Statistique et analyse des données

Jean-Pierre Barthelemy

A propos des partitions centrales sur un ensemble non nécessairement fini

Statistique et analyse des données, tome 2, n° 3 (1977), p. 54-62.

<http://www.numdam.org/item?id=SAD_1977__2_3_54_0>

Statistique et Analyse des Données
3 - 1977

à propos des partitions centrales sur un ensemble non nécessairement fini

BARTHELEMY Jean - Pierre

Ecole Nationale Supérieure de Chronométrie et Micromécanique de Besançon
25030 BESANÇON CEDEX

Résumé : Dans cet article, on examine ce que devient la notion de partition centrale lorsque le support n'est plus un ensemble fini.

Summary : In this paper we studie the central partitions when the support set is not necessary finite.

INTRODUCTION

En 1965, S. Régnier a introduit la notion de partition centrale sur un ensemble fini X. Une telle partition représente, en quelque sorte, un "résumé à distance minimum" d'une famille de partitions (classifications).

Nous nous proposons d'examiner ce que devient cette notion lorsque X n'est pas nécessairement fini. Cette généralisation n'est nullement gratuite : que l'on songe aux cartes de géographie, aux "découpages" d'une portion du sol effectués par les naturalistes (à la suite de relevés) etc...

§ 1. $\mathcal{A}$ .PARTITIONS

1.1. Définition

Soit X un ensemble et soit $\mathcal{A}$ une σ-algèbre de parties de X.

Une $\mathcal{A}$-partition de X est une partition de type dénombrable P de X (i.e. formée d'un nombre au plus dénombrable de classes) telle que toute classe $A \in P$ est un élément de $\mathcal{A}$.

Notons $\wedge$ et $\vee$ les opérations inf et sup du treillis Π des partitions de X. Il est clair, si P et Q sont des $\mathcal{A}$-partitions de X, que $P \wedge Q$ et $P \vee Q$ sont aussi des $\mathcal{A}$-partitions.

L'ensemble $\Pi(\mathcal{A})$ des $\mathcal{A}$-partitions de X forme donc un sous-treillis de Π.

Par ailleurs, à chaque partition P de X est associée une relation d'équivalence eq(P). Il est clair que eq(P), lorsque P est une $\mathcal{A}$-partition, est un élément de la σ-algèbre produit $\mathcal{A} \otimes \mathcal{A}$.

1.2. Définition de la quasi-distance D entre les $\mathcal{A}$-partitions.

Rappelons qu'une quasi-distance d vérifie tous les axiomes d'une distance, sauf le premier remplacé par :

- $d(x,y) \geqslant 0$ et $d(x,x) = 0$

Soit μ une mesure positive finie définie sur $\mathcal{A}$ et soit d une quasi-distance bornée sur $\mathcal{A}$. Il est clair que pour tout couple (P, Q) de $\mathcal{A}$-partitions la famille

$$\big(\mu(A \cap B)\, d(A,B)\big)_{(A,B) \in P \times Q}$$ est sommable.

On pose alors

$$D(P,Q) = \sum_{A \in P,\, B \in Q} \mu(A \cap B)\, d(A,B)$$

et on montre sans difficulté que :

Proposition 1 : D est une quasi-distance sur $\Pi(\mathcal{A})$.

Nous nous plaçons désormais dans le cas particulier où la quasi-distance d est définie par :

$$d (A,B) = \mu(A \Delta B) = \mu (A \cup B) - \mu (A \cap B)$$

On dira alors que deux $\mathcal{A}$- partitions P et Q sont <u>égales presque partout</u> (notation P = Q p.p.) lorsque D (P,Q) = 0 ; cela signifie que pour tout $A \in P$ et pour tout $B \in Q$ on a,

soit $\mu (A \cap B) = 0$,

soit $\mu (A \Delta B) = 0$.

On obtient ainsi une relation d'équivalence.

De même, on dira que P est <u>presque contenue</u> dans Q (notation $P \leqslant Q$ p.p.) lorsque pour tout $A \in P$, il existe $B \in Q$ telle que $\mu (A - A \cap B) = 0$.

On vérifie aisément que l'égalité presque partout est compatible avec la structure de treillis de $\Pi (\mathcal{A})$, que si P = Q p.p. et $Q \leqslant R$ p.p. ou $P \leqslant Q$ p.p. et Q = R p.p., alors $P \leqslant R$ p.p. et que P = Q p.p. si et seulement si $P \leqslant Q$ p.p. et $Q \leqslant P$ p.p. . De même, si $P \leqslant Q$ p.p. il existe Q' tel que $P \leqslant Q'$ et Q' = Q p.p. .

1.3. <u>Deux autres définitions de D.</u>

Si P est une $\mathcal{A}$-partition il est clair que la famille $(\mu (A)^2)_{A \in P}$ est sommable. Pcsons alors :

$$m (P) = \sum_{A \in P} \mu(A)^2$$

m est une fonction numérique croissante sur le treillis $\Pi (\mathcal{A})$ compatible avec l'égalité presque partout (plus précisément, si $P \leqslant Q$ p.p. et $m (P) \geqslant m (Q)$ alors P = Q p.p.).

On montre facilement que :

<u>Proposition 2</u> : <u>Pour tout</u> $P, Q \in \Pi (\mathcal{A})$,

$$D (P, Q) = m (P) + m (Q) - 2 m (P \wedge Q).$$

Par ailleurs, si l'on note μ^2 la mesure produit $\mu \otimes \mu$, on obtient :

<u>Proposition 3</u> : <u>Pour</u> P, Q $\in \Pi$ ($\mathcal{A}$) ,

$$D(P, Q) = \mu^2 (eq (P) \Delta eq (Q)).$$

Preuve : Partons des deux relations d'équivalence $\mathcal{R}$ = eq (P) et $\mathcal{S}$ = eq (Q). Il vient, en notant $\chi_{\mathcal{R}}$ la fonction caractéristique de $\mathcal{R}$ et $\mathcal{R}$ (x) la classe d'équivalence de x $\in$ X modulo $\mathcal{R}$:

$$\mu^2 (\mathcal{R} \Delta \mathcal{S}) = \int_{X \times X} | \chi_{\mathcal{R}} (x,y) - \chi_{\mathcal{S}} (x,y) | d (\mu(x) \otimes \mu(y)) =$$

$$\int_X d\mu(x) \left[\int_{\mathcal{S}(x)} (1 - \chi_{\mathcal{R}} (x,y)) d\mu(y) + \int_{X-\mathcal{S}(x)} \chi_{\mathcal{R}}(x,y) d\mu(y) \right]$$

(le théorème de Fubini étant clairement vérifié) :

$$\int_X \mu (\mathcal{S}(x) \Delta \mathcal{R}(x)) d \mu (x) = \sum_{A\in P,\ B\in Q} \mu(A \cap B) \ \mu (A \Delta B).$$

<u>§2. LE PRINCIPE DE PARETO POUR LES $\mathcal{A}$-PARTITIONS</u>

<u>2.1. La notion de transfert</u>

Cette notion a été introduite par Régnier ([3]), c.f. aussi [2] , [1]), nous allons la généraliser quelque peu :

Soit P une partition de X, A $\in$ P une classe de P et A' un sous-ensemble de X tel que A' $\not\subset$ A. On considère la partition P [A', A] dont les classes sont :

- A $\cup$ A'
- B - B $\cap$ A' pour B $\in$ P, B $\neq$ A et B $\not\subset$ A'.

On dit que P [A', A] est obtenue par <u>transfert</u> de A' en A.

<u>Lemme</u> : <u>Soient</u> P = { A_1, ... , A_p } , Q = { B_1, ... , B_q } <u>deux</u> <u>$\mathcal{A}$-partitions</u>. <u>Soit</u> C $\in$ $\mathcal{A}$ <u>tel que</u> :

μ (C - C $\cap$ B_1) = 0, μ (C $\cap$ A_1) $\neq$ 0, μ (C $\cap$ A_2) $\neq$ 0 et

μ (C $\cap$ A_j) = 0 <u>pour</u> j > 2.

Il existe alors un nombre réel U tel que :

$$D(P,Q) - D(P[C, A_1], Q) = 2\mu(C \cap A_2)^2 + \mu(C \cap A_2)\ U,$$

$$D(P,Q) - D(P[C, A_2], Q) = 2\mu(C \cap A_1)^2 - \mu(C \cap A_1)\ U.$$

Preuve : Il est clair qu'on obtient les égalités suivantes (en posant $C_1 = C \cap A_1$, $C_2 = C \cap A_2$) :

$$\mu((A_1 \cup C_2) \cap B_1) = \mu(A_1 \cap B_1) + \mu(C_2),$$

$$\mu((A_2 - C_2) \cap B_1) = \mu(A_2 \cap B_1) - \mu(C_2),$$

$$\mu((A_1 \cup C_2)\ \Delta\ B_1) = \mu(A_1\ \Delta\ B_1) - \mu(C_2),$$

$$\mu((A_2 - C_2)\ \Delta\ B_1) = \mu(A_2\ \Delta\ B_1) + \mu(C_2),$$

et pour $k > 1$:

$$\mu((A_1 \cup C_2) \cap B_k) = \mu(A_1 \cap B_k),\ \mu(A_2 - C_2) \cap B_k) = \mu(A_2 \cap B_k),$$

$$\mu((A_1 \cup C_2)\Delta\ B_k) = \mu(A_1\ \Delta\ B_k) + \mu(C_2),$$

$$\mu((A_2 - C_2)\ \Delta\ B_k) = \mu(A_2\ \Delta\ B_k) - \mu(C_2).$$

Or :

$$D(P,Q) - D(P[C,A_1],Q) = \sum_{k=1}^{q} \Big[\mu(A_1 \cap B_k)\ \mu(A_1\ \Delta\ B_k) -$$

$$((A_1 \cup C_2) \cap B_k)\ \mu((A_1 \cup C_2)\ \Delta\ B_k) + \mu(A_2 \cap B_k)\ \mu(A_2 \Delta\ B_k) -$$

$$\mu((A_2 - C_2) \cap B_k)\ \mu((A_2 - C_2)\ \Delta\ B_k).$$

En utilisant les égalités ci-dessus, on trouve que cette expression vaut :

$$\mu(C_2)^2 + \mu(C_2)\Big[\mu(A_1 \cap B_1) - \mu(A_1\ \Delta\ B_1)\Big] + \mu(C_2)^2 +$$

$$\mu(C_2)\Big[\mu(A_2\Delta\ B_1) - \mu(A_2 \cap B_1)\Big] + \mu(C_2)\sum_{k=2}^{q}\Big[\mu(A_2 \cap B_k) -$$

$$\mu(A_1 \cap B_k)\Big] =$$

$$2\mu(C_2)^2 + \mu(C_2)\Big[\mu(A_1 \cap B_1) - \mu(A_1 \Delta B_1) + \mu(A_2 \Delta B_1) - \mu(A_2 \cap B_1) + \sum_{k=2}^{q} (\mu(A_2 \cap B_k) - \mu(A_1 \cap B_1))\Big] =$$

$$2\,\mu(C_2)^2 + \mu(C_2)\,.\,U\,.$$

Pour calculer $D(P,Q) - D(P[C,A_2], Q)$, il suffit de changer C_2 et C_1 et d'échanger A_1 et A_2.
D'où le résultat.

2.2. Le principe de Pareto

Soit $(\lambda_i)_{i \in \mathbb{N}}$ une suite de nombres réels >0 choisie une fois pour toutes. Pour chaque m.uple $\mathbb{Q} = (Q_1,\dots,Q_m)$ de A-partitions et pour chaque A-partition P, posons :

$$\Delta(P,\mathbb{Q}) = \sum_{i=1}^{m} \lambda_i\, D(P,Q_i)$$

<u>Proposition 4</u> : <u>Soit</u> $\mathbb{Q} = (Q_1,\dots,Q_m) \in \Pi(A)^m$ <u>et soit</u> P <u>une</u> A-<u>partition telle que</u> $\bigwedge_{i=1}^{m} Q_i$ <u>ne soit pas presque contenue dans</u> P. <u>Il existe alors au moins une</u> A-<u>partition</u> P' <u>telle que</u> :

$$\Delta(P',\mathbb{Q}) < \Delta(P,\mathbb{Q}).$$

Preuve : Si $\bigwedge_{i=1}^{m} Q_i$ n'est pas presque contenue dans P, il est clair qu'il existe une classe $D \in \bigwedge_{i=1}^{m} Q_i$ et deux classes A_1 et A_2 de P telles que $\mu(D \cap A_1) \neq 0$ et $\mu(D \cap A_2) \neq 0$.
On pose alors $C = D \cap (A_1 \cup A_2)$ ($\mu(C) \neq 0$) et pour chaque partition Q_i on est sous les hypothèses du lemme de telle sorte que, pour chaque indice i, il existe un nombre réel U_i vérifiant

$$D(P,Q_i) - D(P[C,A_1], Q_i) = 2\,\mu(C_2)^2 + \mu(C_2)U_i\,,$$

$$d(P,Q_i) - D(P[C,A_2], Q_i) = 2\,\mu(C_1)^2 - \mu(C_1)U_i$$

(toujours pour $C_1 = C \cap A_1$, $C_2 = C \cap A_2$)

Il suffit alors, pour obtenir le résultat, de poser :

$$P' = P\,[C, A_1] \quad \text{lorsque} \quad \sum_{i=1}^{m} \lambda_i\, U_i > 0 \;,$$

$$P' = P\,[C, A_2] \quad \text{lorsque} \quad \sum_{i=1}^{m} \lambda_i\, U_i < 0$$

On dit (c.f. [3] , [2]) qu'une $\mathcal{A}$ -partition P est <u>centrale</u> pour $\mathbb{Q} = (Q_1,\ldots,Q_m) \in \Pi\,(\mathcal{A})^m$ lorsque :

$$\Delta\,(P, \mathbb{Q}) = \inf_{R \in \Pi(\mathcal{A})} \Delta\,(R, Q)\ .$$

<u>Corollaire (principe de Pareto)</u> : <u>Si</u> $P \in \Pi\,(\mathcal{A})$ <u>est centrale pour</u> $\mathbb{Q} = (Q_1,\ldots,Q_m)$, $\bigwedge_{i=1}^{m} Q_i$ <u>est presque contenue dans</u> P.

<u>§ 3. APPLICATION : PARTITIONS CENTRALES POUR UNE FAMILLE DE PARTITIONS DE TYPE FINI.</u>

<u>3.1. Le problème des partitions centrales</u>

Désignons par $C\,(\mathbb{Q})$ l'ensemble des partitions centrales pour $\mathbb{Q} \in \Pi\,(\mathcal{A})^m$. Dans le cas où X n'est pas fini, rien ne permet d'affirmer que $C\,(\mathbb{Q})$ est non vide [en effet, le quotient de $\Pi\,(\mathcal{A})$ par l'égalité presque partout n'est pas compact pour la métrique induite par D ; en outre des questions, plus subtiles (convergence faible), de la théorie de l'intégration ne s'appliquent pas]. Toutefois, la proposition 4 permet de résoudre ce problème des partitions centrales dans un cas particulier important dans la pratique.

3.2. Existence des partitions centrales pour une famille de partitions de type fini

Soit $R \in \Pi(A)$; désignons par $\Pi(A;R)$ l'ensemble de A - partitions contenant R ; $\Pi(A;R)$ est un sous-treillis de $\Pi(A)$ isomorphe ou treillis Π_R de toutes les partitions de R (considérée, comme partout dans ce texte, comme l'ensemble de ses classes) [Pour $R = \{A_1,\dots,A_r\}$ cet isomorphisme associe, par exemple, à la A-partition

$\{A_1 \cup A_2, A_3 \cup A_4 \cup A_5,\dots\}$ de X la partition

$\{\{A_1, A_2\}, \{A_3, A_4, A_5\},\dots\}$ de R].

Si la partition R est de <u>type fini</u> (i.e. formée d'un nombre fini de classes) le treillis $\Pi(A;R)$ est fini (puisque Π_R l'est).

<u>Proposition 5 : Toute famille (finie) de A-partitions de type fini admet au moins une partition centrale</u>.

Preuve : Soit $\mathcal{Q} = (Q_1,\dots,Q_m)$ un m-uple de A-partitions de type fini.

Posons $R = \bigwedge_{i=1}^{m} Q_i$ (R est de type fini) et soit $P \in \Pi(A;R)$ telle que $\Delta(P,\mathcal{Q}) = \min_{S \in \Pi(A;R)} \Delta(S,\mathcal{Q})$.

Si P n'est pas centrale pour $\mathcal{Q}$, il existe une A-partition P' telle que $\Delta(P',\mathcal{Q}) < \Delta(P,\mathcal{Q})$, en vertu de la démonstration de la proposition 4 (R étant de type fini) on peut toujours supposer $R \leqslant P'$ p.p.. Cela veut dire qu'il existe P'' telle que $R \leqslant P''$ et $P'' = P'$ p.p. Donc $\Delta(P'',\mathcal{Q}) = \Delta(P',\mathcal{Q})$, ce qui contredit la définition de P.

La démonstration de cette proposition nous indique en outre que pour la recherche des partitions centrales (dans le cas "type fini") on pourra utiliser n'importe quel algorithme (l'algorithme des transferts par exemple) construit pour le cas fini, en se plaçant dans le treillis Π_R (qu'on pourra encore, éventuellement, simplifier, en débarassant la partition R de ses classes de mesure nulle).

Signalons pour terminer que les $\mathcal{A}$-partitions de type fini se rencontrent dans la pratique (et c'est ce qui a motivé ce travail) : cartes de géographie, divers découpages d'une portion de sol effectués à la suite d'observations ou de relevés (botanique, zoologie, écologie, etc...). Dans ces cas, la σ-algèbre $\mathcal{A}$ sera le plus souvent la tribu de Borel d'un compact de $\mathbb{R}^2$; le choix d'une mesure μ dépendra de chaque problème traité (par exemple, sur une carte de géographie $\mu(A)$ peut être la superficie de A, le nombre d'habitants de A, etc...).

-=-=-=-=-=-=-=-=-

REFERENCES

[1] J.P. BARTHELEMY, "Sur les éloignements symétriques et le principe de Pareto", Math. Sci. hum., n° 56, 1976, p. 97 - 125.

[2] J.C. LERMAN, Les bases de la classification automatique, Paris, Gauthier-Villars, 1970.

[3] S. REGNIER, "Sur quelques aspects mathématiques des problèmes de classification automatique", I.C.C., Bulletin 4, 1965, p. 175 - 191.