LES CAHIERS DE L'ANALYSE DES DONNÉES

J.-P. BENZÉCRI

Sur la proportion des paires de voisins réciproques pour une distribution uniforme dans un espace euclidien

Les cahiers de l'analyse des données, tome 7, n° 2 (1982), p. 185-188

<http://www.numdam.org/item?id=CAD_1982__7_2_185_0>

Les Cahiers de l'Analyse des Données
Vol. VII - 1982 - n° 2 - p. 185-188

SUR LA PROPORTION DES PAIRES DE VOISINS RÉCIPROQUES POUR UNE DISTRIBUTION UNIFORME DANS UN ESPACE EUCLIDIEN [PROP. RECIP.]

par J. P. Benzécri (1)

Rappel de la définition des voisins réciproques

Soit I un ensemble fini muni d'une distance d : on dit que i et i' sont deux voisins réciproques au sein de I si on a simultanément

$$d(i,i') = \inf \{ d(i,i") \mid i" \in I - \{i\} \} ;$$

$$d(i,i') = \inf \{ d(i',i") \mid i" \in I - \{i'\} \}.$$

Autrement dit s'il n'y a pas d'élément de I (autre que i) plus proche de i que ne l'est i' ; ni d'élément de I (autre que i') plus proche de i' que ne l'est i.

Les paires de voisins réciproques pouvant être agrégés en un seul parcours du tableau des distances (cf. [*PROG. C.A.H. RECIP.*] p.195, il importe de connaître la fréquence des paires de voisins réciproques pour apprécier les mérites de l'algorithme des voisins réciproques.

En particulier on peut considérer le cas modèle suivant : l'ensemble I est construit par tirages, au hasard, successifs indépendants suivant une distribution uniforme sur un ouvert borné (e.g. un cube) de l'espace euclidien de dimension n. Il est clair, vu le caractère local du problème que pour Card I tendant vers l'infini, la forme du domaine ouvert n'intervient pas.

Dans la présente note on étudiera d'abord le cas modèle n = 1 : où les points de l'ensemble I sont tirés suivant une loi uniforme sur un segment de droite (e.g. (0,1)). Puis le cas général, avec la limite $n \to \infty$.

2 *Etude d'une distribution uniforme sur la droite*

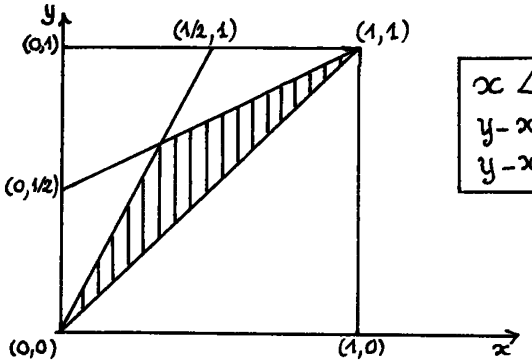
Notons Card I = m : l'ensemble I peut être considéré comme un point (distribué uniformément) du cube $(0,1)^m$: plus précisément on doit identifier deux points du cube dont les m coordonnées ne diffèrent que par une permutation, et qui définissent un même I. On peut encore supposer que :

$$x(1) \leq x(2) \leq \ldots\ldots \leq x(m)$$

Cherchons la probabilité pour que x(r) et x(r+1) soient des voisins réciproques (sous l'hypothèse que $2 \leq r \leq m-2$: i.e. x(2) n'est point premier, ni x(r+1) dernier). Il est clair que cet événement ne

(1) *Professeur de statistique. Université Pierre et Marie Curie.*

dépend que des coordonnées des quatre points de rang r-1, r, r+1, r+2 car seul le point r-1 peut entrer en concurrence avec r+1 pour être le plus proche voisin de r (et de même r+2 avec r pour r+1). On calculera donc la probabilité conditionnelle, x(r-1) et x(r+2) étant fixés, que x(r) et x(r+1) soient voisins réciproques : cette probabilité qui ne dépend pas des coordonnées x(r-1) et x(r+2) ni de r sera aussi la probabilité pour que deux éléments consécutifs soient voisins réciproques.



Pour la commodité du dessin on a noté x(r-1) = 0 ; x(r+2) = 1 (ce qui n'est qu'un changement d'échelle sans effet sur le calcul des probabilités en vue) ; et noté x(r) = x ; x(r+1) = y. Sur le triangle défini par $x < y$, la condition de voisinage réciproque délimite un sous triangle (hachuré sur la figure) qui en est le 1/3. Telle est la probabilité que x(r) et x(r+1) soient voisins réciproques, si $2 \leq r \leq m-2$; dans les cas r = 1 et r = m-1 on trouve sans peine que cette probabilité est 1/2.

Donc l'espérance mathématique du nombre des paires agrégées en un parcours est pour le modèle étudié égal au tiers du nombre m des points de I. En moyenne on aura, après ce parcours, m/3 individus isolés et m/3 paires constituées par agrégation de deux voisins réciproques.

3 *Etude d'une distribution uniforme en dimension n ≥ 2*

m points sont disposés au hasard dans un ouvert de volume V de l'espace euclidien de dimension n par tirages aléatoires indépendants. Pour calculer la proportion moyenne de paires de voisins réciproques, nous commencerons par déterminer la loi de la distance ρ d'un point M à son plus proche voisin M' ; puis nous calculerons la probabilité conditionnelle p(ρ) (fonction de ρ) pour que M soit aussi le point du nuage le plus proche de M'. En intégrant p(ρ) par rapport à la loi de ρ on aura la possibilité pour qu'un point ait un voisin réciproque, probabilité qui n'est autre que le double du rapport à m du nombre moyen de paires de voisins réciproques.

3.1 Loi de la distance d'un point à son plus proche voisin : Notons $B(n)R^n$ (ou, en bref, $B\,R^n$) le volume d'une boule de rayon R dans l'espace de dimension n. La probabilité F(ρ) que la distance étudiée soit supérieure à ρ, se calcule (à la limite $m \to \infty$ comme la probabilité qu'aucun des (m-1) autres points ne tombe dans la boule de rayon ρ ayant pour centre un premier point ; soit :

$$F(\rho) = (1 - (B(n)\,\rho^n/V))^{m-1}$$

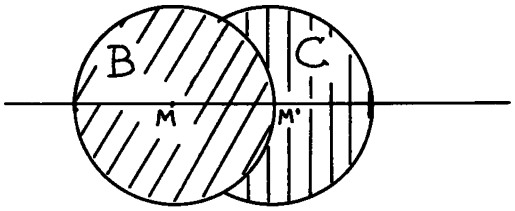
soit à la limite et en notant d = V/m la densité moyenne du nuage de points :

$$F(\rho) = \exp(-\rho^n B(n)\, d)$$

D'où pour la densité de la loi de répartition de ρ entre 0 et l'infini :

$$f(\rho) = B(n).d.n.\rho^{n-1}.\ \exp\ (-\rho^n B(n)\, d).$$

3.2 *Probabilité conditionnelle $p(\rho)$ en fonction de $\rho = (M, M')$, que M soit le plus proche voisin de son plus proche voisin M'* :



Sur la figure on a noté C la partie de la boule de centre M' et de rayon ρ extérieure à la boule B de centre M et de rayon ρ. Les volumes de B et C sont respectivement :

$$\mathrm{Vol}(B) = B(n)\ \rho^n\ ;\quad \mathrm{Vol}(C) = C(n)\ \rho^n\ ;$$

où B(n) et C(n) s'expriment par des intégrales sur le calcul desquelles nous reviendrons. Ceci posé la probabilité conditionnelle $p(\rho)$ cherchée est celle qu'aucun des m-2 points (autres que M et M') ne tombe dans C ; (sous la condition complémentaire, négligeable à la limite qu'aucun point ne tombe dans B). D'où :

$$p(\rho) = (1 - (C(n)\rho^n/(V - B(n)\rho^n)))^{m-2}\ ;$$

soit à la limite (en posant comme ci-dessus $d = V/m$) :

$$p(\rho) = \exp(-\rho^n\ C(n)\ d)\ .$$

3.3 *Probabilité qu'un point ait un voisin réciproque* : On doit calculer une intégrale :

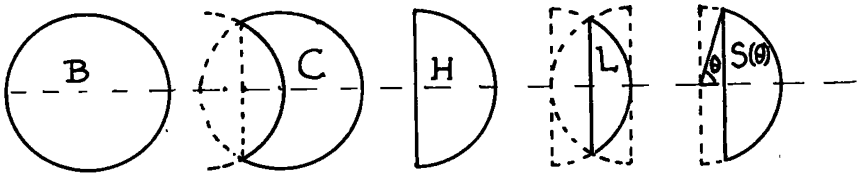
$$\begin{aligned}
&\int\{f(\rho)\ p(\rho)d\rho \mid \rho \in (0,\infty)\}\\
&= \int B(n).d.n.\rho^{n-1}.\ \exp(-\rho^n B(n)d)\ \exp(-\rho^n C(n)d)\,d\rho\\
&= \int B(n).d.n.\rho^{n-1}.\ \exp(-\rho^n (B(n) + C(n))\,d)\,d\rho\\
&= \int (B(n)/(B(n) + C(n)))\,(B(n) + C(n))\,d.n.\rho^{n-1}\exp(-\rho^n (B(n)+C(n))d)d\rho\\
&= B(n)/(B(n) + C(n))\ .
\end{aligned}$$

Dans le cas particulier, déjà traité, où $n = 1$, on a :

$$B = 2\ ;\quad C = 1\ ;$$

d'où $B/(B+C) = 2/3$. La probabilité qu'un point ait un voisin réciproque est 2/3 ; soit pour le taux des paires la valeur 1/3 déjà trouvée au § 1. Pour $n \geq 2$ il faut revenir au calcul de B(n) et de C(n).

3.4 Calculs de volume



La figure définit des notations qu'il est inutile de commenter. On a :

$$B = 2H = 2S(\pi/2) \ ;$$
$$C = 2(H-L) = 2(S(\pi/2) - S(\pi/3)) \ ;$$
$$B/(B+C) = S(\pi/2)/(2S(\pi/2) - S(\pi/3)) \ ;$$
$$= 1/(2-(S(\pi/3)/S(\pi/2))) \ ;$$

Un calcul classique donne pour $S(\theta)$:

$$S(\theta) = B(n-1) \int \{\sin^n t \ dt \mid t \in (0,\theta)\},$$

(où $B(n-1)$ est comme ci-dessus, le volume de la boule de rayon 1 en dimension n-1). Faisons le calcul pour n = 2

$$S(\theta) = 2 \int \sin^2 t \ dt$$
$$= \int (1-\cos 2t) dt = \theta - (\sin 2\theta/2) \ ;$$
$$S(\pi/2) = \pi/2 \ ; S(\pi/3) = (\pi/3) - (\sqrt{3}/4) \approx 0{,}614 \ ; \text{ d'où :}$$
$$B/(B+C) \approx 0{,}621 < 2/3.$$

Quand n tend vers l'infini, $S(\pi/3)$ tend à être négligeable vis-à-vis de $S(\pi/2)$ et donc $B \sim C$ et $B/(B+C) \approx 1/2$.

4 Conclusion : Efficacité de l'algorithme des voisins réciproques

Ainsi quand n varie de 0 à l'infini le nombre de paires de voisins réciproques agrégés (en moyenne) au premier parcours varie de m/3 à m/4.

Ce qui arrive aux itérations suivantes est plus difficile à apprécier car, les points restants après agrégation n'étant pas en situation quelconque ils ne rentrent pas nous semble-t-il dans un modèle aléatoire simple. Si toutefois chaque parcours du tableau des distances réduisait véritablement d'environ (1/4) le nombre des sommets, on pourrait évaluer comme suit en fonction du nombre m des individus le nombre d des distances à calculer :

$$\Delta d/\Delta m \approx (m^2/2)/(m/4) = 2m$$

(car en bref à chaque parcours on calcule $m^2/2$ distances et m varie de m/4...) d'où par intégration

$$d \approx m^2 \quad \text{(pour une autre évaluation, cf. p. 226)}$$

On sait que l'algorithme de base qui effectue une agrégation par parcours, donne $d = m^3/6$. Nos évaluations, même approximatives montrent en tout cas de façon certaine l'efficacité de l'algorithme des voisins réciproques ; efficacité que confirme l'expérimentation sur les données réelles.