L'exploration dans les espaces de grande dimension

Alain Franc & Olivier Coulaud

INRA BioGeCo & INRIA Pleiade ; INRIA Hiepacs

Journées "Mexico", Bordeaux
13 novembre 2018

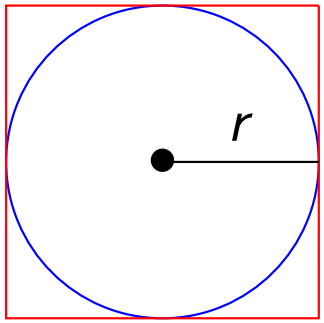
Motivations

- Nous avons l'habitude d'une géométrie à 3 dimensions
- sur laquelle est basée notre "intuition"
- mais ...
- dans les grandes dimensions ... cela se passe autrement ....
- $\Longrightarrow$ ce qui mène à de mauvaises "intuitions"

Dans l'exposé

- Certaines de ces "contre-intuitions" seront présentées (volumes)
- Des résultats contre-intuitifs sur les fonctions dans ces espaces
- une application à la réduction de la dimension
- une (brève) application à la biodiversité

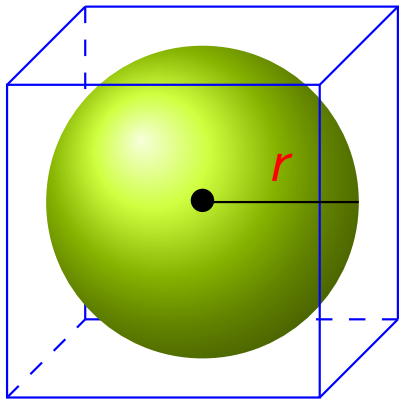
Géométrie simple à deux dimensions



Aires

cercle (incrit)	$A_{\mathrm{S}}(r)$	πr^2
carré	$A_{\mathrm{C}}(2r)$	$4r^2$
ratio	$A_{\mathrm{S}}(r)/A_{\mathrm{C}}(2r)$	$\pi/4$

Géométrie simple à 3 dimensions



Volumes

sphère	$V_{\mathrm{S}}(r)$	$\frac{4}{3}\pi r^3$
cube	$V_{\mathrm{C}}(2r)$	$8r^3$
ratio	$V_{\mathrm{S}}(r)/V_{\mathrm{C}}(2r)$	$\pi/6$

Géométrie de l'hypercube

Volume

$$\text{Vol} = a^n = (2r)^n, \qquad a = 1$$

Diagonale

En dimension n, la longueur de la diagonale d'un cube de coté 1 est

$$\Delta_n = \sqrt{\underbrace{1^2 + \ldots + 1^2}_{n \text{ fois}}} = \sqrt{n}$$

Observation

Dans un cube de coté $a = 1$ en dimension n il existe une ligne droite de longueur $\sqrt{n}$

Géométrie de l'hypersphère

Volume

$$\text{Vol} = V_n\, r^n, \qquad V_n = \frac{\pi^{n/2}}{\Gamma\left(1+\frac{n}{2}\right)} = \frac{\pi^m}{m!} = \frac{1}{\sqrt{\pi n}}\left(\sqrt{\frac{2e\pi}{n}}\right)^n$$

$$\text{avec} \quad n = 2m, \qquad m! = \sqrt{2\pi m}\left(\frac{m}{e}\right)^m$$

Observations

- Le volume d'une sphère de rayon 1 tend vers zéro ni n tend vers ∞
- Le rayon d'une sphère de volume 1 dans $\mathbb{R}^n$ est

$$\rho_n = \sqrt{\frac{n}{2e\pi}} \quad \longrightarrow \infty$$

Géométrie et mesures à n dimensions

Volume du cube

$$\text{volume } = (2r)^n = 2^n r^n$$

Volume de la sphère

$$\text{volume } = V_n r^n, \qquad V_n = \frac{1}{\sqrt{\pi n}} \left(\sqrt{\frac{2e\pi}{n}} \right)^n$$

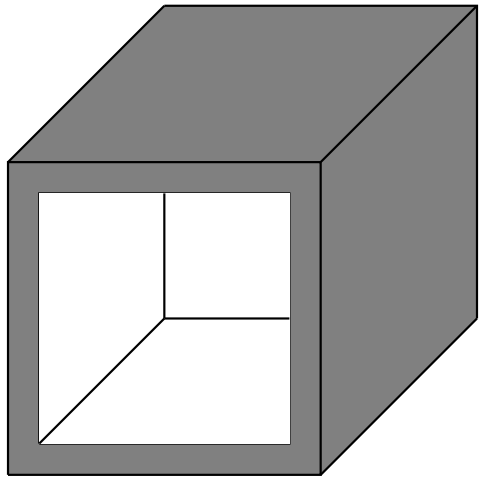
Ratio

$$\Omega_n = \frac{1}{\sqrt{\pi n}} \left(\sqrt{\frac{e\pi}{2n}} \right)^n \longrightarrow 0$$

Conséquence

Tous les points sont à l'extérieur de la sphère inscrite

Sur la peau d'un hyercube ...



Emboitement

$$\begin{array}{lll} \text{cube I} & a = 1 & \text{Vol} = 1 \\ \text{cube II} & a = 1 - \epsilon & \text{Vol} = (1-\epsilon)^n \end{array}$$

Observation

Tous les points sont sur la peau de l'hypercube $((1-\epsilon)^n \to 0)$

Loi gaussiennes

Distribution gaussienne à n variables

$$\begin{aligned} p(x) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp -\frac{x_i^2}{2} \\ &= \frac{1}{\sqrt{(2\pi)^n}} \exp -\frac{1}{2} \sum_i x_i^2 \end{aligned}$$

Queue de distribution

On peut (simplement) montrer que, $\forall\, r > 0$

$$\mathbb{P}\{\|x\| \leq r\} < \frac{2}{n} \left(r \sqrt{\frac{e}{n}} \right)^n \quad \longrightarrow 0$$

Observation (oxymore)

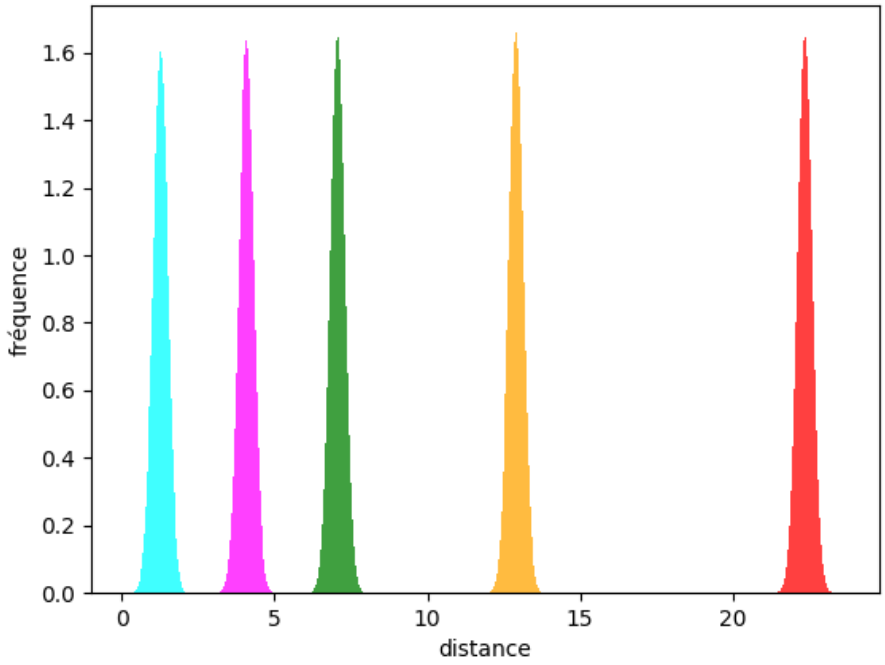
L'essentiel de la distribution d'une loi gausienne en grande dimension est concentrée dans la queue de distribution

Résumé des observations

Tout est à la **périphérie** de l'espace disponible ...(se le rappeler!)

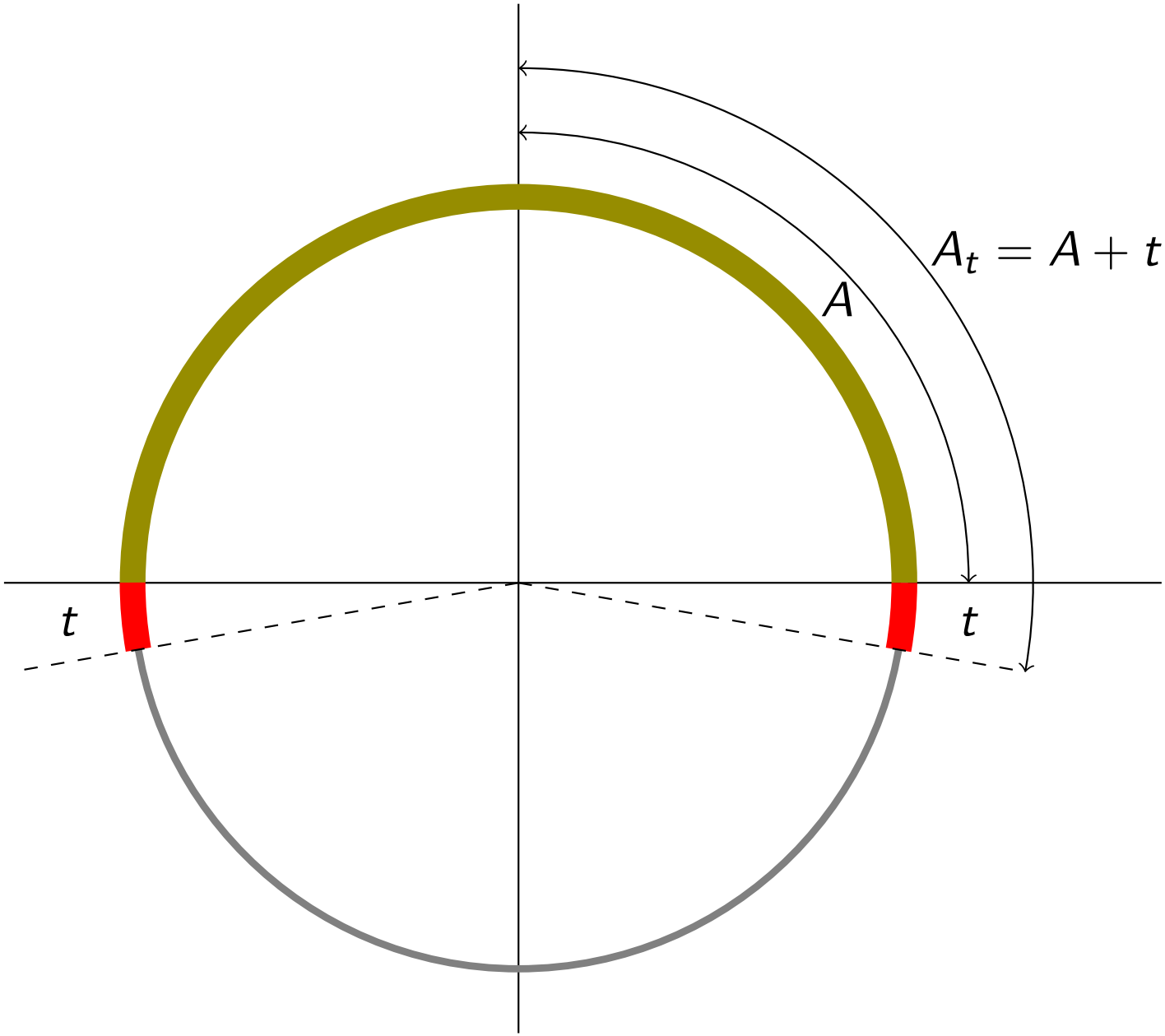
- le volume d'un hypercube de coté 1 dans $\mathbb{R}^n$ reste 1
- le volume de la sphère de rayon 1 dans $\mathbb{R}^n$ tend vers zéro
- le ratio entre le volume de la sphère inscrite dans le cube et du cube tend vers zéro
- un nuage de points aléatoires dans la sphère est sur la peau de la sphère en grande dimension
- un nuage de points aléatoires dans le cube est sur la peau du cube en grande dimension (l'essentiel des points est dans les coins)
- l'essentiel des points d'une gaussienne multivariée est à grande distance du centre (moyenne pour chaque dimension) = à la périphérie de la distribution

Distribution des distances (nuages aléatoires)



Histogramme des distances deux à deux de nuages de $m = 3000$ points aléatoires (loi uniforme dans le cube $[0,1]^n$) ; $n = 10$; $n = 100$; $n = 300$; $n = 1000$; $n = 3000$

Une esquisse de ce qui va suivre



Concentration de la mesure sur la sphère

- Sphère dans $\mathbb{R}^{n+1}$

$$\mathbb{S}^n = \{x \in \mathbb{R}^{n+1} \ : \ \|x\| = 1\}$$

- A est une calotte sphérique de mesure $\frac{1}{2}$ (l'hémisphère Nord si $n = 2$)
- alors (Lévy, 1919, 1951)

$$\mu(A_t) \geq 1 - \left(\exp - \frac{(n-1)t^2}{2}\right)$$

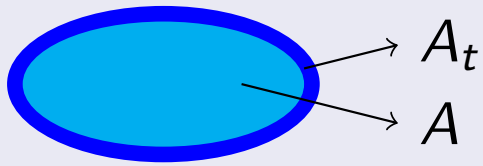
Observation

- Si on choisit un "équateur", l'essentiel des points sur la sphère est à distance $\frac{1}{\sqrt{n}}$ de cet équateur (90 % à $\frac{2.14}{\sqrt{n}}$). C'est une concentration (lente, en $\frac{1}{\sqrt{n}}$) de la mesure à l'équateur.
- L'essentiel des points est à la périphérie de la calotte.

Regard géométrique

Données

- une mesure μ sur un ensemble ($\int dx$ sur $\mathbb{R}^n$)
- une partie A telle que $\mu(A) = \frac{1}{2}$
- un réel $t > 0$
- $A_t = \{x \in \mathbb{R}^n \ : \ d(x, A) \leq t\}$



Concentration de la mesure

$$\mu(A_t) \geq 1 - \left(\exp -c(n)\frac{t^2}{2v}\right)$$

Concentration de la valeur d'une fonction

Fonction lipschitzienne

Fonction bornée dans ses variations: plus petit C tq

$$\forall\, x, y, \quad |f(x) - f(y)| \leq C|x - y|$$

Concentration de la mesure

- On se donne F une fonction $C-$Lipschitz

$$\mathbb{S}^{n-1} \xrightarrow{F} \mathbb{R}$$

- on note m_{F} la médiane de F
- alors

$$\mu\{x \,:\, |F(x) - m_{\mathrm{F}}| > t\} \leq 2\exp -\frac{(n-1)t^2}{2C^2}$$

Observation

Si n est grand, F est presque partout égale à sa médiane à t près.

Cela mérite un peu d'explication ... (1/3)

Propriétés isopérimétriques

- Dans un plan, le cercle de rayon r est la surface de taille πr^2 qui a le plus petit périmètre
- ou: tout ensemble de surface πr^2 a un périmètre plus grand que $2\pi r$
- cela a été généralisé à la sphère et à l'hypersphère

Pour les calottes sphériques (Lévy, 1951)

- Sur la peau de la sphère, les surfaces de taille donnée et de plus petit périmètre sont les calottes sphériques.
- Toute surface de même taille aura un périmètre supérieur.

Cela mérite un peu d'explication ... (2/3)

L'idée ....

- On prend une sphère $\mathbb{S}^{n-1}$, et une fonction

$$\mathbb{S}^{n-1} \xrightarrow{f} \mathbb{R}$$

- qui soit Lipschitzienne

$$|f(x) - f(y)| \leq d(x, y)$$

- On considère l'ensemble

$$A = \{x \in \mathbb{S}^{n-1} \ : \ f(x) < m\}$$

Alors, $\mu(A) = \frac{1}{2}$

Cela mérite un peu d'explication ... (3/3)

L'idée ....

- On remarque que

$$d(x, A) < t \quad \Longrightarrow \quad |f(x) - m| < t$$

soit

$$A_t = \{x \in \mathbb{S}^{n-1} \ : \ f(x) < m + t|\}$$

- Comme $\text{vol}\, A = \text{vol}\, C = \frac{1}{2}$

$$\text{vol}\, A_t \geq \text{vol}\, C_t \geq 1 - \left(\exp - \frac{(n-1)t^2}{2}\right)$$

- et

$$\mathbb{P}\{f(x) > m + t\} \leq \exp - \frac{(n-1)t^2}{2}$$

Concentration de la mesure

Définition

- X: variable aléatoire, $\mu = \mathbb{E}(X)$, c, v constantes
- il y a concentration de la mesure si

$$\mathbb{P}\{|X - \mu| > t\} \leq \exp -c_n \frac{t^2}{2v}$$

Lien avec les grandes déviations

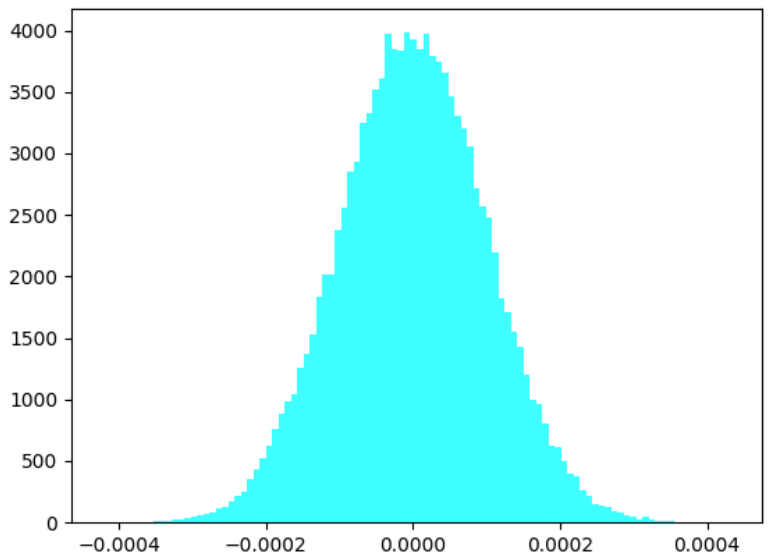
concentration de la mesure	$\forall\, t$
grandes déviations	$t \to \infty$

Talagrand

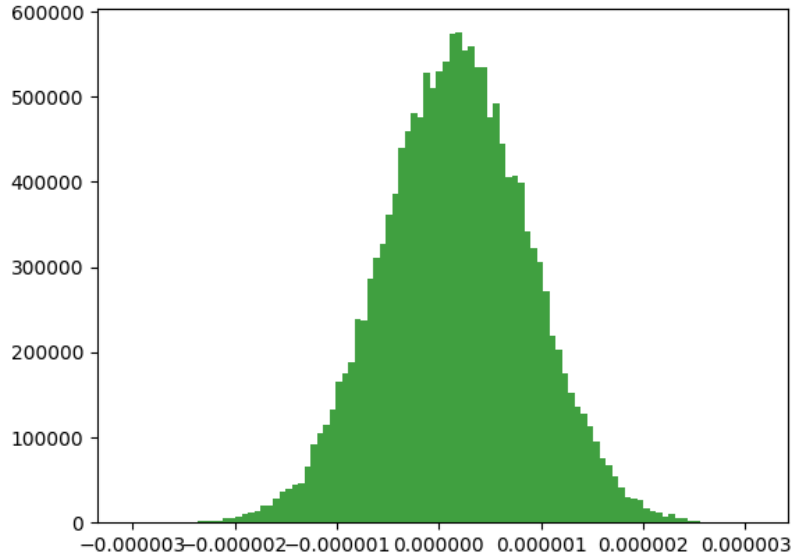
A random variable that smoothly depends on a large number of [almost] independent random variables (but not too much on any of them), is 'essentially' constant, in a 'dimension-free' way.

Comme Saint-Thomas ...

Moyenne



Fonction Polynôme



Histogramme des valeurs d'une fonction sur la sphère, calculé à partir de 50 000 points uniformément répartis sur la peau d'une sphère de $\mathbb{R}^{10\,000}$. A gauche, la moyenne $\frac{1}{n}\sum_i x_i$. A droite: fonction polynôme $f(x) = \frac{1}{2n}\sum_i \alpha_i x_i^2$, avec $\alpha_i \sim \mathcal{N}(0,1)$.

Projection aléatoire

Lemme de Johnson-Lindenstrauss (1984)

- Donnés: $n \in \mathbb{N},\ \epsilon \in [0,1], \quad k \geq \frac{8 \operatorname{Log} n}{\epsilon^2}$
- Alors, $\forall\, X = [x_1, \ldots, x_n], \quad x_i \in \mathbb{R}^n$
- $\exists\, \mathbb{R}^k \subset \mathbb{R}^n, \quad L \quad \text{linéaire} \quad \mathbb{R}^n \longrightarrow \mathbb{R}^k$
- tels que $\forall\, x_i, x_j \in X$

$$(1-\epsilon)\|x_i - x_j\|^2 \leq \|Lx_i - Lx_j\|^2 \leq (1+\epsilon)\|x_i - x_j\|^2$$

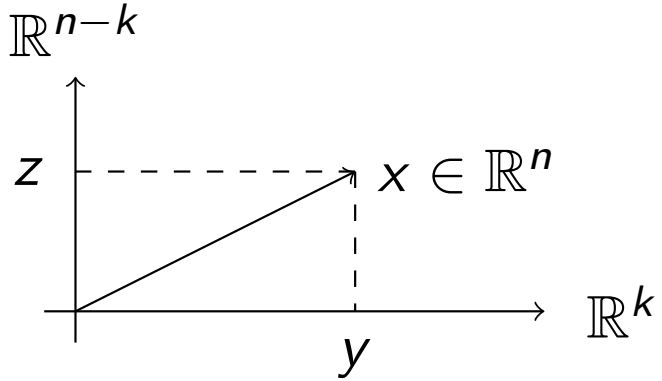
Exemples de conséquences

- plongements lipschitziens
- recherche des plus proches voisins
- MDS par projection aléatoire

Démonstration (Dasgupta & Gupta, 1999, J. L., 1984)

Idée

Un vecteur aléatoire $x \in \mathbb{R}^n$ est fortement concentré autour de sa moyenne



$\|x\| = 1$ (aléatoire sur la spère)

$\mathbb{R}^k$ choisi aléatoirement

$$\mu = \mathbb{E}(\|y\|) = \sqrt{\frac{k}{n}}$$

et on montre que ...(concentration de la mesure)

$$k \geq \frac{24}{3\epsilon^2 - 2\epsilon^3} \operatorname{Log} n \quad \Longrightarrow \quad \begin{cases} \mathbb{P}\{\|y\| \leq (1-\epsilon)\mu\} & \leq & \frac{1}{n^2} \\ \mathbb{P}\{\|y\| \geq (1+\epsilon)\mu\} & \leq & \frac{1}{n^2} \end{cases}$$

Démonstration probabiliste (1/2)

Colonne vertébrale (Johnson & Lindenstrauss, 1984)

Démonstration probabiliste d'un résultat déterministe.

- On se donne $\mathbb{R}^k$ et on définit

$$Lx = \sqrt{\frac{n}{k}}\, y$$

- soit $\mathcal{C}_{ij}(\mathbb{R}^k) \equiv \mathcal{C}_{ij}$ la condition

$$(1-\epsilon)\|x_i - x_j\| \leq \|Lx_i - Lx_j\| \leq (1+\epsilon)\|x_i - x_j\|$$

- On vient de montrer: si $\mathbb{R}^k$ est choisi au hasard

$$\forall\, i \neq j, \quad \mathbb{P}\{C_{ij} \text{ vrai}\} \geq 1 - \frac{2}{n^2}$$

- On note $\mathcal{C}(\mathbb{R}^k) = \mathbb{P}\left\{\forall\, i \neq j,\, \mathcal{C}_{ij} \text{ vrai}\right\} = \mathbb{P}\left\{\bigcap_{i\neq j} \mathcal{C}_{ij}\right\}$

Démonstration probabiliste (2/2)

Colonne vertébrale (Johnson & Lindenstrauss, 1984)

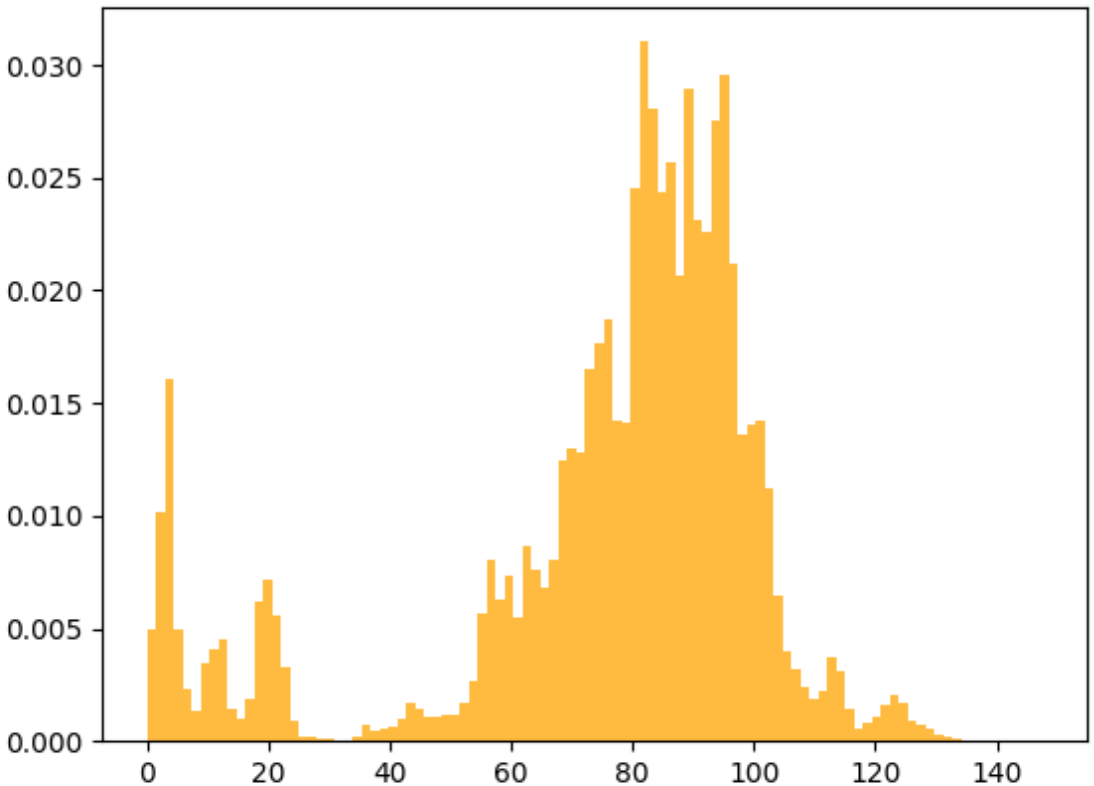
- Alors

$$\begin{aligned}
\mathbb{P}\{\mathcal{C}(\mathbb{R}^k) \text{ vrai}\} &= \mathbb{P}\{\textstyle\bigcap_{i\neq j} \mathcal{C}_{ij}\} \\
&= \mathbb{P}\left\{\overline{\bigcup \overline{\mathcal{C}_{ij}}}\right\} \\
&= 1 - \mathbb{P}\left\{\bigcup \overline{\mathcal{C}_{ij}}\right\} \\
&\geq 1 - \sum_{i\neq j} \mathbb{P}\{\overline{\mathcal{C}_{ij}}\} \\
&\geq 1 - \frac{n(n-1)}{2}\frac{2}{n^2} \\
&= \frac{1}{n} \\
&> 0
\end{aligned}$$

- et

$$\mathbb{P}\{\mathcal{C}(\mathbb{R}^k) \text{ vrai}\} > 0 \quad \Longrightarrow \quad \exists\ \mathbb{R}^k \quad \text{tq} \quad \mathcal{C}(\mathbb{R}^k) \text{ vrai}$$

Un jeu de données réelles



Histogramme des distances deux à deux entre 20 867 séquences. Distance d'alignement local, marqueur 18S-v4. Echantillon environnemental de protistes marins, prélevé dans le Bassin d'Arcachon.

MDS

Problème

Etant donné	un tableau de distances D $(n \times n)$
	avec $D[i,j] = d(i,j)$
	une dimension $k < n$
trouver	un nuage de n points $x_i \in \mathbb{R}^k$
tels que	$\|x_i - x_j\| \approx d(i,j)$ de façon optimale
	$\sum_{i<j} (\|x_i - x_j\| - d(i,j))^2$ minimal

Solution (Torgerson, 1958)

Calculer	la matrice de Gram G à patrir de D
	$(g_{ij} = \langle x_i, x_j \rangle,$ matrice $n \times n)$
Calculer	la SVD de G: $G = U \Sigma V^{\mathrm{T}}$
Résultat	$X = U \Sigma^{1/2}$

ACP & SVD

ACP

$$\begin{aligned} C &= A^{\mathrm{T}}A \\ CV &= V\Lambda \quad (Cv = \lambda v) \\ X &= AV \\ U &= X\Lambda^{-1/2} \end{aligned}$$

SVD

$$\begin{aligned} A &= U\Sigma V^{\mathrm{T}} \\ A^{\mathrm{T}}A &= V\Lambda V^{\mathrm{T}}, \quad \Lambda = \Sigma^2 \\ X &= AV \\ &= U\Sigma \end{aligned}$$

Stratégie

- Il faut réaliser la SVD d'une matrice $n \times n$
- Cela revient à développer son ACP
- on associe à A le nuage de n points de ses lignes dans $\mathbb{R}^n$
- il existe un sous-espace de dimension $k \approx \frac{\mathrm{Log}\, n}{\epsilon^2}$ où la projection $\widetilde{A}$ de A respecte les distances (la forme)
- on réalise l'ACP/SVD de $\widetilde{A}$

SVD par projection aléatoire (Halko & al., 2011)

Heuristique, rang fixé

On projette la matrice A sur un sous-espace aléatoire de dimension k

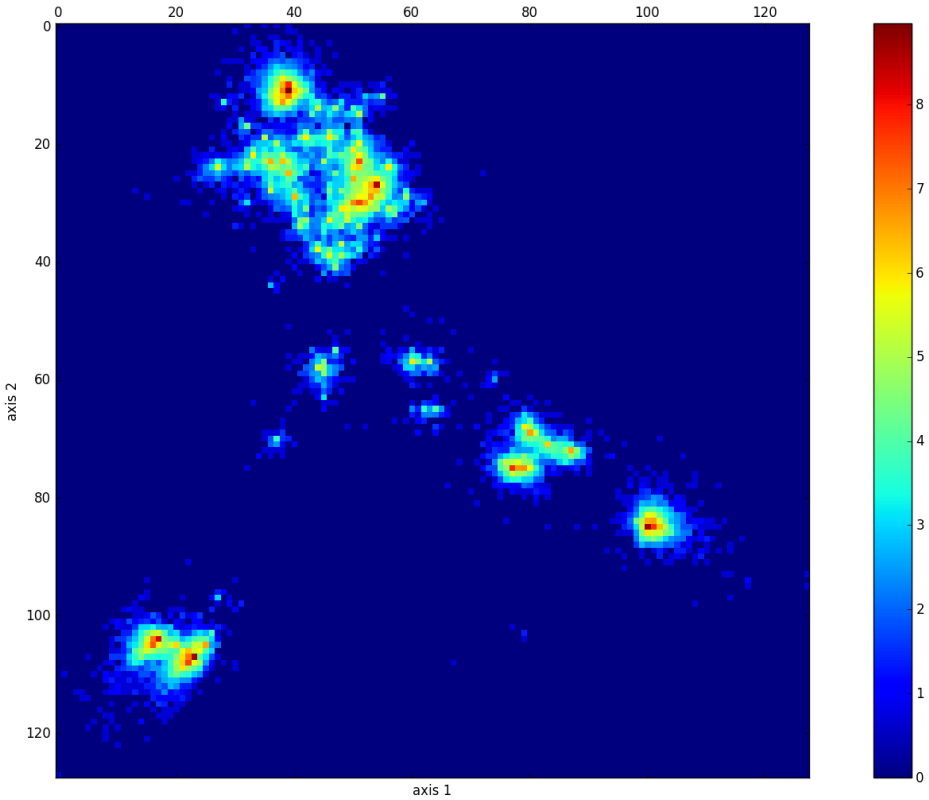
Algorithme

Donné	A	matrice à analyser	$m \times n$
Construire	Ω	matrice aléatoire (gaussienne)	$n \times k$
Calculer	$Y = A\Omega$	projection des lignes de A	$m \times k$
Calculer	$Y = QR$	Décomposition QR de Y	
	Q		$m \times k$
Calculer	$B = Q^{\mathrm{T}} A$	Projection des colonnes de A	$k \times m$
Décomposer	$B = U_{\mathrm{B}} \Sigma V^{\mathrm{T}}$	SVD de B	
Calculer	$\widetilde{A} = QB$	approximation de A	
Résultat	$A \approx U \Sigma V^{\mathrm{T}}$	$U = QU_{\mathrm{B}}$	

Observation

Cela revient à trouver une matrice Q telle que $QQ^{\mathrm{T}} \approx \mathbb{I}_m$

Caractérisation de la biodiversité



Projection sur les axes 1 & 2 d'une MDS à partir des distances deux à deux de 120 000 séquences d'un échantillon d'eucaryotes dans une grotte (donnéees: D. Fontaneto). La couleur d'un pixel indique le logarithme du nombre de séquences projetées sur ce pixel.