Statistique et analyse des données

Jacques Dauxois
Jeanne Fine-Fontan
Alain Pousse

Échantillonnage en segmentation. Étude de la convergence

Statistique et analyse des données, tome 4, n° 3 (1979), p. 45-53.
<http://www.numdam.org/item?id=SAD_1979__4_3_45_0>

Statistique et Analyse des Données
3 - 1979 pp. 45-53

ECHANTILLONNAGE EN SEGMENTATION
ETUDE DE LA CONVERGENCE

Jacques DAUXOIS, Jeanne FINE-FONTAN, Alain POUSSE

Laboratoire de Statistique et Probabilités
E.R.A. - C.N.R.S. n° 591
Université Paul Sabatier - 31077 TOULOUSE Cedex

1 - ECHANTILLON STATISTIQUE ET ECHANTILLON DE BASE EN ANALYSE DES DONNEES.

Soit $(\Omega, \mathcal{A}, P)$ un espace probabilisé et X une v.a. définie sur $(\Omega, \mathcal{A}, P)$ et à valeurs dans un espace probabilisable $(\Omega', \mathcal{A}')$. On suppose que $\mathcal{A}$ est une tribu qui contient le singleton $\{x\}$, pour tout x de Ω : ceci est notamment le cas si Ω est fini ou dénombrable et si $\mathcal{A}$ est la tribu de ses parties, ou si Ω est un espace topologique muni de sa tribu borélienne.

On obtient, de façon classique, un échantillon statistique de taille n associé à X à partir de l'espace probabilisé $(\Omega, \mathcal{A}, P)^{\otimes \mathbb{N}^*}$. Pour tout n de $\mathbb{N}^*$ et tout i de $I_n = \{1, \ldots, n\}$, on note Π_i la ième projection canonique : $\omega = \{\omega_j\}_{j \in \mathbb{N}^*} \to \omega_i$. Ces applications $(\Pi_i)_{i \in I_n}$ sont des v.a. à valeurs dans $(\Omega, \mathcal{A})$, indépendantes et de même loi P. On pose :

$$\forall\, i \in I_n \qquad X_i = X \circ \Pi_i$$

Les v.a. $(X_i)_{i \in I_n}$ forment un échantillon statistique, c'est-à-dire sont indépendantes et de même loi P_X que X. On a par ailleurs, pour tout ω de $\Omega^{\mathbb{N}^*}$:

$$X_i(\omega) = X(\omega_i)$$

Si Ω_n^ω ($n \in \mathbb{N}^*$, $\omega \in \Omega^{\mathbb{N}^*}$) est le sous-ensemble de Ω image de la suite finie $(\omega_1, \ldots, \omega_n)$ et si on le munit de la tribu de ses parties, on voit que la valeur en ω de la v.a. $X_i (i \in I_n)$ est effectivement la valeur prise en ω_i par la restriction X^n de X à Ω_n^ω, ce qui légitime la pratique courante de l'Analyse des Données.

Cependant, comme ni P, ni même P_X ne sont connues, on probabilise $(\Omega_n^\omega, \mathcal{P}(\Omega_n^\omega))$ par la probabilité $\mu_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{\omega_i}$, où δ_x est la mesure de Dirac au point x (μ_n affecte donc à chaque "individu" ω_i un "poids" égal à sa fréquence d'apparition). Cela revient implicitement à considérer la probabilité $\nu_n^\omega = \frac{1}{n} \sum_{i=1}^{n} \delta_{\omega_i}$ sur $(\Omega, \mathcal{A})$. Or, pour tout A de $\mathcal{A}$:

$$\nu_n^\omega(A) = \frac{1}{n}\sum_{i=1}^{n} 1_A(\omega_i) = \frac{1}{n}\sum_{i=1}^{n} 1_A \circ \Pi_i(\omega)$$

La suite $\{1_A \circ \Pi_i\}_{i\in\mathbb{N}^*}$ est une suite de v.a. réelles, indépendantes, de même loi et intégrables (d'espérance mathématique P(A)), donc, d'après la loi forte des grands nombres, $\nu_n^\omega(A)$ converge presque sûrement vers P(A). La probabilité image de μ_n^ω par X^n, qui est aussi l'image de ν_n^ω par X, converge donc de la même façon vers P_X.

On remarque qu'on peut sans difficulté considérer, non un espace probabilisé $(\Omega,\mathcal{A},P)$, mais une structure statistique $(\Omega,\mathcal{A},\mathcal{P})$.

La formalisation ci-dessus a été implicitement utilisée en [3] pour établir la convergence d'analyses factorielles obtenues par échantillonnage.

2 - LA SEGMENTATION

En Analyse des Données, la méthode de segmentation répond au problème suivant : soit un échantillon de taille n sur lequel sont observées p variables statistiques qualitatives $X_1,\ldots,X_p$, et une variable statistique Y qui peut être qualitative ou quantitative uni- ou multidimensionnelle. On cherche à "expliquer" Y au moyen des autres variables. La segmentation conduit d'une part à une partition de l'échantillon en "segments", chacun le plus homogène possible, et les plus différenciés deux à deux, et d'autre part à un arbre dichotomique dont chaque branche met en évidence une hiérarchie entre les variables explicatives $X_1,\ldots,X_p$, la première apparue (donc la plus haute dans l'arbre) étant considérée comme la plus liée au critère Y, et ainsi de suite.

Soient donc p v.a. $X_1,\ldots,X_p$ sur $(\Omega,\mathcal{A},P)$, "qualitatives" (c'est-à-dire que X_i est à valeurs dans un ensemble fini E_i, muni de la tribu de ses parties), et une v.a. Y de $(\Omega,\mathcal{A},P)$ dans $(\Omega',\mathcal{A}')$. A chaque étape, on va considérer l'une des v.a. $(X_i)_{i=1,\ldots,p}$, que l'on notera X. Soit $\mathcal{B}$ la sous-tribu de $\mathcal{A}$ engendrée par X. En suivant [1] et [2], on se place sur $L^2(\Omega,\mathcal{A},P)$ pour définir la segmentation.

Cas où Y est qualitative

Ω' est alors fini, et $\mathcal{A}'$ est $\mathcal{P}(\Omega')$. Soit $\mathcal{C}$ la sous-tribu de $\mathcal{A}$ engendrée par Y.
Par définition, une étape de segmentation est la recherche de la v.a. de Bernoulli X' (X' ne prend donc que deux valeurs distinctes) de $L^2(\Omega,\mathcal{B},P)$, centrée et normée, la plus proche de $L^2(\Omega,\mathcal{C},P)$.

Comme $\mathcal{B}$ est finie, il n'existe qu'un nombre fini de v.a. de Bernoulli dans $L^2(\Omega,\mathcal{B},P)$, donc le problème admet toujours une solution.

La méthode est actuellement utilisée souvent sans aucune contrainte sur les variables explicatives (quant au nombre de leurs modalités, notamment) ; il nous a donc paru intéressant d'étudier ici la stabilité dans le cadre le plus général.

Cas où Y est quantitative

La définition est la même que ci-dessus, en remplaçant $L^2(\Omega,\mathcal{C},P)$ par le sous-espace F_Y de $L^2(\Omega,\mathcal{A},P)$ des v.a. (centrées) combinaisons linéaires des composantes centrées de Y.

On peut considérer qu'on a ainsi défini la "segmentation exacte", qu'on ne peut obtenir que si P est connue (ou au moins $P_{(X,Y)}$).

On travaille en Analyse des Données à partir d'un échantillon $\omega_1,\ldots,\omega_n$. On peut alors utiliser les notations du paragraphe 1, que l'on applique aux v.a. X,Y ou Z = (X,Y). La segmentation est alors obtenue comme ci-dessus en remplaçant $(\Omega,\mathcal{A},P)$ par $(\Omega_n^\omega, \mathcal{P}(\Omega_n^\omega), \mu_n^\omega)$, X par X^n et Y par Y^n. Elle est implicitement regardée comme une "segmentation approchée par échantillonnage" de la segmentation exacte. On se propose d'examiner ici le bien-fondé de ce point de vue, par une étude de la convergence lorsque la taille n de l'échantillon augmente indéfiniment.

3 - CAS D'UN CRITERE Y QUALITATIF.

A chaque étape de la segmentation exacte, soit X' une v.a. de Bernoulli, de $L^2(\Omega,\mathcal{B},P)$, qui engendre la tribu $\mathcal{D}$. On note $\{D_1,D_2\}$ la partition de Ω qui lui est associée. Il correspond à X' une seule v.a. de Bernoulli U' centrée et normée qui engendre $\mathcal{D}$ (au signe près). La tribu $\mathcal{C}$ est engendrée par une partition $\{C_1,\ldots,C_q\}$ de Ω, et on note :

$$p_{jk} = P[D_j \cap C_k] \qquad j = 1,2 \qquad k = 1,\ldots,q$$

d'où, de façon classique :

$$p_{.k} = p_{1k} + p_{2k} \qquad\qquad p_{j.} = \sum_{k=1}^{q} p_{jk}$$

Soit d la distance de U' à $L^2(\Omega,\mathcal{C},P)$, et ρ l'unique coefficient de corrélation canonique de l'analyse canonique de $\mathcal{D}$ et $\mathcal{C}$ (cf. [3] p. 79 à 81 ; on sait qu'on peut alors supposer, sans perte de généralité, que les $p_{.k}$ et $p_{j.}$ sont non nuls). On a :

$$d^2 = 1 - \rho^2$$

et, d'après [4] ou [3] :

$$1 + \phi^2 = \sum_{i=o}^{1} \rho_i^2 = 1 + \rho^2 = \sum_{j=1}^{2}\sum_{k=1}^{q} \left(\frac{p_{jk}}{p_{j.}p_{.k}}\right)^2 p_{j.}\, p_{.k}$$

d'où :

$$d^2 = 2 - \sum_{j=1}^{2}\sum_{k=1}^{q} \frac{p_{jk}^2}{p_{j.}p_{.k}}$$

On peut remarquer que d^2 ne dépend que des tribus $\mathcal{D}$ et $\mathcal{C}$, et ne dépend donc de X', v.a. de Bernoulli initiale, que par l'intermédiaire de $\mathcal{D}$.

Pour la segmentation approchée, on considère la restriction X'^n de X' à Ω_n^ω ; X'^n est mesurable par rapport à la tribu $\mathcal{B}^n$ engendrée par la restriction X^n de X à Ω_n^ω. Il lui correspond la partition $\{D_1^n, D_2^n\}$ de Ω_n^ω, où $D_j^n = D_j \cap \Omega_n^\omega$ (j=1,2), et on pose : $C_k^n = C_k \cap \Omega_n^\omega$ (k = 1,...,q) ; la restriction Y^n de Y à Ω_n^ω engendre alors la tribu $\mathcal{C}^n$ associée à la partition $\{C_k^n\}_{k=1,\ldots,q}$.

On note :

$$\Pi_{jk}^{(n,\omega)} = \mu_n^\omega[D_j^n \cap C_k^n] = \nu_n^\omega[D_j \cap C_k]$$

et donc : $\Pi_{j.}^{(n,\omega)} = \nu_n^\omega(D_j)$ $\qquad$ $\Pi_{.k}^{(n,\omega)} = \nu_n^\omega[C_k]$

D'après le paragraphe 1, pour tout couple (j,k), et pour presque tout ω de $\Omega^{\mathbb{N}^*}$:

$$\lim_{n\to\infty} \Pi_{j,k}^{(n,\omega)} = P[\, D_j \cap C_k \,] = p_{jk}$$

$$\lim_{n\to\infty} \Pi_{j.}^{(n,\omega)} = p_{j.} \qquad \lim_{n\to\infty} \Pi_{.k}^{(n,\omega)} = p_{.k}$$

Comme $p_{j.}$ et $p_{.k}$ sont non nuls, il existe n_o tel que $\Pi_{j.}^{(n,\omega)}$ et $\Pi_{.k}^{(n,\omega)}$ sont non nuls, presque sûrement, pour $n \geq n_o$. Si V'_n est une v.a. de Bernoulli centrée et normée associée à X'^n, sa distance $d_{n,\omega}$ à $L^2(\Omega_n^\omega, \mathcal{B}^n, \mu_n^\omega)$ vérifie, comme ci-dessus :

$$d^2_{n,\omega} = 2 - \sum_{j=1}^{2} \sum_{k=1}^{q} \frac{[\Pi_{jk}^{(n,\omega)}]^2}{\Pi_{j.}^{(n,\omega)}\, \Pi_{.k}^{(n,\omega)}}$$

On en déduit, pour presque tout ω de $\Omega^{\mathbb{N}^*}$:

$$\lim_{n\to+\infty} d_{n,\omega} = d$$

En considérant toutes les sous-tribus $\mathcal{D}$ de Bernoulli de $\mathcal{B}$, qui sont en nombre fini (on peut les indexer par i variant dans I), on obtient toutes les sous-tribus $\mathcal{D}^n$ de Bernoulli de $\mathcal{B}^n$. Donc, pour presque tout ω, il existe $n_1(\omega)$ tel que :

$$\forall\ n \geq n_1(\omega) \qquad \inf_{i\in I} d^i_{n,\omega} = \inf_{i\in I} d^i$$

La sous-tribu $\mathcal{D}$ "optimale" induit alors sur Ω_n^ω la sous-tribu $\mathcal{D}^n$ "optimale".

On voit ainsi que, pour presque tout ω, il existe $n_1(\omega)$ tel que, pour tout n supérieur ou égal à $n_1(\omega)$, la segmentation sur l'échantillon "coïncide" avec la segmentation exacte.

4 - CAS D'UN CRITERE Y QUANTITATIF UNIDIMENSIONNEL

Comme précédemment, à chaque étape de la segmentation exacte, on note X' une v.a. de Bernoulli centrée normée de $L^2(\Omega, \mathcal{B}, P)$, et $\{D_1, D_2\}$ la partition de Ω associée à la tribu $\mathcal{D}$ engendrée par X'.

Soit Y une v.a. réelle sur $(\Omega, \mathcal{A}, P)$, de carré intégrable. Le sous-espace F_Y est alors de dimension 1, et engendré par $Y - E(Y)$. Soit W le vecteur unitaire de F_Y égal à $\frac{Y-E(Y)}{\|Y-E(Y)\|}$.

Minimiser la distance de X' à F_Y est équivalent à maximiser la valeur absolue du coefficient de corrélation de X' et W. Or, X' s'écrit :

$$X' = a\, 1_{D_1} + b\, 1_{D_2}$$

avec, comme X' est centrée et normée :

$$a = \sqrt{\frac{P(D_2)}{P(D_1)}} \quad , \quad b = -\sqrt{\frac{P(D_1)}{P(D_2)}} \, .$$

Le coefficient de corrélation est donc :

$$< X', W > = a \int_{D_1} W\, dP + b \int_{D_2} W\, dP$$

W est centrée, d'où :

$$\int_\Omega W\, dP = \int_{D_1} W\, dP + \int_{D_2} W\, dP = 0$$

et ainsi :

$$< X' , W > = \left(\sqrt{\frac{P(D_2)}{P(D_1)}} + \sqrt{\frac{P(D_1)}{P(D_2)}} \right) \int_{D_1} W \, dP = \frac{1}{\sqrt{P(D_1)P(D_2)}} \int_{D_1} W \, dP$$

Pour expliciter la segmentation approchée, on reprend les notations du paragraphe 3. Soit V^n la v.a. $\dfrac{Y^n - E_n(Y^n)}{\| Y^n - E_n(Y^n) \|_n}$, où E_n est l'intégrale sur Ω_n^ω pour la probabilité μ_n^ω, et $\| \cdot \|_n$ la norme de $L^2(\Omega_n^\omega, \mathcal{P}(\Omega_n^\omega), \mu_n^\omega)$. On cherche la partition $\{\Delta_1^n , \Delta_2^n\}$ de Ω_n^ω qui maximise, comme ci-dessus, la valeur absolue de :

$$\frac{1}{\sqrt{\mu_n^\omega(\Delta_1^n)\mu_n^\omega(\Delta_2^n)}} \int_{\Delta_1^n} V^n \, d\mu_n^\omega = I(\Delta_1^n, \Delta_2^n)$$

A chaque partition $\{D_1, D_2\}$ de Ω, on associe $D_1^n = D_1 \cap \Omega_n^\omega$ et $D_2^n = D_2 \cap \Omega_n^\omega$.

D'après le paragraphe 1, pour i = 1 ou 2 :

$$\mu_n^\omega(D_i^n) = \nu_n^\omega(D_i) \qquad \text{et} \qquad \lim_{n \to +\infty} \nu_n^\omega(D_i) = P(D_i) \quad \text{presque sûrement .}$$

De plus

$$E_n(Y^n) = \int_{\Omega_n^\omega} Y^n \, d\mu_n^\omega = \int_{\Omega} Y \, d\nu_n^\omega = \frac{1}{n} \sum_{i=1}^{n} Y \circ \pi_i(\omega)$$

Comme Y appartient à $L^1(\Omega, \mathcal{A}, P)$, on déduit de la loi forte des grands nombres la convergence presque sûre de $E_n(Y^n)$ vers E(Y). En outre :

$$\| Y^n - E_n(Y^n) \|_n^2 = \operatorname{var}_n Y_n = E_n [Y_n^2] - [E_n(Y_n)]^2 ,$$

qui converge presque sûrement (loi forte appliquée à Y_n^2 cette fois) vers

$$E(Y^2) - [E(Y)]^2 = \| Y - E(Y) \|^2$$

Enfin, la loi forte appliquée à $Y \, 1_{D_1}$ établit la convergence presque sûre de $\int_{D_1^n} Y^n d\mu_n^\omega = \int_{D_1} Y \, d\nu_n^\omega$ vers $\int_{D_1} Y \, dP$. Comme $I(D_1^n, D_2^n)$ s'écrit :

$$I(D_1^n, D_2^n) = \frac{1}{\sqrt{\mu_n^\omega(D_1^n)\mu_n^\omega(D_2^n)}} \cdot \frac{1}{\| Y^n - E_n(Y^n) \|_n} \left[\int_{D_1^n} Y^n \, d\mu_n^\omega - E_n(Y^n) \, \mu_n^\omega(D_1^n) \right]$$

on en déduit la convergence presque sûre, lorsque n augmente indéfiniment, de $I(D_1^n, D_2^n)$ vers

$$\frac{1}{\sqrt{P(D_1)P(D_2)}} \int_{D_1} W \, dP .$$

De la même façon qu'au paragraphe 3, on voit donc que, si $\{D_1,D_2\}$ est la partition optimale de la segmentation exacte, la segmentation approchée, pour n suffisamment grand, conduit presque sûrement à la partition $\{D_1^n,D_2^n\}$, où D_i^n est la trace de D_i sur Ω_n^ω (i=1,2).

5 - CAS D'UN CRITERE QUANTITATIF MULTIDIMENSIONNEL.

Dans le cadre du paragraphe précédent, on peut voir que la distance d de X' à F_Y vérifie :

$$\|Y - E(Y)\|^2 \; d^2 = \int_{D_1} (Y-\bar{Y}_1)^2 \, dP + \int_{D_2} (Y-\bar{Y}_2)^2 \, dP$$

où

$$\bar{Y}_1 = \frac{1}{P(D_1)} \int_{D_1} Y \, dP \quad \text{et} \quad \bar{Y}_2 = \frac{1}{P(D_2)} \int_{D_2} Y \, dP$$

En effet, si Z est Y-E(Y), on a :

$$\int_{D_1} (Y-\bar{Y}_1)^2 \, dP = \int_{D_1} (Z-\bar{Z}_1)^2 \, dP = \int_{D_1} Z^2 \, dP - \frac{1}{P(D_1)} \left(\int_{D_1} Z \, dP\right)^2$$

d'où, comme Z est centrée

$$\|Z\|^2 . \; d^2 = \int_{\Omega} Z^2 \, dP - \frac{1}{P(D_1)P(D_2)} \left(\int_{D_1} Z \, dP\right)^2$$

c'est-à-dire, puisque $Z = W \|Z\|$:

$$d^2 = [\; 1 - < X', W >^2 \;] \quad , \text{ d'où le résultat.}$$

C'est la généralisation de cette expression qui va être utilisée dans le cas d'un critère multidimensionnel.

Soit, pour q dans $\mathbb{N}^*$, $Y = (Y_1,\ldots,Y_q)$ une v.a. sur $(\Omega,\mathcal{A},P)$, à valeurs dans $(\mathbb{R}^q, \mathcal{B}_{\mathbb{R}^q})$ où $\mathcal{B}_{\mathbb{R}^q}$ est la tribu borélienne de $\mathbb{R}^q$. On suppose que chaque composante Y_ℓ de Y est de carré intégrable. Etant donné la définition de F_Y, on ne restreint pas la généralité en supposant Y centrée.

Soit Λ la matrice des covariances de Y ; on suppose $Y_1,\ldots,Y_q$ linéairement indépendantes dans $L^2(\Omega,\mathcal{A},P)$ (cas auquel on se ramène systématiquement par suppression éventuelle de certaines variables qui, étant combinaisons linéaires des autres, n'apportent aucune information supplémentaire pour la segmentation) ; cela implique que Λ est inversible. On munit alors $\mathbb{R}^q$ de la "métrique de Mahalanobis", de matrice Λ^{-1}, et on sait (cf [1]) que minimiser la distance de X' à F_Y équivaut à minimiser sur $\{D_1,D_2\}$, partition de Ω, l'expression :

$$D^2 = \int_{D_1} \|Y-\bar{Y}_1\|^2_{\Lambda^{-1}} \, dP + \int_{D_2} \|Y-\bar{Y}_2\|^2_{\Lambda^{-1}} \, dP$$

où $\bar{Y}_j$ est l'élément de $\mathbb{R}^q$: $\frac{1}{P(D_j)} \int_{D_j} Y \, dP \qquad (j=1,2)$

La segmentation approchée conduit à minimiser sur $\{\Delta_1^n, \Delta_2^n\}$, partition de Ω_n^ω :

$$D_n^2 = \int_{\Delta_1^n} \|Y^n - \bar{Y}_1^n\|^2_{\Lambda_n^{-1}} \, d\mu_n^\omega + \int_{\Delta_2^n} \|Y^n - \bar{Y}_2^n\|^2_{\Lambda_n^{-1}} \, d\mu_n^\omega$$

où, pour $j = 1,2$:

$$\bar{Y}_j^n = \frac{1}{\mu_n^\omega(\Delta_j^n)} \int_{\Delta_j^n} Y^n \, d\mu_n^\omega$$

et où Λ_n est la matrice des covariances de Y^n (matrice que l'on suppose provisoirement inversible). Si on note m_{ij} (resp. m_{ij}^n) $(1 \le i \le q, \ 1 \le j \le q)$ le terme général de Λ^{-1} (resp. Λ_n^{-1}), on a:

$$\|Y - \bar{Y}_1\|^2_{\Lambda^{-1}} = \sum_{i,j=1}^{q} m_{ij} \, (Y - \bar{Y}_1)_i \, (Y - \bar{Y}_1)_j$$

d'où :

$$D^2 = \sum_{i,j=1}^{q} m_{ij} \left[\int_{D_1} (Y - \bar{Y}_1)_i \, (Y - \bar{Y}_1)_j \, dP + \int_{D_2} (Y - \bar{Y}_2)_i \, (Y - \bar{Y}_2)_j \, dP \right]$$

et, de la même façon :

$$D_n^2 = \sum_{i,j=1}^{q} m_{ij}^n \left[\int_{\Delta_1^n} (Y^n - \bar{Y}_1^n)_i \, (Y^n - \bar{Y}_1^n)_j \, d\mu_n^\omega + \int_{\Delta_2^n} (Y^n - \bar{Y}_2^n)_i \, (Y^n - \bar{Y}_2^n)_j \, d\mu_n^\omega \right]$$

Or, pour $k = 1,2$, si $D_k^n = D_k \cap \Omega_n^\omega$:

$$\int_{D_k^n} (Y^n - \bar{Y}_k^n)_i \, (Y^n - \bar{Y}_k^n)_j \, d\mu_n^\omega = \int_{D_k} Y_i \, Y_j \, d\nu_n^\omega - (\bar{Y}_k^n)_i \int_{D_k} Y_j \, d\nu_n^\omega - (\bar{Y}_k^n)_j \int_{D_k} Y_i \, d\nu_n^\omega + \nu_n^\omega(D_k) (\bar{Y}_k^n)_i \, (\bar{Y}_k^n)_j$$

$$= \int_{D_k} Y_i \, Y_j \, d\nu_n^\omega - \nu_n^\omega(D_k) \, (\bar{Y}_k^n)_i \, (\bar{Y}_k^n)_j$$

avec, d'après la loi forte des grands nombres (Y_i et Y_j appartiennent à L^2, donc $Y_i \, Y_j$ est intégrable, ainsi que $Y_i \, 1_{D_k}$, $Y_j \, 1_{D_k}$ et $Y_i \, Y_j \, 1_{D_k}$) :

$\forall \ i,j = 1,\ldots,q$ $\quad \lim_{n\to\infty} \int_{D_k} Y_i \, Y_j \, d\nu_n^\omega = \int_{D_k} Y_i \, Y_j \, dP$ $\quad$ presque sûrement

$\forall \ i = 1,\ldots,q$ $\quad \lim_{n\to\infty} \int_{D_k} Y_i \, d\nu_n^\omega = \int_{D_k} Y_i \, dP$ $\quad$ presque sûrement

$\lim_{n\to\infty} \nu_n^\omega(D_k) = P(D_k)$ $\quad$ presque sûrement

$$\forall\ i = 1,\dots,q \qquad \lim_{n\to\infty} (\bar{Y}_k^n)_i = \lim_{n\to\infty} \frac{1}{\nu_n^\omega(D_k)} \int_{D_k} Y_i \, d\nu_n^\omega = \frac{1}{P(D_k)} \int_{D_k} Y_i \, dP = (\bar{Y}_k)_i \quad \text{presque sûrement}$$

ce qui entraine :

$$\forall\ i,j = 1,\dots,q \qquad \lim_{n\to\infty} \int_{D_k^n} (Y^n - \bar{Y}_k^n)_i \, (Y^n - \bar{Y}_k^n)_j \, d\mu_n^\omega = \int_{D_k} (Y - \bar{Y}_k)_i \, (Y - \bar{Y}_k)_j \, dP \quad \text{presque sûrement.}$$

Par ailleurs, le terme général de Λ_n est :

$$\lambda_{ij}^n = < Y_i^n , Y_j^n >_n = \int_{\Omega_n^\omega} Y_i^n \, Y_j^n \, d\mu_n^\omega = \int_{\Omega} Y_i \, Y_j \, d\nu_n^\omega$$

d'où : $$\lim_{n\to\infty} \lambda_{ij}^n = \int_{\Omega} Y_i \, Y_j \, dP = \lambda_{ij} \quad \text{presque sûrement .}$$

Le déterminant de Λ^n converge donc vers celui de Λ, qui est non nul puisque Λ est inversible. Donc, pour n suffisamment grand, le déterminant de Λ^n est non nul, et Λ^n est inversible (ce qui justifie l'hypothèse faite plus haut) ; comme on l'obtient par composition d'applications continues à partir des λ_{ij}^n, chaque terme m_{ij}^n de Λ_n^{-1} converge donc presque sûrement vers le terme correspondant m_{ij} de Λ^{-1}. On en déduit :

$$\lim_{n\to+\infty} D_n^2 = D^2 \quad \text{presque sûrement.}$$

Il y a un nombre fini de partitions $\{D_1, D_2\}$ de Ω qui conduisent, par restriction, à toutes les partitions $\{D_1^n, D_2^n\}$ de Ω_n^ω ; ceci démontre, comme au § 3., mais ici dans le cas d'un critère Y multidimensionnel, que, pour presque tout ω, il existe encore $n_1(\omega)$ tel que, pour tout n supérieur ou égal à $n_1(\omega)$, la segmentation sur l'échantillon coïncide avec la segmentation exacte.

6 - CONCLUSION

On a établi, dans tous les cas, la convergence par échantillonnage, et donc la stabilité pour un échantillon de taille suffisamment grande, de la méthode de segmentation, à chacune des étapes. Cependant, il demeure qu'une modification minime à une étape peut avoir des répercussions très importantes sur les étapes suivantes. Il faut noter qu'on obtient ici presque sûrement, à partir d'un certain rang n_o, non une approximation, mais précisément la restriction à Ω_n^ω de la variable X et de la dichotomie $\{D_1, D_2\}$ correspondant à la segmentation exacte : cela parce qu'il y a un nombre fini de v.a. explicatives X_i, ayant chacune un nombre fini de modalités. Pour un échantillon de taille supérieure à n_o, le risque de perturbation des étapes suivantes est donc nul, presque sûrement.

BIBLIOGRAPHIE

[1] A. BACCINI et A. POUSSE - Segmentation aux moindres carrés : un aspect synthétique - Revue de Statistique Appliquée. 1975. Vol. XXIII N°3.

[2] A. BACCINI - Aspect synthétique de la segmentation et traitement de variables qualitatives à modalités ordonnées.
Thèse de 3ème cycle - Université Paul Sabatier - Toulouse - 1975.

[3] J. DAUXOIS et A. POUSSE - Les analyses factorielles en Calcul des Probabilités et en Statistique : essai d'étude synthétique.
Thèse - Université Paul Sabatier - Toulouse - 1976.

[4] H.O. LANCASTER - The Chi-squared Distribution - Wiley Publications in Statistics - 1969.