UFR des mathématiques

TRAVAIL ENCADRE DE RECHERCHE

Concentration de la mesure et réduction de dimension

NAIT ABDELLAH ABDELOUAHAB
BESSAHA NASSIM
ELHAJJ RIM

Encadrant :
M.Adrien Hardy

MAI 2019

Table des matières

1 Introduction

Parmi les nombreuses techniques algorithmiques utilisant des plongements métriques, une des techniques à la fois les plus simples conceptuellement mais également les plus riches en application est celle de la réduction de dimension. Elle consiste à trouver, étant donnée n points dans un espace métrique, un plongement de ces n points dans le même espace métrique mais de dimension plus faible, le tout avec une distorsion raisonnable. Il est clair (mais délicat à formaliser et prouver) que c'est difficile à réaliser pour tous les espaces à n points possibles, sinon on pourrait plonger tout l'espace de haute dimension dans un espace bien moins riche. Cette difficulté est contournée par l'usage de méthodes probabilistes, qui fournissent un plongement correct dans une majorité des cas, mais se "trompent" dans d'autres. Dans le domaine des algorithmes d'approximation, l'intérêt est essentiel : une partie très importante des opérations réalisées sur les points d'un espace métrique ont une dépendance au moins linéaire en dimension. Et si celle-ci est importante ,la complexité en pâtit. Quitte à perdre une précision sur l'algorithme qui est comme on va le voir, extrêmement raisonnable ($1 + \epsilon$ pour tout ϵ), on peut réduire la dimension de façon logarithmique, et ainsi grandement accélérer la rapidité d'exécution de toutes les opérations usuelles. En un certain sens, la dimension de réduction peut être vue comme une technique géométrique de compression de données, dont on connaît l'importance en pratique.

Dans ce projet, on présente le cas de la réduction de dimension, en étudiant le lemme de Johnson-Lindenstrauss qui donne de très bons résultats, puis on discute de quelques applications et enfin des difficultés qui apparaissent lorsque l'on cherche à généraliser cette technique à d'autres espaces.

2 Préliminaires et notions de base

Dans ce projet on utilise les bases de probabilités suivantes :
Soit X une variable aléatoire définie sur l'espace probabilisé $(\Omega, \mathscr{F}, \mathbb{P})$.
Si X est dans $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$ et $\mathrm{X} \in \mathrm{L}_1$, alors l'espérance de X est définie par

$$\mathbb{E}(X) = \int_{-\infty}^{+\infty} X(w)\, d\,\mathbb{P}(w).$$

Si X est une variable aléatoire discrète réelle alors

$$\mathbb{E}(X) = \sum_{w \in \Omega} X(w)\mathbb{P}(w).$$

Si $X \geq 0$, alors $\mathbb{E}(X) = \int_0^{+\infty} \mathbb{P}(X \geq x)dx$.

Si $X \in L_2$, alors sa variance est

$$Var(X) = \mathbb{E}(X - \mathbb{E}(X))^2.$$

La fonction génératrice des moments est définie par :

$$\text{Pour tout } t \in \mathbb{R}, \qquad \mathbb{M}_X(t) = \mathbb{E}(\exp(t\,X)).$$

On introduit ainsi la norme notée par $\|X\|_p = (\mathbb{E}|X|^p)^{\frac{1}{p}}, \qquad p \in (0, \infty)$.

La covariance de deux variables aléatoires X et Y est

$$cov\,(\,X, Y\,) = \mathbb{E}\,(\,X - \mathbb{E}\,(\,X\,))\,(\,Y - \mathbb{E}\,(\,Y\,))\, = (X - \mathbb{E}\,(X), Y - \mathbb{E}\,(Y))_{L^2}.$$

Si X est à valeur dans $(\mathbb{R}^d, \mathscr{B}(\mathbb{R}^d))$, on dira que X est un vecteur aléatoire.
Si $X = {}^t(X_1, ..., X_d)$ avec $X_j \in L^1$ pour tout j, on définit son espérence comme le vecteur des espérences de ses entrées, $\mathbb{E}(X) = {}^t(\,\mathbb{E}\,[X_1]\,, ..., \mathbb{E}\,[X_d])$, et sa matrice de covariance par

$$\sum_X = [\,Cov(X_i, X_j)]_{i,j=1}^d$$

si $X_j \in \mathrm{L}^2$ pour tout j. Noter $\sum_X$ est une matrice symétrique semi-dénie positive.

Markov et Tchebychev sont les premières inégalités de la concentration qu'on va abordé dans ce projet.

Proposition 2.1. Soit X une variable aléatoire positive et intégrable alors pour tout t strictement positive,

$$\mathbb{P}(\,X \geq t\,) \leq \frac{\mathbb{E}\,(X)}{t}\ .$$

Si g est une fonction positive strictement croissante et $\mathbb{E}\left(g\left(X\right)\right) < \infty$, alors

$$\mathbb{P}\left(X \geq t\right) = \mathbb{P}\left(g\left(X\right) \geq g\left(t\right)\right) \leq \frac{\mathbb{E}\left(g\left(X\right)\right)}{g\left(t\right)} .$$

En choisissant $g(x) = (x - \mathbb{E}\left(x\right))^2$ on obtient une conséquence connue de l'inégalité de Markov, l'inégalité de Bienaymé-Tchebychev qui offre une convergence quadratique en t et qui definie la concentration de X autour de sa moyenne.

Corollaire 2.2. (Inégalité de Markov et Tchebychev)
Soit X une variable aléatoire admettant une espérance $\mathbb{E}(X)$ et une variance $Var(X)$ alors pour tout t strictement positive,

$$\mathbb{P}\left(\mid X - \mathbb{E}(X) \mid \geq t\right) \leq \frac{Var\left(X\right)}{t^2} .$$

Remarque 2.3. On peut obtenir mieux que ce dernier si $\mathbb{E}\left(X^{k}\right) \leq \infty$, alors

$$\mathbb{P}\left(\mid X - \mathbb{E}\left(X\right) \mid \geq t\right) \leq \frac{\left|\mathbb{E}\left(X - \mathbb{E}\left(X\right)\right)\right|^k}{t^k} .$$

3 Concentration de la somme des variables aléatoires

Les variables bornées forment une classe importante de variables aléatoires. L'inégalité de Hoeffding va nous permettre de contrôler des déviations de variables bornées sans aucune hypothèse sur la loi.

Proposition 3.1. (L'inégalité de Hoeffding)
Soit une suite $(X_k)_{1\leq k\leq n}$ de variables aléatoires indépendantes vérifiant pour 2 suites $(a_k)_{1\leq k\leq n}$, $(b_k)_{1\leq k\leq n}$ de nombres réels telle que : $\forall k \in \mathbb{N}, \qquad a_k \leq X_k \leq b_k$.
On pose $S_n = X_1 + X_2 + ... + X_n$, alors

$$\begin{aligned}
\mathbb{P}\left(S_n - \mathbb{E}\left(S_n\right) \geq t\right) &\leq \exp\left(\frac{-2\,t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right) \\
\mathbb{P}\left(S_n - \mathbb{E}\left(S_n\right) \leq -t\right) &\leq \exp\left(\frac{-2\,t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right) \\
\mathbb{P}(\,|\, S_n - \mathbb{E}\left(S_n\right)\,|\geq t) &\leq 2\,\exp\left(\frac{-2\,t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right).
\end{aligned}$$

Remarque 3.2. (Borne de Chernof)
L'hypothèse fondamentale dans l'inégalité de Hoeffding est l'aspect bornée des variables aléatoires. On peut parfois obtenir des bornes exponentielles explicites, en supposant seulement que la variable X admet des moments exponentielles, c'est-à-dire que $\mathbb{E}\left(e^{t\,X}\right) < \infty$ pour t ≥ 0. Dans ce cas, pour tout réel a et tout t > 0, on applique l'inégalité de Markov à $e^{t\,X}$, on a :

$$\mathbb{P}\left(X \geq a\right) = \mathbb{P}\left(e^{t\,X} \geq e^{t\,a}\right) \leq \frac{\mathbb{E}\left(e^{t\,X}\right)}{e^{t\,a}}\ .$$

de même pour tout t < 0 :

$$\mathbb{P}\left(X \leq a\right) = \mathbb{P}\left(e^{t\,X} \geq e^{t\,a}\right) \leq \frac{\mathbb{E}\left(e^{t\,X}\right)}{e^{t\,a}}\ .$$

En particulier :

— $\mathbb{P}(X \geq a) = \mathbb{P}(e^{t\,X} \geq e^{t\,a}) \leq \inf_{t>0} \frac{\mathbb{E}\left(e^{t\,X}\right)}{e^{t\,a}}$.

— $\mathbb{P}(X \leq a) = \mathbb{P}(e^{t\,X} \geq e^{ta}) \leq \inf_{t<0} \frac{\mathbb{E}\left(e^{t\,X}\right)}{e^{t\,a}}$.

Dans l'inégalité de Hoeffding si S_n se réduit à une seule variable, on pose $S_n = X_1$, avec $\mathbb{E}(X_1) = 0$ et $a \leq X_1 \leq b$ l'inégalité devient

$$\mathbb{P}(|S_n| \geq t) \leq 2\,\exp\left(\frac{-2\,t^2}{(b-a)^2}\right).$$

Cette inégalité implique que toutes les variables aléatoires bornées appartiennent à une classe plus large de variables dites variables sous gaussiennes qui sera le sujet du paragraphe suivant.

3.1 La concentration sous gaussienne

Les variables aléatoires sous-gaussiennes constituent l'une des principales familles de variables aléatoires jouissant d'une propriété d'intégrabilité forte en probabilité, les fameuses variables aléatoires gaussiennes font une partie de cette famille.

Nous présentons les principales propriétés de cette classe de variables aléatoires.

Proposition 3.3. Soit X une variable aléatoire, les propriétés suivantes sont équivalentes :

i. $\mathbb{P}\,(\,|X| \geq t\,) \;\leq 2\;\exp\,(\,-t^2/K_1^2\,), \qquad \forall\, t \geq 0.$

ii. $\|X\|_p \;\leq\; K_2\,\sqrt{p}, \qquad \forall\, p \geq 1.$

iii. $\mathbb{E}\,(\,\exp\,(\,X^2/K_3^2\,)\,) \;\leq\; 2.$

De plus, si $\mathbb{E}\,(\,X\,) = 0$ alors (i), (ii) et (iii) sont équivalentes à :

iv. Si $\;\mathbb{E}\,(X) = 0$ alors $\mathbb{E}\,(\exp(\lambda X)\,) \;\leq\; \exp\,(\,\lambda^2 K_4^2)$.

Démonstration : $(i) \Longrightarrow (ii)$
Pour une variable X positive : $\mathbb{E}(X) = \int_a^\infty \mathbb{P}\,(\,X \geq t\,)\,\mathrm{d}t$.

$$\begin{aligned}
\mathbb{E}\,(\,|X|^p\,) &= \int_a^\infty \mathbb{P}(\,|X|^p \geq u\,)\,\mathrm{d}u \\
&= \int_a^\infty \mathbb{P}\,(\,|X| \geq t\,)\,p\,t^{p-1}\,\mathrm{d}t \qquad\qquad (\text{changement de variable } u = t^p) \\
&\leq \int_a^\infty 2\exp\,(\frac{-t^2}{K_1})\,p\,t^{p-1}\,\mathrm{d}t \;=\; p\,\sqrt{K_1}^{\,p}\,\Gamma(\frac{p}{2}\,) \qquad (\text{changement de variable } t^2 = s) \\
&\leq p\,\sqrt{K_1}^{\,p}\;(\,\frac{p}{2}\,)^{\frac{p}{2}} \qquad (\,\Gamma\,(p) < p^p\,)
\end{aligned}$$

Donc, on déduit :

$$\begin{aligned}
\mathbb{E}(\,|X|^p\,)^{\frac{1}{p}} &\leq K_1\,p^{\frac{1}{p}}(\,\frac{p}{2}\,)^{\frac{1}{2}} \\
&= \underbrace{K_1\,\frac{\sqrt[p]{p}}{\sqrt{2}}}_{=K_2}\,\sqrt{p}.
\end{aligned}$$

$(ii) \Longrightarrow (iii)$
Supposons que la propriété (ii) est vérifiée, on va montrer d'abord que (ii) implique

$$\mathbb{E}(\,\exp\,(\,\lambda^2\,X^2\,)\,) \;\leq\; \exp(K_3^2\lambda^2), \quad \text{pour tout} \quad |\lambda| \leq \frac{1}{K_3}. \tag{3.1}$$

Qui est plus générale que la propriété (iii), et à l'aide d'un développement de Taylor

$$\begin{aligned}\mathbb{E}(\exp(\lambda^2 X^2)) &= \mathbb{E}\left(1+\sum_{n\geq 1}\frac{\mathbb{E}((\lambda^2 X^2)^n)}{n!}\right)\\ &= 1+\sum_{n\geq 1}\frac{\lambda^{2n}\,\mathbb{E}(X^{2n})}{n!}.\end{aligned}$$

En utilisant la propriété (ii) : $\mathbb{E}(X^{2n}) \leq (2K_2 n)^n$ et la formule de striling : $(n! \geq (\frac{n}{e})^n)$ on trouve :

$$\begin{aligned}\mathbb{E}(\exp(\lambda^2 X^2)) &\leq 1+\sum_{n\geq 1}\frac{(2\lambda^2 K_2 n)^n}{(\frac{n}{e})^n}\\ &= \sum_{n\geq 0}(2\lambda^2 K_2 n)^n)\\ &= \frac{1}{1-2e\lambda^2 K_2}\end{aligned}$$

La série $\sum_{n\geq 0}(2\lambda^2 K_2 n)^n)$ converge si $2\lambda^2 K_2 e \leq 1$,
et on sait que $\frac{1}{1-x} \leq \exp(x) \quad \forall x \in [0, \frac{1}{2}]$, alors

$$\begin{aligned}\mathbb{E}(\exp(\lambda^2 X^2) &\leq \exp(4e\lambda^2 K_2) \qquad \text{pour tout } |\lambda| \leq \frac{1}{2\sqrt{e\,K_2}}.\\ &\leq \exp(\lambda^2 K_3^2) \qquad \text{avec } K_3 = 2\sqrt{e\,K_2}\end{aligned}$$

ce qui implique la propriété (iii), en prenant $K_3 = \frac{1}{2\sqrt{e}}$.

$(iii) \Longrightarrow (i)$

$$\begin{aligned}\mathbb{P}(|X| \geq t) &= \mathbb{P}(X^2 \geq \exp(t^2))\\ &\leq \exp(-t^2)\mathbb{E}(X^2) \qquad (\text{ Inégalité de Markov })\\ &\leq 2\exp(-t^2) \qquad (\text{ propriété } (iii) \text{ avec } K_3 = 1)\end{aligned}$$

donc (i) est prouvé avec $K_1 = 1$.

$(iii) \Longrightarrow (iv)$
On a $\exp(X) \leq X + \exp(X^2)$ est vrai pour tout X.

$$\begin{aligned}\mathbb{E}(\exp(\lambda X)) &\leq \mathbb{E}(\lambda X + \exp(\lambda^2 X^2))\\ &\leq \mathbb{E}(\exp(\lambda^2 X^2))\\ &\leq \exp(\lambda^2 K_3^2) \qquad si\ |\lambda| \leq \frac{1}{K_3}\end{aligned}$$

Donc la propriété (iv) est vérifiée avec $K_3 = K_4$ si $|\lambda| \leq \frac{1}{K_3}$.
Si $|\lambda| \geq \frac{1}{K_3}$, on va utiliser $\lambda X \leq \lambda^2 + X^2$ qui est vrai $\forall X$ et $\forall \lambda$

$$\begin{aligned}
\mathbb{E}(\exp(\lambda X)) &\leq \exp(\lambda^2)\,\mathbb{E}(\exp(X^2)) \\
&\leq 2\exp(\lambda^2) \qquad \text{(propriété (iii) avec } K_3 = 1) \\
&\leq \exp(2\lambda^2) \qquad (\text{car } |\lambda| \geq \frac{1}{K_3})
\end{aligned}$$

D'où la propriété (iv) est vérifiée avec $K_4 = \sqrt{2}$, d'où le résultat.

$(iv) \Longrightarrow (i)$
Supposons que propriété (iv) est vérifiée, et soit $\lambda \geq 0$, donc à l'aide de Markov :

$$\begin{aligned}
\mathbb{P}(X \geq t) &\leq \exp(-\lambda t)\,\mathbb{E}(\exp(\lambda X)) \\
&\leq \exp(-\lambda t + \lambda^2 K_4^2) \qquad (\text{Propriété } (iv)) \\
&\leq \exp(\frac{-t^2}{4K_4^2}) \qquad (\text{minimisation de } g(\lambda) = -\lambda t + \lambda^2 K_4^2).
\end{aligned}$$

On refait la même démonstration pour $\mathbb{P}(-X \geq t)$.

Donc : $\mathbb{P}(|X| \geq t) \leq \exp(\frac{-t^2}{4K_4^2})$,
d'où la propriété (i), avec $K_1 = 2K_4$.

Définition 3.4. Une variable aléatoire qui satisfait l'une des propriétés précédentes est appelée sous gaussienne.

Remarque 3.5. La meilleure valeur de la constante K_3 est dite norme sous gaussienne qui est définie de la façon suivante :

$$\|X\|_{\psi_2} = \inf\{\, t > 0\,,\, \mathbb{E}\exp(X^2/t^2) \leq 2\,\}.$$

La constante K_i dans chacune des inégalités est égale à $\|X\|_{\psi_2}$ à constante près :
c.à.d :

$$\text{Pour tout } C_i > 0, \qquad K_i = C_i\,\|X\|_{\psi_2}.$$

Reprenons la proposition 3.3. en terme de norme sous-gaussienne. Il indique que chaque variable aléatoire sous-gaussienne X satisfait les limites suivantes :

$$\begin{gathered}
\mathbb{P}(|X|) \leq 2\exp(-C_1 t^2/\|X\|_{\psi 2}^2) \qquad \forall t \geq 0 \\
\|X\|_p \leq C_2 \|X\|_{\psi 2}\sqrt{p} \qquad \forall\, p \geq 1 \\
\mathbb{E}(exp(X^2/\|X\|_{\psi 2}^2)) \leq 2.
\end{gathered}$$

Si $\mathbb{E}(X) = 0$, alors

$$\mathbb{E}(\exp(\lambda X)) \leq \exp(C_3\, \lambda^2\, \|X\|_{\psi 2}\,), \qquad \forall \lambda \in \mathbb{R}.$$

Pour prouver qu'une fonction est une norme, elle doit satisfaire les 3 propriétés suivantes :

i. $p(\mathbf{u} + \mathbf{v}) \leq p(\mathbf{u}) + p(\mathbf{v})$
ii. $p(a\mathbf{v}) = |a| p(\mathbf{v})$
iii. Si $p(\mathbf{v}) = 0$, et $\mathbf{v} = 0$.

Nous allons maintenant montrer chacune de ces propriétés pour la norme sous-gaussienne $\|\cdot\|_{\psi_2}$.

i. Soit $f(x) = e^{x^2}$, on a

$$\begin{aligned} f\left(\frac{|X+Y|}{a+b}\right) &\leq f\left(\frac{|X|+|Y|}{a+b}\right) \\ &\leq \frac{a}{a+b} f\left(\frac{|X|}{a}\right) + \frac{b}{a+b} f\left(\frac{|Y|}{b}\right) \quad \text{Inégalité de Jensen (fonction convexe)} \\ f\left(\frac{|X+Y|}{a+b}\right) &\leq \frac{a}{a+b} f\left(\frac{|X|}{a}\right) + \frac{b}{a+b} f\left(\frac{|Y|}{b}\right) \\ f\left(\frac{|X+Y|}{\|X\|_{\psi_2} + \|Y\|_{\psi_2}}\right) &\leq \frac{a}{a+b} 2 + \frac{b}{a+b} 2 = 2 \quad \textit{prenant} \quad \text{a= } \|X\|_{\psi_2} \text{ et } b = \|Y\|_{\psi_2} \end{aligned}$$

donc $\|X\|_{\psi_2} + \|Y\|_{\psi_2}$ est dans l'ensemble $t > 0 : \exp(X^2/t^2) \leq 2$, et la preuve est complète..

ii. On a

$$\begin{aligned} \|a\, X\|_{\psi_2} &= \inf\{\, t > 0 : \exp(-(a\, X)^2/t^2) \leq 2\,\} \\ &= \inf\{\, a\, u > 0 : \exp(-X^2/t^2) \leq 2\,\} \quad \textit{prenant} \quad \text{t=a u} \\ &= a\, \|X\|_{\psi_2}. \end{aligned}$$

iii. On a

$$\begin{aligned} &\|X\|_{\psi_2} = 0 \\ \Rightarrow &\inf t > 0 : \exp(-X^2/t^2) \leq 2 = 0 \\ \Rightarrow &\exp(-X^2/t^2) \leq 2 \\ \Rightarrow &\exp(-X^2/t^2) \leq 2 \quad \text{Inégalité de Jensen} \\ \Rightarrow &X^2 \leq -t^2 \log 2 \\ \Rightarrow &X^2 \leq \lim_{t \to 0} -t^2 \log 2 \\ \Rightarrow &X^2 \leq 0 \\ \Rightarrow &X = 0. \end{aligned}$$

Remarque 3.6. La constante K_1 dans chacune des inégalités est égale à $\|X\|_{\psi_2}$, à une constante près,
c.à.d : $K_i = C_i \|X\|_{\psi_2} \quad et \quad C_i > 0$

— $X \hookrightarrow N(0,1)$ est une variable sous gaussienne.

— $X \hookrightarrow N(\mu, \sigma)$ est une variable sous gaussienne.

— $X \hookrightarrow Bern(p)$ est une variable sous gaussienne.

Chaque variable aléatoire bornée est sous gaussienne, par contre les variables aléatoires qui suivent une loi de poisson, exponentielle ou cauchy ne sont pas des sous gaussiennes.

Proposition 3.7. (**La somme des variables sous gaussiennes**)

Soit $X_1, ..., X_N$ des variables aléatoires sous gaussiennes indépendantes, telle que $\mathbb{E}(X_i) = 0$, alors :
$S = \sum_{i=1}^{N} X_i$ est sous gaussienne et $\| \sum_{i=1}^{N} X_i \|_{\psi 2}^2 \leq C \sum_{i=1}^{N} \|X_i\|_{\psi 2}^2$
avec C une constante positive.

Démonstration : On calcule la fonction génératrice des moments de S

$$\begin{aligned}
\mathbb{E}(\exp(\lambda \sum_{i=1}^{N} X_i)) &= \prod_{i=1}^{N} \mathbb{E}(\exp(\lambda X_i)) \\
&\leq \underbrace{\prod_{i=1}^{N} \exp(C\lambda^2 \|X_i\|_{\psi 2}^2)}_{=\exp(\lambda^2 K^2)} \qquad \text{(d'aprés la propriété (iv) de la proposition 3.3.).} \\
\mathbb{E}(\exp(\lambda S)) &\leq \exp(\lambda^2 K^2) \qquad \text{avec } K^2 = C \sum_{i=1}^{N} \|X_i\|_{\psi 2}^2.
\end{aligned}$$

D'où, d'après la dernière propriété de la proposition 3.3, S est sous gaussienne et $\|S\|_{\psi_2} \leq C' K$, avec C' une constante positive.

3.2 La concentration sous exponentielle

Les distributions sous-exponentielles sont une famille spéciale de distributions à borne lourde. Le nom provient d'une de leurs propriétés, que leurs bornes diminuent plus lent que la loi exponentielle, et plus lent que celle se la distribution sous gaussienne.

Proposition 3.8. Soit X une variable aléatoire, les propriétés suivantes sont équivalentes :

i. $\mathbb{P}(|X| \geq t) \leq 2 \exp(-t/K_1) \qquad , \forall t \geq 0.$

ii. $\|X\|_p \leq K_2\, p, \qquad \forall p \geq 0.$

iii. $\mathbb{E}\,(\,\exp\,(\,|X|/K_3\,)\,) \leq 2.$

De plus, si $\mathbb{E}\,(\,X\,) = 0$ alors (i), (ii) et (iii) sont équivalentes à :

iv. $\mathbb{E}\,(\,\exp(\lambda X)\,) \leq \exp\,(\,\lambda^2\, K_4^2\,), \qquad \forall\, |\lambda| \leq \frac{1}{K_4}.$

Définition 3.9. Une variable aléatoire qui satisfait l'une des propriétés précédentes est appelée sous exponentielle.

Remarque 3.10. La meilleure valeur de la constante K_3 est dite norme sous exponentielle qui est définie de la façon suivante :

$$\|X\|_{\psi_1} = \inf\{t > 0, \mathbb{E}\,\exp\,(\,|X|\,/t\,) \leq 2\}.$$

Les distributions sous-gaussiennes et sous-exponentielles sont liées. Tout d'abord, toutes les variables aléatoires sous-gaussiennes sont des variables aléatoires sous-exponentielles. Deuxièmement, le carré d'une variable aléatoire sous-gaussienne est sous-exponentielle.

Proposition 3.11. Une variable X est sous gaussienne si et seulement si X^2 est sous exponentielle en plus : $\|X^2\|_{\psi 1} = \|X\|_{\psi 2}^2$.

Démonstration : $\|X^2\|_{\psi 1}$ est le minimum des nombres $t > 0$ qui satisfait $\mathbb{E}\,(\,\exp\,(\,X^2/t\,)\,) \leq 2$, alors que $\|X\|_{\psi 2}$ est le minimum des nombres $L > 0$ qui satisfait $\mathbb{E}\,(\,\exp(X^2/L^2\,)\,) \leq 2$, on posons $t = L^2$, on trouve le résultat.

Proposition 3.12. Soit X et Y deux variables aléatoires indépendantes et sous gaussiennes, alors XY est sous exponentielle et $\|X\,Y\|_{\psi 1} \leq \|X\|_{\psi 2}\,\|Y\|_{\psi 2}$.

Démonstration : Soit $\|X\|_{\psi 2} = K$ et $\|X\|_{\psi 2} = L$,
donc, il faut montrer que $\mathbb{E}(\exp(\frac{|X\,Y|}{K\,L})) \leq 2$, on va s'en servir de l'inégalité de Young : $a\,b \leq \frac{a^2}{2} + \frac{b^2}{2} \quad \forall\, a, b \in \mathbb{R}$.

$$\begin{aligned}
\mathbb{E}(\exp(\frac{|X\,Y|}{K\,L})) &\leq \underbrace{\mathbb{E}(\exp(\frac{X^2}{2\,K^2} + \frac{Y^2}{2\,L^2}))}_{\mathbb{E}(\exp(\frac{X^2}{2\,K^2})\,\exp(\,\frac{Y^2}{2\,L^2}))} \\
&\leq \frac{1}{2}\,\mathbb{E}(\exp(\frac{X^2}{K^2}) + \exp(\,\frac{Y^2}{L^2})) \qquad (\text{ par l'inégalité de Young }) \\
&\leq \frac{1}{2}\,\mathbb{E}(2 + 2) = 2.
\end{aligned}$$

D'où le résultat.

3.2.1 Théorème de Bernstein

Théorème 3.13. Soit $X_1, ..., X_n$ des variables aléatoires sous exponentielles indépendantes, et $\mathbb{E}(X_i) = 0$. $\forall t \geq 0$, on a

$$\mathbb{P}\{|\sum_{i=1}^{n} X_i| \geq t\} \leq 2\exp[-c\min(\frac{t^2}{\sum_{i=1}^{n}\|X\|_{\psi_1}}, \frac{t}{\max\|X\|_{\psi_1}})].$$

Démonstration : Soit $S = \sum_{i=1}^{n} X_i$

$$\mathbb{P}(S \geq t) \leq \exp(-\lambda t)\prod_{i=1}^{n}\mathbb{E}(\exp(\lambda X_i)) \tag{3.2}$$

Pour un choix de λ très petit tel que

$$0 \leq \lambda \leq \frac{\tilde{c}}{\max\|X\|_{\psi_1}} \tag{3.3}$$

$\tilde{c} > 0$, et en appliquant la dernière propriété de la variable aléatoire sous exponentielle on a :

$$\mathbb{E}(\exp(\lambda X_i) \leq \exp(C\lambda^2\|X_i\|_{\psi_1}^2)$$

En remplaçant dans (3.2) :

$$\mathbb{P}(S \geq t) \leq \exp(-\lambda t + C\lambda^2\sigma^2) \quad \text{avec } \sigma^2 = \sum_{i=1}^{n}\|X_i\|_{\psi_1}^2.$$

Soit $g(\lambda) = -\lambda t + C\lambda^2\sigma^2$, minimisant la fonction g qui satisfait (3.3), on trouve $\mathbb{P}(S \geq t) \leq \exp(-c\min(\frac{t^2}{\sigma^2}, \frac{t}{\max\|X_i\|_{\psi 1}}))$
avec $c = \min(\frac{1}{4C}, \frac{\tilde{c}}{2})$.

On refait la même preuve pour $-X_i$, on obtient :

$$\mathbb{P}(-S \geq t) \leq \exp(-c\min(\frac{t^2}{\sigma^2}, \frac{t}{\max\|X_i\|_{\psi 1}})),$$

d'où le résultat.

4 Lemme de Johnson Lindestrauss

On s'intéresse à la question suivante : étant donné des points dans un espace euclidien de grande dimension, est-il possible de les envoyer linéairement dans un espace de petite dimension sans trop modifier les distances entre ces points ?

Le lemme de Johnson-Lindenstrauss répond à cette interrogation et il établit qu'un ensemble de points dans un espace euclidien de grande dimension peut être plongé dans un espace de plus petite dimension, avec une faible distorsion, c'est-à-dire en préservant les distances de façon approchée. Avant d'énoncer le théorème, nous allons voir quelques notions dont on aura besoin pour le démontrer.

4.1 Isotropie

Définition 4.1. un vecteur aléatoire X est dit isotropique si

$$\mathbb{E}\,(\,X\,X^T\,) = \mathbb{I}_n.$$

Remarque 4.2. $X \hookrightarrow N\,(0,1)$ est isotropique car $\mathbb{I}_n = \mathbb{E}\,(\,X\,X^T\,) - \underbrace{\mathbb{E}\,(X)^2}_{=0}$.

Toute variable aléatoire X de variance strictement positive peut se transformer en une variable Y centrée réduite :

$$Y = \frac{X - \mathbb{E}\,(X)}{\sqrt{Var(X)}}$$

qui est isotropique

Et tout vecteur aléatoire de matrice de variance-covariance Σ inversible peut se transformer en un vecteur centré isotropique Z :

$$Z = (\Sigma)^{-1/2}(X - \mathbb{E}(X))$$

Proposition 4.3. $X \in \mathbb{R}^n$ un vecteur aléatoire est dit isotropique si et seulement si :

$$\mathbb{E}\,(\,<X\,,\,x>^2\,) = \|x\|_2^2, \qquad \forall\, x \in \mathbb{R}^n.$$

Démonstration :

$$\begin{aligned} X \text{ est isotropique} &\iff \mathbb{E}\,(\,X\,X^T\,) = \mathbb{1}_n \\ &\iff x^T\,\mathbb{E}\,(\,X\,X^T\,)\,x \;=\; \|x\|_2^2 \\ &\iff \mathbb{E}\,(\,<X, x>^2\,) = \|x\|_2^2. \end{aligned}$$

4.2 Centrage

Propositon 4.4. Soit X une variable aléatoire sous exponentielle alors $X - \mathbb{E}(X)$ l'est aussi et $\|X - \mathbb{E}\,(X)\|_{\psi_1} \leq C\|X\|\psi_1$.

Démonstration : Pour une constante $\|a\|_{\psi 1} = C_1 \mid a \mid$ avec $C_1 = \ln 2$,
$\|\mathbb{E}(X)\|_{\psi 1} = C_1 \mid \mathbb{E}(X) \mid \leq \mathbb{E}(\mid X \mid)$ par inégalité de Jensen, la preuve de la proposition est

$$\begin{aligned}
\|X - \mathbb{E}(X)\|_{\psi_1} &\leq \|X\|_{\psi 1} + \| \mathbb{E}(X)\|_{\psi 1} \\
&\leq \|X\|_{\psi 1} + C_1 \, |\, \mathbb{E}(X)| \\
&\leq \|X\|_{\psi 1} + C_1 \; \mathbb{E}(|X|) \\
&\leq \|X\|_{\psi 1} + C_1 \, \| \mathbb{E}(X)\|_1 \\
&\leq \|X\|_{\psi 1} + C_1 \|X\|_{\psi 1} \qquad (\text{car } X \text{ est sous gaussienne}) \\
&\leq C' \, \|X\|_{\psi 1},
\end{aligned}$$

avec $C' = 1 + C$ d'où le résultat.

4.3 Théorème de Johnson Lindestrauss

Etant donné un ensemble $\mathscr{X}$ de N points dans $\mathbb{R}^n$ et $\epsilon \in]0,1[$.
On définit la matrice de projection P tel que :

$$P = \frac{1}{\sqrt{m}} \begin{pmatrix} X_1^T \\ \vdots \\ X_m^T \end{pmatrix} \in \mathrm{R}^{m \times n}$$

avec $X_1, ..., X_m$ sont des vecteurs de $\mathbb{R}^n$ sous gaussiennes centrées isotropiques.

Si m vérifie $m \geq C\,\epsilon^{-2}\log(N)$, alors on a pour tout $x, y \in \mathscr{X}$:

$$\mathbb{P}\Big((1-\epsilon)\|x-y\|_2 \leq \|Px-Py\|_2 \leq (1-\epsilon)\|x-y\|_2\Big) \geq 1 - 2\exp(-c\epsilon^2 m)$$

où C, c des constantes positives .

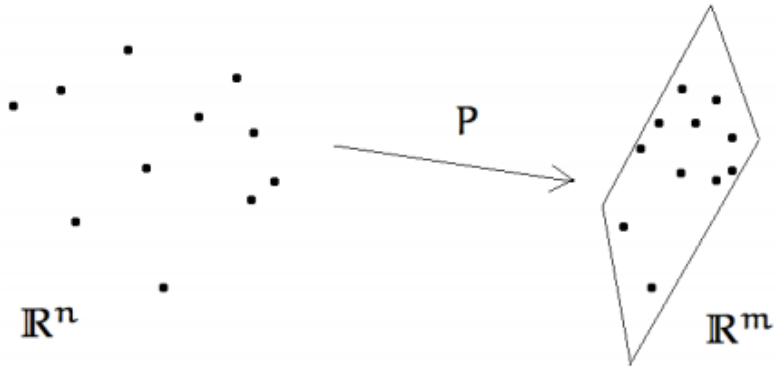

FIGURE 1 – Projection aléatoire de N points de données

Le théorème indique qu'après avoir fixé un niveau d'erreur ϵ, on peut projeter un ensemble de points d'un espace euclidien (quelle que soit sa dimension n) vers un espace euclidien plus petit, tout en modifiant la distance entre deux points par un facteur de $1 \pm \epsilon$ (on conserve approximativement les distances entre les points). La dimension de l'espace image ne dépend que de l'erreur ϵ et du nombres de points N. Étant donné que la dimension est très grande, on peut réaliser une réduction significative de la dimension, qui a des applications en analyse de données et en informatique.

Remarque 4.5. Pour la constante C, les mathématiciens Frankl et Maehara ont montré que $C \simeq 9$[2], et dans la preuve de Dasgupta et Gupta $C \simeq 8$ [3] .

Remarque 4.6. Par la suite on s'intéresse au cas particulier où les X_i sont des vecteurs aléatoires gaussiennes centrées et réduites, c'est à dire que les composantes de X_i, suivent la normale centrée réduite $N(0,1)$, alors dans ce cas on peut calculer la constante c. En effet :
Si $X \hookrightarrow N(0,1)$, alors $Y = X^2 \hookrightarrow \chi_2$ et $M_Y(t) = (1-2t)^{-\frac{1}{2}}$.

$$\begin{aligned} \Longrightarrow \|X\|_{\phi 2} &= \inf\{\, t>0 \quad , \mathbb{E}((\exp(\frac{X^2}{t^2})) \le 2\,\} \\ &= \inf\{\, t>0 \quad , M_y(\frac{1}{t^2}) \le 2\,\} \\ &= \inf\{\, t>0 \quad , (1-\frac{2}{t^2})^{-\frac{1}{2}} \le 2\,\}. \end{aligned}$$

En résolvant $(1-\frac{2}{t^2})^{-\frac{1}{2}} = 2 \implies t = \frac{2\sqrt{2}}{\sqrt{3}}$ $d'o\grave{u}$ $\|X\|_{\psi 2} = \frac{2\sqrt{2}}{\sqrt{3}}$.
On sait que $\|X\,Y\|_{\psi 1} \le \|X\|_{\psi 2}\,\|Y\|_{\psi 2}$ en cas où X et Y sont des variables aléatoires sous gaussiennes indépendantes.
Donc, en prenant $Y = 1$, on a

$$\begin{aligned} \|Y\|_{\phi 2} &= \inf\{\, t>0 \quad , \mathbb{E}((\exp(\frac{1}{t^2})) \le 2\,\} \\ &= \frac{1}{\sqrt{\log(2)}} \quad (\text{ en résolvant } \exp(\frac{1}{t^2}) = 2\,) \end{aligned}$$

d'où $\|X\|\psi_1 \le \frac{1}{\sqrt{\log(2)}}\,\frac{2\sqrt{2}}{\sqrt{3}} = \frac{2\sqrt{2}}{\sqrt{3\,\log(2)}}$, finalement on trouve $c = \frac{2\sqrt{2}}{\sqrt{3\,\log(2)}}$.

Démonstration : Théorème de Johnson Lindestrauss

Par la linéarité de P,
$P\,x - P\,y = P\,(x-y)$, en divisant par $\|$x-y$\|_2$, on trouve

$$(1-\epsilon) \le \|\mathrm{P\,z}\|_2 \le (1+\epsilon) \quad \forall z \in T \text{ avec T} = \{\ \tfrac{x-y}{\|x-y\|_2} : x,y \in X \quad et \quad x \ne y\ \}.$$

Passant au carré et en utilisant $(1+\epsilon) \le (1+\epsilon)^2$ et $(1-\epsilon) \ge (1-\epsilon)^2$, nous trouvons

$$(1-\epsilon)\ \le \|\mathrm{P\,z}\|_2^2 \le\ (1+\epsilon), \quad \forall\, z \in T.$$

Les cordonnées de P z $= \frac{1}{\sqrt{m}}$A z sont $\frac{1}{\sqrt{m}}\,\langle X_i, z\,\rangle$, donc notre inégalité devient

$$|\ \tfrac{1}{\sqrt{m}}\textstyle\sum_{i=1}^{n}\langle X_i, z\rangle^2 - 1\ | \le\ \epsilon, \quad \forall z \in T.$$

Posons $Y_i = \langle X_i, z\rangle^2 - 1$, on accepte que les Y_i sont indépendants, et elles sont centrés car elles sont isotropiques. En effet, d'après la proposition 4.3. on a :

$$\begin{aligned} \mathbb{E}\,(< X_i, z >^2 -1) &= \mathbb{E}\,(< X_i, z >^2) - 1 \\ &= \|z\|^2 - 1 \\ &= 0. \end{aligned}$$

Les X_i sont des sous gaussiennes, alors $<X_i, z>$ est sous gaussienne. Donc $< X_i, z >^2 -1$ est sous exponentielle, d'où on peut appliquer le lemme de Berstein.

$$\mathbb{P}\left(\left|\frac{1}{m}\sum_{i=1}^{n}(< X_i, z >^2 -1)\right|\right) \leq 2\exp\left[-c\min\left(\frac{\epsilon^2}{\sum_{i=1}^{n}\|\frac{1}{m} < X_i, z >^2 -1\|_{\psi_1}^2}, \frac{\epsilon}{\max\|\frac{1}{m} < X_i, z >^2 -1\|_{\psi 1}}\right)\right].$$

Soit $K = \| < X_i, z > \|_{\psi 1}$ (acceptant $K \geq 1$), en utilisant la proposition 4.4, on a :

$$\| < X_i, z >^2 -1\|_{\psi 1} \leq C\,K^2 \Longrightarrow \sum_{i=1}^{n}\| < X_i, z >^2 -1\|_{\psi_1}^2 \leq m\,C\,K^4$$

$$\text{et} \quad \max\|\frac{1}{m} < X_i, z >^2 -1\|_{\psi 1} \leq C\,K^2.$$

Donc

$$\mathbb{P}\left(\left|\frac{1}{m}\sum_{i=1}^{n}(< X_i, z >^2 -1)\right| \geq \epsilon\right) \quad \leq \quad 2\,\exp\left(-c\,m\min(\frac{\epsilon^2}{C\,K^4}, \frac{\epsilon}{C\,K^2})\right).$$

On a $C \geq 1, car\ C = 1 + C_1\ et\ K^4 > K^2 \quad et \quad \epsilon^2 \leq \epsilon \quad \forall\,\epsilon \in [0,1]$, alors

$$\mathbb{P}\left(\left|\frac{1}{m}\sum_{i=1}^{n}(< X_i, z >^2 -1)\right| \geq \epsilon\right) \leq \quad 2\exp\left(-c\,m\,\epsilon^2\right).$$

En prenant l'union sur tous les z possibles :

$$\begin{aligned}\mathbb{P}(\max_z|\frac{1}{m}\sum_{i=1}^{n}(< X_i, z >^2 -1) \geq \epsilon \quad &\leq \quad \sum_{z\in T}\mathbb{P}(|\frac{1}{m}\sum_{i=1}^{n}(< X_i, z >^2 -1)| \geq \epsilon) \\ &\leq \quad |T|\,2\exp(-c\,m\,\epsilon^2).\end{aligned}$$

Par définition de T, $|T| \leq N^2$, donc si on choisit $m \geq C\,\epsilon^{-2}\log(N)$ avec une constante C strictement positive la probabilité de préserver la distance entre les points, après la projection en utilisant le lemme de Lindestrauss est très proche de 1-$\frac{1}{N}$, donc tant que N le nombre de points est grand, tant que la probabilité tend vers 1.

5 Implémentation de la méthode de Johnson Lindestrauss

On fera l'implémentation en trois étapes :

Partie 1 : Soit N le nombre de points issu de nos données qu'on souhaite projeter et une erreur relative ϵ, que nous pouvons tolérer, on peut alors calculer la dimension minimale m de l'espace dans lequel on va projeter nos données tel que les distances soient préserver à notre erreur relative prés.

D'après le théorème :

On sait que : $\mathbb{P}\left((1-\epsilon)\|x-y\|_2 \leq \|Px-Py\|_2 \leq (1-\epsilon)\|x-y\|_2\right) \geq 1-2\exp(-c\epsilon^2 m)$ dès que $m \geq C\,\epsilon^{-2}\log(N)$ avec $\epsilon \in]0,1[$.

Et, en fixant $\alpha = 2\exp(-c\epsilon^2 m)$ comme un seuil de signification qu'on peut choisir arbitrairement 1 % (pour réussir a préserver les distances avec une probabilité de 0.99), donc pour choisir m la dimension de l'espace cibler, il faut respecter les deux contraintes suivantes (**avec ϵ fixé**) :

$$\begin{cases} \epsilon^2 m & \geq -\frac{1}{c}\log(\frac{\alpha}{2}) \\ \epsilon^2 m & \geq C\,\log(N) \end{cases}$$

Donc $m \geq \max(-\frac{1}{c\,\epsilon^2}\log(\frac{\alpha}{2}), \frac{C\,\log(N)}{\epsilon^2})$.

On prend m la plus petite valeur qui vérifie cette dernière inégalité.

Si on a 1 million de données ($N = 1000000$), et on souhaite conserver les distances avec une distorsion $\epsilon = 0.01$ près, alors m est supérieur à un million. Si on augmente epsilon à 0.1, on tombe sur m=11052 ce qui est plus raisonnable à $\epsilon = 0.4$ où on trouve $m = 700$. Sur le graphique en dessous on peut observer l'évolution de m en fonction de ϵ avec $N = 1000000$ et $C = 9$.

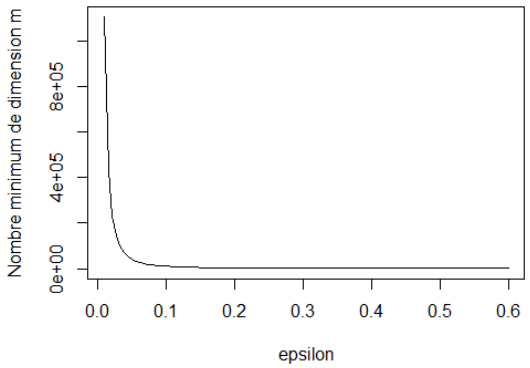

FIGURE 2 – Dimension m en fonction d'epsilon

Partie 2 : On a une matrice de données de N points dans un espace de dimension n, nous pouvons explicitement créer une matrice P de projection telle que la distance entre toute paire de ligne de X ne change qu'avec un facteur de $(1 \pm \epsilon)$ après la projection.
La matrice P est donné par

$$P = \frac{1}{\sqrt{m}} A$$

A est une matrice aléatoire dont les lignes sont des vecteurs sous gaussiennnes. Ici, on s'intéresse au cas des variables aléatoires gaussiennes qui est un cas particulier des variables sous gaussiennes, donc la matrice de projection P sera une matrice de variables aléatoires gaussiennes centrées réduites.

On a montré dans notre cas que $c = \frac{2\sqrt{2}}{\sqrt{3\ \log(2)}}$.

Partie 3 : On crée ensuite l'algorithme qui compare les distances à ϵ fixé en première partie afin de trouver le vrai ϵ (la vraie distorsion) qu'on appelle epsilon optimale . C'est la vraie distorsion avec laquelle les distances se préservent après la projection, qui vérifie pour chaque couple de point (x, y) la contrainte suivante :

$$\left| \frac{\|Px - Py\|}{\|x - y\|} - 1 \right| \leq \epsilon,$$

donc on prend $\epsilon_{opt} = \max \left| \frac{\|Px-Py\|}{\|x-y\|} - 1 \right|$.
On arrive alors à projeter notre base de données en connaissant la vraie distorsion.

En cherchant la distorsion sur une base de données de 10000 individus et 15000 variables (en générant une matrice gaussienne de taille 10000*15000), on trouve la distorsion égale à 0.312753, le temps d'exécution prend beaucoup de temps, environ 6 heures.

Remarque 5.1. Ici, chaque fois qu'on relance le programme, la réalisation de P change, donc on remarque des cas où la distorsion est un peu grande (supérieure à 4), mais en général elle est raisonnable (entre 0.2 et 0.35).

Pour l'application on s'intéresse à l'algorithme des k-means, l'une des techniques d'apprentissage non supervisée, qui vise à regrouper un ensemble d'objets de manière à ce que les objets du même cluster se ressemblent davantage que les objets des autres clusters.

Description de l'algorithme k-means Étant donné un ensemble de points $(x_1, x_2, ..., x_n)$, on cherche à partitionner les n points en k ensembles $S = \{S_1, S_2, \ldots, S_k\}$ $(k \leq n)$ en minimisant la distance entre les points à l'intérieur de chaque partition :

$$\underset{S}{\operatorname{argmin}} \sum_{i=1}^{k} \sum_{x_j \in S_i} \|x_j - c_i\|^2$$

où c_i est le barycentre des points dans S_i.

- Choisir k points qui représentent la position moyenne des partitions $m_1^1, ..., m_k^1$ initiales (initialisé au départ).
- Répéter jusqu'à ce qu'il y ait convergence :
  - assigner chaque observation à la partition la plus proche $S_i^t = \{x_j : \|x_j - c_i^t| \ \leq \|x_j - c_{i^*}^t\| \quad \forall i^* = 1, ..., k\}$.
  - mettre à jour la moyenne de chaque cluster : $c_i^{t+1} = \frac{1}{|S_i^t|} \sum_{x_j \in S_i^t} x_j$.

L'idée est très simple, on choisit k individus au hasard, qui constituent les centroïdes initiaux. On fait passer toutes les observations, on affecte chaque observation au centroïde qui lui est le plus proche, les centroïdes sont alors remis à jour. Puis, on réitère le passage de tous les individus jusqu'à ce que la solution soit stable.
k-means est basé sur le calcule des distances entre les individus, donc il sera un bon exemple pour étudier la projection aléatoire de Johnson-Lindestrauss .

L'objectif dans notre cas est d'appliquer l'algorithme des k-means à notre base de données, avant et après la projection aléatoire de Johnson-Lindestrauss, ensuite comparer les clusters obtenus dans les deux cas (avant et après la projection), en souhaitant garder les mêmes groupes après avoir diminuer la dimension.

Par la suite, on travaille sur une base de données de taille N=10000 et n=15000.
On applique l'algorithme des k-means sur la base de données et on calcule $\frac{N_1}{N}$ la proportion des individus du premier cluster et $\frac{N_1'}{N}$ celle du même cluster après la projection aléatoire de Johnson-Lindestrauss. On souhaite avoir $\frac{N_1'}{N} \simeq \frac{N_1}{N}$, on note $E' = \frac{|\, N_1' - N_1 \,|}{N}$ la différence des deux proportions. Pour conclure que le cluster contient presque les mêmes points avant et après la projection, il faut que E soit très proche de zéro. On peut voir E comme l'erreur commise sur le nombre d'individus dans le premier cluster par le lemme de Johnson-Lindestrauss.

On refait l'opération pour 100 réalisation de P et on note E la moyenne des erreurs E' commises dans chaque realisation.

Dans un premier temps, on fixe N le nombre d'individus et on fait varier n le nombre des variables.

Le graphe suivant trace cette erreur en fonction du nombre de variables étudiées

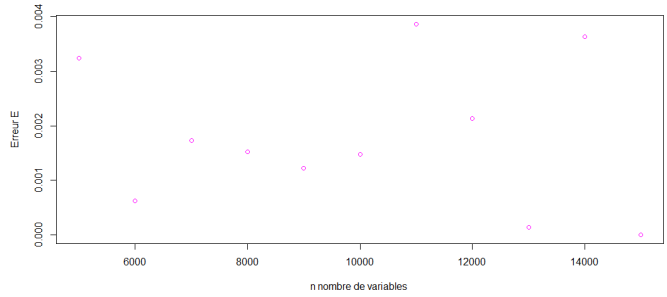

FIGURE 3 – l'erreur E en fonction du nombre des variables n

On remarque que l'erreur est faible, entre 0 et 0.004, ce qui signifie que les clusters gardent presque la même concentration d'individus, mais on ne voit pas l'influence de n, le nombre de variables sur E car les points du graphe sont repartis aléatoirement.

On note que l'étude de cette erreur dépend des réalisations aléatoires de P (matrice de projection), chaque fois qu'on relance le programme, il affiche un graphe différent mais généralement l'erreur reste faible.

Dans un premier temps, on a fixé N et on a fait varier n pour voir la dépendance de n en fonction de E. Cette fois, on fixe le nombre de variables n et on fait varié le nombre d'individus N. Le graphe suivant trace cette erreur en fonction du nombre d'individus étudiés

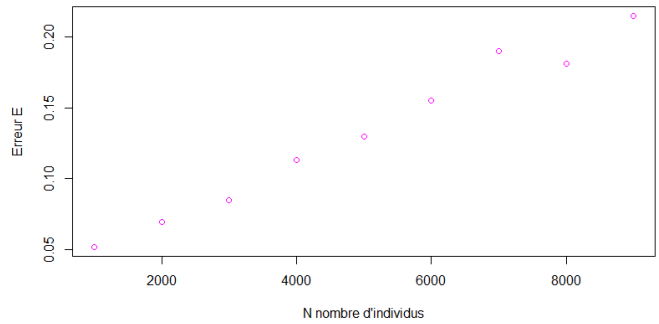

FIGURE 4 – L'erreur E en fonction du nombres d'individus N

Dans ce cas, E se situe entre 0 et 0.2. On remarque que l'erreur donnée en fonction du nombre d'individus N, est plus forte que l'erreur qu'on a déterminée en fonction du nombre de variables n. Remarquons une dépendance entre E et N car l'erreur augmente quand le nombre d'individus augmente.

Après avoir lancé le programme plusieurs fois, on conclut que, généralement, la proportion des individus reste la même dans chaque cluster avant et après la projection. Toutefois, grâce à la projection, on réduit le temps de calcul de façon très significative : de l'ordre de quatre fois moins longtemps.

6 Annexe

On génère une matrice avec rnorm qu'on considère comme notre base de données. On a déjà travaillé avec le jeu de données MicroMass [3] qui contient 1300 variables et 931 individus, et vu que la dimension de la base de données n'est pas assez importante, on arrive pas a trouvé des résultats surprenants.

On fixe la constance C et la tolérance ϵ et c pour calculer m la dimension d'espace après la projection.

```
data<-replicate(15000,rnorm(10000,mean = 0, sd=10))
N=nrow(data)
C=8
epsilon=0.2
m=ceiling((C*log(N))/epsilon^2)
```

On trace l'évolution de dimension m en fonction de epsilon.

```
fonc = function(epsilon) { ceiling((C*log(N))/epsilon^2)}
curve(fonc,xlab = "epsilon",
      ylab = "Nombre minimum de dimension",0.01,0.6)
```

Fonction 6.1. La fonction matrice-project génère une matrice de variables aléatoires gaussiennes centrées réduites.

```
matrice_project=function(k,d){

  A=replicate(d,rnorm(k,mean = 0, sd=1))
  P=1/sqrt(k)*A
  return(P)
}
matrice_project(m,n)
```

Fonction 6.2. La fonction dist calcule la distance entre chaque couple de points qu'on stocke dans une matrice M. La partie triangulaire inférieure de la matrice contient les distances projetées, et la partie triangulaire supérieure de M contient les distances entre les individus avant la projection.

```
M=array(data=c(0),dim =c(N,N))
dist=function(data){

  for (i in 1:(N-1)) {
    for(j in (i+1):N){
```

```
      M[j,i]<-norm(P%*%data[i,]-P%*%data[j,], type = c("2"))
      M[i,j]<-norm( data[i,]- data[j,], type = c("2"))
    }
  }

  return(M)
  }
```

Fonction 6.3. La matrice de projection aléatoire P, devrait conserver les distances avec une distorsion ϵ. La fonction num-point permet de voir le nombre de points qui ne vérifient pas $\left|\frac{\|Px-Py\|}{\|x-y\|}-1\right| \leq \epsilon$.

```
num=0
num_point=function(matrice,epsilon){


  for (i in 1:(N-1)) {
    for(j in (i+1):N){

      if (abs( matrice[j,i] / matrice[i,j]- 1) > epsilon)
        num = num + 1
    }
  }

  return(num)
}
num_point=(dist(data),epsilon)
```

Fonction 6.4. La fonction eps-optim récupère la valeur de la distorsion pour laquelle les distances se préservent avec une probabilité de 99 %.

```
H=array(c(0),dim = c(N(N+1)/2,1))
k=0
eps_optim=function(matrice){


  for (i in 1:(N-1)) {
    for(j in (i+1):N){
```

```
        H[k]<-abs( matrice[j,i] / matrice[i,j]- 1)

        k=k+1

      }
    }

    return(max(H))
}
eps_optim(dist(data),epsilon)
```

Application de K-means Dans cette partie du programme, on applique l'algorithme de k-means à nos données après et avant la projection, au début on fixe N et on varie n et on trace E en fonction de n, puis on fixe n et on varie N.

```
#Ici on fixe N et on varie n

 k=2
 #temps=array(data=c(0),dim=c(10,1))
 E=array(dat=c(0),dim = c(11,1))
 e=array(dat=c(0),dim = c(100,1))
 nmax=ncol(data)
 l=0
 for(i in seq(500,nmax,by=100) ) {
   #beg.time=Sys.time()
    print(i)
    u=array(data=c(0),dim=c(k,1))
    v=array(data=c(0),dim=c(k,1))
    dat=data[,1:1:i]
  for(j in 1:50){
     P=matrice_project(m,i)
     new_data=dat%*%t(P)
     km_old=kmeans(dat,k,
          iter.max =100,algorithm = "Lloyd")
     km_new=kmeans(new_data,k,
          iter.max =100,algorithm = "Lloyd")
     u=km_new$size
     v=km_old$size
     e[1]=abs(u[1] -v[1])/N
```

```
    }
  E[l]=sum(e)/50
  l=l+1
  #end.time<-Sys.time()
  #temps[i]=round(end.time - beg.time,2)
}
i=seq(5000,15000,by= 1000)
plot(i,E ,xlab= "n nombre d'individus ",
             ylab= "Erreur E",col="6")


# Ici on fixe n et on varie N
#temps=array(data=c(0),dim=c(10,1))
k=2
E=array(dat=c(0),dim = c(10,1))
e=array(dat=c(0),dim = c(10,1))
n=ncol(data)
l=0
for(i in seq(100,N,by=100) ) {
  u=array(data=c(0),dim=c(k,1))
  v=array(data=c(0),dim=c(k,1))
  dat=data[1:1:i,]
  for(j in 1:100){
    m= ceiling((C*log(i))/epsilon^2)
    P=matrice_project(m,n)
    new_data=dat%*%t(P)
    km_old=kmeans(dat,k,
          iter.max =100,algorithm = "Lloyd")
    km_new=kmeans(new_data,k,
          iter.max =100,algorithm = "Lloyd")
    u=km_new$size
    print(u)
    v=km_old$size
    print(v)
    e[j]=abs(u[1] -v[1])/N
  }
  E[l]=sum(e)/100
  l=l+1
  #T2<-Sys.time()
  #temps[i]=T2-T1
}
i=seq(1000,9000,by=1000)
```

```
plot(i, Ei, xlab = "N nombre d'individus",
                        ylab ="Erreur E",)
```

Bibliographie

[1] Roman Vershynin,High-Dimensional Probability ,University of California Irvine .March 25 2019 https ://www.math.uci.edu/ rvershyn/

[2] P. Frankl and H. Maehara. The Johnson-Lindenstrauss lemma and the sphericity of some graphs. Journal of Combinatorial Theory, Series B, 44(3) :355–362, 1988.

[3] S. Dasgupta and A. Gupta. An Elementary Proof of a Theorem of Johnson and Lindenstrauss. Random Structures and Algorithms, 22(1) :60–65, 2003.

[3] Banque de jeux de données ; UCI Machine Learning Repository ,https ://archive.ics.uci.edu/ml/datasets/MicroMass