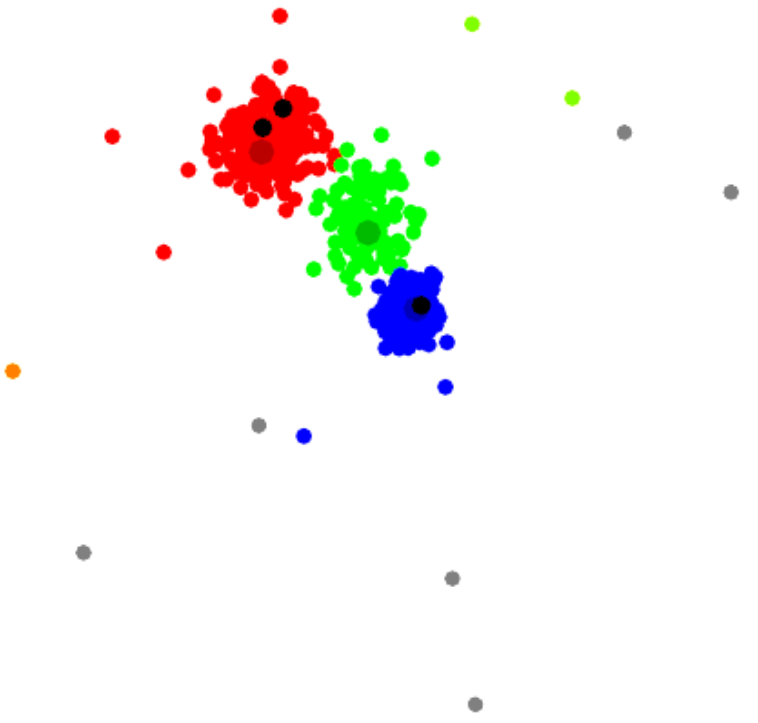
Rapport de stage recherche

Encadrement : Adrien SAUMARD

Structure : ENSAI, Campus de Ker Lann, Bruz
du 15 mars au 24 août 2018



Edouard GENETAY
genetay.edouard@gmail.com

Master 2 de Mathématique Fondamentale et Appliquée
UFR Sciences Nantes, 2017-2018

Remerciements

Je tiens, avant toutes choses, à remercier Adrien SAUMARD de m'avoir proposé un stage et un sujet de thèse alors que nous ne nous connaissions pas plus qu'à travers mon séminaire étudiant que j'ai fait en distantiel. Merci.

Merci également de m'avoir bien encadré pendant ce stage : je remarque, grâce à ce rapport, que j'ai beaucoup progressé, je lui dois en partie.

Merci à Amandine DUBOIS, stagiaire avec laquelle je partage les journées et le bureau. On a traversé les moments stressants de recherche de financement ensemble. Son soutien a été certain.

Enfin merci à Rémi COULAUD, stagiaire également à l'ENSAI, qui a assidument relu mon rapport et conseillé pendant sa rédaction.

Préambule

Ce rapport de stage est non seulement la conclusion de mon année de master 2 de Mathématique Fondamentale et Appliquée, mais aussi la conclusion du master que j'ai eu la chance d'intégrer après mon école d'ingénieur. Il marque du même coup le début du travail de recherche qui m'attend : j'ai la chance de pouvoir continuer la rechercher en thèse sous la direction de Valentin PATILEA et la co-direction de Adrien SAUMARD à l'ENSAI.

Ce rapport fait essentiellement l'état des lieux des concepts et notions importantes de mon futur sujet de thèse. Pour cette raison, je vais consacrer ces pages à décrire et expliquer comment ces notions s'articulent entre elles et dans quel contexte je les rencontre. Pour cette raison, j'écrirai peu de preuve et me contenterai de la dernière partie pour démontrer des résultats intéressants afin de mettre l'accent sur quelques techniques propres à l'étude des queues de distribution des variables aléatoires.

J'espère que le format de ce rapport plaira. Peut-être paraîtra-t-il long aux non probabilistes : sachez que j'ai recherché la clarté des explications pour que tout mathématicien puisse le lire. Peut-être paraîtra-t-il pauvres en preuve aux probabilistes : sachez que c'est surtout des contraintes de temps qui m'ont imposé ce format. Ces 16 semaines de stage se sont partagées en lecture d'article, de livre, en programmation informatique, en recherches personnelles, en séminaires, en recherche de financement et en rédaction de mémoire.

Ce rapport s'organise en 4 parties. La première présente la démarche et les problématiques en statistique théorique à travers la notion d'estimateur : notion transversale, allant des statistiques paramétriques à l'apprentissage statistique. La deuxième partie présente un estimateur particulier et central dans mon futur sujet de thèse : l'estimateur Median-of-Means, aussi appelé MOM. Dans cette partie on ira plus loin que l'estimation de la moyenne présentée en partie 1. La troisième partie décrit l'estimateur de classification le plus classique de l'apprentissage statistique : $k-means$, de sa définition et ses défauts à son amélioration à l'aide de MOM. Enfin, la quatrième et dernière partie présente des preuves incontournables en apprentissage statistique. Les techniques qui y sont présentées sont une partie de ce que j'ai dû maîtriser pour comprendre les articles du domaine. Je devrais, à termes, être autonome et savoir les mettre en oeuvre dans mes propres preuves.

Lecteurs, lectrices, bonne lecture.

1 Introduction à la démarche des statisticiens

Lors de mon stage, j'ai essentiellement cherché à comprendre le cadre des statistiques théoriques, notamment celui de l'apprentissage statistique. Dans cette introduction, je vais décrire les objectifs de la discipline et les moyens à leur disposition pour juger de l'intérêt d'un objet ou d'une méthode.

1.1 Comment comparer des estimateurs statistiques

En statistique inférentielle, on se place dans le cadre d'un modèle et on cherche à estimer les paramètres, à priori inconnus, de ce dernier à partir des observations, pour modéliser concrètement le phénomène en jeu. Pour ce faire, on utilise des fonctionnelles des observations appelées "estimateur". Un estimateur est une fonction qui prend en argument plusieurs variables aléatoires, disons n variables aléatoires (elles représentent les observations), et qui renvoie un objet mathématique : un nombre [1], une loi [2] ou encore des fonctions [3]. Tout l'enjeu des statistiques est de construire des estimateurs de sorte que l'objet renvoyé soit bien ce que l'on cherche à savoir.

De plus, on n'est pas à l'abri d'observer un évènement rare, si bien que l'estimation dans ce cas précis sera biaisée, du fait de la sur-représentation de l'événement rare dans les observations. Donc a priori, il faut beaucoup d'observations du phénomène pour mieux estimer [4], mais cela induit souvent un coup : difficile d'observer, complexité algorithmique trop grande, etc.

Cela va alors pouosser le statisticien à choisir un estimateur plutôt qu'un autre. Le choix se fait en fonction des propriétés théoriques des estimateurs, on peut par exemple citer pèle-mêle : l'exhaustivité, la minimalité, la complétude/liberté, le biais, le risque quadratique, la consistance, la normalité asymptotique, la vitesse de convergence, le maximum de vraisemblance, la robustesse [5], la complexité algorithmique, etc. On va mettre l'emphase ici sur la répartition statistique des valeurs que des estimateurs peuvent prendre autour de leur espérance. On appelle cela "la concentration" des valeurs autour de l'espérance. La concentration est exprimée en termes de probabilité : étant donné un intervalle I centré sur l'espérance, quelle est la probabilité que l'estimateur T tombe dans I sachant qu'il prend en argument n observations distribuées comme X. La concentration n'est pas l'unique propriété mais elle est la propriété centrale dans mon stage.

Prenons quelques lignes pour illustrer ce que l'on vient de dire. Tout estimateur de la moyenne est censé être construit de telle façon à renvoyer une approximation de l'espérance du phénomène. Ce qui signifie que l'on s'attend au moins à ce que l'espérance de l'estimateur vis-à-vis des données soit l'espérance du phénomène. On appelle X la variable aléatoire que l'on observe, $\mathcal{L}$ sa loi et μ son espérance. La quantité que l'on cherche à déterminer à partir des observations [6] $(x_i)_{1\leqslant i\leqslant n}$ de X est μ.

1. C'est le cas pour de la moyenne, la variance ou tout autre paramètre d'une loi
2. c'est le cas en statistique non paramétrique
3. comme dans le cas des fonctions de régression
4. C'est le principe de Monte Carlo : estimer la loi d'un phénomène en faisant tendre le nombre d'observations vers l'infini
5. Pour plus de détails sur ces propriétés des estimateurs, se référer au cours de Frédéric PROÏA [11] et à "Robust Statistics" [5]
6. une observation de X est noté x. Si l'on fait plusieurs observations de X, on suppose souvent que ces dernières sont indépendantes entre elles. C'est pourquoi, sous cette hypothèse d'indépendance entre observations, on considère

On veut alors au minimum avoir un estimateur t de l'espérance de X tel que $\mathbb{E}(T[(X_i)_{1\leqslant i\leqslant n}]) = \mu$. Mais avoir un estimateur d'espérance μ ne suffit pas. Il y a en effet plusieurs façons d'estimer une moyenne μ. Par exemple :

1) à partir de n observations, on peut définir T_1 qui va tirer au hasard de manière uniforme une observation
2) à partir de n observations $(x_i)_{1\leqslant i\leqslant n}$, on peut définir T_2, l'estimateur qui renverra $\frac{1}{n}\sum_{i=1}^{n} x_i$
3) à partir de n observations $(x_i)_{1\leqslant i\leqslant n}$, pour $k \in \mathbb{N}$, on peut définir T_3, l'estimateur qui renverra $\text{Med}\left\{ \frac{1}{|B_i|}\sum_{x\in B_i} x : \ (B_i)_{1\leqslant i\leqslant k} \text{ partition } (x_i)_{1\leqslant i\leqslant n} \right\}$
4) ...

Essayons alors de comparer ces façon d'estimer la moyenne. Si les valeurs possibles des observations sont bornées par M, alors un calcul rapide montre que la concentration dans chacun des cas ci-dessus est au moins tel que :

1) $\mathbb{P}\left(T_1[(X_i)_{1\leqslant i\leqslant n}] \notin I\right) = \mathbb{P}\left(X_1 \notin I\right)$
2) $\mathbb{P}\left(T_2[(X_i)_{1\leqslant i\leqslant n}] \notin I\right) \leqslant 2e^{-\frac{n\lambda(I)^2}{4M^2}}$, où λ est la mesure de Lebesgue
3) $\mathbb{P}\left(T_3[(X_i)_{1\leqslant i\leqslant n}] \notin I\right) \leqslant e^{-\frac{n\lambda(I)^2}{256\sqrt{3}M}}$, où λ est la mesure de Lebesgue

On voit alors que la méthode 1 se fait disqualifier du simple fait que pour cette façon d'estimer, la précision de l'estimation n'augmente pas avec le nombre d'observations. Cela vient du fait que l'on utilise pas l'ensemble de l'information de l'échantillon pour estimer. Enfin, les estimateurs 2 et 3 convergent presque sûrement vers l'espérance de X quand le nombre d'observations augmente. On dit qu'il sont fortement consistants. Pour les comparer plus finement, on peut écrire des inégalités pour choisir celui qui vérifiera la plus fine des inégalités. On peut notamment voir cela en prenant $I := I_n$ telle que $\lambda(I_n) = n^{-1/4}$. Si bien que $I \underset{n\longrightarrow\infty}{\longrightarrow} \{\mu\}$, et que $\mathbb{P}\left(T_{2 \text{ ou } 3}[(X_i)_{1\leqslant i\leqslant n}] \notin I_n\right) \underset{n\longrightarrow\infty}{\longrightarrow} 0$. C'est bon signe. On a la même vitesse de convergence en fonction du nombre n de données. On pourrait continuer à les comparer pour voir s'il reste consistant sur des familles de loi de probabilité plus larges, ou en regardant leur sensibilité aux outliers [7].

L'estimateur T_3 est appelé estimateur median-of-means et est un estimateur optimal de la moyenne (meilleur même que l'estimateur T_2 appelé "estimateur moyenne empirique") pour des variables aléatoires X de variance bornée, d'après l'article *Sub-gaussian means estimators* de Luc Devroye, Matthieu Lerasle, Gabor Lugosi et Roberto I. Oliveira [4].

1.2 Qu'en est-il en apprentissage statistique ?

En apprentissage statistique, il s'agit d'automatiser une tâche. Il convient dans un premier temps

de manière équivalente que l'on a observé plusieurs variables aléatoires $X_1, X_2, \ldots$ indépendantes entre elles et identiquement distribuées (i.i.d.) de même loi que X. Par suite, par cohérence de notation, on note $x_1, x_2, \ldots$ les observations correspondantes

7. outlier = observation aberrante. Notons au passage que nous avons évoqué plus haut les "événements rares" et que l'on parle maintenant "d'outliers". Un événement rare est une observation très peu probable d'un phénomène et un outlier est une observation malencontreuse d'un autre phénomène, différent de ce que l'on pensait observer initialement. Dans la pratique, ces deux cas sont très difficiles à discerner et c'est l'un problème majeur de l'apprentissage statistique

d'analyser la tâche à accomplir pour déterminer si l'on cherche à explorer, expliquer, prévoir ou sélectionner, si l'on en croit la page *apprentissage statistique* de *WikiStat* [12]. Ainsi, guider par ces objectifs, on construit un estimateur adapté. Dans le cadre de mon stage, j'ai exclusivement étudier des estimateurs de classification, plutôt utiles en exploration de données et en prédiction, me semble-t-il.

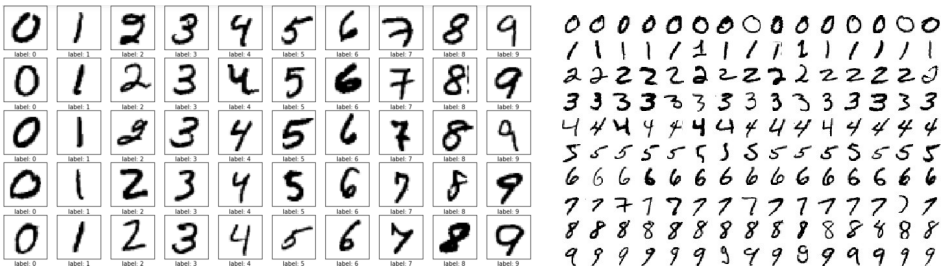


FIGURE 1 – Extraits de la base de données MNIST. A gauche un extrait de données labellisées et à droite un extrait de données non-labellisées.

Prenons une situation du quotidien où l'on rencontre ce problème de classification. On a tous acheté des fruits au supermarché. Petit, nos parents nous apprennent à reconnaître les fruits, aussi bien en photo (vignette sur la balance électronique) qu'en chair et en pépin dans l'étalage. Dans nos esprits, la catégorie de poire est très claire, quelle que soit la situation où l'on rencontre le concept de "poire". C'est pour cela que l'on sait peser ses poires sans se tromper. Depuis quelques années, certains supermarché ont investi dans des balances électronique munit de capteurs afin de prédire la nature de l'article posé sur la balance et de nous faire gagner du temps. Ces balances ont essentiellement besoin d'être capable de proposer à l'utilisateur la vignette correspondant aux fruits : c'est de la prédiction à l'aide d'un algorithme de classification. Il convient alors au préalable d'entraîner sur des exemples un l'algorithme de classification, comme $k-means$, $randomforestclassifier$ ou encore une algorithme de régression logistique. Quand les prédictions de l'algorithme sont suffisamment bonnes [8], on met les balances en service. On peut voir que le problème est le même dans la reconnaissance des chiffres manuscrits (voir Fig.1) [9] [10]. L'apprentissage basé sur des exemples s'appelle "apprentissage statistique".

k-means est un de ces estimateurs de classification. Il fait parti des algorithmes d'apprentissage statistique dit "non supervisé" car il s'appuie sur des données non labellisées comme dans la partie droite de la figure ci-dessus [11]. L'objectif des utilisateurs de $k-means$ est rendre compte de la structure des données en partitionnant l'espace en différentes cellules, tantôt appelées "cellule de voronoï", tantôt appelées "clusters".

8. Typiquement, quand le pourcentage d'erreur satisfait les utilisateurs de l'algorithme

9. Image de gauche de la figure 1 prise sur le site `http://blog.welcomege.com/mnist-database/` consulté le 2 juillet 2018

10. Image de droite de la figure 1 prise sur le site `https://en.wikipedia.org/wiki/MNIST_database` consulté le 2 juillet 2018

11. Un exemple de base de données labellisées étant donné Fig.1 à gauche. L'apprentissage supervisé consiste à entraîner un algorithme à former des catégories à partir de données que l'humain à préalablement catégorisées. La seconde possibilité est l'emploi d'un algorithme qui construira des catégories ex nihilo en explorant les données : on appelle cela l'apprentissage non supervisé

Une idée intuitive et communément admise est de dire qu'un "bon" cluster est un ensemble de points homogènes et tel que la variance dans les cluster est faible. C'est pourquoi *k-means* est un estimateur qui prend en argument les observations dans un espace normé séparable $(\mathbb{X}, \|\cdot\|_p)$, pour $p \geqslant 1$ et renvoie k points de $\mathbb{X}$ tels que la variance de chacun des clusters associé à chacun des k points soit minimale. *k-means* est défini comme suit :

$$kmeans : \begin{cases} \bigcup_{n\in\mathbb{N}} \mathbb{X}^n & \longrightarrow & \mathbb{X}^k \\ (x_i)_{1\leqslant i\leqslant n} & \longmapsto & \underset{(c_j)_{1\leqslant j\leqslant k}\in\mathbb{X}^k}{\operatorname{argmin}} \sum_{i=1}^{n} \min_{1\leqslant j\leqslant k} \|x_i - c_j\|_p^p \end{cases} \tag{1}$$

Les c_j correspondrait aux points centrale du cluster. Ma définition de *k-means* est toutefois abusivement générale : d'ordinaire, $\mathbb{X}$ est $\mathbb{R}^d$ et d est $\|\cdot\|_2$, de sorte que *k-means* renvoie bien la somme des variances [12] dans chaque cluster.

Cette définition permet d'étudier rigoureusement les propriétés de cet estimateur. On voit facilement que la valeur renvoyée par le *k-means* usuel dépend d'avantage des points distants des centres que des points proches des centres. Ce phénomène rend *k-means* sensibles aux "outliers". Ce phénomène est d'autant plus prononcé que p est grand. De plus, l'opérateur argmin est plutôt difficile à mettre en oeuvre dans la pratique du fait des nombreux minima locaux du problème.

Au-delà de la présence des minima locaux, il a été montré par D.Aloise, A.Desphande, P. Hansen et P. Popat en 2007 que le "clustering basé sur la somme de carré" [2] est NP-difficile. Puis, M. Mahajan, P. Nimbhorkar et V.Varadarajan dans [9] en 2012 montrent que le problème que doit résoudre *k-means* dans le plan est NP-difficle. Donc le problème s'annonce difficile même dans un cas particulier où les minima locaux ne font pas obstacle.

La complexité algorithmique est également une propriété importante des estimateurs. Comme on l'a dit dans le paragraphe précédent, la recherche théorique d'estimateur nécessaire à la résolution de divers problèmes s'accompagne de la recherche des procédures de calcul les plus rapides possible. Sans ces recherches d'optimisation computationnelle, un estimateur, aussi parfait qu'il soit, a toutes les chances de ne pas pouvoir être utilisable en pratique à cause de sa complexité algorithmique en temps ou en mémoire. Dans le cas de *k-means*, l'algorithme de base à implémenter pour l'approcher l'estimateur statistique est l'algorithme de Lloyd, décrit ci-dessous.

Algorithme de Lloyd

1 : Choisir k points au hasard jouant le rôle de centroïdes
2 : **Répéter**
3 : - Former k groupes en associant chaque point à son centroïde le plus proche
4 : - Recalculer le centroïde de chaque cluster en minimisant la variance intracluster
5 : **Jusqu'à** critère d'arrêt satisfait

Très récemment, en 2016, Y. Lu [13] et H.H.Zhou [14] ont montré un résultat sur l'algorithme de Lloyd lui-même dans l'article "Statistical and computationnal guarantees of Lloyd's algorithm and its variants" [8]. Il existait depuis plus longtemps des résultats théorique sur l'estimateur mais

12. La norme associé à la variance est la norme $\|\cdot\|_2$
13. prononciation selon un français "lou"
14. prononciation selon un francais "djo"

pas sur l'algorithme qui l'approche. Leur résultat est notamment important parce qu'il donne directement des garanties sur l'objet implémentable. Alors qu'il est difficile de savoir si les résultats théoriques valables pour l'estimateur sont conservés dans l'algorithme [15]. Ils montrent en effet dans un cas très simple à 2 clusters que si l'initialisation de l'algorithme est un peu meilleure que 50% d'erreur, alors l'algorithme de Lloyd converge linéairement pendant $\mathcal{O}(\log(n))$ itérations, avec n le nombre de données, puis exponentiellement vers la valeur de l'estimateur.

2 L'estimateur Median-Of-Means

L'estimateur MOM, acronyme de "Median-of-Means", signifiant "prendre la médiane des moyennes", est un estimateur de la moyenne. Il est l'estimateur T_3 de l'introduction. Sa définition est la suivante :

$$MOM_k : \left\{ \begin{array}{ccc} \bigcup_{n\in\mathbb{N}} \mathbb{R}^n & \longrightarrow & \mathbb{R} \\ (x_i)_{1\leqslant i\leqslant n} & \longmapsto & \text{Médiane}\left[\left\{ \frac{1}{|B_i|} \sum_{x\in B_i} x : (B_i)_{1\leqslant i\leqslant k} \text{ partition } (x_i)_{1\leqslant i\leqslant n} \right\} \right] \end{array} \right. \tag{2}$$

Le premier intérêt de cet estimateur est sa robustesse aux outliers : grâce à l'opérateur "médiane", s'il y a une petite quantité d'outliers, la médiane ne sera pas affectée. Cette robustesse est mesurée par ce que l'on appelle le Breakdown Point. Mon apport personnel à ce jour, sur cet estimateur, concerne cette notion. Le seconde intérêt de MOM est sa presque-sous-gaussiannité. C'est-à-dire que ses valeurs se concentrent autour de sa moyenne dans des proportions semblables à la concentration d'une variable gaussienne. Enfin, le dernier intérêt établi de MOM est son optimalité. En effet, il est intéressant de se demander quelle est la vitesse théorique maximale de convergence d'un estimateur vers sa valeur moyenne lorsque le nombre d'observations tend vers l'infini, et, étant donnée que les phénomènes observés ont telles ou telles propriétés. Dans le cas de MOM, L. Devroye, M. Lerasle, G. Lugosi et R. I. Oliveira ont pu montrer en 2015 dans [4] que dans le cas où la loi sous-jacentes des observations a une variance finie, la vitesse théorique maximale possible est $\frac{1}{\sqrt{n}}$ tout en dépendant d'un paramètre. De plus, ils ont montré que MOM converge à cette même vitesse.

Notons au passage que l'estimateur T_2 de l'introduction dit "estimateur moyenne empirique" converge moins vite que MOM, toute chose égale par ailleurs, notamment à cause de sa sensibilité aux distributions à queue lourde [16]. Puisque *k-means* est construit comme l'estimateur moyenne empirique et que MOM est un moyen d'accroitre les performances dans l'estimation de la moyenne, peut-être est-il possible d'accroitre les performances de classification de *k-means* en s'inspirant de MOM. C'est ainsi qu'est née l'idée de sujet de thèse que m'a proposé Adrien Saumard.

15. Typiquement, les résultats de vitesse de convergence de $k-means$ ne sont plus nécessairement valables pour l'algorithme qui, lui, sera perturbé par les minima locaux

16. Autre façon de dire que les événement rare ont une influence importante sur la moyenne. C'est analogue à sensibilité aux outliers de *k-means*

2.1 Le Breakdown Point de MOM

On peut remarquer si l'on fait la moyenne de 11 valeurs réelles, il suffit de corrompre 1 donnée pour que le résultat de l'estimateur soit arbitrairement loin de la valeur sans corruption. En revanche, la médiane ne souffrira qu'une perturbation bornée : si l'on corrompt l'une des valeurs, au lieu de renvoyer la 6ième valeur, elle donnera soit la 5ième, soit la 6ième, soit la 7ième valeurs parmi les 11 de départs. On voit alors sans mal que l'on ne peut se permettre aucune corruption pour la moyenne empirique et 5 corruptions maximum pour la médiane dans ce cas à 11 valeurs.

Mais par ailleurs, on peut aussi profiter de la randomisation [17]. Dans le cas où l'on estime la moyenne avec T_1 de l'introduction, celui qui voudrait corrompre les données pour biaiser l'estimateur ne saurait quelles données corrompre et devrait alors le faire au hasard. Ainsi, en corrompant au hasard 5 données parmi 11, la probabilité que l'estimateur T_1 donne un résultat arbitrairement loin de la moyenne empirique sans corrumption n'est que de 5/11 et si l'on corrompt 6 observations la probabilité devient de 6/11, et ainsi de suite. On voit ainsi qu'il est aussi important de définir un Breakdown Point associé à une probabilité de biais fini ou infini. On définit ces deux notions comme suit :

Définition. **Breakdown Point Déterministe**
Soit X une variable aléatoire, T un estimateur et $n \in \mathbb{N}$. Le Breakdown Point Déterministe de T pour un échantillon de n observations de X est le nombre maximal, m, d'observations corruptibles en conservant un biais borné par rapport au cas sans corruption. Ceci se formalise comme suit :

En notant,

$$\mathcal{E}_m(X,n) = \left\{ B \subset \bigcup_{j=1}^{n} \bigcup_{k=1}^{m} \{X_j, Y_k\} \;\middle|\; \begin{array}{l} Y_1, \dots, Y_m \text{ quelconques, } X_1, \dots, X_n \overset{i.i.d}{\sim} X \\ \text{et } \mathrm{Card}(B) = n \end{array} \right\}$$

le Breakdown Point Déterministe est :

$$\epsilon_n(T,X) = \frac{1}{n} \max \left\{ m \;\middle|\; \sup_{B \in \mathcal{E}_m(X,n)} |T(B) - T(\{X_i\}_{1 \leqslant i \leqslant n})| < \infty \text{ p.s.} \right\}$$

17. Michaël Launay parle même de "puissance organisatrice du hasard dans `https://www.youtube.com/watch?v=2Wq6H8GMVm0` consulté le 2 juillet 2018

Définition. **Breakdown Point Statistique de seuil R**
Soit X une variable aléatoire, T un estimateur, $n \in \mathbb{N}$ et $R \in [0,1]$. Le Breakdown Point Statistique de seuil R de T pour un échantillon de n observations de X est le nombre maximal, m, d'observations corruptibles en conservant un biais borné par rapport au cas sans corruption avec probabilité au moins égale à $1-R$. Ceci se formalise comme suit :

En notant,

$$\mathcal{E}_m(X,n) = \left\{ B \subset \bigcup_{j=1}^{n} \bigcup_{k=1}^{m} \{X_j, Y_k\} \;\middle|\; \begin{array}{l} Y_1, \dots, Y_m \text{ quelconques, } X_1, \dots, X_n \overset{i.i.d}{\sim} X \\ \text{et } \mathrm{Card}(B) = n \end{array} \right\}$$

le Breakdown Point Statistique est :

$$\epsilon_n(T,X) = \frac{1}{n} \max \left\{ m \;\middle|\; \mathbb{P}\left(\sup_{B \,\in\, \mathcal{E}_m(X,n)} |T(B) - T(\{X_i\}_{1 \leqslant i \leqslant n})| < \infty \right) \right\} \geqslant 1 - R$$

Il découle immédiatement des définitions que :

Propriété. $\epsilon_n(T,X) = \epsilon_n(T,X,0)$ *et* $\forall R, S \in [0,1],\ R \leqslant S,\ \forall n \in \mathbb{N},\ \epsilon_n(T,X,R) \leqslant \epsilon_n(T,X,S)$

On montre facilement que

Propriété. $\epsilon_n(MOM_k, X) = \dfrac{k}{2n}$

J'ai pu montrer pendant mon stage que :

Propriété. $\lim\limits_{\substack{n \longrightarrow \infty \\ n/k \longrightarrow t}} \epsilon_n(MOM_k, X, R) = 1 - R^{1/t}$

L'objectif maintenant est de trouver une inégalité de concentration pour $MOM_{n/2}$ pour avoir un résultat non-asymptotique. Cette situation ressemble à un effet de seuil. J'ai l'intuition que $\epsilon_n(MOM_k, X, R)$ se concentre en fonction de R autour de $\epsilon_n(MOM_k, X, \frac{1}{2})$ avec une vitesse en $\frac{1}{\sqrt{n}}$.

2.2 La sous-gaussiannité de MOM

Dans le cas de l'estimateur moyenne empirique, on connaît tous le théorème centrale limite qui stipule que si un phénomène aléatoire a une variance finie, alors l'estimation de la moyenne de ce phénomène corrigée par la variance suit asymptotiquement une loi normale centrée réduite.

Théorème. ***théorème centrale limite***

Soit X une variable aléatoire de variance $\sigma^2 \in \mathbb{R}$, on sait que sa moyenne existe, on l'appelle μ, et si l'on note X_i la variable aléatoire correspondant à la i-ième observation de X, alors :
$S_n = \dfrac{\frac{1}{n}\sum_{i=1}^{n} X_i - \mu}{\sigma^2}$ *converge en loi vers $\mathcal{N}(0,1)$.*

Dis autrement, cela signifie que : $\forall I \subset \mathbb{R},\ \mathbb{P}(S_n \in I) \underset{n \longrightarrow \infty}{\longrightarrow} \mathbb{P}(N \in I)$, avec $N \sim \mathcal{N}(0,1)$.

Le point important de ce théorème est l'existence de la variance de X. Si le phénomène en question suit une loi telle que la variance n'existe pas, cela ne fonctionne pas. On n'aura pas aussi bien pour MOM mais on peut montrer un résultat similaire, un peu plus faible. D'abord, montrons ce que cela implique pour les intervalles de confiance :

Propriété. *Soit $n \in \mathbb{N}$, on note Φ^{-1} la fonction quantile de la loi normale*[18]*. Si N, $(N_i)_{1 \leqslant i \leqslant n}$ sont indépendantes et suivent une loi normale centrée réduite, alors*
$$\forall \alpha \in \mathbb{R}, \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} N_i \notin \left[\mu - \frac{\Phi^{-1}(\alpha)}{\sqrt{n}}, \mu + \frac{\Phi^{-1}(\alpha)}{\sqrt{n}}\right]\right) \leqslant 2\alpha.$$

Si bien que cela conduit à dire qu'une variable aléatoire se concentre normalement autour de sa moyenne lorsque qu'elle vérifie une telle relation.

Définition. concentration L-sous-gaussienne
Soit Φ^{-1} la fonction quantile de la loi normale. Soit $(X_n)_{n\in\mathbb{N}}$ une suite de variables aléatoires, toutes de moyenne μ, on dit que la suite $(X_n)_{n\in\mathbb{N}}$ se concentre de manière sous-gaussienne autour de sa moyenne s'il existe une constante L telle que
$$\forall \alpha \in \mathbb{R},\ \mathbb{P}\left(X_n \notin \left[\mu - L\frac{\Phi^{-1}(\alpha)}{\sqrt{n}}, \mu + L\frac{\Phi^{-1}(\alpha)}{\sqrt{n}}\right]\right) \leqslant 2\alpha.$$

On vient d'expliquer ce qu'est la concentration sous-gaussienne en toute rigueur. Mais la littérature n'a pas encore définitivement arrêté les définition et on trouve également des articles qui utilisent l'équivalent asymptotique de ϕ^{-1} au lieu de ϕ^{-1} : $\Phi^{-1}(\alpha) \underset{\alpha\to 1^-}{\sim} \sqrt{2\log\left(\frac{1}{1-\alpha}\right)}$.

Définition. concentration L-sous-gaussienne
Soit $(X_n)_{n\in\mathbb{N}}$ une suite de variables aléatoires, toutes de moyenne μ, on dit que la suite $(X_n)_{n\in\mathbb{N}}$ a une concentration sous-gaussienne autour de sa moyenne s'il existe une constante L telle que
$$\forall \alpha \in [0,1[,\ \mathbb{P}\left(|X_n - \mu| > L\frac{\sqrt{2\log\left(\frac{1}{1-\alpha}\right)}}{\sqrt{n}}\right) \leqslant 2\alpha.$$

L'objet $MOM_k((X_i)_{1\leqslant i\leqslant n})$ est une variable aléatoire et on peut écrire une telle relation. Mais étant donné le paramètre k dans sa définition, on ne peut choisir k et la précision α arbitrairement. Il est nécessaire de fixer la précision souhaitée puis d'adapter k. On parle alors d'estimateur α-dépendant et il est possible que toutes les valeurs de α ne soient pas compatibles avec une concentration sous-gaussienne. Pour garder à l'esprit que MOM_k est α-dépendant, on pourra être amené à écrire $MOM_{k,\alpha}$.

18. Cela vient simplement du fait que Φ est la fonction de répartition de la loi normale

Définition. estimateur α-dépendant à concentration L-sous-gaussienne
Soit $n \in \mathbb{N}$. Soit $(X_i)_{1 \leqslant i \leqslant n}$ une famille de n variables aléatoires i.i.d [19] de même loi que X. On note $T_{n,\alpha}(X) := T_{n,\alpha}((X_i)_{1 \leqslant i \leqslant n})$ pour faire court. Soit $T_{n,\alpha}$ un estimateur α-dépendant. On dit que $T_{n,\alpha}$ a une concentration sous-gaussienne autour de son espérance si

$$\exists \alpha_{max} > 0,\ \exists L > 0,\ \forall \alpha \in [0, \alpha_{max}],\ \mathbb{P}\left(|T_{n,\alpha}(X) - \mathbb{E}\left[T_{n,\alpha}(X)\right]| > L \frac{\sqrt{2\log\left(\frac{1}{1-\alpha}\right)}}{\sqrt{n}} \right) \leqslant 2\alpha.$$

D'après l'article de 2015 de L.Devroye, G. Lecué, M. Lerasle et R.I. Oliveira [4], MOM est α-dépendant et sous-gaussien pour toute variable aléatoire de distribution dans $\mathcal{P}_\sigma^M$, l'ensemble des distributions de moment σ bornée par M.

Théorème. *Devroye, Lecué, Lerasle, Oliveira*
Soit $M \in \mathbb{R}$, $n \in \mathbb{N}$, $(X_i)_{1 \leqslant i \leqslant n}$ un ensemble de n variables aléatoires i.i.d. de même loi que X. On note $\mathcal{P}_\sigma^M = \left\{ f \,\middle|\, \int_{\mathbb{R}} |f|^\sigma < M^\sigma \right\}$ l'ensemble des distributions de moment σ bornée par M. On prendra $\sigma > 1$ et pour f donnée dans $\mathcal{P}_\sigma^M$, on note μ_f l'espérance de f. Enfin, on note $\widehat{MOM}_{n,k,\alpha}(X) := \widehat{MOM}_{k,\alpha}((X_i)_{1 \leqslant i \leqslant n})$ pour faire court.

$$\forall \alpha \in \left[0, 1 - e^{1-n/2}\right],\quad \sup_{\substack{f \in \mathcal{P}_\sigma^M \\ X \sim f}} \mathbb{P}\left(|\widehat{MOM}_{n,k,\alpha}(X) - \mu_P| > \left(\frac{4(12M)^{\frac{1}{\sigma-1}} 2\log(\frac{1}{1-\alpha})}{n} \right)^{\frac{\sigma-1}{\sigma}} \right) \leqslant 2\alpha$$

Par exemple, cela signifie que MOM_k est un estimateur α-dépendant $\sqrt{48M}$-sous-gaussien sur l'ensemble $\mathcal{P}_2^M$. Mais l'estimateur moyenne empirique est également sous-gaussien sur $\mathcal{P}_2^M$ d'après le TCL. Le vrai gain dans ce théorème est de conserver cette sous-gaussiannité pour toute distribution ayant au moins un moment strictement supérieur à 1. Par ailleurs, MOM est au centre de l'estimation de la moyenne car un autre résultat de L. Devroye et al. dans [4] stipule que si l'on change un peu l'intervalle de confiance, alors quel que soit l'estimateur que l'on prendra pour estimer l'espérance des variables aléatoires, il en existe une pour laquelle la concentration de l'estimateur ne suit pas l'intervalle de confiance. Ce qui est une façon de voir que MOM est optimal en un certain sens. Ce résultat se formule comme ceci :

19. i.i.d. = indépendantes et identiquement distribuées

Théorème. ***Minoration de la vitesse de convergence des estimateur de la moyenne*** *Soit $M > 0$ et n un entier plus grand que 5. Soit $(X_i)_{1\leqslant i\leqslant n}$ un ensemble de n variables aléatoires i.i.d. de même loi que X. On note $\mathcal{P}_\sigma^M = \left\{ f \,\middle|\, \int_{\mathbb{R}} |f|^\sigma < M^\sigma \right\}$ l'ensemble des distributions de moment σ bornée par M. On prend $1 < \sigma < 2$ et pour f donnée dans $\mathcal{P}_\sigma^M$, on note μ_f l'espérance de f. Enfin, pour tout estimateur T, on note $T_n(X) := T((X_i)_{1\leqslant i\leqslant n})$ pour faire court. Le résultat est le suivant : pour tout estimateur T de l'espérance des variables aléatoires de loi dans $\mathcal{P}_\sigma^M$, on a*

$$\forall \alpha \in \left[0, 1 - 2e^{-n/4}\right], \quad \sup_{\substack{f \in \mathcal{P}_\sigma^M \\ X \sim f}} \mathbb{P}\left(|T_n(X) - \mu_P| > \left(\frac{M^{\frac{1}{\sigma-1}} 2\log(\frac{2}{1-\alpha})}{n} \right)^{\frac{\sigma-1}{\sigma}} \right) \geqslant 2\alpha$$

puisque pour n assez grand, on a $\left[0, 1 - 2e^{-n/4}\right] \subset \left[0, 1 - e^{1-n/2}\right]$, il suffit de comparer la longueur des intervalles de confiance pour $\alpha = 1 - 2e^{-n/4}$. Le calcul donne $\log(\frac{2}{1-\alpha}) = n/4$ et $\log(\frac{1}{1-\alpha}) = n/2 - 1$. On se rend alors compte que prendre un intervalle de confiance 2 fois plus large, ne peut conduire à la sous-gaussiannité uniforme sur $\mathcal{P}_\sigma^M$ pour $1 < \sigma < 2$. D'où l'optimalité de MOM.

On comprend donc pourquoi MOM a actuellement regagné de l'intérêt en statistique et pourquoi on essaie d'adapter les estimateurs qui faisaient intervenir la moyenne empirique pour en améliorer les performances statistiques.

3 k-means et l'algorithme de Lloyd

Dans cette partie, on s'intéresse spécifiquement à l'estimateur $k - means$ et à l'algorithme de Lloyd. On va présenter quelques propriétés importantes de ces objets.

3.1 Illustration du fonctionnement de k-means

Comme ça a été présenté dans l'introduction, l'estimateur le plus classique pour partitionner l'espace à partir des données observées est l'estimateur $k - means$. Le partitionnement se fait via des cellules de Voronoï basées sur la norme $\|\cdot\|_2$. Il vise à partitionner l'espace de telle sorte que la somme des variances dans chaque cellule soit minimale :

$$kmeans : \begin{cases} \bigcup_{n\in\mathbb{N}} \mathbb{X}^n & \longrightarrow & \mathbb{X}^k \\ (x_i)_{1\leqslant i\leqslant n} & \longmapsto & \underset{(c_j)_{1\leqslant j\leqslant k}\in\mathbb{X}^k}{\operatorname{argmin}} \sum_{i=1}^{n} \min_{1\leqslant j\leqslant k} \|x_i - c_j\|_2^2 \end{cases}$$

On peut voir ci-dessous un résultat de $k - means$ sur un exemple [20]. Les centres sont représentés par des croix et les cellules de Voronoï par des aplats de couleurs. On se rend tout de suite compte que certaines cellules sont non bornées. De ce fait, si l'on envoyait à l'infini l'un des points d'une

20. Image issue de `http://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_digits.html` consultée le 2 juillet 2018

telle cellule, la seule solution pour retrouver un minimum du critère $k-means$ (minimisation de la somme des variances intraclusters) est que le centre correspondant le suivent vers l'infini. Ce phénomène n'est pas satisfaisant car le centre ainsi déplacé formera un cluster avec un seul point. C'est ce principe qui rend $k-means$ sensibles aux outliers.

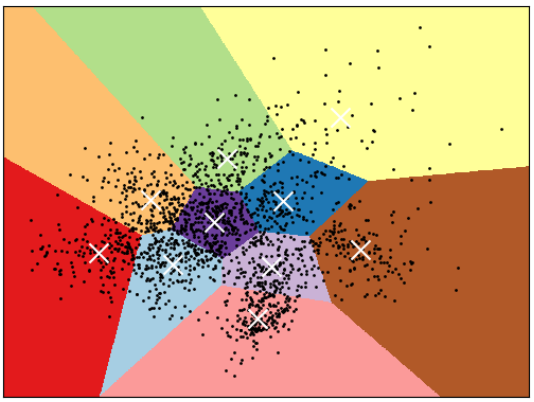


FIGURE 2 – exemple de la partition du plan obtenu à l'aide de k-means avec 10 centres

Un autre problème que l'on rencontre en utilisant $k-means$, pour n'en citer qu'eux, est le problème de position des cellules de Voronoï. En partant de la situation ci-dessous [21], $k-means$ renvoie des centres minimisant la variance intracluster en dépit de la structure des données. L'utilisation de $k-means$ présuppose alors que les données sont structurées en petits groupes convexes, ce qui donne un résultat non satisfaisant quand ce n'est pas le cas. Malgré la banalité de cette remarque, ce travers est particulièrement difficile à détecter en grandes dimensions.

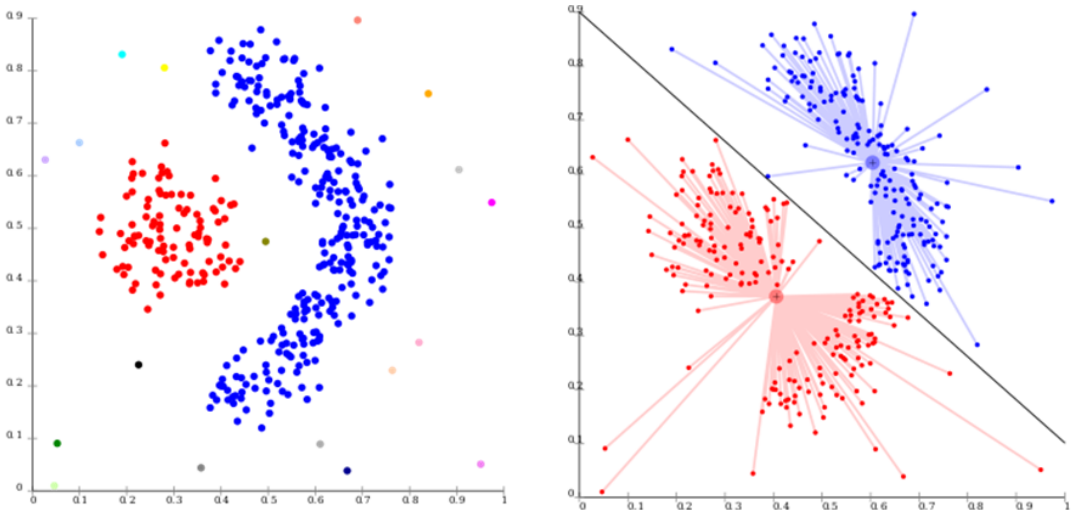


FIGURE 3 – A gauche : une situation de départ déjà clusterisé par l'auteur des données. A droite : le clustering proposé par k-means sur les mêmes données

21. Image de gauche et de droite de la figure 2 prises sur le site `https://en.wikipedia.org/wiki/Cluster_analysis` consulté le 2 juillet 2018

3.2 Approximation de k-means par l'algorithme de Lloyd

On a évoqué le fait que les estimateurs statistiques étaient des objets théoriques qu'il faut réussir à mettre en pratique pour exhiber concrètement l'information. Dans le cas de $k-means$, le premier algorithme proposé par Lloyd en 1953 pour calculer les centres est l'algorithme éponyme. On peut voir ci-dessous l'algorithme et une illustration [22] du processus itératif de l'algorithme de Lloyd.

Algorithme de Lloyd

1 : Choisir k points au hasard jouant le rôle de centroïdes
2 : **Répéter**
3 : - Former k groupes en associant chaque point à son centroïde le plus proche
4 : - Recalculer le centroïde de chaque cluster en minimisant la variance intracluster
5 : **Jusqu'à** critère d'arrêt satisfait

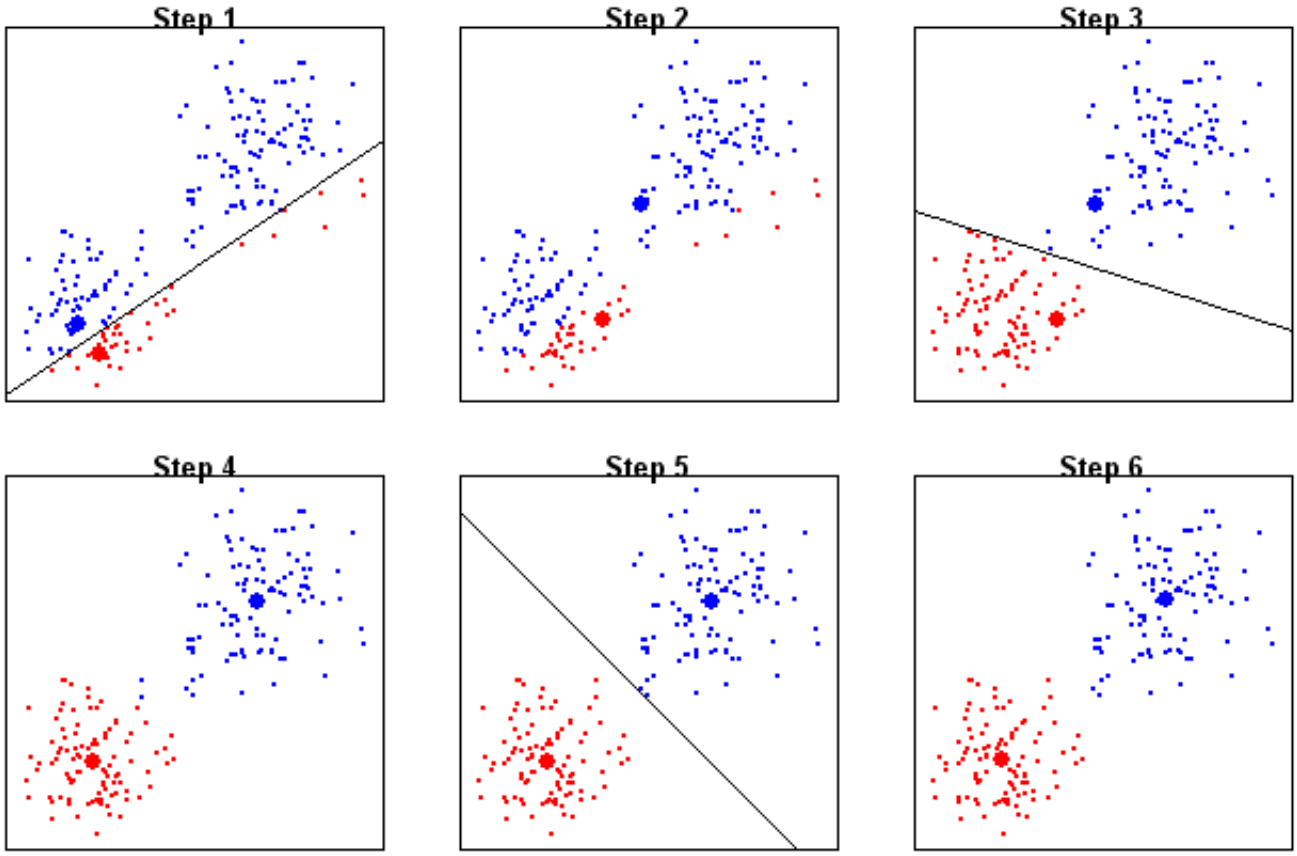


FIGURE 4 – Visualisation des itérations de l'algorithme de Lloyd. Step 1 : initialisation des centres et tracé de la frontière entre les cellules de Voronoï. step 2 : réaffectation des centres dans chaque cellule. Step 3 : tracé de la nouvelle frontière enter les cellules. Step 4 : recalcul des centres. Step 5 : recalcul des cellules. Step 6 : le critère d'arrêt de leur simulation devait être atteint et on sort de la boucle avec ces centres

22. Image issue du site `https://medium.com/@dilekamadushan/introduction-to-k-means-clustering-7c0ebc997e00` consulté le 2 juillet 2018

3.3 Propriété de k-means et de l'algorithme de Lloyd

Les questions que se posent les statisticiens sont par exemple de déterminer la vitesse de convergence de l'estimateur, déterminer la vitesse de convergence de l'algorithme, déterminer le critère qui permet de qualifier la qualité de la convergence.

Pour étudier la convergence de l'estimateur $k-means$, on s'intéresse à la distance théorique maximale qu'on peut garantir entre les centres théoriques et le résultat de l'estimateur en fonction du nombre de point. Pour cela, on suppose qu'il existe des centres de cluster à trouver, puis on quantifie l'écart entre ces centres théoriques et la valeur de l'estimateur :

$$\mathbb{P}\left(\|\text{"centre théorique"} - kmeans((X_i)_{1\leqslant i\leqslant n})\| > r_n(\alpha)\right) \leqslant \alpha$$

Etudier la convergence de l'algorithme de Lloyd, revient à se demander si l'algorithme converge bel et bien vers la valeur de l'estimateur : autrement dit, la mise en oeuvre de $k-means$ est-elle conforme à ce qu'on attend d'elle. Pour ce faire, on regarde si à un nombre d'observations fixé, la limite de l'algorithme quand le nombre d'itérations tend vers l'infini est la valeur de l'estimateur. Puis, pour étudier la vitesse de convergence de l'algorithme, on s'intéresse aux nombre d'itérations nécessaires pour être à une distance donnée de la valeur limite : pour un $\epsilon > 0$ donné, trouver M_{min} tel que $\forall M \geqslant M_{min}$, pour M itérations de l'algorithme, $\|kmeans - \text{"valeur de l'algorithme après M itérations"}\| < \epsilon$.

Les critères investigués par les chercheurs pour qualifier la convergence de l'estimateur ou de l'algorithme sont par exemple : le type de distributions dans chaque cluster, la distance entre les clusters, le rapport "signal sur bruit", le nombre de cluster, etc.

On a vu que $k-means$ minimise la quantité $R((c_j)_{1\leqslant j\leqslant k}) := \sum_{i=1}^{n} \min_{1\leqslant j\leqslant k} \|x_i - c_j\|_2^2$ par rapport aux centres. Cette quantité est une grandeur réelle positive donc il existe nécessairement un minimum à la fonction $(c_1, c_2, \ldots, c_k) \longmapsto R(c_1, c_2, \ldots, c_k)$. Il est donc intéressant de regarder la performance de l'estimateur à travers la distance entre le minimum et le critère obtenu avec l'estimation. Si l'on note $(\hat{c}_1, \hat{c}_2, \ldots, \hat{c}_k)$ les centres obtenus avec l'estimateur et R_{min} le minimum de R, alors il s'agit de regarder la quantité $\ell(\hat{c}_1, \hat{c}_2, \ldots, \hat{c}_k) := R(\hat{c}_1, \hat{c}_2, \ldots, \hat{c}_k) - R_{min}$. C'est la quantité qu'a regardé Clément Levrard et il a démontré en 2015 dans l'article "Non asymptotic bounds for vector quantization in hilbert spaces" [7] la probabilité que ℓ prenne de grandes valeurs décroit exponentiellement vite avec cette valeur. On peut déduire de cette relation une inégalité sur l'espérance de ℓ. Formellement, le théorème est le suivant :

Théorème.
Soit P_M une densité de probabilité à support dans la boule de centre 0 et de rayon M. Soit $(X_i)_{1\leqslant i\leqslant n}$, n vecteurs aléatoires i.i.d. de loi P_M. Soit k le nombre de cluster de $\mathbb{R}^d$. Soit $\hat{c}_n := (\hat{c}_1, \hat{c}_2, \ldots, \hat{c}_k) = kmeans((X_i)_{1\leqslant i\leqslant n})$. Avec ℓ défini ci-dessus, on peut dire que :

$$\exists C^{(1)}_{k,d,P_M}, C^{(2)}_{k,d,P_M} \geqslant 0,\ \forall x \geqslant 0,\ \mathbb{P}\left(\ell(\hat{c}_n) \geqslant \frac{C^{(1)}_{k,d,P_M} + xC^{(2)}_{k,d,P_M}}{n}\right) \leqslant e^{-x}$$

$$\textit{Donc , } \exists C^{(3)}_{k,d,P_M} \geqslant 0,\ \mathbb{E}\left[\ell(\hat{c}_n)\right] \leqslant \frac{C^{(3)}_{k,d,P_M}}{n}$$

Ce théorème signifie aussi que dans un espace normé de dimension d et k clusters, en moyenne, l'écart entre la valeur du critère et la valeur minimale possible décroit inversement proportionnellement au nombre de données. L'article est plus précis en exhibant la forme des constantes et leur dépendance en k, d et M.

Enfin, Y. Lu et H. H. Zhou se sont placés dans un cadre très simple où l'on a deux clusters distribués chacun selon une loi normale multidimensionnelle à symétrie sphérique de variance σ^2 dont l'une admet θ pour vecteur moyen et l'autre admet $-\theta$ pour vecteur moyen. Ils ont pu montrer la vitesse de convergence de l'algorithme de Lloyd ainsi que mettre en exergue un paramètre déterminant dans la précision de l'estimation. Afin de citer ce théorème, il nous faut fixer quelques notations. On indexe à l'aide de la lettre M les itérations de l'algorithme. On essaie de mesurer la capacité de l'algorithme à classer les points dans leur cluster respectif, donc on note A_M la proportion d'observations mal classées à l'itération M. Soit r un paramètre [23] du problème.

Théorème.
Pour un nombre n de données et un ratio "signal sur bruit" r plus grand qu'une même constante C. On peut montrer que si l'initialisation est suffisamment "bonne", c'est-à-dire telle que :

$$A_0 \leqslant \frac{1}{2} - \frac{1}{\sqrt{n}} - \frac{2.56 + \sqrt{\log(r)}}{r}$$

Alors il y a une probabilité $1 - \frac{1}{n^3} - 2\exp\left(-\frac{r^2(1+9d/n)}{3}\right)$ *pour que les deux résultats suivants soient vrais, le premier impliquant le second :*

$$\forall M \geqslant 0, \qquad \mathbb{E}(A_M) \leqslant \frac{1}{r^{M-\lceil \log(n) \rceil}} + 4\exp\left(-\frac{r^2(1+9d/n)}{8}\right)$$
$$\forall M \geqslant \lceil 3\log(n) \rceil, \quad \forall t > 0,\ \mathbb{P}\left(A_M > t\right) \leqslant \frac{1}{tn^2} + 4\exp\left(-\log(t) - \frac{r^2(1+9d/n)}{8}\right)$$

Les conséquences de ce théorème sont simples : si l'on a un nombre de données suffisant par rapport au ratio "signal sur bruit", disons $\frac{r^2(1+9d/n)}{8} \leqslant 2\log(n)$, alors en prenant $t = \exp\left(-\frac{r^2(1+9d/n)}{16}\right)$, on peut montrer que le résultat final est contrôlé par le rapport "signal sur bruit" : $\mathbb{P}\left(A_M > \exp\left(-\frac{r^2(1+9d/n)}{16}\right)\right) \leqslant 5\exp\left(-\frac{r^2(1+9d/n)}{16}\right)$. Et à l'inverse, si la quantité de données n'est pas suffisante par rapport au ratio "signal sur bruit", à savoir $\frac{r^2(1+9d/n)}{8} \geqslant 2\log(n)$, alors du fait que le nombre de données est finie et que A_M est une proportion, on a $A_M \in \{0, \frac{1}{n}, \frac{2}{n}, \ldots, 1\}$, ainsi : $\mathbb{P}\left(A_M > 0\right) = \mathbb{P}\left(A_M \geqslant \frac{1}{n}\right) \leqslant \frac{5}{n}$ d'après le dernier résultat du théorème appliqué pour $t = 1/n$. On peut agréger tous les événements pour dire que si la condition initiale du théorème est vraie avec probabilité $1-\nu$ alors avec probabilité $1 - \nu - \frac{5}{n} - 8\exp\left(-\frac{r^2(1+9d/n)}{3}\right)$, en faisant plus de $\lceil 3\log(n) \rceil$ itérations, on aura une erreur de clustering plus petite que $\exp\left(-\frac{r^2(1+9d/n)}{16}\right)$.

23. Y. Lu et H. H. Zhou appelle se paramètre le ratio "signal sur bruit". Mais on a pas besoin d'entrer dans les détails pour comprendre le théorème

3.4 L'apport pratique de mon stage sur les questions de robustesse et d'amélioration de k-means

Pendant ces quelques mois de travail, j'ai naturellement commencé la bibliographie nécessaire à mon futur sujet de thèse. En plus des théories très générales du livre de Boucheron, Lugosi et Massart [3], des théorèmes très pointus qu'on peut lire dans [4], [7], [6] et [8], j'ai également trouvé le temps de tester les premières idées d'Adrien Saumard sur l'intégration de MOM à $k-means$ pour le rendre plus robuste aux outliers. Adrien m'a soumis une proposition d'algorithme que j'ai implémenté dans mathematica©.

Il faut savoir avant toute chose que les estimateurs robustes ne sont pas très nombreux. Il existe les estimateurs basés sur le "trimming" [24], sur la centralité [25] et sur la norme L^1 qui charge moins les valeurs éloigner du point moyen que les normes L^p pour $p > 1$. Pour faire simple, le trimming et la centralité ignore purement et simplement des données, ce qui n'est pas forcément satisfaisant étant donné que l'on ne sait pas à priori s'il est pertinent de supprimer de l'analyse. La norme L^1 est intéressante en ce qu'elle n'ignore aucune donnée mais tend à donner une poids égale à toutes. Cependant, ce que l'on a décrit pour $k-means$ sur l'isolement d'un cluster en envoyant un point à l'infini reste vrai. On ne peut alors pas vraiment parler de robustesse comme pour MOM car le Breakdown Point statistique de ces estimateurs est nul. MOM_k est un compromis entre la suppression de données et la norme L^1 car en toute rigueur, en ajustant k, MOM varie de l'estimateur de la médiane pour $k = n$ blocs à l'estimateur de la moyenne empirique pour $k = 1$ bloc. Si bien que pour k entre 1 et n, on a un comportement intermédiaire et on laisse la possibilité à des valeurs "extrêmes" d'occuper une proportion non nulle dans le blocs médians de MOM et d'être ainsi représenté.

J'ai exploité ce phénomène, inspiré par l'article de G.Lecué et M.Lerasle [6], pour créer un "score" qui mesure le nombre de fois qu'une donnée a été présente dans le bloc médian de MOM_k en M itértions de l'algorithme. En dimension 1, le résultat a été sans appel : les données les plus centrales ont visité le bloc médian un nombre de fois proportionnel à leur centralité dans leur distributions théoriques et les valeurs "aberrantes" [26] ont été très peu présentent dans le blocs médian. Cette différente a été significative et ce simple score a permis de retrouver leur qualité : aberrante ou pas (voir la figure ci-dessous).

L'algorithme proposé par Adrien Saumard pour classifier les données fonctionne mieux que $k-means$ car il est notablement moins sensible aux minima locaux [27], il n'est pas non plus sensible aux points envoyés à l'infini, et enfin, il permet d'exploiter la même astuce de "score" de façon à évaluer la centralité des données en fonction de leur intéractions avec les autres, sans présupposer de structure étoilé comme dans les méthodes de "profondeur de données". Cet algorithme est présenté sous la figure 5.

24. Conceptuellement, cela revient à ignorer les données les plus isolées.

25. C'est un peu l'opposé du trimming, on ne retient que les données au coeur des nuages de point. A supposer que le nuage de point est un ensemble étoilé. La notion s'appelle la "profondeur de données" en statistique. Cette notion ne s'applique pas aux données de la figure 3 par exemple.

26. Ici, j'entends aussi bien événements rares qu'outliers.

27. Voire insensible mais cela reste à prouver

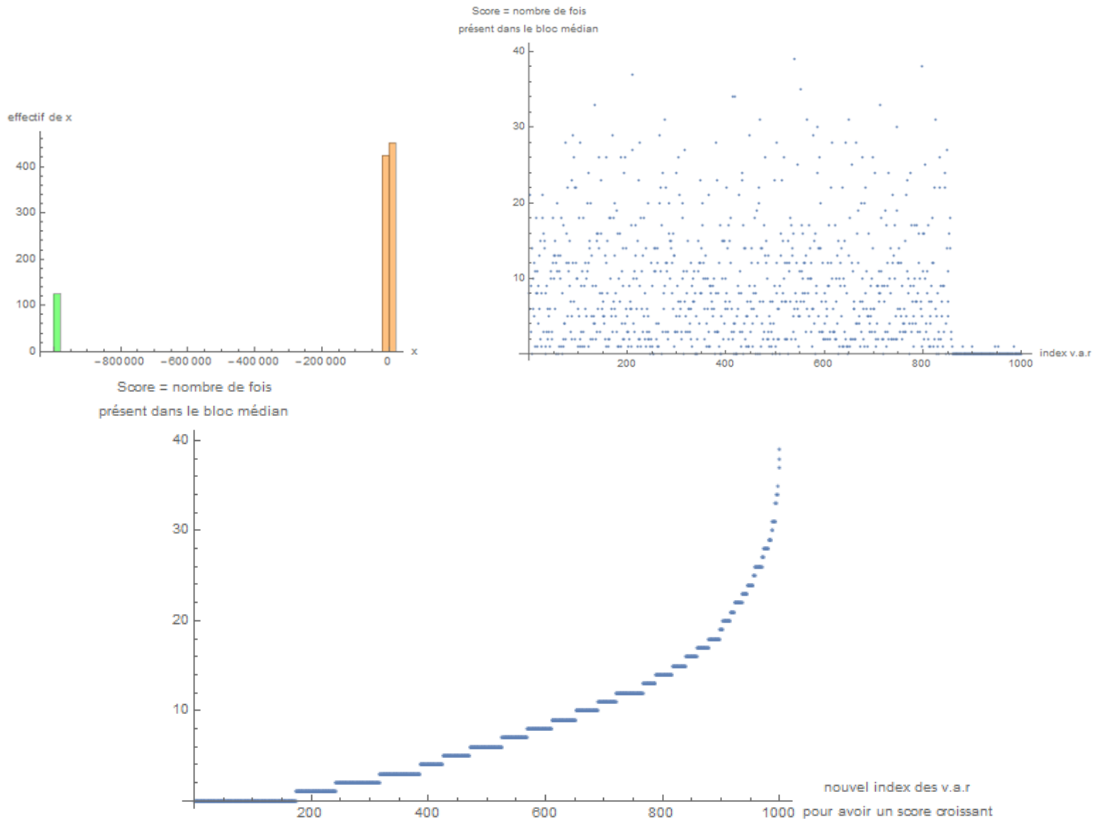


FIGURE 5 – On peut voir sur la figure de gauche la distribution de données que j'ai mis en argument de MOM_2 et à droite on peut lire les scores en ordonnées associés aux index des variables aléatoires réelles numérotées de 1 à 850 pour variables aléatoires gaussiennes centrées réduites et de 851 à 1000 pour les variables aléatoires aberrantes valant presque sûrement -10^6. Le résultat du score est claire : les variables aberrantes ont en grande majorité un score nul. On peut voire sur la troisième figure les scores en fonction des variables aléatoires réindexées de façon à tracer un score croissant. Les 150 variables aberrantes sont toutes dans les 225 variables aléatoires de plus petit score.

Algorithme de Lloyd modifié

1 : Choisir k points $(c_p)_{1\leqslant p\leqslant k}$ jouant le rôle de centroïdes
Répéter
2 : - Associer chaque observation x avec son centroïde c_p le plus proche
3 : - Partitionner aléatoirement les observations en N sous-échantillons $(P_i)_{1\leqslant i\leqslant N}$
Pour chaque P_i
4 : - calculer le barycentre b_{ij} de chacun des clusters "j" de P_i
5 : - calculer $R(P_i) = \sum_{x\in P_i} \min_{1\leqslant j\leqslant k} \|x - b_{ij}\|^2$
6 : - remplacer les centres $(c_p)_{1\leqslant p\leqslant k}$ par les barycentres $(b_{i_0 j})_{1\leqslant j\leqslant k}$ du sous-échantillon P_{i_0} qui réalise la valeur médiane de $(R(P_i))_{1\leqslant i\leqslant N}$
Jusqu'à critère d'arrêt satisfait

L'efficacité de cet algorithme à l'air de parler de lui-même sur l'exemple de j'ai construit pour le tester (voir la figure 6) : j'ai choisi trois clusters principaux dont 2 quasiment joints inscrits dans un carré de 40 cm de côté centré en $(0, 0)$, plus un groupe de 5 données en $(10000, 10000)$ et un ajout de quelques points additionnels autour des clusters principaux pour voir si l'algorithme permet de les détecter comme "non centraux".

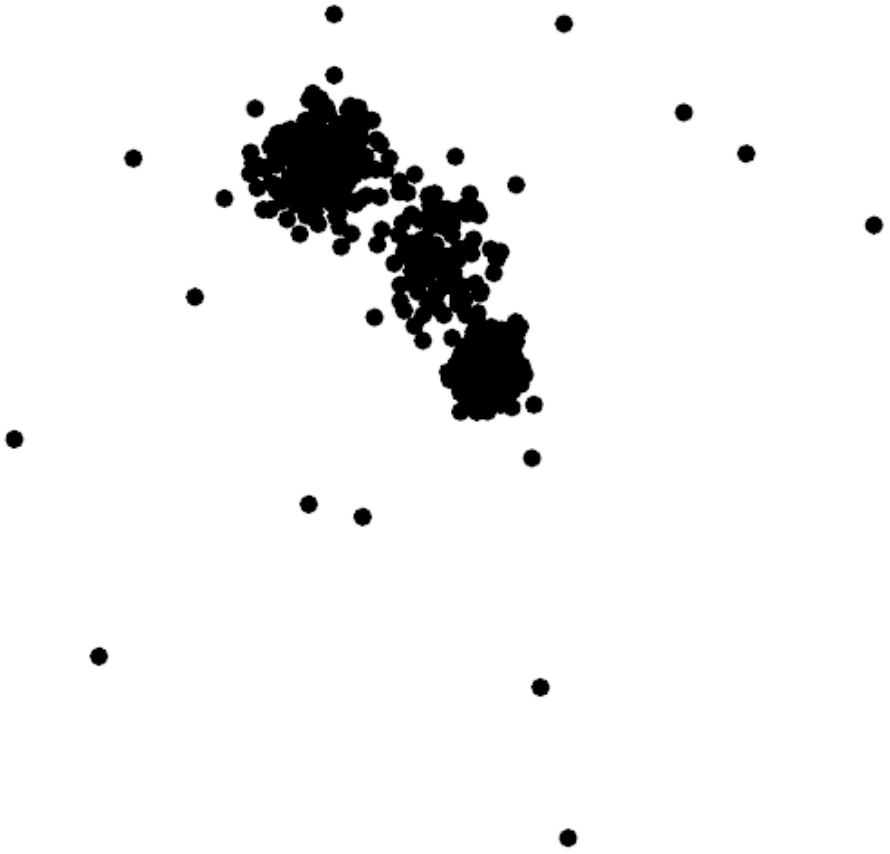


FIGURE 6 – Exemple construit à la main pour tester les propriétés du nouvel algorithme proposé pour améliorer $k - means$. Notez que pour une raison de lisibilité les 5 points placés en $(10000, 10000)$ n'apparaissent pas. Tous les autres points sont dans le carré de côté 40 et centré en $(0, 0)$. Les effectifs sont : 200 observations autour de $(-10, 10)$, 300 observations autour de $(5, -8)$, 100 observations autour de $(1, 1)$, 20 observations autour de $(0, 0)$ avec une grande variances pour faire des points isolés et enfin 5 points en $(10000, 10000)$ pour imiter une corruption.

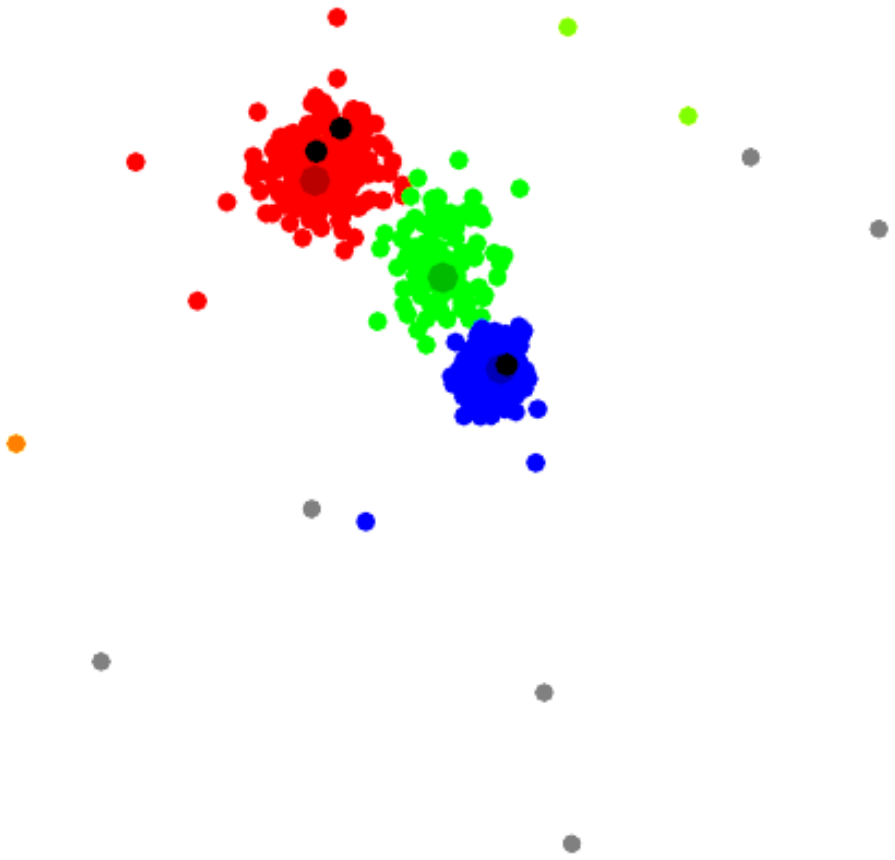


FIGURE 7 – Résultat du nouvel algorithme sur l'exemple construit à la main pour tester ses propriétés. On peut voir 4 éléments différents : les petits points noirs sont les données choisies pour initialiser les centres dans l'algorithme ; les petits points de couleur sont les données classées dans l'une des familles "rouge, "verte" et "bleue" ; les gros points transparents de couleur sont les centres finaux de leur famille et ne sont pas des données mais des points représentatifs des familles ; les petits points de couleurs et jaunies sont des données "non cenrales" au sens du score (score inférieur à 1 sur 200 itérations).

On peut d'abord remarquer que visuellement, la structure des données proposée par l'algorithme est satisfaisante. Malheureusement je n'étais plus en mesure [28] d'utiliser mathematica© pour comparer avec $k-means$, mais le résultat aurait été simple : un cluster en $(10000, 10000)$ et 2 clusters plus ou moins satisfaisant autour de $(0, 0)$. Si l'on s'intéresse maintenant au "score", on voit que certaines données on été jaunies. Les données jaunies sont celles qui ont été présentes au plus 1 fois dans le blocs médian sur 200 itérations de l'algorithme. On voit également que les données les "moins centrales", au sens du "score", sont bien celles que l'on pense intuitivement être les "moins centrale".

Les simulations sont donc très concluantes et Adrien Saumard prévoit que nous écrivions un article sur le sujet. Elles m'ont aussi permis de prendre en main mon sujet de façon plus heuristiques et viennent compléter mon apprentissage théorique présenté en partie dans la section suivante.

28. Licence expirée

4 Elements de théorie du processus empirique

Jusqu'ici nous avons exploré la démarche et les résultats importants de mon stage sans prendre le temps de détailler les preuves ou les théories sous-jacentes. Pourtant il y a bien des théories importantes comme la théorie de l'apprentissage de Vapnik et Chervonenkis, la théorie de la concentration de la mesure ou encore la théorie du processus empirique. J'ai eu l'occasion de faire une introduction à la concentration de la mesure en séminaire étudiant le 22 novembre 2017. Je vais donc me restreindre à des éléments de la théorie du processus empirique.

Ce qu'on appelle en statistique le processus empirique est La variable aléatoire $\sum_{i=1}^{n} X_i$ où chaque X_i est une variable aléatoire de loi $\mathcal{L}$ et indépendante des autres. Dans le théorème précédent de Y. Lu et H. H. Zhou, on a eu besoin de contrôler l'erreur commise entre le vecteur moyen θ et le vecteur moyen estimé $\hat{\theta}$. Or, sans rentrer dans le détail, l'estimation à l'itération M vaut $\hat{\theta}^{(M)} = \frac{1}{n}\sum_{i=1}^{n} \hat{z}_i^{(M)} y_i$, où $\hat{z}_i^{(M)} = \pm 1$ est le label estimé à l'itération M de la données y_i. On reconnait la forme du processus empirique où $X_i = \hat{z}_i^{(M)} y_i / n$.

Ce qu'il est particulièrement intéressant de comprendre est ce que l'on appelle le supremum du processus empirique. On le notera Z. Le supremum du processus empirique est la valeur maximale que peut prendre le processus empirique sur plusieurs instants. Typiquement, si l'on regarde la variable aléatoire $X_i^{(M)}$ qui donne le nombre d'oeufs pondus par la poule "i" le jour "M" dans un poulailler, alors le nombre d'oeufs récoltés le jour "M" est la valeur du processus empirique $\sum_{i=1}^{n} X_i^{(M)}$ à l'instant "M". Il devient intéressant, pour dimensionner son poulailler, de contrôler le nombre maximal d'oeufs possible par jour. Dans ce cas, on pourra regarder le supremun du processus empirique $Z = \sup_{M \in \mathbb{N}^*} \sum_{i=1}^{n} X_i^{(M)}$, où le supremum est pris sur tous les jours possibles. La question à laquelle on chercherait à répondre serait alors "trouver le nombre n de poules pour que la probabilité d'avoir plus de 5 oeufs par jour soit inférieur à 1/52". Une inégalité du type $\mathbb{P}(Z > t) \leqslant \exp(-t)$ nous permettrait de répondre. De manière analogue, Lu et Zhou ont eu besoin de contrôler la norme d'un vecteur aléatoire et le produit scalaire de 2 vecteurs aléatoires, deux situations qui s'inscrivent dans le cadre de l'étude du supremum du processus empirique.

L'objectif de cette partie est de démontrer deux lemmes "techniques" de leur l'article [8] concernant le processus empirique et d'aller plus loin en démontrant ce qu'on appelle en statistique le principe de contraction et le principe de symétrisation. Deux principes fondamentaux, nécessaires pour comprendre les preuves de G. Lecué et M. Lerasle dans leur article de novembre 2017 "Robust machine learning by Median-of-Means : theory and practice" [6].

4.1 L'inégalité de Chernoff

Cette inégalité est fondamentale dans l'étude des queues de distribution car elle donne des bornes notablement moins grossières que les inégalité de Markov ou de Bienaymé-Tchebychev. Pour constater la progression des inégalités, on va toutes les citer :

Propriété. ***inégalité de Markov***
Soit X une variable aléatoire réelle d'espérance finie. On a :

$$\forall t > 0,\ \mathbb{P}\left(|X - \mathbb{E}\left[X\right]| > t\right) \leqslant \frac{\mathbb{E}\left[|X - \mathbb{E}\left[X\right]|\right]}{t}$$

Propriété. ***inégalité de Bienaymé-Tchebychev***
Soit X une variable aléatoire réelle de variance finie. On a :

$$\forall t > 0,\ \mathbb{P}\left(|X - \mathbb{E}\left[X\right]| > t\right) \leqslant \frac{\mathbb{V}\left[|X - \mathbb{E}\left[X\right]|\right]}{t^2}$$

Propriété. ***généralisation de l'inégalité de Markov par toute fonction convexe***
Soit $\varphi : \mathbb{R} \longmapsto \mathbb{R}^+$ une fonction croissante positive. Soit X une variable aléatoire réelle telle que $\mathbb{E}\left[\varphi(X)\right] < \infty$. On a :

$$\forall t > 0,\ \mathbb{P}\left(X > t\right) \leqslant \frac{\mathbb{E}\left[\varphi(X)\right]}{\varphi(t)}$$

Théorème. ***inégalité de Cramér-Chernoff***
Soit X une variable aléatoire réelle pour laquelle il existe des moments exponentiels. C'est-à-dire telle que

$\exists U \subset \mathbb{R}^+$ *ouvert* $, \forall \lambda \in U,\ \mathbb{E}\left[e^{\lambda X}\right] < \infty$. *En notant* $\psi_X(\lambda) = \log \mathbb{E}\left[e^{\lambda X}\right]$,

on a :

$$\forall t > 0,\ \mathbb{P}\left(X > t\right) \leqslant \exp\left(-\sup_{\lambda \in U}\left(\lambda t - \psi_X(\lambda)\right)\right)$$

Démonstration.
Puisque $\mathbb{E}\left[e^{\lambda X}\right] < \infty$ pour $\lambda \in U$, et que $\varphi_\lambda : \mathbb{R} \longmapsto \mathbb{R}^+, y \longmapsto e^y$ est positive croissante, on peut appliquer la propriété adéquate ci-dessus pour avoir :

$$\forall \lambda \in U,\ \forall t > 0,\ \mathbb{P}\left(X > t\right) \leqslant \frac{\mathbb{E}\left[\varphi_\lambda(X)\right]}{\varphi_\lambda(t)} = e^{-\lambda t}\mathbb{E}\left[e^{\lambda X}\right] = \exp\left(-\left[\lambda t - \psi_X(\lambda)\right]\right)$$

Ensuite, prendre l'inf du membre de droite sur U revient à prendre le sup sur U de $\lambda t - \psi_X(\lambda)$ et le résultat s'ensuit. Ce théorème fait apparaître "$\sup_{\lambda \in U} (\lambda t - \psi_X(\lambda))$" car cet objet a été étudié et s'appelle la transformée de Cramér. □

Corollaire. ***inégalité de Cramér-Chernoff pour une variable gaussienne***
Soit N un variable aléatoire de loi normale centrée et de variance ν. On peut calculer que $\sup_{\lambda \geqslant 0} (\lambda t - \psi_X(\lambda)) = \dfrac{t^2}{2\nu}$, *ainsi :*

$$\forall t > 0,\ \mathbb{P}(N > t) \leqslant \exp\left(-\frac{t^2}{2\nu}\right)$$

Et comme N est symétrique, on a $\forall t > 0,\ \mathbb{P}(|N| > t) \leqslant 2\exp\left(-\dfrac{t^2}{2\nu}\right)$

4.2 Preuves techniques utiles au théorème de Y. Lu et H. H. Zhou

On va à présent utiliser l'inégalité de Chernoff pour démontrer deux lemmes de leur article. On démontre ces lemmes car ils font appel à des notions incontournables en concentration de la mesure et apprentissage statistique. Ces notions font partie des premières que j'ai apprises pendant ce stage.Elles portent le nom "d'argument de l'union bornée" et de "chainage". La première consiste simplement à borner le supremum sur S du processus empirique indexé par S en utilisant le fait qu'il existe une borne, pour chaque indice d'indexation, indépendante de S. La seconde technique sert à fournir des bornes sur un espace non-dénombrable en le discrétisant et en contrôlant l'erreur commise par la discrétisation.

Lemme. *Soit ω une variable aléatoire suivant une loi normale en dimension d de symétrie sphérique de variance σ^2. Si l'on indice ω alors cela signifie qu'on en prend une copie indépendante de toutes les autres copies. Soit n un entier naturel et $S \subset \{1, 2, \ldots, n\}$. On pose $Z_S = \sum_{i \in S} \omega_i$, un analogue du processus empirique, pris seulement sur les variables aléatoires d'indice dans S. Z_S est donc une variable aléatoire de $\mathbb{R}^d$ muni de sa norme euclidienne $\|\cdot\|$. En notant $|S|$ le cardinal de S, on a :*

$$\mathbb{P}\left(\|Z_S\| > \sqrt{2\sigma^2 (n + 9d) |S|}\right) \leqslant \exp(-\frac{n}{10})$$

Démonstration.
On a déjà utilisé le fait que $\mathbb{E}\left[e^{\lambda\omega}\right] = e^{\lambda^2\sigma^2/2}$. On utilise dans cette preuve le fait que $\forall \lambda \in [0, 1/2\sigma^2[,\ \mathbb{E}\left[e^{\lambda\omega^2}\right] = \dfrac{1}{\sqrt{1 - 2\lambda\sigma^2}}$. Comme Z_S est la somme de $|S|$ gaussiennes indépendantes, on peut dire que $Z_S \sim \mathcal{N}(0, |S| I_d)$, où I_d est la matrice identité en dimension d. Donc les composantes de Z_S qu'on note $(z_k)_{1 \leqslant k \leqslant d}$ sont indépendantes et sont distribuées selon une $\mathcal{N}(0, |S|)$. D'où $\|Z_S\|^2 = \sum_{k=1}^{d} z_k^2$ et $\forall \lambda \in [0, 1/2|S|[,\ \mathbb{E}\left[e^{\lambda\|Z_S\|^2}\right] = \left(\dfrac{1}{\sqrt{1 - 2\lambda|S|}}\right)^d$.

On sait d'après une des étapes de la démonstration de l'inégalité de Chernoff que $\forall \lambda \in [0, 1/2|S|[,\ \forall t > 0,\ \mathbb{P}\left(\|Z_S\|^2 > t\right) \leqslant \exp\left(-\lambda t - \frac{d}{2}\log(1 - 2\lambda|S|)\right)$.

Grâce à λ et t on peut rendre le membre de droite indépendant de $|S|$ en prenant $\lambda = x/|S|$ et $t = u|S|$: $\forall x \in [0, 1/2[,\ \forall u > 0,\ \mathbb{P}\left(\|Z_S\|^2 > u|S|\right) \leqslant \exp\left(-xu - \frac{d}{2}\log(1 - 2x)\right)$.

C'est à ce moment là que l'on va chercher à utiliser "l'argument de l'union bornée" pour avoir une relation impliquant le maximum des Z_S et se défaire de la dépendance en $|S|$. L'ensemble des sous-ensembles de $\{1, 2, \ldots, n\}$ a un cardinal de 2^n. Donc il ne reste qu'à majorer grossièrement en utilisant le fait que :

$$\mathbb{P}\left(\max_S(Z_S - \sqrt{u|S|}) > 0\right) = \mathbb{P}\left(\bigcup_S\{\|Z_S\| > \sqrt{u|S|}\}\right) \leqslant \sum_S \mathbb{P}\left(\|Z_S\| > \sqrt{u|S|}\right)$$

et que :

$$\mathbb{P}\left(\|Z_S\| > \sqrt{u|S}\right) = \mathbb{P}\left(\|Z_S\|^2 > u|S|\right).$$

En écrivant $2^n = e^{\log(2)n}$, il vient :

$$\forall x \in [0, 1/2[,\ \forall u > 0,\ \mathbb{P}\left(\max_S(Z_S - \sqrt{u|S|}) > 0\right) \leqslant \exp\left(\log(2)n - xu - \frac{d}{2}\log(1 - 2x)\right)$$

La conclusion découle de l'inégalité précédente en particularisant à $x = 0.49$ et $u = 1.62(n + 4d)$:

$$\begin{array}{rcl} \mathbb{P}\left(\max_S(Z_S - \sqrt{u|S|}) > 0\right) & \leqslant & e^{(\log(2) - 1.62\cdot 0.49)n} e^{-(4\cdot 1.62\cdot 0.49 - \log(50)/2)d} \\ & \leqslant & e^{-0.1006n} e^{-1.219d} \\ & \leqslant & e^{-0.1n} \end{array}$$

□

Lemme. *Soit ω une variable aléatoire suivant une loi normale en dimension d de symétrie sphérique de variance σ^2. Si l'on indice ω alors cela signifie qu'on en prend une copie indépendante de toutes les autres copies. On notera v' le vecteur transposé de v. Soit enfin n, un entier naturel. On s'intéresse au rayon spectral ρ_n de la matrice $\sum_{i=1}^n \omega_i\omega_i'$. Le résultat est le suivant :*

$$\mathbb{P}\left(\rho_n > 1.62\left(n + 4d\right)\sigma^2\right) \leqslant \exp(-\frac{n}{10})$$

Démonstration.
Pour plus de concision, on note $A = [\omega_1, \omega_2, \ldots, \omega_n]$ la matrice aléatoire dont les colonnes sont des copies i.i.d. de ω. On note λ les valeurs propres de $\sum_{i=1}^n \omega_i\omega_i'$ et u leurs vecteurs propres normalisés associés. On sait de l'algèbre linéaire que le rayon spectrale qui nous intéresse vérifie :

$$
\begin{aligned}
\rho_n &= \max_{1\leqslant i\leqslant n}(\lambda_i) \\
&= \lambda_{max} \text{ (la plus grande valeur propre)} \\
&= {u_{max}}' \cdot \left[\left(\sum_{i=1}^{n} \omega_i \omega_i' \right) u_{max} \right] \\
&= \sup_{\|a\|=1} \left[a' \left(\sum_{i=1}^{n} \omega_i \omega_i' \right) a \right] \\
&= \sup_{\|a\|=1} \sum_{i=1}^{n} (a'\omega_i)(a'\omega_i)' \text{ or les termes dans la sommes sont des réels} \\
\rho_n &= \sup_{\|a\|=1} \sum_{i=1}^{n} (a'\omega_i)^2 \\
&= \sup_{\|a\|=1} \sum_{j=1}^{n} \left[a'A\right]_j^2 \text{ où l'on somme sur les colonnes car a'A est un vecteur ligne} \\
&= \sup_{\|a\|=1} \|a'A\|_2^2 \\
\rho_n &= \sup_{\|a\|=1} \|Aa\|_2^2
\end{aligned}
$$

Maintenant, on va utiliser le "chaînage" pour remplacer le sup en max. Soit B_1 la sphère unité de $\mathbb{R}^d$ et soit C un ϵ-réseau sur B_1. Comme son nom le suggère, C est un ensemble de points de B_1 tel que l'union de toutes les boules de rayon ϵ centrées en ces points recouvre B_1. De ce fait, pour tout a de norme 1, il existe un vecteur b de C tel que $\|a-b\| \leqslant \epsilon$. Ainsi :

$$
\begin{aligned}
\|Aa\|_2 &\leqslant \|Ab\|_2 + \|A(b-a)\|_2 \\
\|Aa\|_2 &\leqslant \sup_{b\in C} \|Ab\|_2 + \left[\sup_{\|a\|=1} \|Aa\|_2 \right] \|b-a\|_2 \\
\|Aa\|_2 &\leqslant \sup_{b\in C} \|Ab\|_2 + \epsilon \sup_{\|a\|=1} \|Aa\|_2 \\
\sup_{\|a\|=1} \|Aa\|_2 &\leqslant \sup_{b\in C} \|Ab\|_2 + \epsilon \sup_{\|a\|=1} \|Aa\|_2 \\
\sup_{\|a\|=1} \|Aa\|_2 &\leqslant \sup_{b\in C} \|Ab\|_2 + \epsilon \sup_{\|a\|=1} \|Aa\|_2 \\
(1-\epsilon) \sup_{\|a\|=1} \|Aa\|_2 &\leqslant \sup_{b\in C} \|Ab\|_2 \\
(1-\epsilon)^2 \sup_{\|a\|=1} \|Aa\|_2^2 &\leqslant \sup_{b\in C} \|Ab\|_2^2 \\
(1-\epsilon)^2 rho_n &\leqslant \sup_{b\in C} \|Ab\|_2^2
\end{aligned}
$$

On a vu ci-dessus que $\rho_n = \sup_{\|a\|=1} \sum_{i=1}^{n} (a'\omega_i)^2$, donc on va s'intéresser à $\sum_{i=1}^{n} (b'\omega_i)^2$ en vu d'utiliser un "argument d'union bornée". On sait que pour b fixé, puisque ω_i est un vecteur gaussien centré réduit, on a $b'\omega_i \sim \mathcal{N}(0,1)$ et est indépendant des autres copies. Par l'inégalité de Chernoff :

$$
\begin{array}{rcl}
\forall \lambda, t>0,\ \mathbb{P}\left(\sum_{i=1}^n (b'\omega_i)^2 > t\right) & \leqslant & e^{-\lambda t}\mathbb{E}\left[e^{\lambda \sum_{i=1}^n (b'\omega_i)^2}\right] \\
& \leqslant & e^{-\lambda t}\mathbb{E}\left[\prod_{i=1}^n e^{\lambda (b'\omega_i)^2}\right] \\
& \leqslant & e^{-\lambda t}\prod_{i=1}^n \mathbb{E}\left[e^{\lambda (b'\omega_i)^2}\right] \text{ on utilise ensuite l'indépendance} \\
& \leqslant & e^{-\lambda t}\prod_{i=1}^n \frac{1}{\sqrt{1-2\lambda}} \text{ d'après la démonstration précédente} \\
\forall \lambda, t>0,\ \mathbb{P}\left(\sum_{i=1}^n (b'\omega_i)^2 > t\right) & \leqslant & e^{-\lambda t}(1-2\lambda)^{-n/2}
\end{array}
$$

On va pouvoir étudier ρ_n à l'aide d'un "argument d'union bornée" :

$$
\begin{array}{rcl}
\mathbb{P}\left(\rho_n > t\right) & = & \mathbb{P}\left((1-\epsilon)^2\rho_n > (1-\epsilon)^2 t\right) \\
& \leqslant & \mathbb{P}\left(\sup_{b\in C} \|Ab\|_2 > (1-\epsilon)^2 t\right) \\
& \leqslant & \sum_{b\in C} \mathbb{P}\left(\|Ab\|_2 > (1-\epsilon)^2 t\right) \\
& \leqslant & \sum_{b\in C} \mathbb{P}\left(\sum_{i=1}^n (b'\omega_i)^2 > (1-\epsilon)^2 t\right) \\
\forall \lambda, t>0,\ \mathbb{P}\left(\rho_n > t\right) & \leqslant & \sum_{b\in C} e^{-\lambda(1-\epsilon)^2 t}(1-2\lambda)^{-n/2} \\
& \leqslant & |C| e^{-\lambda(1-\epsilon)^2 t}(1-2\lambda)^{-n/2} \\
\forall \lambda, t>0,\ \mathbb{P}\left(\rho_n > t\right) & \leqslant & \exp\left[\log(|C|) - \lambda(1-\epsilon)^2 t - n\log(1-2\lambda)/2\right]
\end{array}
$$

On sait, notamment grâce au lemme 4.1 dans "Empirical processes : theory and applications" de David Pollard publié en 1990 [10] que $|C| \leqslant (1+\frac{2}{\epsilon})^d$. Donc :

$$
\begin{array}{rcl}
\forall \lambda, t, \epsilon>0,\ \mathbb{P}\left(\rho_n > t\right) & \leqslant & \exp\left[\log(|C|) - \lambda(1-\epsilon)^2 t - n\log(1-2\lambda)/2\right] \\
& \leqslant & \exp\left[d\log\left(1+\frac{2}{\epsilon}\right) - \lambda(1-\epsilon)^2 t - n\log(1-2\lambda)/2\right]
\end{array}
$$

Il suffit de particulariser à $\lambda = 1/4$, $\epsilon = 0.05$ et $t = 2n + 18d$ pour avoir le résultat du lemme. On ne fait pas l'application numérique cette fois-ci.

□

4.3 Le principe de symétrisation

Pour achever la partie technique de ce rapport, on va présenter ici un dernier élément du supremum du processus empirique : le principe de symétrisation. Grâce à cette méthode, on peut comparer les situations où l'on a des variables aléatoires distribuées non symétriquement à des situations où elles seraient symétriques. On a nécessairement des inégalités plus large que dans la vraie situations mais au moins on contrôle d'avantage de choses. De plus, ce résultat repose essentiellement sur l'utilisation de copies indépendantes des variables aléatoires initiales.

Théorème. ***principe de symétrisation***

Soit n un entier, $X_1, X_2, \ldots, X_n$ n variables aléatoires centrées indépendantes, non nécessairement identiquement distribuées. On note $X_{i,s}$ la s-ième copie indépendante de la variable aléatoire X_i indexée par $\mathcal{T}$. Enfin, on prendre $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ n variables aléatoires, dites de Rademacher, telle que $\mathbb{P}(\varepsilon_i = 1) = \mathbb{P}(\varepsilon_i = -1) = 1/2$, indépendantes entre elles et indépendantes des X. Le résultat stipule que :

$$\frac{1}{2}\mathbb{E}\sup_{s\in\mathcal{T}}\left|\sum_{i=1}^{n}\varepsilon_i X_{i,s}\right| \;\leqslant\; \mathbb{E}\sup_{s\in\mathcal{T}}\left|\sum_{i=1}^{n} X_{i,s}\right| \;\leqslant\; 2\mathbb{E}\sup_{s\in\mathcal{T}}\left|\sum_{i=1}^{n}\varepsilon_i X_{i,s}\right|$$

Démonstration.

Dans cette preuve on ne démontrera que l'inégalité supérieure car la seconde se démontre exactement de la même façon. Soit $X'_1, X'_2, \ldots, X'_n$ des copies indépendantes des variables aléatoires X. On indexe ces nouvelles variables X' par $\mathcal{T}$ selon la même règle d'indépendance. On peut alors successivement montrer que :

$$\begin{aligned}
\left|\sum_{i=1}^{n} X_{i,s}\right| &\leqslant \left|\sum_{i=1}^{n} X_{i,s} - 0\right| \\
&\leqslant \left|\sum_{i=1}^{n} X_{i,s} - \mathbb{E}\left({X_{i,s}}'|X_{i,s}\right)\right| \text{ par indépendance et centrage des variables aléatoires} \\
&\leqslant \left|\mathbb{E}\left(\sum_{i=1}^{n} X_{i,s} - {X_{i,s}}'|X_{i,s}\right)\right| \text{ par linéarité de l'espérance conditionnelle} \\
&\leqslant \mathbb{E}\left(\left|\sum_{i=1}^{n} X_{i,s} - {X_{i,s}}'\right| |X_{i,s}\right) \text{ par inégalité triangulaire} \\
&\leqslant \sup_{s\in\mathcal{T}}\mathbb{E}\left(\left|\sum_{i=1}^{n} X_{i,s} - {X_{i,s}}'\right| |X_{i,s}\right) \text{ en passant au sup à droite} \\
&\leqslant \mathbb{E}\left(\sup_{s\in\mathcal{T}}\left|\sum_{i=1}^{n} X_{i,s} - {X_{i,s}}'\right| |X_{i,s}\right) \begin{array}{l}\text{vrai car vient de } X \leqslant \sup X \text{ , puis } E(X|B) \leqslant E(\sup X|B) \\ \text{et enfin } \sup E(X|B) \leqslant \sup E(\sup X|B) = E(\sup X|B)\text{"}\end{array} \\
&\leqslant \mathbb{E}\left(\sup_{s\in\mathcal{T}}\left|\sum_{i=1}^{n}\varepsilon_i(X_{i,s} - {X_{i,s}}')\right| |X_{i,s}\right) \text{ car } X_{i,s} - {X_{i,s}}' \sim {X_{i,s}}' - X_{i,s} \text{ du fait } X_i \text{ et } {X_i}' \text{ i.i.d.} \\
&\leqslant \mathbb{E}\left(\sup_{s\in\mathcal{T}}\left|\sum_{i=1}^{n}\varepsilon_i X_{i,s}\right| + \sup_{s\in\mathcal{T}}\left|\sum_{i=1}^{n}\varepsilon_i {X_{i,s}}'\right| |X_{i,s}\right) \text{ par inégalité triangulaire et propriété du sup} \\
&\leqslant \sup_{s\in\mathcal{T}}\left|\sum_{i=1}^{n}\varepsilon_i X_{i,s}\right| + \mathbb{E}\left(\sup_{s\in\mathcal{T}}\left|\sum_{i=1}^{n}\varepsilon_i {X_{i,s}}'\right| |X_{i,s}\right) \text{ car } \mathbb{E}(X_{i,s} + {X_{i,s}}'|X_{i,s}) = X_{i,s} + \mathbb{E}({X_{i,s}}'|X_{i,s}) \\
&\leqslant \sup_{s\in\mathcal{T}}\left|\sum_{i=1}^{n}\varepsilon_i X_{i,s}\right| + \mathbb{E}\left(\sup_{s\in\mathcal{T}}\left|\sum_{i=1}^{n}\varepsilon_i {X_{i,s}}'\right|\right) \begin{array}{l}\text{car } \mathbb{E}({X_{i,s}}'|X_{i,s}) = \mathbb{E}({X_{i,s}}') \\ \text{par indépendance entre } X_i \text{ et } {X_i}'\end{array} \\
&\leqslant \sup_{s\in\mathcal{T}}\left|\sum_{i=1}^{n}\varepsilon_i X_{i,s}\right| + \mathbb{E}\left(\sup_{s\in\mathcal{T}}\left|\sum_{i=1}^{n}\varepsilon_i X_{i,s}\right|\right) \text{ car } X_i \text{ et } {X_i}' \text{ ont même loi}
\end{aligned}$$

Pour achever la preuve, il n'y a plus qu'à prendre le supremum sur $\mathcal{T}$ puis l'espérance. □

Conclusion

Grâce à ce stage, ma culture statistique s'est considérablement accrue par rapport au début. Initialement, je tirais ma seule connaissance des statistiques d'un cours de statistiques inférentielles de M1. Aujourd'hui, je suis sur les rails d'une thèse en apprentissage statistique.

En effet, comme on a pu le voir, je me suis familiarisé avec les statistiques théoriques de l'apprentissage statistique et de la concentration de la mesure. J'ai pu progresser non seulement d'un point de vue théorique comme est témoigne la partie 4, avec le principe de symétrisation, le chaînage et l'argument de l'union bornée, mais aussi et surtout d'un point de vue pratique, à travers la découverte de l'estimateur $k-means$ et des différents algorithmes que j'ai pu implémenter : de $k-means$ à sa version robuste, en passant par MOM, je bénéficie de d'heuristiques riches et fécondes, espérons-le.

Ces 16 semaines m'ont permis de gagner de l'expérience qu'il faudra savoir réexploiter dans mes recherches à venir. Dans un premier temps, notre attention va être portée sur l'article scientifique devant présenter l'adaptation de $k-means$. Puis dans un second temps, j'approfondirai ces notions pendant jusqu'à la fin officiel de mon stage fin août.

Références

[1] Charu C. AGGARWAL, Chandan K. REDDY *Data Clustering, algorithms and applications* CRC Press. 2014

[2] Daniel ALOISE, Amit DESPHANDE, Pierre HAMSON, Preyas POPAS, *NP-Hardness of Euclidean sum-of-squares clustering.* soumis le 20 juillet 2007.

[3] Stéphane BOUCHERON, Gabor LUGOSI, Pascal MASSART *Concentration inequalities, a non asymptotic theory of independance* Oxford University Press. 2013

[4] Luc DEVROYE, Matthieu LERASLE, Gabor LUGOSI, Roberto IMBUZEIRO OLIVEIRA *Sub-gaussian mean estimators* Ann. Statist. **44**(6), 2695-2725 (2016)

[5] Frank R. HAMPEL, Elvezio Mauro Danilo RONCHETTI, Peter J. ROUSSEEUW, Werner A. STAHEL *Robust Statistics, the approach based on influence functions* WILEY. 1986.

[6] Guillaume LECUE, Matthieu LERASLE *Robust machine learning by median-of-means : theory and practice* arXiv preprint, arXiv : 1711.10306 (2017)

[7] Clément LEVRARD *Non-asymptotic bounds for vector quantization in Hilbert spaces* Ann. Statist. **43**(2), 592-619 (2015)

[8] Yu LU, Harrison HUIBIN ZHOU *Statistical and computational guarantees of Lloyd's algorithm and its variants.* Yale University, 8 décembre 2016.

[9] Meena MAHAJAN, Prajakta NIMBHPORK, Kasturi VARADARAJAN *The planar k-means problem is NP-Hard.* Theoretical Computer Science **442**(2012) 13-21

[10] David POLLARD *Empirical processes : theory and applications.* NSF-CBMS Regional Conference Series in Probability and Statistics, Vol. 2, 1990, `http://www.jstor.org/stable/4153175`

[11] Frédéric PROÏA *Statistique Inférentielle.* Polycopié d'un cours dispensé en master 1 de mathématique fondamentale et appliquée à Angers en 2016-2017.

[12] Apprentissage Machine/statistique `http://wikistat.fr/pdf/st-m-Intro-ApprentStat.pdf` consulté le 2 juillet 2018