{k_j^*} \rangle &\geq -2c'\sqrt{\log(n)}\|\mu_{k_i^*} - \mu_{k_j^*}\| \ ; \\ |\|E_i - E_j\|^2 - 2p| &\leq 4c\left(\sqrt{p\log(n)} + \log(n)\right)\ . \end{aligned}$$

On se restreint à l'événement de probabilité $1 - \frac{2}{n^2}$ sur lequel ces deux inégalités sont satisfaites. Pour $i \neq j \in [1,n]$,

- Si $k_i^* = k_j^*$, alors $\|X_i - X_j\|^2 = \|E_i - E_j\|^2 \leq 2p + 4c\left(\sqrt{p\log(n)} + \log(n)\right)$,
- Si $k_i^* \neq k_j^*$, alors $\|X_i - X_j\|^2 \geq 2p - 4c\left(\sqrt{p\log(n)} + \log(n)\right) - 4c'\sqrt{\log(n)}\|\mu_{k_i^*} - \mu_{k_j^*}\| + \|\mu_{k_i^*} - \mu_{k_j^*}\|^2$.

Donc, si $\Delta^2 \geq c_1\left(\log(n) + \sqrt{p\log(n)}\right)$, pour c_1 une constante numérique choisie assez grande, on a, pour tout $i \neq j$,

- si $k_i^* = k_j^*$, alors $\|X_i - X_j\|^2 \leq 2p + \frac{\Delta^2}{3}$,
- si $k_i^* \neq k_j^*$, alors $\|X_i - X_j\|^2 \geq 2p + \frac{2\Delta^2}{3}$.

Ainsi, pour tout $i \neq j$ tel que $k_i^* = k_j^*$ et $i' \neq j'$ tel que $k_{i'}^* \neq k_{j'}^*$, on a, avec probabilité au moins $1 - \frac{2}{n^2}$, que

$$\|X_i - X_j\|^2 < \|X_{i'} - X_{j'}\|^2\ . \tag{2}$$

En d'autres termes, Algorithme 1 choisira toujours, si cela est possible, de fusionner des groupes qui intersectent le même groupe de G^*. Par récurrence sur $t \in [0, n-K]$, on déduit de cela que $G^{(t)}$ est une sous-partition de G^*, ie que chaque groupe de $G^{(t)}$ est une sous-ensemble d'un groupe de G^* .

Initialisation: La partition $G^{(0)} = \{1\}, \ldots, \{n\}$ est en effet une sous-partition de G^*.

Etape de récurrence: Soit $t \in [0, n-K-1]$ et supposons que $G^{(t)}$ est une sous-partition de G^* et prouvons qu'il en va de même pour $G^{(t+1)}$. Puisque $t \leq n-K-1$, $|G^{(t)}| \geq K+1$. Donc, il existe au moins deux groupes de $G^{(t)}$ qui sont des sous-groupes d'un même groupe de G^*. L'Equation (2) implique que l'Algorithme 1 choisira de fusionner deux tels groupes. Donc, $G^{(t+1)}$ sera aussi une sous-partition de G^*. Cela conclut la récurrence

En particulier, la partition $G^{(n-K)}$ est une sous-partition de G^*. Puisque $|G^{(n-K)}| = K$, on a bien $G^{(n-K)} = G^*$. Cela conclut la preuve de la Proposition 1. □

2.2 Critère K-means

Une manière classique et moins naïve est de regarder le citère K-means d'une partition. Pour une partition $G = G_1, \dots, G_K$, ce critère est définit par

$$\mathrm{Crit}(G) = \sum_{k=1}^{K} \sum_{a \in G_k} \|X_a - \sum_{b \in G_k} \frac{X_b}{|G_k|}\|^2. \tag{3}$$

Ce critère quantifie les déviations des points autour du centre de leur cluster respectif. Dans le cas du modèle de mélange de gaussiennes isotropiques,

2.2.1 L'exact K-means estimator

Regardons dans un premier temps la partition $\hat{G}$ qui minimise le crittère 3. Dans le cas du modèle de mélange de gaussiennes isotropiques, il s'agit du maximum de vraisemblance. Cette partition $\hat{G}$ n'est pas calculable en temps polynomial car il faut calculer le critère de toutes les partitions, ce qui nécessite un nombre exponentiel de calculs. Cependant, il est interressant de regarder les garanties théoriques de cet estimateur, notamment car il est optimal, au sens minimax du terme. Nous allons préciser ce que cela signifie en Section 3.

L'analyse de cet estimateur a été l'un des travaux de mon année de prédoctorat avec Christophe Giraud et Nicolas Verzelen. Vous pouvez la trouver dans `https://arxiv.org/abs/2402.18378`. La proposition suivante donne une borne de l'erreur de clustering en fonction du signal-to noise ratio s^2.

Proposition 2. *Il existe des constantes numériques c, c' et c'' tels que, si*

$$s^2 \geq c \log(K) \ ,$$

alors, avec probabilité plus grande que $1 - \frac{c'}{n^2}$, on a

$$err(\hat{G}, G^*) \leq \exp^{-c'' s^2} \ .$$

En particulier, cette borne implique une borne pour le recouvrement exact de la partition G^*. Si $s^2 \gtrsim \log(n)$, alors $\mathbb{P}[\hat{G} = G^*] \geq 1 - \frac{c'}{n^2}$. En terme de la séparation Δ, on obtient un recouvrement exact de la partition G^* lorsque $\Delta^2 \gtrsim \log(n) + \sqrt{\frac{p}{n} K \log(n)}$. En grande dimension, cela est bien mieux que le seuil $\log(n) + \sqrt{p \log(n)}$ atteint par l'algorithme de hierarchical clustering.

2.2.2 Une formulation alternative du critère K-means

Nous allons formuler de manière matricielle le critère K-means. Cela est utile à la fois pour analyser l'estimateur exact K-means, mais aussi pour introduire un estimateur basé sur la relaxation convexe de l'exact K-means.

Pour $G = G_1, \dots, G_K$ une partition de $[1, n]$, on défintit la matrice $B \in \mathbb{R}^{n \times n}$ par

$$B_{ij} := \sum_{k \in G_k} \mathbf{1}_{i,j \in G_k} \frac{1}{|G_k|} \ .$$

Développons alors le critère K-means 3,

$$
\begin{aligned}
Crit(G) =& \sum_{k=1}^{K} \sum_{a \in G_k} \|X_a - \sum_{b \in G_k} \frac{X_b}{|G_k|}\|^2 \\
=& \sum_{k=1}^{K} \frac{1}{|G_k|} \sum_{a,b \in G_k} \|X_a - X_b\|^2 \\
=& \sum_{k=1}^{K} \frac{1}{|G_k|} \sum_{a,b \in G_k} \|X_a\|^2 + \|X_b\|^2 - 2\langle X_a, X_b \rangle \\
=& 2 \sum_{a \in [1,n]} \|X_a\|^2 - 2 \sum_{a,b \in [1,n]} B_{ab} X X^T_{a,b} \ .
\end{aligned}
$$

Ainsi, minimiser le critère K-means revient à maximiser le produit scalaire $\langle B, XX^T \rangle$ sur l'ensemble des matrices B associées à une partition G. Cet ensemble est explicité dans la proposition suivante, que l'on trouve dans [4].

Proposition 3. *L'ensemble des matrices B associées à une partition G s'écrit*

$$
\mathcal{B} = \{B \in S_n(\mathbb{R})^+ : \ B_{ij} \geq 0, \mathrm{Tr}(B) = K, B1 = 1, B^2 = B\} \ .
$$

Ainsi, minimiser le critère K-means revient à trouver la matrice $\hat{B} \in \mathrm{argmax}_{B \in \mathcal{B}} \langle B, XX^T \rangle$.

2.2.3 Relaxation Convexe

A partir de cela, nous pouvons maintenant introduire un algorithme de relaxation convexe, définit dans [4] et analysé dans [5]. Remarquons que le critère à maximiser $\langle XX^T, B \rangle$ est linéaire. Ainsi, on est en mesure de la maximiser sur un ensemble convexe. L'idée est donc de relacher les contraintes définissant l'ensemble $\mathcal{B}$ afin d'obtenir un ensemble convexe.

La seule contrainte qui n'est pas convexe dans la définition de $\mathcal{B}$ ci-dessus est $B^2 = B$. On introduit alors

$$
\mathcal{C} := \{B \in S_n(\mathbb{R})^+ : \ B_{ij} \geq 0, \mathrm{Tr}(B) = K, B1 = 1\} \ ,
$$

qui est alors un ensemble convexe, en tant qu'intersection d'ensembles convexes.

La solution du problème relaxé est

$$
\hat{B}^{SDP} \in \underset{B \in \mathcal{C}}{\mathrm{argmax}} \langle XX^T, B \rangle \ .
$$

Il n'existe pas forcément une partition G telle que $\hat{B}^{SDP}$ soit associé à G. Il est donc nécessaire de faire une étape en plus afin d'avoir un clustering des données à la fin. Nous passons cet étape sous silence ici car elle est plutôt complexe. Notons cependant $\hat{G}^{SDP}$ la partition obtenue à la fin.

Proposition 4. *Il existe des constantes numériques c, c' et c'' tels que, si $s^2 \geq cK$, alors, avec probabilité supérieure à $1 - \frac{c'}{n^2}$, on a:*

$$
err(\hat{G}^{SDP}, G^*) \leq e^{-c'' s^2} \ .
$$

On obtient une décroissance exponentielle similaire à celle obtenue pour l'estimateur exact K-means. Seulement, ici on a besoin de $s^2 \gtrsim K$ au lieu de $s^2 \gtrsim \log(K)$.

2.2.4 Idée des preuves des propositions 2 et 4

La matrice $\hat{B}$, (respectivement $\hat{B}^{SDP}$), maximise le critère $\langle XX^T, B\rangle$ sur l'ensemble $\mathcal{B}$, (respectivement $\mathcal{C}$). Ainsi, nous avons *de facto* $\langle XX^T, \hat{B}-B^*\rangle \geq 0$ (respectivement $\langle XX^T, \hat{B}^{SDP}-B^*\rangle \geq 0$.) Dans la suite, nous allons écrire $\hat{B}$ indifféremment de $\hat{B}^{SDP}$, la différence résidant dans le controle des termes de bruit. En utilisant la décomposition $X = A\mu + E$, on obtient

$$\langle A\mu(A\mu)^T, B^* - \hat{B}\rangle \leq \langle A\mu E^T + E(A\mu)^T, \hat{B} - B^*\rangle + \langle EE^T - pI_n, \hat{B} - B^*\rangle \ .$$

Que cela soit pour l'analyse de l'estimateur exact K-means ou pour l'analyse de la relaxation convexe, nous suivons la stratégie (i)minorer le terme de signal $\langle A\mu(A\mu)^T, B^* - \hat{B}\rangle$ en fonction de l'erreur de clustering $err(\hat{G}, G^*)$ (ii) majorer uniformément avec grande probabilité sur l'ensemnble $\mathcal{B}$ (respectivement l'ensemble $\mathcal{C}$) le terme de bruit croisé $\langle A\mu E^T + E(A\mu)^T, \hat{B} - B^*\rangle$ (iii) majorer uniformément avec grande probabilité sur l'ensemnble $\mathcal{B}$ (respectivement l'ensemble $\mathcal{C}$) le terme de bruit quadratique $\langle EE^T - pI_n, \hat{B} - B^*\rangle$.

Les points (i) et (ii) sont traités de la même manière pour l'exact K-means et le SDP. La différence se situe dans le point (iii). En effet, pour l'exact K-means, on utilise le fait que l'espace $\mathcal{B}$ soit discret et on effectue des inégalités d'Hanson-Wright 1 combiné avec des bornes d'union. Cependant, pour le programme SDP, il est nécessaire de contrôler la norme opératueur de la matrice $EE^T - pI_n$. De là provient la différence entre les deux bornes obtenues.

2.3 Algorithme de Lloyd

En pratique, les statisticiens utilisent l'algorithme de Lloyd qui est une procédure itérative. Le principe est découpler l'estimation de la partition et l'estimation des centres des clusters.

Algorithm 2: Algorithme de Lloyd

Data: $X_1, \ldots, X_n, G^{(0)}$
$t \leftarrow 0$;
while $t < maxiter$ **do**
  $\mu_k^{(t)} = \frac{1}{|G_k^{(t)}|}\sum_{i\in G_k^{(t)}} X_i$;
  $G_k^{(t+1)} = \{i; k \in \operatorname{argmin} \|X_i - \mu_k^{(t)}\|^2\}$;
  $t \leftarrow t+1$
end
Result: La partition $G^{(t)}$.

Cet algorithme est très peu couteux computationnellement parlant. Cependant, il possède deux contraintes majeures;

- il est nécessaire d'avoir une initialisation qui soit suffisemment bonne pour garantir la convergence de l'algorithme
- il souffre de biais même dans le cas où les covariances sont toutes égales à l'identité. Pour contourner ceci, on peut remplacer dans l'étape de clustering $\operatorname{argmin} \|X_i - \mu_k^{(t)}\|^2$ par $\operatorname{argmin}(e_i - \mathbf{1}_{G_k^{(t)}})^T(XX^T - pI_n)(e_i - \mathbf{1}_{G_k^{(t)}})$.

3 Seuil informationnel

Dans la section précedente, nous avons vu que l'estimateur exact K-means $\hat{G}$ avait des proporiétés statistiques très satisfaisantes. En réalité, ces propriétés sont optimales, tant pour la reconstruction

partielle de G^* que pour la reconstruction exacte.

Soit $\bar{\Delta} > 0$. Posons $\Theta_{\bar{\Delta}}$ l'ensemble des K-uplet $\mu_1, \dots, \mu_K$ tels que $\frac{1}{2}\min_{k\neq l} \|\mu_k - \mu_l\|^2 \geq \bar{\Delta}^2$. Dénotons aussi $\mathcal{G}$ l'ensemble des partitions de $[1, n]$ qui sont presque équilibrés ($n/K - 1 \geq |G_k^*| \leq n/K + 1$ par exemple). Nous allons établir des bornes minimax à la fois pour la reconstruction partielle et pour la reconstruction exacte.

Pour la reconstruction partielle, le risque minimax est

$$\inf_{\hat{G}} \sup_{G\in\mathcal{G}} \sup_{\mu\in\Theta_{\bar{\Delta}}} \mathbb{E}_{\mu,G}[err(\hat{G}, G^*)] \ .$$

Pour la reconstruction exacte, le risque minimax est

$$\inf_{\hat{G}} \sup_{G\in\mathcal{G}} \sup_{\mu\in\Theta_{\bar{\Delta}}} \mathbb{P}_{\mu,G}[\hat{G} \neq G^*] \ .$$

Grace à la proposition 2, nous savons que lorsque $\bar{\Delta}^2 \gtrsim \log(n) + \sqrt{\frac{p}{n} K \log(n)}$, alors

$$\inf_{\hat{G}} \sup_{G\in\mathcal{G}} \sup_{\mu\in\Theta_{\bar{\Delta}}} \mathbb{P}_{\mu,G}[\hat{G} \neq G^*] \to 0 \ .$$

Il en va de même pour le risque minimax de la reconstruction partielle lorsque $\bar{\Delta}^2 \gtrsim \log(K) + \sqrt{\frac{p}{n} K \log(K)}$.

Pour obtenir des bornes inférieures sur de telles quantités, l'outil principal est le lemme de Fano. Celui ci s'appuie sur des bornes de la divergences KL entres différentes distribution. Voici son énoncé. On rappelle que, étant donné deux probabilités $\mathbb{P} << \mathbb{Q}$, la divergence de Kullback Leibler entre $\mathbb{P}$ et $\mathbb{Q}$ est définit par

$$KL(\mathbb{P}, \mathbb{Q}) := \mathbb{E}_{\mathbb{Q}}\left[\log \frac{d\mathbb{P}}{d\mathbb{Q}}\right] \ .$$

Proposition 5. *Soit $(\mathbb{P}_j)_{j\in[1,N]}$ des probabilités sur un même espace $\mathcal{X}$ et $\mathbb{Q}$ une probabilité sur cet espace $\mathcal{X}$ telle que $\mathbb{P}_j << \mathbb{Q}$. On a*

$$\min_{\hat{j}:\mathcal{X}\to[1,N]} \frac{1}{N}\sum_{j=1}^{n} \mathbb{P}_j[\hat{j}(X) \neq j] \geq 1 - \frac{1 + \frac{1}{N}\sum_{j=1}^{n} KL(\mathbb{P}_j, \mathbb{Q})}{\log(N)} \ .$$

Dans cette section, nous allons détailler la stratégie pour obtenir une borne inférieure sur le risque minimax pour la reconstruction exacte. Cette méthode est assez classique. Nous allons cependant passer sous silence les détails techniques de calcul. Soit $\hat{G}$ un estimateur et $S = \{G^{(0)}, \dots, G^{(N)}\} \subset \mathcal{G}$. On définit $\hat{j} = j$ si $\hat{G} = G^{(j)}$ et $\hat{j} = 0$ sinon. Alors, pour tout $G^{(j)} \in S$ et $\mu \in \Theta_{\bar{\Delta}}$, $\mathbb{P}_{\mu,G^{(j)}}[\hat{G} = G^{(j)}] \leq \mathbb{P}_{\mu,G^{(j)}}[\hat{j} = j]$. Ainsi, il est suffisant de minorer $\inf_{\hat{j}} \frac{1}{N+1}\sum_{j=0}^{N} \mathbb{P}_{\mu,G^{(j)}}[\hat{j} \neq j]$. Pour cela, notre stratégie est de trouver un ensemble $S \subset \mathcal{G}$ et des vecteurs $\mu_1, \dots, \mu_K \in \Theta_{\bar{\Delta}}$ tel que:

- Le cardinal de S soit suffisemment grand.
- On puisse borner uniformément $KL(\mathbb{P}_{\mu,G}, \mathbb{P}_{\mu,G^{(0)}})$ pour $G \in S$.

Prenons $G^{(0)} \in \mathcal{G}$ et prenons S l'ensemble des partitions obtenues en permutant deux points de deux groupes différents consécutifs de $G^{(0)}$. Alors, $\log(|S|)$ est de l'ordre de $\log(n)$, à constante multiplicative près. Pour conclure en utilisant le lemme de Fano, il suffit de borner $KL(\mathbb{P}_{\mu,G}, \mathbb{P}_{\mu,G^{(0)}})$ par $c\log(n)$, avec c que l'on peut devra choisir assez petit.

Supposons que $\bar{\Delta}^2 \lesssim \log(n) + \sqrt{\frac{p}{n} K \log(n)}$ et distingons deux cas afin de borner $KL(\mathbb{P}_{\mu,G}, \mathbb{P}_{\mu,G^{(0)}})$.

3.1 Cas $\bar{\Delta}^2 \lesssim \log(n)$

Il faut choisir les vectueurs $\mu_1, \dots, \mu_K \in \Theta_{\bar{\Delta}}$. Prenons e un vectuer unitaire quelquonque et définissons $\mu_k = k\bar{\Delta}e$. Alors, pour une partition $G \in S$, seul deux points auront une distributions marginales différentes de la distribution marginale correspondante à $G^{(0)}$ et, la distance entre les centres des gaussiennes correpondantes sera de $\bar{\Delta}$. En utilisant l'additivité des divergences KL de variables indépendantes et la divergence Kl entre deux gaussiennes de même variance $KL(\mathcal{N}(\mu_1, I_p), \mathcal{N}(\mu_2, I_p)) = \frac{\|\mu_1 - \mu_2\|^2}{2}$, on obtient

$$KL(\mathbb{P}_{\mu,G}, \mathbb{P}_{\mu,G^{(0)}}) = \bar{\Delta}^2 \lesssim \log(n) \ .$$

3.2 Cas $\log(n) \lesssim \bar{\Delta}^2 \lesssim \sqrt{\frac{p}{n} K \log(n)}$

Dans ce cas là, nous allons choisir aléatoirement les $\mu_1, \dots, \mu_K$ selon la distribution ρ suivante. Les μ_k seront tirés indépendemment et uniformément sur l'hypercube $\mathcal{E} := \{-\varepsilon, \varepsilon\}^p$, où $\varepsilon := \sqrt{\frac{2}{p}}\varepsilon$. La condition $\log(n) \lesssim \bar{\Delta}^2 \lesssim \sqrt{\frac{p}{n} K \log(n)}$ implique $p \gtrsim \frac{n}{K} \log(n)$ et donc, avec grande probabilité, $\mu_1, \dots, \mu_K \in \Theta_{\bar{\Delta}}$.

Ensuite, quelques passages techniques permettent de borner la diverence KL sous ce choix de distribution ρ, pour toute parition $G \in S$ par

$$KL(\mathbb{P}_{\mu,G}, \mathbb{P}_{\mu,G^{(0)}}) \lesssim \log(n) \ .$$

3.3 Conclusion

Ces bornes permettent d'obtenir la proposition suivante qui est une borne inférieure informationelle.

Proposition 6. *Il existe des constantes numérique c, C et K_0 tels que, si $\bar{\Delta}^2 \leq c\left(\log(n) + \sqrt{\frac{p}{n} K \log(n)}\right)$, alors*

$$\inf_{\hat{G}} \sup_{G \in \mathcal{G}} \sup_{\mu \in \Theta_{\bar{\Delta}}} \mathbb{P}_{\mu,G}[\hat{G} \neq G^*] \geq C \ .$$

Or, si $\bar{\Delta}^2 \gtrsim \log(n) + \sqrt{\frac{p}{n} K \log(n)}$, l'estimateur exact K-means est capable de retrouver exactement toute partition G avec grande probabilité. Ainsi, nous avons montré qu'à constante multiplicative près, l'estimateur exact K-means est optimal pour la reconstruction exacte, au sens minimax du terme.

En réalité, il en va de même pour la reconstruction partielle en dimension $p \gtrsim \log(K)$. On peut en effet montrer que si $\bar{\Delta}^2 \lesssim \log(K) + \sqrt{\frac{p}{n} K \log(K)}$, le risque minimax pour la recontruction partielle est minorée par une constante numérique C.

4 Vers un gap computationnel

Dans cette section, nous négligeons tous les termes multiplicatifs logarithmiques dans nos ordres de grandeur.

L'estimateur exact K-means est optimal. On a un seuil informationnel pour la reconstruction de la partition G^* de l'ordre de $\Delta^2 \gtrsim 1 + \sqrt{\frac{pK}{n}}$. Cet estimateur n'est malheureusement pas calculable en temps polynomial.

En grande dimension, on ne connait aucun estimateur calculable en temps polynomial qui atteigne ce seuil-ci. Au lieu de $\Delta^2 \gtrsim \sqrt{\frac{pK}{n}}$, les meilleurs seuils atteints par des algorithmes calculables en

temps polynomial, comme le hierarchical clustering ou le programme de relaxation SDP, sont au moins de l'ordre $\Delta^2 \gtrsim \sqrt{p}\left(1 \wedge \sqrt{\frac{K^2}{n}}\right)$.

Ainsi, on conjecture qu'il existe un gap statistique/computationnel en grande dimension ($p \geq \frac{n}{K^2}$). C'est à dire que l'on pense qu'il existe une région, $\sqrt{\frac{pK}{n}} \lesssim \Delta^2 \lesssim \sqrt{p}\left(1 \wedge \sqrt{\frac{K^2}{n}}\right)$, dans laquelle il existe des estimateurs capables de retrouver la partition G^* avec grande probabilité mais dont aucun d'entre eux ne soit calculable en temps polynomial.

Il existe différentes approches afin d'étayer ce genre de phenomènes. L'une d'entre elles est de considérer certaines classes d'algorithmes et de montrer des bornes inférieures sur ces classes. La méthode dite de faible degré consiste à exhiber un régime dans lequel aucun polynome de faible degré (degré logarithmique) n'est capable d'estimer mieux que trivialement la partition cachée.

Dans `https://arxiv.org/abs/2402.18378`, nous nous basons sur cette méthode afin d'obtenir une borne inférieure. On obtient qu'en grande dimension ($p \geq n$), aucun estimateur polynomial de degré $\log(n)^2$ n'est capable d'estimer mieux que trivialement si les points 1 et 2 sont dans le même groupe lorsque $\Delta^2 \lesssim \sqrt{p}\left(1 \wedge \sqrt{\frac{K^2}{n}}\right)$.

Ce genre de résultat ne permet pas de prouver rigoureusement un gap computationnel. Cependant, il permet de donner des éléments de réponse qui vont dans ce sens là.

References

1. Ordovás-Pascual, I. & Almeida, J. S. A fast version of the k-means classification algorithm for astronomical applications. *Astronomy and Astrophysics* (2014).
2. Herwig, R., Albert Poustka, C. M., Christophe Bull, H. L. & O'Brien, J. Large-Scale Clustering of cDNA-Fingerprinting Data. *National Library of Medecine* (1999).
3. Giraud, C. *Introduction to high-dimensional statistics* xvi+255 (CRC Press, Boca Raton, FL, 2015).
4. Peng, J. & Wei, Y. Approximating K-means-type Clustering via Semidefinite Programming. *SIAM J. on Optimization* **18,** 186–205 (Feb. 2007).
5. Giraud, C. & Verzelen, N. Partial recovery bounds for clustering with the relaxed Kmeans. *arXiv:1807.07547* (2018).