MATHÉMATIQUES ET SCIENCES HUMAINES

S. RÉGNIER

Sur quelques aspects mathématiques des problèmes de classification automatique

Mathématiques et sciences humaines, tome 82 (1983), p. 31-44

<http://www.numdam.org/item?id=MSH_1983__82__31_0>

Math. Sci. hum. (21e année, n°82, *1983*, p.31-44)

SUR QUELQUES ASPECTS MATHEMATIQUES DES PROBLEMES DE CLASSIFICATION AUTOMATIQUE *

S. REGNIER

Maison des Sciences de l'Homme, Centre de Calcul

2. ASPECT STATISTIQUE

II. 1. Deux problèmes de convergence.

Dans la plupart des applications, l'ensemble E des objets à classer n'est qu'un échantillon observé dans une population E' plus vaste, virtuellement observable, et le système de descripteurs $a = (a_1, a_2 \ldots a_p)$ qui représente E dans F est défini a priori sur E' tout entier

$$a : E' \qquad F = F_1 \times F_2 \ldots F_p.$$

Dans des cas plus exceptionnels**, on peut envisager d'augmenter le nombre p des critères formels

$$a_h : E \longrightarrow F_h$$

et l'ensemble des images

$$F = F_1 \times F_2 \ldots \times F_p$$

est virtuellement plongé dans un ensemble F' plus vaste. Nous allons préciser et étudier le problème de la convergence :

des partitions centrales

a) lorsque $F \longrightarrow F'$ (E fixe fini)

b) lorsque $E \longrightarrow E'$ (F fixe fini)

---

* texte de Nov 66 non publié. Suite de l'article paru dans ICC, 1965.

** Exemple : n objets sont classés par p examinateurs a_h, travaillant séparément selon des critères propres à chacun, et extraits au hasard d'une vaste population d'examinateurs possibles. Voir Réf. (3).

a) Les p critères $a_h : E \to F_p$ définissent, on l'a vu, autant de partitions S_h, et les partitions centrales X sont celles qui miniminisent la fonction

$$H_p(X) = \frac{1}{p} \sum_{1}^{p} D^2 (X, sh)$$

(D^2 = cardinal de la différence symétrique).

Nous supposerons que les sh sont des éléments aléatoires indépendants, de même loi L sur P (E) = ensemble des partitions de E (fini). Ils définissent une loi de probabilité empirique L_p et les variables aléatoires:

$$L_p (x) = \text{Proportion de } [S_h = x]$$

convergent presque sûrement (et même presque complétement sûrement) vers

$$L (x) = \text{Probabilité } [s_h = x] \text{ pour tout h .}$$

Par suite :

$$H_p (x) = \sum_{y \in P (E)} D^2 (x,y) \; L_p (y) ,$$

fonction continue des variables $L_p (y)$

converge presque sûrement vers $H (x) = \sum D^2 (x,y) \; L(y) \qquad y \in P(E)$.

On définit naturellement les partitions centrales de toute répartition L sur P(E) comme celles qui rendent H(x) minima, et le centre comme l'ensemble

$$C = \{ x \text{ tel que : } \forall y \quad H(x) \leq H(y) \}$$

non vide puisque E est fini.

Les centres C_p des lois L_p sont définis de la même façon. Nous allons démontrer que

$$\lim_{p \to \infty} Pr (C_p \subset C) = 1$$

$$Pr [\text{lim sup } C_p = C] = 1$$

b) Supposons inversement que les objets à classer

$$E = \left\{ O_1, O_2 \quad . \quad . \quad O_i \ldots O_n \right\}$$

soient une famille de n éléments aléatoires indépendants de même loi L' dans un ensemble E' éventuellement infini.

Pour étudier la convergence des partitions centrales de E quand n tend vers l'infini, nous allons les représenter par des partitions de l'ensemble fixe des <u>images virtuelles</u>

$$F = F_1 \times F_2 \quad . \quad . \quad F_p \qquad \text{(fini)} \cdot$$

La représentation générale :

$$a : \quad E' \longrightarrow F$$

sera supposé mesurable, pour la loi L' et la mesure triviale de F. On a vu en I.6 que chaque partition centrale de E est compatible avec a et définit de ce fait une partition X' de l'ensemble des "<u>images réelles</u>" : a (E) $\subset$ F. On peut associer à X' toutes les partitions X de F qui ont même restriction sur a (E), et la matrice booléenne de X

$$f,g \in F \times F \longrightarrow \begin{cases} X_{fg} = 1 & \text{si f et g sont dans la même classe} \\ = 0 & \text{sinon} \end{cases}$$

devra rendre minima la fonction numérique positive

$$Q\ (X-S) = \sum_{f \text{ et } g \in F} (X_{fg} - S_{fg})^2 \quad n_f\ n_g$$

pour minimiser la somme :

$$\sum_{1}^{p} p_h\ Q\ (X-S^h) = Q\ (X-S) + \sum p_h\ Q\ (S-S^h)$$

en désignant par :

- n_f , le nombre d'objets O_i ayant f pour image
- S_h , la matrice de la partition canonique de F définie par la projection :

$$F \longrightarrow F_h$$

- et S la matrice de similarité : $\sum_{h=1}^{h} p_h \quad S_h$

(les p_h sont des poids de somme 1 associés aux espaces facteurs F_h).

Ainsi sont définies les partitions centrales de

F = ensemble des images virtuelles

relativement à la pondération n_f définie par

$a : E \longrightarrow F$.

Maintenant les images

$f_i = a(O_i)$ sont des éléments aléatoires de F (fini), indépendants et de même loi L = image de L' par a

$$\begin{cases} L(f) = Pr\left(f_i = f\right) = L'\left(\bar{a}^1(f)\right), \\ \text{pour tout f pris dans F} \end{cases}$$

Elles définissent sur F une loi de probabilité empirique :

$$L_n(f) = \frac{n_f}{n}$$

qui converge presque sûrement vers L (f) quand n tend vers l'infini.

$$H_n(X) = \frac{Q(x-s)}{n^2} = \sum_{fg} (X_{fg}-S_{fg})^2 \quad L_n(F) \quad L_n(g)$$

étant fonction continue des variables $L_n(f)$ converge presque sûrement vers

$$H(X) = \sum_{f \text{ et } g \in F} (X_{fg} - S_{fg})^2 \quad L(f) \quad L(g) \quad \text{fonction positive}$$

définie sur P (F), comme H_n.

Soient alors c_n et C les ensembles où H_n et H atteignent leur borne inférieure :

$$C = \left\{ x : \quad H(x) \leq H(y) \quad \forall \quad Y \in P(F) \right\}$$

C_n = idem en H_n . Ils sont non vides puisque P (F) est fini

A tout élément de c_n correspond une partition centrale de l'échantillon E. Par définition, tout X pris dans C constituera de même une partition centrale de F relativement à la loi image L (f) , et X est l'image par a d'une partition X' de E' : les X'- classes dans E' seront les images réciproques par a des X- classes dans F.

Ainsi sont définies les partitions centrales d'un ensemble E' éventuellement infini, muni d'une probabilité L' et de p descriptions mesurables, à valeurs dans des espaces F_h finis - elles-mêmes munies au besoin de poids p_h de somme 1.

Les correspondances employées sont utilement résumées par le tableau suivant:

étant donné $a : E \longrightarrow F$

l' ensemble des partitions de F soit	P (E)	injection connonique d'inclusion
contient	↑	
l'ensemble des . . . compatibles avec a . .	P (E/a)	
auquelles correspond de façon biunivoque	↕	bijection
l'ensemble des partitions de a (E) $\subset$ F . . .	P [a(E)]	surjection définie par la restriction
qui est l'image de	↑	
l'ensemble des partitions de F	P (F)	⟶

pour l'opérateur de restriction de F à a (E)

Ici nous montrerons seulement que, quand $n \longrightarrow \infty$:

$$\lim \Pr \left[c_n \subset c \right] = 1$$

$$\Pr \left[\limsup c_n \subset C \right] = 1 .$$

II.2 Centre et valeurs centrales d'ordre m d'une loi de probabilité L sur un espace métrique X

Le problème a) mérite d'être posé dans le cadre général suivant :

Considérons un espace de probabilité ($\underline{X}$, B, L) muni d'une distance d (x,y) telle que les applications $y \longrightarrow d(x,y)$ de $\underline{X}$ dans R soient mesurables pour tout x dans X (telle, autrement dit, que toutes les boules soient mesurables)

On peut définir avec FRECHET [1] , $\forall\, m > 0$,

a) le m^e moment absolu par rapport au point $x \in \underline{X}$

$$H(x) = \int d^m(x,y) \quad L(dy) \quad \leq + \infty$$

Nous supposerons H (x) partout fini (*).

b)le centre d'ordre m

$$C = \left\{ x \;\middle|\; H(x) \leq H(y) \text{ pour tout } y \right\}$$

quand C n' est pas vide, ses éléments s'appelleront valeurs centrales d'ordre m.

Exemples

1- Si $\underline{X}$ est fini, ou compact pour la métrique d, C est non vide.

2- Si $\underline{X}$ est vectoriel sur R,et si la métrique est Euclidienne. L'unique valeur centrale d'ordre 2 est constituée par l'espérance mathématique :

$$E(X) = \int x\, L(dx)$$

en vertu de l'identité de KOENIG :

$$H(x) = H\left[E(X)\right] + D^2\left[x, E(x)\right]$$

Problème de convergence.

Considérons un échantillon de la loi L :

p variables indépendantes $X_i \in \underline{X}$, de loi L; elles définissent une loi empirique sur $\underline{X}$: $L_p = \frac{1}{p} \sum \delta(X_i)$ et par suite, des moments :

$$H_p(x) = \int d^m(x,y) \quad L_p(d_y) = \frac{1}{p} \sum_1^p d^m(x, X_i)$$

(*) Sinon l'inégalité triangulaire permet de voir que H (X) est partout infini .

et un centre C_p où H_p atteint borne inférieure. $H_p(x)$ est la moyenne des p variables $d^m(x, X_i)$, indépendantes de même loi. Par suite, quand $p \longrightarrow \infty$ $H_p(x)$ converge presque sûrement vers l'espérance mathématique finie

$$H_p(x) \longrightarrow H(x) \text{ (fini) p.s.}$$

Nous allons en déduire que quand $p \longrightarrow \infty$,

pour $\underline{X}$ dénombrable :

$$\limsup C_p \subset C \qquad \text{presque sûrement}$$

pour $\underline{X}$ fini : $\lim P\left[C_p \subset C\right] = 1$ et $\limsup C_p = C$ p.s.

II. 3 Centres d'une suite convergente de fonctions.

Appellons centre de toute fonction numérique : $H : \underline{X} \longrightarrow R$ l'ensemble C éventuellement vide où H atteint sa borne inférieure.

$$C = \left\{ x \in \underline{X} \;\middle|\; H(x) \leq H(y) \quad \forall\ y \in \underline{X} \right\}.$$

Si une suite de fonctions H_p tend vers H en chaque point de X, on voit facilement que :

$\lim \sup_{n \longrightarrow \infty} C_p \subset C$ $\quad p > n$ | et que si $\underline{X}$ est fini, $C_p \subset C$ pour p assez grand.

En effet si $H(x) > H(y)$, pour p assez grand, $H_p(x) > H_p(y)$, et la suite des C_p qui contiennent x ne peut être infinie, d'où :

$x \notin \limsup C_p$. Si $\underline{x}$ fini la suite des C_p qui contiennent au moins 1 $x \notin C$ est également finie.

Ces 2 résultats sont les seules propriétés générales dans le domaine des fonctions certaines, comme le montrent les contre-exemples suivants :

$\underline{x}$ = les rationnels compris entre 0 et 1

1) $H_p(x) = \sup\ (0,\ x - 1/p)$

alors $C_p = \left[0, \frac{1}{p}\right] \supset C = \left\{0\right\}$, inclusion stricte pour tout p

2) $H_p(x) = \frac{x}{p}$ $\qquad C = \underline{X} = [0,1]$ et $C_p = \left\{0\right\}$

on voit aussi facilement que la suite C_p n'est pas en général convergente, et peut même être quelconque : (C_p) etant donnée on notera B_p les fonctions indicatrices et on posera : $H_p(x) = -\frac{1}{p} B_p(x)$, par exemple.

Fonctions aléatoires à limite certaine

Si les fonctions H_p sont aléatoires, la limite H restant une fonction certaine, on a d'abord un résultat indépendant du cardinal de $\underline{X}$:

LEMME

Si $x \in \underline{X} - C$

1) $H_p \longrightarrow H$ en probabilité pour tout $x \Rightarrow \begin{cases} \text{Si l'événement } x \in C_p \text{ est mesurable} \\ \Pr\left[x \in C_p\right] \longrightarrow 0 \end{cases}$

2) Si $H_p \longrightarrow H$ presque sûrement pour tout x,

$$\Pr\left[\limsup_p c_p \ni x\right] = 0$$

En effet

$x \notin C$ entraine

$\exists\, y, a : H(y) = H(x) - 2a$ et $a > 0$

alors $H_p(y) - H_p(x) = H_p(y) - h(y) + H(x) - H_p(x) - 2a$

(1) et $\begin{cases} \sup_p \left[H_p(y) - H_p(x)\right] \leq \sup_n \left|H_p(y) - H(y)\right| \\ \qquad + \sup_n \left|H_p(x) - H(x)\right| - 2a \end{cases}$

par suite

1) $x \in C_p \Rightarrow H_p(y) - H_p(x) \geq 0 \quad \forall\, y$, et en particulier :

$\Rightarrow \begin{cases} H_p(y) - H(y) \geq a \\ H_p(x) - H(x) \geq a \end{cases}$ ou bien

Les probabilités de ces deux événements : $1_p(x)$ et $c_p(y)$, tendent vers 0 lorsque $H_p(x)$, $H(x)$ en probabilité pour tout x, par suite :

$\Pr(x \in C_p) \leq 1_p(x) + 1_p(y) \longrightarrow 0$

2) $x \in \limsup c_p$ signifie : la suite des c_p qui contiennent x est infinie, soit :

$$\forall\, n\ \exists\, p > n \qquad \forall\, y\ H_p(y) \geq H_p(x)$$

et entraine : $\forall\, y\ \forall\, n\ \exists\, p > n \quad H_p(y) \geq H_p(x)$ soit

$$\forall\, y \quad \limsup \left[H_p(y) - H_p(x)\right] \geq 0$$

maintenant, les événements mesurables :

$$E_n(y) : \sup_{p > n} \left[H_p(y) - H_p(x)\right] \geq 0$$

forment une suite décroissante : $E_{n+1} \Rightarrow E_n$

si bien que :

$$\Pr\left[\lim E_n\right] = \lim P\left(E_n\right) \text{ , pour tout y.}$$

Dans le cas particulier où $H(y) = H(x) - 2a$

il résulte de (1) que :

$$\Pr\left[\sup_{p>n}\left[H_p(y) - H_p(x)\right] \geq 0\right]$$

$$\leq \Pr\left\{\sup_{p>n}\left|H_p(y) - H(y)\right| \geq a\right\}$$

$$+ \Pr\left\{\sup_{p>n}\left|H_p(x) - H(x)\right| \geq a\right\}$$

et ces deux probabilités tendent vers 0 si

$$H_p \longrightarrow H \qquad \text{presque sûrement.}$$

Par conséquent $$\Pr\left\{\limsup\left[H_p(y) - H_p(x)\right] \geq 0\right\} = 0$$

pour le point y choisi , et à fortiori

$$\Pr\left[x \in \limsup c_p\right] = 0$$

- De ce lemme résulte immédiatement :

1) Si $H_p \longrightarrow H$ en probabilité et $\underline{X}$ fini

$$\Pr\left[c_p \subset C\right] \longrightarrow 1$$

2) Si $H_p \longrightarrow H$ presque sûrement et $\underline{X}$ dénombrable

$$\Pr\left[\limsup c_p \subset C\right] = 1$$

En effet dans 1) l'événement complémentaire s'écrit :

$$A_p : \exists\ x \notin c \qquad x \notin C_p$$

$$\Pr\left[A_p\right] \leq \sum_{x \in \underline{X} - C} \Pr(x \in c_p) \quad \text{(somme finie)}$$

et l'on a vu que ces probabilités tendent vers 0

ainsi $$\Pr\left[A_p\right] \longrightarrow 0$$

Dans 2) l'événement complémentaire s'écrit :

$$A : \exists\ x \notin C \qquad x \in \limsup c_p$$

$$\Pr\ A \leq \sum_{x \in \underline{X} - c} \left[\Pr\ x \in \limsup\ c_p\right] \text{ (somme dénombrable)}$$

et l'on a vu que toutes ces probabilités sont nulles .

Les deux restrictions sur le cardinal de X sont indispensables. En effet :

1) En reprenant le 1er contre-exemple ci-dessus, où $\underline{X}$, dénombrable = les rationnels de $[0,1]$

$$c_p \supset c \quad \forall\ p \quad \text{et}$$

$$\Pr\left[C_p \subset C\right] = 0 \quad \forall\ p$$

2) Soit U une variable répartie uniformément sur l'ensemble $\underline{X} = [0,1]$ continu .

La fonction

$$H_p(x) = \begin{cases} 1 & \text{pour } x = 0 \text{ ou } U \\ 0 & \text{sinon} \end{cases}$$

et presque sûrement égale à

$$H(x) = \begin{cases} 1 & \text{pour } x = 0 \\ 0 & \text{sinon} \end{cases}$$

On a alors $c_p = \{0, U\}$ quelque soit p

$c = \{0\}$ et par suite

$$\Pr\left[\limsup c_p \subset c\right] = \Pr\{U = 0\} = 0$$

II. 4 Inclusion réciproque du centre d'une loi L.

Dans le cas des moments aléatoires d'une loi empirique

$$H_p(x) = \frac{1}{p} \sum d^m(x, X_i) = \sum_y d^m(x,y)\ L_p(y)$$

et pour $\underline{X}$ fini, l'indépendance des variables X_i dans $\underline{X}$ va entrainer presque sûrement l'inclusion réciproque :

$$Pr\left[C \subset \lim\sup C_p \right] = 1$$

Le centre C étant ici fini, il suffit d'établir que pour tout $x \in C$

$$Pr\left[x \in \lim\sup c_p \right] = \lim_{n \to \infty} Pr\left[x \in \bigcup_{p>n} c_p \right] = 1$$

Cette dernière probabilité, étant non-décroissante doit être constamment égale à 1.

C'est dire que l'événement complémentaire :

$$B_n^x \iff \forall\ p > n \qquad \exists\ y \qquad H_p(y) < H_p(x)$$

doit être presque impossible.

Posons $\forall\ y,\ S_p(y) = p\left[H_p(y) - H_p(x)\right] = \sum_1^p \left[d^m(y,X_i) - d^m(x,X_i)\right]$

C'est une somme de variables $D_i(y) = d^m(y, X_i) - d^m(x, X_i)$

mutuellements indépendantes et de même loi, bornées par:

$B = Max\ d^m(x, X)$, borne finie puisque $\underline{X}$ est fini

Les moyennes sont :

$M(y) = H(y) - H(x) \geq 0$ (puisque $x \in C$)

nulles seulement si $y \in C$

Les moments du second ordre :

$E\left[D(y), D(y')\right]$ sont également finis si bien que les variances et covariances vont former une matrice V finie

En vertu du théorème central limite, le vecteur S_p de $R^{\underline{X}}$ défini par les variables $S_p(y)$ suit asymptotiquement une loi normale :

et $\dfrac{S_p - pM}{\sqrt{p}}$ équivaut à $T = N\left[0, V\right] \in R^{\underline{X}}$ de moyenne 0, et variance V

L'événement $B_n^{\times}$ signifie

$$\forall\ p > n \qquad \exists\ y \qquad S_p(y) < 0$$

Considérons alors une sous-suite infinie arbitraire : $p_k > n$

Les vecteurs :

$$Z_k = S_{p_{k+1}} - S_{p_k}$$

sont somme de $v_k = p_{k+1} - p_k$ vecteurs D_i indépendants, et sont mutuellement indépendants.

$B_n^{\times}$ entraîne

$$\forall\ k \quad \exists\ y \qquad Z_k(y) < -S_{p_k}(y) \leq B.p_k$$

Les événements :

$$B_k : \exists\ y \qquad Z_k(y) \leq B\ p_k$$

sont mutuellement indépendants; par suite :

$$\Pr\ B_n^{\times} \leq \Pr\left[\forall\ k\ B_k\right] = \prod_{k=1}^{\infty} \Pr\left[B_k\right]$$

Pour certaines suites p_k, ce produit infini va être nul. Il suffit que

$$\lim \Pr\left[B_k\right] < 1$$

or : $$1 - \Pr\left[B_k\right] = \Pr\left[\forall\ y \qquad Z_k(y) > Bp_k\right]$$

si $v_k \longrightarrow \infty$ la loi limite du vecteur

$$\frac{Z_k - v_k M}{\sqrt{v_k}}$$ est celle de T, et

$1 - \Pr(B_k)$ a limite que :

$$\Pr\left[\forall\ y \qquad T(y) > \frac{Bp_k}{\sqrt{v_k}} - M(y)\sqrt{v_k}\right]$$

si $\dfrac{\sqrt{v_k}}{p_k} = a > 0$. C'est-à-dire que $p_{k+1} = p_k(1 + a^2 p_k)$

les second membres ont pour limites :

$-\infty$ si $M(y) > 0$ et

$B_{/a}$ si $M(y) = 0$ c'est à dire si y pris dans C

Par suite

$$1 - \lim \Pr\ B_k = \Pr\left[\ \forall\ y \in C \qquad T(y) > B/_a\ \right]$$

cette probabilité est en général positive, et

$$\Pr\left[B_n^x = 0\right]$$

Exceptionnellement $\lim \Pr\left[B_k\right] = 1$ si

$\forall\ y \in C$, $T(y) = 0$ presque sûrement, par suite :

$D_i(y) = 0$ p.s. et $d(x, X) = d(y, X)$ presque sûrement.

La loi L serait concentrée dans " l'axe médiateur " du centre c. Alors

$H_p(y) =$ constante pour $y \in C$.

Quand C_p coupe c, $c_p \supset C$. Comme c_p est non vide :

$$\Pr\left[\lim \sup\ c_p > c\right] = 1 \quad \text{et de plus}$$

$$\Pr\left[C \subset C_p\right] \longrightarrow 1.$$

Tout se passe comme si C était réduit à un seul point.

Remarque $E_p = \Pr\left[C \subset C_p\right]$ peut tendre vers 0 :

Prenons une loi L uniforme sur un ensemble à 2 éléments : $\underline{X} = \{x,y\}$

$H(x) = H(y) = \frac{1}{2}\ d^m(x,y)$, donc $C = \underline{X}$. $C_p = \underline{X}$ seulement si

$H_p(x) = H_p(y)$ et $L_p(x) = L_p(y)$. $E_p = 0$ pour p impair,

et : $E_p = \left(\frac{1}{2}\right)^{2m} C_{2m}^{m}$ pour $P = 2m$.

$E_p \longrightarrow 0$ quand $m \longrightarrow \infty$

Dans cet exemple : $\lim \sup\ c_p = C$ et

$\lim \inf\ c_p = \emptyset$ presque sûrement.

Références :

[1] FRECHET M., *Généralités sur les probabilités. Variables aléatoires*, chapitre III, dans le § "Valeurs typiques d'ordre positif", p.89, *in Traité du calcul des probabilités et de ses applications* par E. Borel. tome I, les principes de la théorie des probabilités. Fascicule III. Premier livre., Paris, Gauthier-Villars, 1950.

[2] (1) "Les éléments aléatoires de nature quelconque dans un espace distancié", *Annales de l'Institut Henri Poincaré*, vol. XIV, 1948,pp. 215-310.
(2) "L'intégrale abstraite d'une fonction abstraite d'une variable abstraite et son application à la moyenne d'un élément aléatoire de nature quelconque", *Revue Scientifique*, 82^{e} année, 1944, pp. 483-512.

[3] "Positions typiques d'un élément aléatoire de nature quelconque", *Ann. Ec. Norm. Sup.*, t.LXV, 1948, pp.211-237.