REVUE DE STATISTIQUE APPLIQUÉE

ALAIN BACCINI
ALAIN POUSSE

Segmentation aux moindres carrés : un aspect synthétique

Revue de statistique appliquée, tome 23, n° 3 (1975), p. 17-35
<http://www.numdam.org/item?id=RSA_1975__23_3_17_0>

SEGMENTATION AUX MOINDRES CARRÉS : UN ASPECT SYNTHÉTIQUE (1)

Alain BACCINI et Alain POUSSE
Laboratoire de statistique
Université Paul Sabatier – Toulouse

INTRODUCTION

Nous considérons ici la segmentation comme une technique de partition d'une population par dichotomies successives au moyen de variables explicatives et *par référence à un critère,* ou variable à expliquer. Nous n'envisagerons donc pas les méthodes de segmentation s'apparentant à la classification par l'absence de variable à expliquer (comme par exemple [18]).

On se propose dans ce cadre de donner une vue synthétique de la segmentation en l'envisageant sous un aspect nouveau. On s'intéresse principalement à la segmentation dite "aux moindres carrés".

Divers problèmes se trouvent posés dès qu'on aborde ce type de segmentation : le choix de la métrique (pourquoi la distance du χ^2 ?), l'utilisation des variables explicatives (une à une, et non simultanément), . . . La présentation envisagée ici, outre son aspect plus général, permet d'apporter à ces questions, déjà développées par ailleurs, de nouveaux éléments de réponse.

Au lieu de se placer sur des espaces d'individus et d'utiliser des critères propres à chacun de ces espaces, on se place sur l'espace des variables ; cela permet de retrouver la plupart des méthodes usuelles tout en les présentant dans un cadre homogène. Cela permet en outre d'en étudier de façon plus complète les propriétés, d'obtenir une définition plus originale faisant un lien avec la régression, et de proposer diverses approximations.

Après avoir présenté la nature du problème abordé, puis rappelé l'essentiel des méthodes les plus courantes, on exposera le point de vue permettant de traiter la segmentation sous un angle plus global, et l'on montrera alors comment sont ainsi synthétisées les principales techniques, avant de terminer par les approximations proposées.

I – PRESENTATION DU PROBLEME ABORDE

1.1 Données et notations :

On dipose d'un ensemble I de n individus, ou unités statistiques, i

$$(i \in I = \{1 , 2 , \ldots , n\}) ,$$

---

(1) Texte remis en Mars 1974, révisé en Mars 1975.

à chacun desquels est associé un poids $p_i > 0 \left(\sum_{i \in I} p_i = 1 \right)$, et sur lesquels on observe d'une part une variable statistique Y à valeurs dans un ensemble E et pouvant être soit quantitative (dans ce cas E est l'ensemble R^q , $q \in N^*$) soit qualitative (E est alors un ensemble fini, ordonné ou pas, non contenu dans R), et d'autre part une suite de p variables statistiques X_1 , X_2 , . . . , X_p toutes qualitatives. Soit X l'une quelconque de ces variables ; on note m le nombre de modalités de X.

La variable statistique Y est appelée "variable à expliquer" (ou parfois critère), et les variables statistiques X "variables explicatives" (ou parfois facteurs).

A chaque X est associée une partition $\mathcal{P}$ de I égale à $\{G_1 , G_2 , \ldots , G_m\}$ (m variant avec X) où G_k , $k \in K = \{1 , 2 , \ldots , m\}$, est l'ensemble des individus de I présentant la $k^{ème}$ modalité de X.

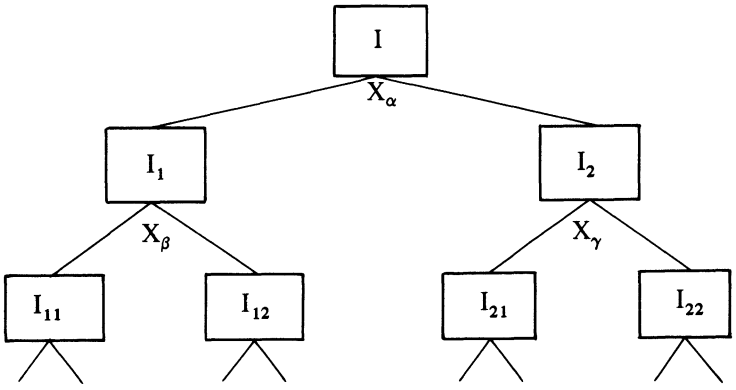
1.2. Le problème abordé et la procédure générale utilisée :

La segmentation a pour but d'"expliquer" la variable statistique Y au moyen des variables statistiques X, en utilisant les observations de chacune de ces variables sur I. Pour cela elle consiste à rechercher :

– une partition de I en sous-ensembles qui soient entre eux "les plus différents possible" (condition 1), et chacun "le plus homogène possible" (condition 2), ceci relativement à Y et en un sens que l'on précisera ultérieurement ; on convient d'obtenir cette partition par dichotomies successives de I, chacune étant induite par une dichotomie des différentes modalités de l'une des variables explicatives X. ;

– simultanément, un classement des variables explicatives X par ordre décroissant "d'influence" sur Y.

La procédure de segmentation opérant par dichotomies successives de I permet d'obtenir un "arbre dichotomique" de la façon suivante :



I_1 : ensemble des individus de I présentant certaines modalités de X_α ;

I_2 : ensemble des individus de I présentant les autres modalités de X_α ;

I_{11} : ensemble des individus de I_1 présentant certaines modalités de X_β ;

I_{12} : ensemble des individus de I_1 présentant les autres modalités de X_β ;

et ainsi de suite . . .

On notera que la même variable X peut se trouver sur plusieurs branches différentes et revenir plusieurs fois sur la même branche (cf. pour exemple [13], [5], etc. . .).

Plus une variable explicative est haut placée dans l'arbre dichotomique, plus elle est interprétée comme influente sur Y.

La fixation d'une règle d'arrêt permet d'obtenir la partition cherchée lorsque la procédure s'est terminée sur chacune des branches de l'arbre dichotomique. En pratique cette règle doit s'appliquer assez vite pour éviter d'obtenir un arbre trop complexe et difficile à interpréter.

Remarque : Diverses règles d'arrêt sont utilisées, que nous ne discuterons pas ici. On peut citer, parmi les plus courantes, les suivantes :

a) On peut arrêter la procédure sur un sous-ensemble I′ de I si la dichotomie optimale $\{I'_1, I'_2\}$ de ce sous-ensemble est telle que l'effectif de l'élément de cardinal minimum $(\inf(\text{card}(I'_1), \text{card}(I'_2)))$ soit inférieur à un seuil n_0 fixé a priori (le choix de n_0 dépendant de l'ordre de grandeur de n).

b) La règle d'arrêt peut être liée à une distance. Par exemple on calcule, lorsque Y est qualitative, la distance du χ^2 (cf. [3]) entre les deux groupes de modalités de X permettant de former I'_1 et I'_2, et on poursuit les dichotomies tant que cette distance est assez grande (dépasse le quotient par n de la borne de la table du χ^2 correspondant à un seuil que l'on se fixe a priori).

c) Certaines règles sont fondées sur un indice mesurant la perte relative d'"information", liées au test de FISHER-SNEDECOR (cf. [17]).

d) Il existe encore diverses autres règles d'arrêt plus ou moins simples (cf. par exemple [13], p. 429).

Pour chaque niveau de segmentation (correspondant à une ligne du schéma ci-dessus), et pour chaque sous-ensemble de I figurant à ce niveau, la méthode de recherche de la dichotomie optimale de cet ensemble est toujours la même : déterminer la variable explicative et la dichotomie de ses modalités optimisant un certain indice qui représente "l'influence" des variables explicatives sur la variable à expliquer. Pour cette raison, la recherche d'une dichotomie optimale sera par la suite toujours décrite sur l'ensemble I lui-même.

Remarque : Bien qu'elle soit parfois utilisée dans ce but, la segmentation n'est pas une méthode essentiellement prédictive. Elle s'attache surtout à expliquer une variable statistique par un certain nombre de caractères qualitatifs. L'obtention d'une partition de I dont chaque sous-ensemble soit le plus différencié possible relativement à Y peut certes être utilisé pour faciliter une éventuelle prédiction, mais ce n'est pas le but premier de la méthode. On trouvera une illustration de cette remarque en 4.4. .

II – RESUME DES METHODES USUELLES

On expose ici les méthodes de segmentation aux moindres carrés, qui sont les plus développées et les plus courantes, ainsi que la segmentation par analyse discriminante, qui nous permettra d'illustrer la différence entre segmentation et prédiction. Citons aussi pour mémoire la segmentation par la théorie de l'information (cf. [11] par ex., chap. II.2.).

Toutes ces méthodes sont fondées sur le principe des dichotomies successives présenté plus haut, leur différence résidant dans la nature du critère de sélection de la dichotomie optimale à chaque niveau.

2.1. La segmentation aux moindres carrés :

Y étant à valeurs dans E quelconque (cf. 1.1.) ; on choisit une représentation des individus dans un espace euclidien H, grâce à une application $f : I \to H$ qui factorise à travers Y : $f = g \circ Y$, où g est une application de E dans H déterminée suivant la nature de Y. On note d la métrique sur H.

Dans H, on définit le barycentre $\bar{A}$ de toute partie A de f(I) par :

$$\bar{A} = \frac{1}{P(A)} \sum_{i \in I_A} p_i \, f(i), \quad \text{où} \quad I_A = f^{-1}(A), \quad \text{et où} \quad P(A) = \sum_{i \in I_A} p_i .$$

Ceci permet de considérer sur les parties de f(I) :

– l'indice intragroupe D(A) de toute partie A, qui est le moment d'inertie de A par rapport à son barycentre $\bar{A}$:

$$D(A) = \sum_{i \in I_A} p_i \, d^2[f(i), \bar{A}]$$

– L'indice intergroupe $\mathcal{D}(A_1, A_2)$ de deux parties A_1 et A_2, qui est la somme des moments d'inertie de $\bar{A}_1$ et $\bar{A}_2$ affectés des poids $P(A_1)$ et $P(A_2)$ par rapport au barycentre $\overline{A_1 \cup A_2}$ de $A_1 \cup A_2$:

$$\mathcal{D}(A_1, A_2) = P(A_1)\, d^2\,[\bar{A}_1, \overline{A_1 \cup A_2}] + P(A_2)\, d^2\,[\bar{A}_2, \overline{A_1 \cup A_2}] .$$

Le théorème de KOENIG-HUYGENS permet alors de remarquer que pour tout couple (A_1, A_2) de parties disjointes de f(I), on a :

$$D(A_1) + D(A_2) + \mathcal{D}(A_1, A_2) = D(A_1 \cup A_2) . \qquad (1)$$

On retient alors, parmi toutes les dichotomies $\{I_1, I_2\}$ de I correspondant à une dichotomie de $\{G_1, \ldots, G_m\}$ (cf. 1.1.), celle pour laquelle $\mathcal{D}[f(I_1), f(I_2)]$ est maximum, ce qui répond à la condition 1 de 1.2., ou, ce qui revient au même d'après (1), celle pour laquelle $D[f(I_1)] + D[f(I_2)]$ est minimum, ce qui répond en moyenne à la condition 2 de 1.2. .

Le développement technique de cette procédure va dépendre de la nature de la variable à expliquer Y :

a) *La méthode E.L.I.S.E.E.* (Exploration des Liaisons et Interactions par Segmentation d'un Ensemble Expérimental) :

Cette méthode s'applique dans le cas où la variable à expliquer Y est qualitative (nous la supposerons à q modalités). Chaque individu est ici affecté du même poids $\frac{1}{n}$, et H est l'espace R^q muni de la métrique du χ^2 (cf. [2]), admettant sur la base canonique la matrice $D = \text{diag}\left(\frac{1}{P_1}, \ldots, \frac{1}{P_q}\right)$, où P_ℓ est la somme des poids des individus prenant la ℓ-ième modalité de Y. L'application f fait correspondre à un individu prenant la ℓ-ième modalité de Y ($\ell \in L$, $L = \{1, 2, \ldots, q\}$) le ℓ-ième vecteur de cette base de R^q.

X étant une variable explicative quelconque à m modalités, on note $n_{k\ell}$ ($k \in K$, $K = \{1, 2, \ldots, m\}$, $\ell \in L$) le nombre d'individus de I prenant la modalité k de X et la modalité ℓ de Y, et :

$$n_{.\ell} = \sum_{k=1}^{m} n_{k\ell}\ ,\ n_{k.} = \sum_{\ell=1}^{q} n_{k\ell}\ ;\ \text{(on a donc : } \sum_{\ell=1}^{q} n_{.\ell} = \sum_{k=1}^{m} n_{k.} = n\text{)}.$$

Parmi toutes les partitions de I en deux sous-ensembles I_1 et I_2, avec

$$I_1 = \bigcup_{k \in K_1} G_k\ ,\ I_2 = \bigcup_{k \in K_2} G_k\ ,\ K_1 \cap K_2 = \phi\ ,\ K_1 \cup K_2 = K \qquad (2)$$

on cherche celle qui maximise $\mathcal{D}(I_1, I_2)$. En posant :

$$n_{K_1\ell} = \sum_{k \in K_1} n_{k\ell}\ ,\ n_{K_2\ell} = \sum_{k \in K_2} n_{k\ell}\ ,\ n_{K_1.} = \sum_{k \in K_1} n_{k.}\ ,\ n_{K_2.} = \sum_{k \in K_2} n_{k.}\ ,$$

il vient (cf. [3]) :

$$\mathcal{D}(I_1, I_2) = \frac{n_{K_1.} n_{K_2.}}{n} \sum_{\ell=1}^{q} \frac{1}{n_{.\ell}} \left(\frac{n_{K_1\ell}}{n_{K_1.}} - \frac{n_{K_2\ell}}{n_{K_2.}}\right)^2 = \Phi^2\ ,$$

où Φ^2 a été défini par Cramer et Lancaster (cf. [12]).

b) *La méthode A.I.D.* : Celle-ci s'applique dans le cas où Y est quantitative unidimensionnelle : H est alors R, et chaque individu i de I est représenté par le point Y(i) affecté du poids p_i.

On cherche encore la partition $\{I_1, I_2\}$ de I définie comme en (2) rendant minimum :

$$D(I_1) + D(I_2) = \sum_{i \in I_1} p_i (Y(i) - \bar{Y}_{K_1})^2 + \sum_{i \in I_2} p_i (Y(i) - \bar{Y}_{K_2})^2$$

où $\bar{Y}_{K_j} = \frac{1}{P_j} \sum_{i \in I_j} p_i\ Y(i)$ pour $j = 1,2$ avec $P_j = \sum_{i \in I_j} p_i$ (cf. [3]).

c) *Généralisation de la méthode A.I.D. au cas où Y est multidimensionnelle :* H est dans ce cas R^q muni de la métrique de MAHALANOBIS (la dimension de Y est $q > 1$) de matrice Λ^{-1} sur la base canonique, où Λ est la matrice des covariances de $Y = (Y_1, Y_2, \ldots, Y_q)$. On note $\|\ .\ \|$ la norme de cet espace. Chaque individu i de I est représenté par le point Y(i) de R^q affecté du poids p_i (cf. [3]).

La partition $\{I_1 , I_2\}$ cherchée minimise alors :

$$D(I_1) + D(I_2) = \sum_{i \in I_1} p_i \, ||Y(i) - \overline{Y}_{K_1}||^2 + \sum_{i \in I_2} p_i \, ||Y(i) - \overline{Y}_{K_2}||^2 .$$

Remarque : La recherche pratique de la dichotomie optimale $\{I_1 , I_2\}$ de I consiste dans tous les cas exposés ci-dessus à calculer les indices appropriés pour chacune des dichotomies possibles.

2.2. La segmentation par analyse discriminante :

Introduite par BELSON (cf. [1]), puis par [19], cette méthode "S.P.A.D." s'applique seulement au cas où Y est qualitative à deux modalités. On construit une dichotomie des modalités de chacune des variables explicatives X en faisant correspondre à sa k-ième modalité la modalité ℓ de Y pour laquelle $\frac{n_{k\ell}}{n_{.\ell}}$ est maximum ($k \in K$; $\ell \in L$; $L = \{1 , 2\}$). Séparant alors les modalités de X auxquelles on a fait correspondre la première modalité de Y de celles auxquelles on a fait correspondre la deuxième, on obtient une dichotomie $\{K_1 , K_2\}$ de ces modalités induisant une dichotomie $\{I_1 , I_2\}$ de I.

[19] associe alors à chaque variable explicative X un indice de discrimination d défini par :

$$1 - \frac{1}{2}\left[\frac{1}{n_{.2}} \sum_{k \in K_1} n_{k2} + \frac{1}{n_{.1}} \sum_{k \in K_2} n_{k1}\right] ,$$

et retient celle des dichotomies de I associée à la variable explicative pour laquelle d est maximum.

III – AUTRE APPROCHE DE LA SEGMENTATION

On considère l'espace probabilisé $(I , \mathcal{P}(I) , P)$, où :

– $\mathcal{P}(I)$ est la tribu des parties de I ;

– P est la probabilité sur $(I , \mathcal{P}(I))$ définie par : $P[\{i\}] = p_i , i \in I$.

On fait ici de la statistique descriptive, et l'on n'introduit cet espace probabilisé associé à I que par commodité mathématique. Il ne s'agit donc pas d'une structure statistique telle qu'on l'entend en statistique inférentielle.

Toute variable aléatoire réelle X de $L^2(I , \mathcal{P}(I) , P)$ (que l'on désignera par L^2) se décomposant de façon unique sur les n indicatrices 1_i des $\{i\}$ sous la forme $X = \sum_{i=1}^{n} X(i) \, 1_i$, L^2 est un espace vectoriel réel de dimension n qui sera identifié à R^n. La métrique de cet espace euclidien est définie par le produit scalaire :

$$\forall (X_1 , X_2) \in L^2 \times L^2 : < X_1 , X_2 > = \int X_1 X_2 \, dP = \sum_{i=1}^{n} p_i \, X_1(i) \, X_2(i) ,$$

auquel est associée la norme : $||X|| = \sqrt{<X , X>}$, et a donc pour matrice dans la base canonique : $D_p = \text{diag}(p_1 , \ldots , p_n)$.

C'est sur cet espace L^2, identifié à R^n muni de la métrique D_p, que l'on se placera dorénavant pour aborder la segmentation.

Soit $\mathcal{B}$ la sous-tribu de $\mathcal{P}(I)$ engendrée par une v.a. (variable aléatoire) X qualitative à m modalités. C'est une tribu sur l'ensemble fini I, qui est engendrée par la partition $\{G_1, \ldots, G_m\}$ de I. Toute v.a. réelle $\mathcal{B}$-mesurable appartient à $L^2(I, \mathcal{B}, P)$ et peut alors s'écrire comme une combinaison linéaire des indicatrices des G_k. On notera X^k l'indicatrice de G_k, $(k \in K)$, v.a. appelée parfois "indicatrice de la k-ième modalité de X". $\{X^1, \ldots, X^m\}$ forme une base de $L^2(I, \mathcal{B}, P)$, par rapport à laquelle on repèrera toute v.a. $\mathcal{B}$-mesurable.

Remarque : On dit parfois que l'on a "représenté" la variable statistique qualitative X par la variable statistique quantitative $(X^1, \ldots, X^m)$, alors qu'on a simplement choisi ici une base de $L^2(I, \mathcal{B}, P)$. En effet, les v.a. qualitatives étant à valeurs dans des espaces probabilisables quelconques (non réels), on ne s'intéresse pas aux valeurs qu'elles prennent, faute de pouvoir les utiliser. Le seul élément que l'on peut considérer est la tribu engendrée. Si l'on change la représentation à ce niveau, on obtient une autre base de $L^2(I, \mathcal{B}, P)$, et donc simplement un changement de repérage des v.a. qui le composent.

Si Y est qualitative à q modalités, on considère de même les q v.a. indicatrices des éléments de la partition $H_1, \ldots, H_q$ engendrée par Y sur I. On note $\{Y^1, \ldots, Y^q\}$ ces indicatrices.

On se place donc maintenant dans le cas où Y est quantitative de dimension q $(q \in N^*)$, auquel on a ramené le cas précédent, et qui constitue ainsi le cas le plus général. Y est alors $(Y_1, \ldots, Y_q)$. Désignant par X l'une quelconque des variables explicatives, soit F_{X_0} (resp. F_{Y_0}) le sous-espace vectoriel de R^n engendré par $X^1, X^2, \ldots, X^m$ (resp. $Y_1, \ldots, Y_q$ et $e = \sum_{i=1}^{n} 1_i$). (F_{X_0} est encore $L^2(I, \mathcal{B}, P)$); Le vecteur e appartenant ainsi à la fois à F_{X_0} (car $\sum_{k=1}^{m} X^k = e$) et à F_{Y_0}, on appelle F_X (resp. F_Y) le sous-espace vectoriel de F_{X_0} (resp. F_{Y_0}) orthogonal à e ; F_X (resp. F_Y) représente donc le sous-espace vectoriel des variables centrées de F_{X_0} (resp. F_{Y_0}).

On cherche alors une v.a. de BERNOULLI de F_X (c'est-à-dire du type

$$X' = a \sum_{k \in K_1} X^k + b \sum_{k \in K_2} X^k ,$$

avec :

$$(a, b) \in R^2 , a \neq b , K_1 \cap K_2 = \phi , K_1 \cup K_2 = K)$$

qui soit normée et de distance à F_Y minimum.

Ce problème admet au moins une solution puisque les v.a. X' normées sont en nombre fini. Ainsi, à chaque variable explicative X_j se trouve associée une (au moins) v.a. particulière X'_j dont on connaît la distance à F_Y ; on retient alors celle des X'_j, soit X'_{j_0}, minimisant cette distance. Il lui correspond une dichotomie de I "respectant" les modalités de X_{j_0} (c'est-à-dire ne séparant jamais 2 individus présentant la même modalité de X_{j_0}).

On obtient ainsi par itérations successives un arbre dichotomique comme décrit en 1.2..

Remarque : Dans le cas où la minimisation de la distance entre X' et F_Y ne donne pas une solution unique, on devrait théoriquement considérer sur le même plan les différentes dichotomies de I correspondant aux différentes v.a. de BERNOULLI X' obtenues, et continuer la procédure sur chaque dichotomie séparément, ce qui conduirait à plusieurs arbres dichotomiques différents et par conséquent à plusieurs partitions optimales différentes de I.

En pratique, il est préférable de ne conserver que celle de ces dichotomies minimisant la distance de X' à F_Y à l'étape précédente (l'ensemble des individus ayant été modifié).

On va développer maintenant cette présentation dans chaque cas particulier, suivant la nature de Y, et montrer qu'elle englobe les diverses méthodes de segmentation aux moindres carrés.

IV – DEVELOPPEMENTS TECHNIQUES ET COMPARAISON AUX METHODES USUELLES

Pour des raisons de commodités techniques, on commence ici par développer le cas où Y est une v.a. qualitative, on étudie ensuite celui où elle est quantitative unidimensionnelle, pour terminer par le cas où elle est quantitative multidimensionnelle, ce cas constituant, comme on l'a déjà vu, le cas le plus général.

4.1. La variable à expliquer Y est qualitative :

Pour toute variable explicative X, considérons ici les deux sous-espaces vectoriels F_X et F_Y de $L^2(I, \mathcal{P}(I), P)$.

Soit X' une v.a. de BERNOULLI normée de F_X, et soit λ le coefficient de corrélation canonique unique de l'analyse canonique de X' et de F_Y. La distance d de X' à F_Y vérifie :

$$d^2 = 1 - \lambda^2.$$

Or, on peut ici écrire (cf. [4] p. 140, et plus généralement [12]) :

$$\Phi^2 = \frac{\chi^2}{n} = \lambda^2 , \qquad \text{où la quantité } \chi^2 \text{ est calculée entre } X' \text{ et } Y.$$

Comme d'après 2.1.a., on a :

$$\mathcal{D}(I_1 , I_2) = \Phi^2 , \qquad \text{on déduit :} \qquad \mathcal{D}(I_1 , I_2) + d^2 = 1 .$$

Il est par conséquent équivalent de maximiser $\mathcal{D}(I_1 , I_2)$ ou de minimiser d^2 ; on voit donc que la recherche de la v.a. de BERNOULLI centrée et normée la plus proche de F_Y va conduire à la même dichotomie $\{I_1 , I_2\}$ de I que la méthode E.L.I.S.E.E..

La procédure exposée ici a l'avantage de remplacer les notions d'indice intragroupe et d'indice intergroupe (qui ne sont pas des distances) par celles de *distance* entre éléments de $L^2(I, \mathcal{P}(I), P)$ ce qui permet d'utiliser des outils mathématiques plus efficaces. Il apparaît ainsi plus naturel, comme dans beaucoup de méthodes d'analyse des données, de raisonner sur les variables que

sur les individus, la transition se faisant ici commodément par les tribus engendrées.

4.2. La variable à expliquer Y est quantitative unidimensionnelle :

Y est ici une v.a. réelle définie sur $(I, \mathcal{P}(I), P)$, appartenant donc a $L^2(I, \mathcal{P}(I), P)$. F_Y est le sous-espace vectoriel de L^2 de dimension 1 engendré par la seule v.a. Y-E(Y). La nature particulière de Y permet de compléter dans ce cas les considérations générales sur le problème de la segmentation.

a) *Aspects complémentaires de la segmentation :* Comme il a déjà été vu, on cherche, pour chacune des v.a. X, la v.a. X' réelle de BERNOULLI, $\mathcal{B}$-mesurable et normée, la plus proche de F_Y dans $L^2(I, \mathcal{P}(I), P)$. Il est équivalent de chercher à minimiser la distance de X' à Y-E(Y) ou la distance de X' à F_Y, cela revenant dans les 2 cas à maximiser la corrélation entre X' et Y-E(Y).

En notant $\mathcal{C}$ la sous-tribu de $\mathcal{B}$ engendrée par X', cela revient à chercher une sous-tribu $\mathcal{C}$ de $\mathcal{B}$ qui soit de BERNOULLI (engendrée par un élément de $\mathcal{B}$ autre que I ou ϕ), et qui minimise :

$$||Y - E(Y) - E^{\mathcal{C}}(Y - E(Y))|| = ||Y - E^{\mathcal{C}}(Y)||.$$

En effet, pour $\mathcal{C}$ donnée, la v.a. $\mathcal{C}$-mesurable la plus proche de Y est $E^{\mathcal{C}}(Y)$, espérance conditionnelle à $\mathcal{C}$ de Y.

Cela revient encore à chercher la fonction réelle mesurable f telle que f(X), qui engendre $\mathcal{C}$ donc ne prend sur I que deux valeurs distinctes, rende $||Y - E^{f(X)}(Y)||$ minimum, donc rende le rapport de corrélation de Y et f(X) maximum (cf. [9] chap. 6).

La tribu $\mathcal{C}$ conduit à une partition $\{I_1, I_2\}$ de I, et on retient finalement la partition et la v.a. pour laquelle $||Y - E^{\mathcal{C}}(Y)||$ est le plus petit. Par itérations successives, on obtient alors l'arbre dichotomique considéré en 1.2..A chacune de ses branches est associée une suite croissante de sous-tribus de $\mathcal{P}(I)$. Le choix a priori d'une règle d'arrêt (qui, mathématiquement, permet de définir un temps d'arrêt-cf. [14] chap. II) permet de cesser, sur chaque branche, les itérations. Cette remarque, qui apparaît naturellement ici, n'est pas propre à ce cas particulier, mais demeure valable quelle que soit la nature de Y.

Remarque : On pourrait chercher la fonction réelle mesurable de tous les X_j rendant $||Y - E^{f(X_1, \ldots, X_p)}||$ minimum, avec ou sans contraintes du type BERNOULLI, ou autres.

Ce type de problème, de portée beaucoup plus générale, ne sera pas abordé ici.

b) *Explicitation et équivalence à la méthode A.I.D. :* Si $\mathcal{C}$ est la sous-tribu de BERNOULLI de $\mathcal{D}$ engendrée par la partition $\{I_1, I_2\}$ de I associée à la partition $\{K_1, K_2\}$ de K, la v.a. $\mathcal{C}$-mesurable la plus proche de Y est :

$$E^{\mathcal{C}}(Y) = \bar{Y}_{K_1} \sum_{k \in K_1} X^k + \bar{Y}_{K_2} \sum_{k \in K_2} X^k$$

(avec les notations introduites en 2.1. b)).

D'après a), on cherche la sous-tribu $\mathcal{C}$, ou, ce qui revient au même, K_1 et K_2, rendant $||Y - E^{\mathcal{C}}(Y)||$ minimum.

Posant $W = Y - E(Y)$, on a : $E^{\mathcal{C}}(W) = E^{\mathcal{C}}(Y) - E(Y)$, et par conséquent :

$$Y - E^{\mathcal{C}}(Y) = W - E^{\mathcal{C}}(W)$$

Comme d'autre part $||W - E^{\mathcal{C}}(W)||^2 + ||E^{\mathcal{C}}(W)||^2 = ||W||^2$, il est donc équivalent de minimiser $||Y - E^{\mathcal{C}}(Y)||$ ou de maximiser $||E^{\mathcal{C}}(W)||^2$. $E^{\mathcal{C}}(W)$ s'écrivant $\overline{W}_{K_1} \sum_{k \in K_1} X^k + \overline{W}_{k_2} \sum_{k \in K_2} X^k$ et étant centrée, puisque W l'est, on peut écrire : $\overline{W}_{K_2} = - \overline{W}_{K_1} \frac{P_1}{P_2}$, d'où l'on déduit :

$$||E^{\mathcal{C}}(W)||^2 = \overline{W}^2_{K_1} \times P_1 + \overline{W}^2_{K_2} \times P_2 = \overline{W}^2_{K_1} \times \frac{P_1}{P_2} = - \overline{W}_{K_1} \cdot \overline{W}_{K_2} .$$

Si on considère, sur l'axe réel, les $\overline{W}_k$, moyennes pondérées de W sur G_k, on sait qu'alors la dichotomie optimale est de la forme :

$$K_1 = \{k \, ; \, \overline{W}_k < C\}, \, K_2 = \{k \, ; \, \overline{W}_k \geqslant C\} \, , \, C \in R \, .$$

On cherche donc C (d'où a priori $m - 1$ dichotomies possibles) associé à la partition $\{K_1 , K_2\}$ qui maximise $- \overline{W}_{K_2} \cdot \overline{W}_{K_2}$, ou encore qui maximise $OA_1 \cdot OA_2$, où A_1 (resp. A_2) est le barycentre de $\{\overline{W}_k \, ; \, k \in K_1\})$, (resp. de $\{\overline{W}_k \, ; \, k \in K_2\}$) d'où la méthode pratique de dichotomisation.

Il est clair que la dichotomie obtenue est invariante par translation ou homothétie sur Y.

La v.a. $E^{\mathcal{C}}(Y)$ obtenue minimise $||Y - E^{\mathcal{C}}(Y)||^2$. Or, en explicitant :

$$||Y - E^{\mathcal{C}}(Y)||^2 = \sum_{i \in I_1} p_i (Y(i) - \overline{Y}_{K_1})^2 + \sum_{i \in I_2} p_i (Y(i) - \overline{Y}_{K_2})^2$$

C'est là la somme des indices intragroupes de I_1 et I_2 (cf. 2.1.b)), et l'on retrouve donc, sous une nouvelle présentation, la méthode A.I.D..

4.3. La variable à expliquer Y est quantitative multidimensionnelle :

A chaque étape, on considère une v.a. qualitative X à m modalités, qui engendre la sous-tribu $\mathcal{B}$ de $\mathcal{P}(I)$, et la v.a. Y à valeurs dans $(R^q , \mathcal{B}_{R^q})$, où $\mathcal{B}_{R^q}$ désigne la tribu des boréliens de R^q. Donc $Y = (Y_1 , \ldots , Y_q)$, et pour tout $\ell \in L = \{1 , 2 , \ldots , q\}$, Y_ℓ appartient à $L^2(I , \mathcal{P}(I) , P)$. On suppose également que chacune des Y_ℓ est centrée et normée. On verra par la suite (cf. 4.3. B-a)) que cette hypothèse n'enlève rien de sa généralité à cette étude, la segmentation étant invariante par cette transformation.

La méthode générale exposée en III. nous conduit à considérer le sous-espace F_Y de L^2 engendré par les v.a. $Y_1 , \ldots , Y_q$. Ce sous-espace est de dimension q, $Y_1 , \ldots , Y_q$ étant linéairement indépendantes (on aura éventuellement supprimé certaines v.a.). On cherche alors la v.a. de BERNOULLI X', engendrant la sous-tribu de BERNOULLI $\mathcal{C}$ de $\mathcal{B}$, qui soit normée et centrée, et qui soit la plus proche de F_Y. On va étudier tout d'abord un cas particulier, auquel on se ramènera dans le cas général.

A $-\{Y_\ell\}_{\ell \in L}$ *forme une base orthonormée de* F_Y :

Soit Y' la projection orthogonale de X' sur F_Y. Y' est la somme des projections orthogonales de X' sur chacun des supports de Y_ℓ, donc :

$$Y' = \sum_{\ell=1}^{q} < X' , Y_\ell > Y_\ell .$$

Or, la projection orthogonale de Y_ℓ sur le support de X' étant $E^{\mathcal{C}}(Y_\ell)$, il vient :

$$Y' = \sum_{\ell=1}^{q} ||E^{\mathcal{C}}(Y_\ell)|| \, Y_\ell .$$

On cherche X', donc $\mathcal{C}$, rendant $||X' - Y'||$ minimum. Or :

$$||X' - Y'||^2 + ||Y'||^2 = ||X'||^2 = 1 .$$

Cela revient donc à chercher $\mathcal{C}$, sous-tribu de BERNOULLI de $\mathcal{B}$, qui maximise :

$$||Y'||^2 = \sum_{\ell=1}^{q} ||E^{\mathcal{C}}(Y_\ell)||^2 .$$

a) technique de recherche de $\mathcal{C}$:

On a vu en 4.2. que Y_ℓ étant centrée,

$$||E^{\mathcal{C}}(Y_\ell)||^2 = - \bar{Y}^\ell_{K_1} \cdot \bar{Y}^\ell_{K_2} ,$$

où $\bar{Y}^\ell_{K_1}$ (resp. $\bar{Y}^\ell_{K_2}$) désigne la moyenne pondérée de Y_ℓ sur I_1 (resp. $I_2 = I - I_1$), partie de I qui engendre $\mathcal{C}$. D'où :

$$||Y'||^2 = - \sum_{\ell=1}^{q} \bar{Y}^\ell_{K_1} \bar{Y}^\ell_{K_2} .$$

La recherche de $\mathcal{C}$ se ramène donc à la recherche de la dichotomie $\{K_1 , K_2\}$ de K qui minimise :

$$S = \sum_{\ell=1}^{q} \bar{Y}^\ell_{K_1} \bar{Y}^\ell_{K_2} .$$

Comme en 4.2., on calculera S pour chaque dichotomie $\{K_1 , K_2\}$ séparant en deux sous-ensembles de $H = R^q$ le nuage des centres de gravité des éléments $Y(G_k)$ de la partition engendrée par X sur Y (I).

b) *équivalence à la généralisation de la méthode A.I.D.* (cf. 2.1.c)) :

D'après 4.2., il vient :

$$||E^{\mathcal{C}}(Y_\ell)||^2 = 1 - D_\ell(I_1) - D_\ell(I_2)$$

où $D_\ell(I_1)$ (resp. $D_\ell(I_2)$) est l'indice intragroupe de la projection sur Y_ℓ du nuage $Y(I_1)$ (resp. $Y(I_2)$) de points de H (H, identifiable au dual de F_Y, est appelé généralement "espace des individus" (cf. [4] tome 2 chap. 7)).

On cherche $\mathcal{C}$, donc I_1 et I_2, qui minimise :

$$D = \sum_{\ell=1}^{q} [D_\ell(I_1) + D_\ell(I_2)] ,$$

soit encore :

$$D = \sum_{\ell=1}^{q} \left[\sum_{i\in I_1} p_i[Y_\ell(i) - \bar{Y}^\ell_{K_1}]^2 + \sum_{i\in I_2} p_i[Y_\ell(i) - \bar{Y}^\ell_{K_2}]^2 \right] ,$$

ou :

$$D = \sum_{i\in I_1} p_i \Big(\sum_{\ell=1}^{q} [Y_\ell(i) - \bar{Y}^\ell_{K_1}]^2 \Big) + \sum_{i\in I_2} p_i \Big(\sum_{\ell=1}^{q} [Y_\ell(i) - \bar{Y}^\ell_{K_2}]^2 \Big).$$

Comme H est muni dans ce cas de la métrique euclidienne de matrice unité (de façon générale, de la métrique de MAHALANOBIS ; cf. 2.1. c)), il vient :

$$D = \sum_{i\in I_1} p_i \|Y(i) - \bar{Y}_{K_1}\|^2 + \sum_{i\in I_2} p_i \|Y(i) - \bar{Y}_{K_2}\|^2 = D(I_1) + D(I_2) .$$

On cherche donc $\mathcal{C}$, ou encore $\{I_1 , I_2\}$, qui minimise $D(I_1) + D(I_2)$, somme des indices intragroupes de I_1 et I_2. On aboutit donc à la même dichotomie que la généralisation de la méthode A.I.D. au cas multidimensionnel.

c) présentation équivalente de la segmentation :

On a vu au 4.2. que, pour Y réelle, la segmentation est, à chaque étape, la recherche de la sous-tribu de BERNOULLI $\mathcal{C}$ de $\mathcal{B}$ rendant $\|Y - E^{\mathcal{C}}(Y)\|$ minimum. On peut reprendre cette définition à condition, Y appartenant ici à $(L^2)^q$, de définir une métrique sur $(L^2)^q$.

Dans le cas le plus général, on définira la métrique de $(L^2)^q$ associée à une métrique de matrice A de l'espace H, en posant pour tout couple (U , V) de v.a. centrées de $(L^2)^q$:

$$<U , V>_{(L^2)^q , A} = \sum_{i=1}^{n} p_i < U(i) , V(i) >_A = E[<U , V>_A]$$

d'où :

$$\|U\|^2_{(L^2)^q , A} = \sum_{i=1}^{n} p_i \|U(i)\|^2_A = E[\|U\|^2_A] .$$

Comme ici H est muni de la métrique euclidienne de matrice unité, il vient :

$$\|Y - E^{\mathcal{C}}(Y)\|^2_{(L^2)^q , I_q} = E\left[\sum_{\ell=1}^{q} [Y_\ell - E^{\mathcal{C}}(Y_\ell)]^2 \right]$$

$$= \sum_{\ell=1}^{q} E([Y_\ell - E^{\mathcal{C}}(Y_\ell)]^2) = \sum_{\ell=1}^{q} ||Y_\ell - E^{\mathcal{C}}(Y_\ell)||^2$$

et, comme

$$||Y_\ell - E^{\mathcal{C}}(Y_\ell)||^2 + ||E^{\mathcal{C}}(Y_\ell)||^2 = ||Y_\ell||^2 = 1 ,$$

il vient :

$$||Y - E^{\mathcal{C}}(Y)||^2_{(L^2)^q, I_q} = q - \sum_{\ell=1}^{q} ||E^{\mathcal{C}}(Y_\ell)||^2 .$$

Donc chercher $\mathcal{C}$ qui, d'après a), maximise $\sum_{\ell=1}^{q} ||E(Y_\ell)||^2$ est équivalent à chercher $\mathcal{C}$, sous-tribu de BERNOULLI de $\mathcal{B}$, qui minimise $||Y - E^{\mathcal{C}}(Y)||_{(L^2)^q, I_q}$.

Remarque : $E^{\mathcal{C}}(Y)$ désigne ici la v.a. à valeurs dans R^q de composantes

$$(E^{\mathcal{C}}(Y_1), \ldots, E^{\mathcal{C}}(Y_q))$$

généralisation à $(L^2)^q$ de l'espérance conditionnelle de v.a. réelle. (cf. [16]).

$B - \{Y_\ell\}_{\ell \in L}$ forme une base non nécessairement orthonormée de F_Y :

Soit $\{Z_\ell\}_{\ell \in L}$ une base orthonormée de F_Y, et M la matrice de passage de $\{Y_\ell\}$ à $\{Z_\ell\}$. En confondant Y et Z avec les matrices colonnes de leurs composantes, on peut écrire $Z = MY$.

a) Le sous-espace F_Y engendré par $\{Y_\ell\}_{\ell \in L}$ étant confondu avec celui engendré par $\{Z_\ell\}_{\ell \in L}$, la segmentation faite en prenant Z au lieu de Y comme critère conduit au même résultat, puisque dans chaque cas on cherche X', v.a. de BERNOULLI de F_X la plus proche du même F_Y. De façon générale, la segmentation est invariante par transformation linéaire, bijective de la variable Y, ainsi d'ailleurs que par translation, l'espace F_Y étant engendré par les v.a. $Y_\ell - E(Y_\ell)$, $\ell \in L$.

D'après A, on cherche donc $\mathcal{C}$, sous-tribu de BERNOULLI de $\mathcal{B}$, qui maximise $\sum_{\ell=1}^{q} ||E^{\mathcal{C}}(Z_\ell)||^2$.

b) La généralisation de la méthode A.I.D. représente les individus par les points Y(i) dans H, qui est alors muni de la métrique de MAHALANOBIS de matrice Λ^{-1}, où Λ est la matrice des covariances de Y. Cela revient à munir l'espace vectoriel des v.a., qui est le dual du précédent, isomorphe à R^q et rapporté à la base $\{Y_\ell\}_{\ell \in L}$, de la métrique de matrice Λ.

Soient U et V deux v.a. de cet espace vectoriel, représentées par les matrices colonnes R et T dans la base $\{Y_\ell\}_{\ell \in L}$ et R' et T' dans la base $\{Z_\ell\}_{\ell \in L}$.

Alors : $$R = {}^tM\,R' , T = {}^tM\,T'$$

et : $$<U , V> = {}^tR.\Lambda.T = {}^tR'\,M\,\Lambda\,{}^tM\,T'.$$

Donc la métrique considérée a pour matrice, dans la base $\{Z_\ell\}_{\ell \in L}$, $M\Lambda {}^tM$, qui est la matrice des covariances de Z, c'est-à-dire I_q, matrice unité. Dans la base $\{Z_\ell\}_{\ell \in L}$, c'est donc la métrique euclidienne de matrice unité.

D'après A, on a donc :

$$\sum_{\ell=1}^{q} ||E^{\mathcal{C}}(Z_\ell)||^2 = D(I_1) + D(I_2) .$$

On retrouve donc encore la généralisation de la méthode A.I.D.. On établit de plus son invariance par transformations affines bijectives du critère Y.

c) La norme de $(L^2)^q$ associée à la métrique de matrice Λ^{-1} de H est définie, si N est la matrice colonne représentant la v.a. W, par :

$$||W||^2_{(L^2)^q} = E\,[{}^tN\,\Lambda^{-1}\,N] .$$

D'où, en confondant encore les v.a. avec les matrices colonnes de leurs composantes aléatoires :

$$||Y - E^{\mathcal{C}}(Y)||^2_{(L^2)^q,\,\Lambda^{-1}} = E\,[{}^t[Y - E^{\mathcal{C}}(Y)]\,\Lambda^{-1}\,[Y - E^{\mathcal{C}}(Y)]] ;$$

or : $$Z = MY ,$$

d'où : $$E^{\mathcal{C}}(Z) = M\,E^{\mathcal{C}}(Y) ,$$

et donc

$$||Y - E^{\mathcal{C}}(Y)||^2_{(L^2)^q,\,\Lambda^{-1}} = E\,[{}^t[Z - E^{\mathcal{C}}(Z)]\,{}^tM^{-1}\,\Lambda^{-1}\,M^{-1}\,[Z - E^{\mathcal{C}}(Z)]] ,$$

et comme $M\,\Lambda\,{}^tM = I_q$,

$$||Y - E^{\mathcal{C}}(Y)||^2_{(L^2)^q,\,\Lambda^{-1}} = E\,[{}^t(Z - E^{\mathcal{C}}(Z))\,(Z - E^{\mathcal{C}}(Z))] = ||Z - E^{\mathcal{C}}(Z)||^2_{(L^2)^q,\,I_q} .$$

Donc, d'après A, chercher $\mathcal{C}$ qui maximise $\sum_{\ell=1}^{q} ||E^{\mathcal{C}}(Z_\ell)||^2$ est équivalent à chercher $\mathcal{C}$ qui minimise

$$||Z - E^{\mathcal{C}}(Z)||^2_{(L^2)^q,\,I_q} , \quad \text{donc} \quad ||Y - E^{\mathcal{C}}(Y)||_{(L^2)^q,\,\Lambda^{-1}} .$$

D'où le résultat :

Dans le cas le plus général, la segmentation est donc la recherche de la sous-tribu de BERNOULLI $\mathcal{C}$ de $\mathcal{B}$ qui minimise $||Y - E^{\mathcal{C}}(Y)||_{\Lambda^{-1}}$ dans $(L^2)^q$.

Remarque 1 : Dans le cas où Y est qualitative, on a remarqué que la segmentation ne dépend que de la tribu $\mathcal{A}$ engendrée par Y. $\{Y^\ell\}_{\ell \in L}$ étant l'ensemble des indicatrices associées à cette tribu (indicatrices introduites au III), on considère la v.a. $\mathcal{Y}$ de $(L^2)^q$ de composantes $(Y^1, \ldots, Y^q)$. Comme F_Y, défini au III, est le sous-espace vectoriel des v.a. centrées engendré par $(Y^1, \ldots, Y^q)$, on voit que la segmentation obtenue en prenant pour critère Y est analogue à celle obtenue en prenant pour critère $\mathcal{Y}$. Elle s'obtient, à chaque étape, en cherchant la sous-tribu de BERNOULLI $\mathcal{C}$ de $\mathcal{B}$ qui minimise $||\mathcal{Y}_1 - E^{\mathcal{C}}(\mathcal{Y}_1)||$ dans $(L^2)^{q-1}$, où $\mathcal{Y}_1$ est obtenue à partir de $\mathcal{Y}$ en supprimant Y^q par exemple (pour obtenir des variables linéairement indépendantes) et en centrant et normant

les autres Y^ℓ ($\ell = 1, \ldots, q-1$). On peut donc ainsi généraliser au cas d'un critère qualitatif la définition donnée en c) ci-dessus de la segmentation.

Remarque 2 : Cette définition fait clairement apparaître la différence entre segmentation et régression : la régression est la recherche, une sous-tribu $\mathcal{C}$ étant donnée (tribu engendrée par une v.a. X en général), de la v.a. de $L^2(\mathcal{C})$ la plus proche d'une v.a. donnée Y, ce qui conduit à $E^{\mathcal{C}}(Y)$. Lorsqu'on ne considère que la régression linéaire (par exemple en analyse des données), $L^2(\mathcal{C})$ est remplacé par un autre sous-espace vectoriel (le sous-espace engendré par les composantes de X), mais le problème reste le même. En segmentation, par contre, c'est la tribu $\mathcal{C}$ qu'on cherche, qui minimise $||Y - E^{\mathcal{C}}(Y)||$ sous certaines contraintes.

4.4. Remarques sur la méthode S.P.A.D. :

Cette méthode ne traite que le cas particulier où Y est une variable statistique qualitative, avec $q = 2$. Soient Y^1 et Y^2 les deux indicatrices obtenues, et $n_{k\ell}$ ($\ell = 1, 2$; $k = 1, \ldots, m$) le nombre d'individus, à chaque étape, présentant la k-ième modalité de X de la ℓ-ième modalité de Y. On peut prendre comme variable génératrice de F_Y : $U = n_{.2}\, Y^1 - n_{.1}\, Y^2$, d'où :

$$\bar{U}_k = \frac{n_{.1}\, n_{.2}}{n_{k.}} \left(\frac{n_{k1}}{n_{.1}} - \frac{n_{k2}}{n_{.2}} \right) .$$

La v.a. X′ de BERNOULLI de F_X qui conduit à la dichotomie optimale pour la méthode présentée ci-dessus est obtenue en ordonnant les $\bar{U}_k$ sur la droite réelle et en les séparant de façon appropriée (cf. 4.2. b)). Comme U est centrée, le barycentre des $\bar{U}_k$ est l'origine. Par ailleurs, d'après l'expression des $\bar{U}_k$, la méthode S.P.A.D. revient à les séparer en le groupe des $\bar{U}_k$ positifs et celui des $\bar{U}_k$ négatifs (cf. 2.2.). Or, comme nous allons le voir sur un exemple, ce n'est pas là en général la dichotomie à laquelle conduit la méthode proposée ici (donc à laquelle conduit la méthode E.L.I.S.E.E.).

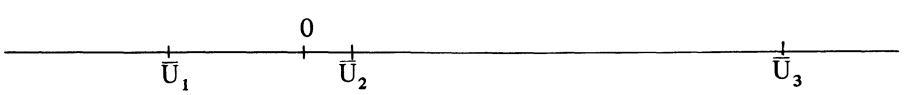
Exemple : Soit X à 3 modalités, et le tableau d'effectifs suivant :

Y \ X	1	2	3	Total
1	245	55	200	500
2	455	45	0	500
Total	700	100	200	1 000

Nous obtenons, à une homothétie près de rapport $\frac{1}{50}$:

$$\bar{U}_1 = -3, \quad \bar{U}_2 = 1, \quad \bar{U}_3 = 10 ;$$

d'où le schéma :



La méthode S.P.A.D. conduit donc à la partition : modalité 1 de X d'une part, modalités 2 et 3 d'autre part.

Or, si on revient au critère défini en 4.2. b), il vient ici pour la dichotomie proposée par la méthode S.P.A.D. :

$$\overline{OA}_1 = -3\,,\ \overline{OA}_2 = 7\ ,\ \text{d'où}\quad -\overline{OA}_1 \,.\, \overline{OA}_2 = 21$$

Pour l'autre dichotomie possible, on obtient :

$$\overline{OA}'_1 = -\frac{20}{8}\,,\ \overline{OA}'_2 = 10\ ,\ \text{d'où}\quad -\overline{OA}'_1 \,.\, \overline{OA}'_2 = 25$$

C'est donc la seconde dichotomie qui est optimale, regroupant les modalités 1 et 2 de X d'une part, la modalité 3 de l'autre.

Du point de vue de la segmentation, les modalités 1 et 2 de X ont, relativement à Y, un comportement plus proche que 2 et 3, ce qui n'est plus le cas dans une optique purement prédictive.

Pour revenir au cas général, on peut remarquer que la méthode proposée dans cette étude regroupe les modalités de X les plus proches au sens de la métrique du χ^2 (la métrique sur L^2 correspondant ici, comme en analyse des correspondances, à la métrique du χ^2 -cf. [4] tome 3, chap. 14).

La méthode S.P.A.D. est donc optimale comme méthode de prévision (au sens BAYESIEN – cf. [19] Chap. A2), mais non comme méthode de segmentation.

Cela rejoint diverses remarques de BELSON. (cf. [1]).

V – APPROXIMATIONS POSSIBLES DE LA METHODE PROPOSEE .

Ces approximations ont pour but de fournir une simplification de la méthode exposée aux paragraphes III et IV, essentiellement pour les cas où des données trop importantes ne permettraient pas de traiter cette méthode avec des moyens limités en calcul automatique, encore que son traitement numérique ne présente généralement pas de difficulté sur les ordinateurs actuels.

Le principe de ces approximations est de remplacer la variable à expliquer Y par un "résumé optimal" en fonction de X. Elles peuvent ainsi permettre de traiter le cas où Y est quantitative multidimensionnelle en ne disposant, par exemple, que des programmes traitant le cas d'une variable à expliquer unidimensionnelle.

5.1. Première approximation :

Elle se fait en deux étapes :

a) Y étant à valeurs dans R^q (cas le plus général), pour chaque variable explicative X, on cherche le couple (X^*, Y^*) de $F_X \times F_Y$, tel que :

Y^* soit à distance minimum de F_X, avec $||Y^*|| = 1$;

. X^* soit la projection orthogonale de Y^* sur F_X.

(X^* , Y^*) apparaît ainsi comme le premier couple de variables canoniques de $F_X \times F_Y$, à cela près de X^* n'est pas normée.

b) Pour chaque v.a. X_j , $j \in \{1, 2, \ldots, p\}$, on cherche la v.a. de BERNOULLI normée X'_j de F_X la plus proche de Y_j^*, ou ce qui revient au même de X_j^* (il faut en effet remarquer que Y^* varie avec j). Ainsi, on associe encore à toute v.a. X_j une v.a. de BERNOULLI X'_j dont on connaît la distance à Y_j^*, et l'on conserve celle des X'_j minimisant cette distance, ainsi que la dichotomie de I qui lui est associée.

La partition de I obtenue de cette façon ne sera en général qu'une approximation de la partition optimale, mais elle s'obtiendra de façon plus simple en utilisant à chaque étape du processus d'abord le premier niveau (puisque seul intervient ici le premier couple de variables) d'une procédure de type analyse canonique (il s'agira en fait d'une analyse factorielle discriminante si Y est quantitative, et d'une analyse des correspondances si elle est qualitative, (cf. [4], tome 3, chap. 13 et 14)), et en utilisant ensuite la méthode de segmentation dans le cas le plus simple : celui d'une variable à expliquer unidimensionnelle.

5.2. Approximations plus fines :

Le principe est le même qu'en 5.1., mais pour chaque v.a. X, on considère comme approximation de la v.a. Y les ν premières variables canoniques de F_Y , Y_1^* , Y_2^* , . . . , Y_ν^* (ν étant choisi en fonction de l'"information" que l'on désire conserver), et non plus la première seulement. Appelant alors F_{Y*} le sous-espace vectoriel de R^n engendré par ces variables, on est ramené à chaque étape à la méthode générale en remplaçant F_Y par F_{Y*}.

5.3. Précisions sur ces approximations :

En fonction de la nature de la variable à expliquer Y, on peut préciser les propriétés des approximations présentées ci-dessus.

a) Y est qualitative à deux modalités, ou quantitative unidimensionnelle : il est clair que l'approximation présentée en 5.1. et la méthode générale sont identiques chaque fois que F_Y est de dimension 1. C'est évidemment le cas si Y est quantitative unidimensionnelle, et ça l'est aussi si Y est qualitative à deux modalités. En effet, F_{Y0} est alors engendré par Y^1 et Y^2, indicatrices telles que $Y^1 + Y^2 = e$, et donc F_Y sous-espace vectoriel de F_{Y0} orthogonal à e est de dimension 1.

b) Y est qualitative à q modalités ($q > 2$) : au lieu d'appliquer la méthode générale à F_X et F_Y, l'approximation présentée en 5.2. conduit à considérer F_X et F_{Y*}, où F_{Y*} est le sous-espace vectoriel de F_Y engendré par les ν premières variables canoniques obtenues par analyse canonique classique de X et Y. Si P est uniforme sur $I\left(p_i = \frac{1}{n}, \forall\, i \in I\right)$, étant donnée la nature des v.a. intervenant, cette analyse est une analyse des correspondances. Or on peut remarquer qu'elle est équivalente à l'analyse en composantes principales de $E^X(Y)$ (cf. [15]), qui est ici représentée par la matrice ayant pour colonnes les centres de gravité g_k des sous-ensembles $Y(G_k)$ de H.

La méthode générale revient à utiliser le nuage complet des centres de gravité, alors que l'approximation utilise un "résumé optimal" de ce nuage (sa projection sur le sous-espace vectoriel de H engendré par les ν premiers axes principaux). On obtient donc un résultat moins bon en général (si $\nu < q$) mais nécessitant un calcul plus simple. On peut de plus dans chaque cas évaluer (grâce aux valeurs propres obtenues dans l'analyse des correspondances) la perte d'information due au "résumé" utilisé.

c) Y est quantitative multidimensionnelle : comme dans le cas précédent, on considère F_X et F_Y^*.

$Y_1^*, Y_2^*, \ldots, Y_\nu^*$ sont ici obtenues par l'analyse factorielle discriminante de X et Y, correspondant dans ce cas à l'analyse canonique. Cela revient encore à l'analyse en composantes principales de $E^X(Y)$, toujours représentée par la matrice ayant pour vecteurs colonnes les centres de gravité g_k. On peut donc comme ci-dessus évaluer la perte d'"information" due au "résumé" utilisé.

CONCLUSION

En se plaçant dans l'espace $L^2(I, \mathcal{P}(I), P)$, on a donc obtenu deux présentations équivalentes opérant une synthèse de la segmentation, soit à partir de la distance d'une v.a. de BERNOULLI de L^2 au sous-espace engendré par la variable à expliquer, soit en cherchant une tribu de BERNOULLI $\mathcal{C}$ telle que le sous-espace $L^2(I, \mathcal{C}, P)$ soit le plus proche de la variable à expliquer.

Ce point de vue semble permettre, de plus, d'aborder efficacement des domaines dépassant le cadre de la segmentation, en abandonnant les dichotomies successives, ou en utilisant conjointement les variables explicatives, en profitant aussi des possibilités de simplification offertes par les approximations possibles ; ce pourraît être le cas en particulier pour la typologie, la classification . . . et même pour un certain nombre de problèmes qu'aucune méthode courante d'analyse des données ne permet de traiter de façon satisfaisante.

BIBLIOGRAPHIE

[0] BACCINI A. et POUSSE A. – Point de vue unitaire de la segmentation. Quelques conséquences. *C.R.A.S.* Tome 280 série A p 241 (Janvier 1975).

[00] BACCINI A. Aspect synthétique de la segmentation et traitement de variables qualitatives à modalités ordonnées. Thèse de 3ème cycle. Laboratoire de Statistique. Université Paul Sabatier Toulouse (Mars 1975).

[1] BELSON W.A. Matching and prediction on the principle of biological classification. *Applied Statistics* tome VIII p. 65-75 (1959).

[2] BENZECRI J.P. – L'analyse des données. Dunod (1973).

[3] BOUROCHE J.M. et TENENHAUS M. – Quelques méthodes de segmentation. *R.I.R.O.* Juin 1970.

[4] C.E.E.E. – Analyse des données multidimensionnelles, tomes 2 et 3 (1972).

[5] CELLARD J.C., LABBE B. et SAVITSKY G. – Le programme E.L.I.S.E.E., présentation et applications. *METRA* vol. 6 N° 3 sept. 1967.

[6] CRAMER H. – Mathematical methods of Statistics. Princeton (1945)

[7] DEMPSTER A.P. – Elements of continuous multivariate analysis. Addison-Wesley Publishing Co. (1969).

[8] EDWARDS A.W.F. and CAVALLI-SFORZA L.L. – A method for cluster analysis. *Biometrics* vol. 21 N° 2 (Juin 1965).

[9] FOURGEAUD C. et FUCHS A. – Statistique.Dunod (1967)

[10] FREEDMAN H.P. and RUBIN J. – On some invariant criteria for grouping data. *J.A.S.A.* p. 1159-78 (déc. 1967)

[11] HUGUES M. – Segmentation et typologie : deux techniques du marketing moderne. Bordas (1970).

[12] LANCASTER H.O. – The Chi-Squared distribution. Wiley (1969)

[13] MORGAN J.N. and SONQUIST J.A. Problems ins the analysis of survey data, a proposal – *J.A.S.A.* Vol. 58 N° 302, June 1963.

[14] NEVEU J. Martingales à temps discret – Masson (1972)

[15] POUSSE A. Sur l'analyse canonique considérée comme une analyse en composantes principales. *C.R.A.S.* Paris Janvier 1973

[16] POUSSE A. – Analyse canonique et analyse en composantes principales. Publications du Laboratoire de Statistique Université Paul Sabatier Toulouse N° 01-1973.

[17] VO KHAC Kh. et NGHIEM Ph. T. – Etude sur les aspects théoriques et pratiques de la segmentation aux moindres carrés. *R.I.R.O.* N° 8 (1968)

[18] WILLIAMS W.T. and LAMBERT J.M. – Multivariate methods in plant ecology. *Journal of Ecology*, N° 47 p. 830101 (1959).

[19] PROGRAMMES de structuration des données – Annexe II : segmentation et typologie *S.E.M.A.* (Septembre 1970)