STATISTIQUE ET ANALYSE DES DONNÉES

JACQUES DAUXOIS
ALAIN POUSSE

Essai de synthèse sur quelques problèmes mathématiques de l'analyse des données

Statistique et analyse des données, tome 1, n° 1 (1976), p. 47-67
<http://www.numdam.org/item?id=SAD_1976__1_1_47_0>

© Association pour la statistique et ses utilisations, 1976, tous droits réservés.

L'accès aux archives de la revue « Statistique et analyse des données » implique l'accord avec les conditions générales d'utilisation (http://www.numdam.org/conditions). Toute utilisation commerciale ou impression systématique est constitutive d'une infraction pénale. Toute copie ou impression de ce fichier doit contenir la présente mention de copyright.

NUMDAM
Article numérisé dans le cadre du programme
Numérisation de documents anciens mathématiques
http://www.numdam.org/

ESSAI DE SYNTHÈSE SUR QUELQUES PROBLÈMES MATHÉMATIQUES DE L'ANALYSE DES DONNÉES

par

Jacques DAUXOIS et Alain POUSSE (*)

Ces dernières années, d'assez nombreuses recherches ont été entreprises, essentiellement en France, pour dégager les mécanismes fondamentaux des analyses dites factorielles et aborder certaines généralisations. On se propose dans cet article de préciser une démarche contribuant à cette évolution d'une branche importante de l'analyse des données et de présenter quelques-uns des résultats qui en sont issus. Le point de vue global que nous développons permet à la fois d'établir des liens entre plusieurs travaux récents, et de donner divers aperçus originaux.

(*) Laboratoire de Statistique . Université Paul Sabatier . Toulouse .

Les analyses factorielles ont été développées initialement dans un cadre probabiliste précis (Hotelling, Anderson, ...), à propos d'un nombre fini de v.a. réelles, et en ne faisant intervenir que des combinaisons linéaires de ces v.a. ; cela permet de qualifier de telles analyses de "linéaires". Sous des formes diverses (analyses en composantes principales, des correspondances, canoniques, discriminantes), elles ont été introduites en statistique descriptive comme méthode d'analyse des données ([4],[9]).

L'application de telles méthodes à des problèmes débordant leur domaine habituel d'utilisation semblerait intuitivement pouvoir être fructueuse ; toutefois elle présuppose qu'un certain nombre d'études préliminaires théoriques soient effectuées. Certaines de ces études ne sont que des généralisations plus ou moins aisées de choses connues, d'autres au contraire sont nouvelles et ne se sont développées que récemment.

Il apparaît que le choix d'un langage et d'un outil bien adaptés, pour une étude mathématique plus générale des divers problèmes que l'on peut envisager, est indispensable tant sur le plan théorique que pour les utilisateurs éventuels. Quelles que soient les possibilités de généralisation envisagées, on débouche de façon naturelle sur des problèmes d'analyse fonctionnelle ; certains s'expriment par ailleurs simplement en usant de façon plus systématique du langage probabiliste.

On peut dégager, à partir de là, un certain nombre de questions. Il n'est évidemment pas possible d'y répondre complètement ici ; on se propose simplement d'essayer de montrer, en développant sommairement certains exemples précis, dans quel sens on peut envisager des éléments de réponse, et quelle démarche a été la nôtre. Pour chaque partie, les références bibliographiques permettront au lecteur d'avoir plus de détails. Les principales de ces questions sont les suivantes :

a) La contrainte de linéarité imposée par les méthodes classiques peut dans certains cas apparaître comme un peu artificielle, et de plus génératrice d'une perte inutile d'information. Un des premiers problèmes que l'on peut envisager est donc la définition d'analyses "non-linéaires".

b) Les diverses analyses factorielles ont été, à l'origine, définies comme des méthodes "pas à pas". Ainsi présentées, elles sont soumises à des conditions assez exigeantes d'existence ; on peut, au moins pour certaines d'entre elles (analyse canonique) donner une définition générale possédant le double avantage d'assurer dans tous les cas l'existence de l'analyse, et de fournir un moyen "global" de l'obtenir.

c) Pour étudier de nombreux phénomènes aléatoires, on est souvent amené à chercher une description de ceux-ci. Or une telle étude directe est souvent mal commode, voire même la plupart du temps impossible. Il est , en effet , évident qu'une étude quelconque ne pourra se faire généralement qu'au travers d'un échantillon ; les analyses factorielles qu'il peut fournir permettent d'obtenir une représentation ou une visualisation qui ne sont qu'une description des données, et non du phénomène étudié. Une telle représentation est néanmoins couramment considérée dans les interprétations comme une description "approchée" du phénomène. Il est donc nécessaire de pouvoir formuler cette notion d'approximation en termes mathématiques. Or cela suppose que l'on soit capable d'une part de définir et d'étudier une analyse factorielle directe, d'autre part de la comparer à celle obtenue à partir d'un échantillon de taille n. Pour ce faire, on est conduit à la définition des analyses

factorielles dans un cadre aussi large que possible, et en particulier à envisager des analyses factorielles linéaires ou non-linéaires de fonctions aléatoires mesurables à valeurs dans un espace de Hilbert réel séparable. Pour un type d'analyse factorielle donné, définir des modes de convergence permet alors d'étudier l'évolution, suivant la taille de l'échantillon, des représentations obtenues au moyen des données, donc de préciser la notion de "représentation approchée". On peut également s'intéresser à la taille d'un échantillon nécessaire pour une bonne représentation. On est donc conduit à un double problème : de définition et de convergence.

d) La définition d'analyses non linéaires amène à l'analyse spectrale d'un opérateur intégral dont les données fournissent une estimation du noyau. Le traitement numérique est possible par discrétisation des variables qui interviennent, puisqu'on se ramène ainsi à du calcul matriciel et à une analyse des correspondances (éventuellement généralisée). Pour pouvoir apprécier la qualité de l'analyse obtenue (par rapport à l'analyse non linéaire cherchée), il faut étudier : sur le plan statistique, l'effet de la discrétisation et celui de l'échantillonnage (d'où encore des problèmes de convergence) ; sur le plan numérique, l'effet des approximations nécessaires au calcul (autre problème de convergence, qui intervient constamment - et notamment en c) - , mais qui, n'étant pas spécifique à une étude statistique, restera en marge de l'étude présente (on peut voir [6] à ce sujet, à propos d'analyse en composantes principales)).

e) Une autre question concerne plus particulièrement la notion d'analyse canonique. Celle-ci se définit facilement pour deux ensembles finis de v.a.r. ; elle possède de plus des propriétés très remarquables.

Il est évidemment intéressant de chercher à étendre la notion d'A.C. à k (k > 2) ensembles finis de v.a.r. . Malheureusement on s'aperçoit très vite qu'il n'est pas possible d'espérer une généralisation (linéaire ou non linéaire) ; de nombreuses extensions sont en effet envisageables, aucune d'entre elles ne pouvant être, a priori, considérée comme meilleure que les autres. Le cas concret traité peut seul permettre de juger de l'intéret de telle ou telle méthode. Toutefois certaines des généralisations semblent avoir de "bonnes" propriétés, en liaison avec l'analyse en composantes principales.

1 - ANALYSES LINEAIRES ET NON-LINEAIRES : On expose ici le cas, typique, de l'analyse canonique.

1.1. Variables statistiques et langage probabiliste :

Etant donné un couple de variables statistiques $X = (X_1,\ldots,X_p)$ et $Y = (Y_1,\ldots,Y_q)$ à valeurs dans $\mathbb{R}^p$ et $\mathbb{R}^q$ respectivement, et leurs observations sur n individus dont chacun est muni d'un "poids" $p_i > 0$ ($\sum_1^n p_i = 1$), on peut identifier l'espace $\mathbb{R}^n$ muni de la métrique des poids de matrice $\mathrm{diag}(p_i)$ à l'espace $L^2(\Omega, \mathcal{A}, P)$ où Ω est l'ensemble fini $\{1,2,\ldots,n\}$ ($n \in \mathbb{N}^*$), $\mathcal{A}$ est la tribu des parties de Ω et P est la probabilité sur $\{\Omega, \mathcal{A}\}$ telle que : $\forall i \in I, P(\{i\}) = p_i$. L'analyse canonique de X et Y est fondée sur la comparaison de F_X et F_Y , ensembles des combinaisons linéaires des $\{X_i\}_{i=1\ldots p}$ et $\{Y_j\}_{j=1\ldots q}$ respectivement. Exposée en ces termes, l'analyse canonique de variables statistiques est analogue à celle précisée ci-dessous pour les variables aléatoires. De plus, ce langage probabiliste se prête bien à une synthèse, une clarification et à partir de là à l'élaboration de techniques nouvelles dans de nombreux secteurs de l'analyse des données. (cf. par exemple, en ce qui

concerne la segmentation, [3] , [2] et [11.2] , et en ce qui concerne les analyses factorielles classiques [11.1].)

1.2. Analyse canonique linéaire de deux v.a. :

Soit $X = \{X_i\}_{i=1,2...p}$ et $Y = \{Y_j\}_{j=1,2...q}$ deux v.a. définies sur l'espace probabilisé $(\Omega, \mathcal{A}, P)$ à valeurs respectivement dans $\mathbb{R}^p$ et $\mathbb{R}^q$ munis de leurs tribus boréliennes. On note $L^2(\Omega, \mathcal{A}, P)$ (= L^2) l'espace de Hilbert des (classes de) v.a.r. de carré P-intégrable muni du produit scalaire usuel :

$$\Phi : (f,g) \in (L^2)^2 \longrightarrow \langle f,g \rangle = \int_\Omega fgdP$$

Pour chaque i de $I = \{1,2...p\}$ et chaque j de $J = \{1,2...q\}$, on suppose que X_i et Y_j sont de carré P-intégrable ; on note F_X (resp. F_Y) le sous-espace de L^2 engendré par les v.a.r. $\{X_i\}_{i\in I}$ (resp. $\{Y_j\}_{j\in J}$) et 1_Ω. Faire l'analyse canonique de X et Y consiste à chercher dans une première étape un couple (f,g) de $F_X \times F_Y$ tel que $\|f\| = \|g\| = 1$ et que le produit scalaire $\langle f,g \rangle$ soit maximum. On voit immédiatement que le couple cherché est $f_o = g_o = 1_\Omega$. Dans la deuxième étape, on cherche le (ou un) couple (f,g) de $(F_X \ominus \text{vect}\{f_o\}) \times (F_Y \ominus \text{vect}\{g_o\})$ vérifiant $\|f\| = \|g\| = 1$, et tel que le produit scalaire $\langle f,g \rangle$ soit maximum ; les v.a. cherchées sont centrées, et donc leur corrélation est maximale. Un tel couple étant déterminé, on poursuit l'opération autant que possible sous les contraintes d'orthonormalité habituelles.

1.3. Analyse canonique non-linéaire :

Dans 1.2., on cherche à chaque étape des v.a. f et g qui sont respectivement des combinaisons linéaires des X_i ($i \in I$) et des Y_j ($j \in J$). De manière plus générale, on peut chercher $(f,g) \in (L^2)^2$ sous la forme $(\alpha \circ X, \beta \circ Y)$ où α (resp. β) est une application mesurable quelconque de

$(\mathbb{R}^p, \mathcal{B}_{\mathbb{R}^p})$ (resp $(\mathbb{R}^q, \mathcal{B}_{\mathbb{R}^q})$) dans $(\mathbb{R}, \mathcal{B}_{\mathbb{R}})$. Si on désigne par $\mathcal{B}$ (resp. $\mathcal{C}$) la tribu complétée engendrée par X (resp. Y), on est ainsi conduit à la recherche de f dans $L^2(\mathcal{B})$ et g dans $L^2(\mathcal{C})$ normées et maximisant Φ.

Il est clair qu'à la première étape, un tel couple (f_o, g_o) existe toujours : il suffit de prendre $f_o = g_o = 1_\Omega$. On peut alors chercher s'il existe (f_1, g_1) de $[L^2(\mathcal{B}) \ominus \text{vect}\{f_o\}] \times [L^2(\mathcal{C}) \ominus \text{vect}\{g_o\}]$ (donc un couple de v.a. centrées de $L^2(\mathcal{B})$ et $L^2(\mathcal{C})$ respectivement), avec $\|f\| = \|g\| = 1$, maximisant Φ. Si un tel couple peut être trouvé, on poursuit dans $[L^2(\mathcal{B}) \ominus \text{vect}\{f_o, f_1\}] \times [L^2(\mathcal{C}) \ominus \text{vect}\{g_o, g_1\}]$, et ainsi de suite.

On peut noter que la notion importante (parce qu'elle est intrinsèque) pour le non-linéaire est celle de tribu complétée engendrée ; et on est ainsi conduit à la notion d'analyse canonique de deux sous-tribus complètes $\mathcal{B}$ et $\mathcal{C}$, quelconques, de $\mathcal{A}$, que l'on peut encore qualifier d'analyse canonique de $L^2(\mathcal{B})$ et $L^2(\mathcal{C})$. Une telle analyse définie "pas à pas" ainsi qu'il est indiquée ci-dessus n'existe pas toujours ; on vérifie que moyennant la compacité de l'opérateur $E^{\mathcal{B}} \circ E^{\mathcal{C}}$ (où $E^{\mathcal{B}}$ est l'opérateur "espérance conditionnelle à $\mathcal{B}$ " c'est-à-dire le projecteur orthogonal de L^2 sur $L^2(\mathcal{B})$), elle existe, et est fournie par l'analyse spectrale de $E^{\mathcal{B}} \circ E^{\mathcal{C}}|_{L^2(\mathcal{B})}$ (ou $E^{\mathcal{C}} \circ E^{\mathcal{B}}|_{L^2(\mathcal{C})}$).

Une telle analyse canonique, si elle existe, conduit, lorsque les espaces $L^2(\mathcal{B})$ et $L^2(\mathcal{C})$ sont séparables, à une décomposition "privilégiée" de $L^2(\mathcal{B})$ et $L^2(\mathcal{C})$ en sommes directes.

Il faut noter enfin que cette présentation permet de définir aisément la notion d'analyse canonique pour deux v.a. à valeurs dans des espaces probabilisés quelconques, et qu'elle met en évidence diverses propriétés (par exemple, l'invariance de l'analyse canonique de deux v.a. par transformations bijectives bimesurables).

On peut trouver des compléments sur l'A.C. non linéaire dans [5.1.] ou [5.2.] ,[8.1.] ou [8.2.], et un résumé dans [12] .

1.4. Analyse canonique de deux sous-espaces fermés d'un espace de Hilbert :

L'A.C. linéaire et l'A.C. non-linéaire conduisent à la même démarche, l'une à partir des sous-espaces fermés F_X et F_Y de L^2 , l'autre à partir des sous-espaces fermés $L^2(\mathcal{B})$ et $L^2(\mathcal{C})$ de L^2. Elles entrent donc toutes deux dans le cadre suivant :

Soit H un espace de Hilbert réel séparable muni d'un produit scalaire noté $\langle .,. \rangle$, H_i (i=1,2) deux sous-espaces fermés de H. On peut comme en 1.2. ou 1.3. définir "pas à pas" une analyse canonique de H_1 et H_2 , permettant d'englober les cas 1.1., 1.2., 1.3. . Toutefois, comme il a été remarqué, une telle analyse n'est pas toujours définie. On peut se débarasser des questions d'existence en utilisant la remarque faite en 1.2., et présenter la recherche de l'analyse canonique de H_1 et H_2 comme celle de représentations conjointes privilégiées $\mathcal{H}_1$ et $\mathcal{H}_2$, de H_1 et H_2 respectivement, sous forme d'intégrales hilbertiennes relativement à une même mesure. De telles représentations peuvent être obtenues en faisant l'analyse spectrale de l'opérateur auto-adjoint $A = (P_1+P_2-I)^2$, où P_i est le projecteur orthogonal de H sur H_i . Cette analyse spectrale permet d'obtenir une représentation isométrique de H par une intégrale hilbertienne $\mathcal{H}$. Ayant remarqué que, pour i = 1,2 : $V_i = A_{|H_i} = P_i o P_{3-i|H_i}$ est un opérateur auto-adjoint de H_i dans lui-même, on peut en déduire l'existence des représentations cherchées $\mathcal{H}_1$ et $\mathcal{H}_2$. On peut ainsi donner une définition générale de l'A.C. des deux sous-espaces H_1 et H_2 .

Dans le cas où le spectre de A est dénombrable, on obtient, au lieu de représentations, des décompositions en somme directe de H_1 et H_2 . Un cas fort important est alors celui où est assurée la compacité de l'opérateur $(P_1+P_2-I)|_{\overline{H_1+H_2}}$ (ou, ce qui est équivalent, celle de l'un quelconque des

opérateurs V_1 , V_2 , ou P_1 o P_2), qui entraine la fermeture dans H de H_1+H_2. Cette condition est assurée, en particulier, si l'un des espaces H_i est de dimension finie ; elle l'est donc si $H = L^2$, $H_1 = F_X$ et $H_2 = F_Y$.

Cette présentation générale de l'A.C. conduit à des applications en Calcul des Probabilités et en Analyse des Données. Elle permet en outre de formaliser diverses pratiques intuitives, par la définition d'étapes intermédiaires entre le linéaire et le non-linéaire dont le paragraphe suivant constitue un exemple.

On peut trouver le détail de cette question dans [5.4.].

1.5. Analyse canonique semi-linéaire :

C'est par définition l'analyse canonique de $H_1 = \sum_{i=1}^{p} L^2(\mathcal{B}_i)$ et $H_2 = \sum_{j=1}^{g} L^2(\mathcal{C}_j)$ dans L^2 , où $\mathcal{B}_i$ (resp. $\mathcal{C}_j$) est la sous-tribu complétée de $\mathcal{A}$ engendrée par X_i (resp. Y_j). Cela revient à chercher, à chaque étape, le couple (f,g) où f et g sont de la forme : $f = \sum_{i=1}^{p} \alpha_i \text{ o } X_i$,
$g = \sum_{j=1}^{p} \beta_j \text{ o } Y_j$ (α_i et β_j , fonctions numériques mesurables).

On peut voir qu'une partie de ce qui est appelé analyse non-linéaire en [8.2.] est, selon cette définition, une analyse semi-linéaire.

1.6. Analyse des correspondances :

Outre son intérêt pratique, maintes fois souligné ([4]), l'intérêt théorique du cas particulier qui fait l'objet de ce paragraphe vient de ce qu'il constitue à la fois une analyse linéaire et non-linéaire (cf. [5.1.]).

Supposons que les tribus $\mathcal{B}_1$ et $\mathcal{B}_2$ engendrées par les v.a. X_1 et X_2 n'ont qu'un nombre fini d'éléments. Alors, pour j=1,2 , $\mathcal{B}_j$ est engendrée par une partition $\{B_j^k\}_{k=1\ldots p_j}$ de Ω, et $L^2(\mathcal{B}_j)$, qui est de dimension finie,

est encore le sous-espace vectoriel F_j de L^2 engendré par les v.a.r. indicatrices $\{1_{B_j^k}\}_{k=1\ldots p_j}$. Il est clair qu'ici la notion d'analyse canonique non linéaire de $\mathcal{B}_1$ et $\mathcal{B}_2$ et celle d'analyse canonique linéaire de $Y_1 = (1_{B_1^1},\ldots,1_{B_1^{p_1}})$ et $Y_2 = (1_{B_2^1},\ldots,1_{B_2^{p_2}})$ coïncident.

La densité f de (X_1,X_2) par rapport au produit des marges est

$\frac{p_{ij}}{p_{i.}p_{.j}}$, où $p_{ij} = P[B_1^i \cap B_2^j]$, $p_{i.} = P[B_1^i]$, $p_{.j} = P[B_2^j]$. Si

$(\{\rho_k\}_{k\in I}$, $\{f_k\}_{k\in I}$, $\{g_k\}_{k\in I})$ est l'analyse canonique de X_1 et X_2 ,

il vient : $\frac{p_{ij}}{p_{i.}p_{.j}} = \sum_{k\in I^*} \rho_k f_k(i) g_k(j)$, où $f_k(i)$ [resp. $g_k(j)$]

désigne la valeur prise par f_k sur B_1^i (resp. g_k sur B_2^j).

Si X_1 et X_2 sont des variables statistiques, on reconnaît là l'analyse des correspondances, présentée dans une approche différente par [4].

Cette analyse, à la fois linéaire et non-linéaire, joue un rôle important, car, en dehors des cas où un traitement direct est possible (dont on donne plusieurs exemples en [5.3.]), c'est à elle qu'on ramène, par discrétisation, toutes les analyses non linéaires exigeant un traitement numérique.

2 - ANALYSES FACTORIELLES DE FONCTIONS ALEATOIRES , ET DUALITE .

2.1. Généralisation du schéma de dualité

On va montrer comment on peut généraliser de façon progressive les schémas de dualité introduits pour les analyses factorielles classiques (cf. [9]). Cette généralisation permet une clarification des problèmes, et amène à mieux comprendre les liens entre diverses recherches antérieures. Elle débouche de façon naturelle sur les analyses de fonctions aléatoires. L'exemple de l'analyse en composantes principales linéaires semble à ces égards particulièrement significatif.

Soit $X = (X_1 \ldots X_p)$ une v.a. définie sur $(\Omega, \mathcal{A}, P)$, à valeurs dans $(\mathbb{R}^p, \mathcal{B}_{\mathbb{R}^p})$; on suppose que chaque X_i $(i \in I)$ est une v.a. de L^2 centrée. L'espace des variables est donc L^2 que l'on peut identifier à son dual topologique ; celui des individus est encore $E = \mathbb{R}^p$ muni de la métrique M (on note E^* son dual algébrique). Dans le cas où Ω est $\{\omega_1 \ldots \omega_n\}$, on note X la matrice de terme général $X_i(\omega_j)$ (cf.[9]). L'application X peut donc être définie, sur L^2, comme étant l'unique application vérifiant :

$$X : 1_{\omega_j} \longmapsto \frac{1}{p_j} E[X 1_{\omega_j}] = \begin{cases} \frac{1}{p_j} E[X_1 \, 1_{\omega_j}] = \frac{1}{p_j} \int_{\{\omega_j\}} X_1 \, dP = X_1(\omega_j) \\ \vdots \\ \frac{1}{p_j} E[X_p \, 1_{\omega_j}] = X_p(\omega_j) \end{cases}$$

d'où, en posant $U = X \circ D_p$:

$$U : 1_{\omega_j} \rightsquigarrow E[X \, 1_{\omega_j}]$$

L'application U qui généralise celle du cas classique pour Ω quelconque peut être définie par :

pour tout $A \in \mathcal{A}$ $\qquad U : 1_A \rightsquigarrow E(X.1_A) = \begin{cases} E(X_1.1_A) \\ E(X_2.1_A) \\ \hline E(X_p.1_A) \end{cases}$

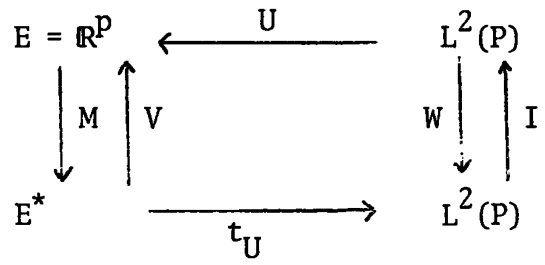
et de là par le procédé habituel (passage à la limite croissante, etc ...)

$$U : f \in L^2(P) \rightsquigarrow E(Xf) = \begin{cases} E(X_1 f) \\ \vdots \\ E(X_p f) \end{cases}$$

La transposée de U, application de E^* dans L^2, est alors définie par :

$${}^tU : u \in E^* \longrightarrow \sum_{i=1}^{p} u_i \, X_i$$

On a alors comme schéma associé :



$W = {}^tU \circ M \circ U$

$V = U \circ {}^tU$

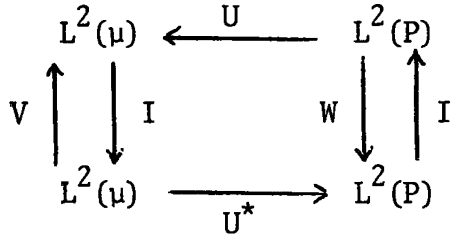
Cette première extension du schéma de dualité de l'A.C.P. classique permet alors d'envisager simplement le schéma de dualité pour l'A.C.P. linéaire d'une fonction aléatoire mesurable.

2.2. Analyse en composantes principales linéaires d'une fonction aléatoire mesurable :

$(\Omega, \mathcal{A}, P)$ et $(T, \mathcal{E}, \mu)$ sont deux espaces probabilisés donnés.

Soit $X = (X_t)_{t \in T}$ une fonction aléatoire mesurable réelle supposée de carré $P \otimes \mu$-intégrable et centrée. (On pourraît supposer seulement que μ est une mesure positive bornée, et que X est à valeurs dans un espace de Hilbert réel séparable).

On peut identifier $L^2(P)$ (resp. $L^2(\mu)$) à son dual topologique, et associer à X le schéma suivant :



I opérateur identité

$W = U^* \circ U$, $V = U \circ U^*$

où $U : f \in L^2(P) \longrightarrow E(Xf) \in L^2(\mu)$ apparaît comme un opérateur de Hilbert-Schmidt et $U^* : u \in L^2(\mu) \rightsquigarrow \int_T X_t(.)\, u(t)\, d\mu(t)$, adjoint de U, est donc aussi de Hilbert-Schmidt. Il en résulte que V (resp. W) est un opérateur de covariance (donc nucléaire) de noyau : $(t,t') \rightsquigarrow E(X_t, E_{t'})$ (resp. $(\omega,\omega') \rightsquigarrow \int_T X_t(\omega)\, X_t(\omega')\, d\mu(t)$).

On définit l'analyse en composantes principales linéaire de X relativement à μ comme l'A.C.P. de l'opérateur U, c'est-à-dire la recherche de l'élément normé de $L^2(\mu)$ dont le transformé par U^* soit de norme maximum, et itération sous contraintes d'orthogonalité.

Cette A.C.P. est obtenue par l'analyse spectrale de l'opérateur V, ou W. On obtient, comme dans le cas fini, les valeurs principales λ_i, les composantes principales f_i et les facteurs principaux u_i, qui permettent d'écrire la décomposition de X :

$$X(\omega,t) = \sum_{i\in I} \sqrt{\lambda_i}\ u_i(t)\ f_i(\omega).$$

Le schéma montre que V et W jouent des rôles pratiquement symétriques. L'opérateur W a été introduit et utilisé pour l'A.C.P. par Y. Escoufier, qui a montré en [7] qu'il est de Hilbert-Schmidt. L'opérateur V est utilisé par J.C. Deville, dans le cadre de l'analyse harmonique ([6]). Il apparaît clairement grâce au schéma que les problèmes traités par ces auteurs sont symétriques l'un de l'autre : Escoufier étudie l'échantillonnage sur T, donc son $L^2(\mu)$ dépend de la taille n de l'échantillon (et est de dimension n), alors que $L^2(P)$ est invariant; il ne peut donc travailler, pour étudier la convergence, que sur W; Deville étudie l'échantillonnage sur Ω et doit nécessairement travailler sur V, pour la même raison. Ce qui a été fait par chacun sur l'un de ces opérateurs peut s'appliquer à l'autre, dans le problème "dual".

On peut trouver les définitions et l'étude des analyses factorielles d'opérateurs et de fonctions aléatoires dans [5.5.].

3 - k-ANALYSES CANONIQUES

On a vu en 1. que toutes les analyses canoniques (linéaires ou non-linéaires) entraient dans le cadre de l'A.C. de deux sous-espaces fermés d'un espace de Hilbert réel séparable. On se propose d'étudier divers types

de généralisation pour k (k > 2) sous-espaces fermés.

Les notations utilisées sont les suivantes : H est un espace de Hilbert réel séparable. On pose $K = \{1,2,\ldots,k\}$ ($k \in N^*$). Pour chaque i de K, H_i est un sous-espace fermé de H, P_i est le projecteur orthogonal de H sur H_i , et Σ_i est la sphère unité de H_i . $\langle .,. \rangle$ est le produit scalaire sur H, et $\|.\|$ la norme associée. H' désigne la somme hilbertienne des H_i , munie du produit scalaire $\langle .,. \rangle_H$, habituel, c'est-à-dire $H' = \bigoplus_{i=1}^{k} H_i$ et pour $u = (u_i)_{i \in K} \in H'$ et $v = (v_i)_{i \in K} \in H'$, $\langle u,v \rangle_{H'} = \sum_{i=1}^{k} \langle u_i, v_i \rangle$.

Si, suivant [5.4.], on cherche une analyse canonique (A.C.) de k sous-espaces toujours définie, on est naturellement conduit à chercher une représentation isométrique de H en une intégrale hilbertienne obtenue à partir de l'opérateur $\left[\left(\sum_{i=1}^{k} P_i - I \right) \Big/ \sqrt{\sum_{i=1}^{k} H_i} \right]^2$. L'objet de la première partie de [5.6.] est d'envisager cette généralisation sous le nom d'A.C.1 .

Si on se borne à la notion d'A.C. compacte, les modes de généralisation sont alors beaucoup plus nombreux, et conduisent en général à des résultats différents.

L'A.C. compacte de deux sous-espaces H_1 et H_2 peut être obtenue indifféremment :

- Soit en faisant l'analyse spectrale d'opérateurs (cf. [5.4.]) : en particulier de $(P_1+P_2-I)/_{H_1+H_2}$ et $P_1/_{H_2}$.

- Soit en donnant une méthode "pas à pas". C'est ce qui est fait classiquement en analyse canonique "linéaire", et en A.C. "non-linéaire" dans [5.1] pour un cas particulier (méthode aisément transposable au cas de deux sous-espaces H_1 et H_2). Le principe de la méthode est le suivant :

. Dans une première étape, on cherche à maximiser $\langle f,g \rangle$ pour $(f,g) \in \Sigma_1 \times \Sigma_2$.

. On poursuit sous des contraintes d'orthonormalité, de type a) ou b)

a) Si (f_j,g_j), élément de $\Sigma_1 \times \Sigma_2$, est le jème couple canonique :

(i) $\langle f_\ell,f_j\rangle = \delta_{\ell j}$, pour $\ell \leq j$

(ii) $\langle g_\ell,g_j\rangle = \delta_{\ell j}$, pour $\ell \leq j$

(iii) $\langle f_\ell,g_j\rangle = \langle g_\ell,f_j\rangle = 0$, pour $\ell \leq j$.

b) Si (f_j,g_j), élément de $H_1 \times H_2$, est le jème couple canonique :

$$\langle f_j,f_\ell\rangle + \langle g_j,g_\ell\rangle = \delta_{j\ell} \; ; \quad \text{pour } \ell \leq j.$$

On vérifie que ces deux types de contraintes conduisent à la même A.C. compacte.

On remarque aisément que maximiser $\langle f,g\rangle$ sur $\Sigma_1 \times \Sigma_2$ est encore équivalent à maximiser $\langle f,g\rangle^2$ sur $\Sigma_1 \times \Sigma_2$, ou encore à minimiser, toujours sur $\Sigma_1 \times \Sigma_2$, le déterminant de Gram de (f,g) :

$$\begin{vmatrix} 1 & \langle f,g\rangle \\ \langle f,g\rangle & 1 \end{vmatrix} = 1 - \langle f,g\rangle^2$$

Les rappels précédents montrent que des extensions possibles de l'A.C. au cas de k sous-espaces peuvent être obtenues :

1 - Soit en considérant des <u>opérateurs</u>, supposés compacts, généralisant les opérateurs compacts indiqués ci-dessus. On est donc conduit par exemple :

a) à considérer $\left[\left(\sum_{i=1}^{k} P_i - I\right)\Big/\overline{\sum_{i=1}^{k} H_i}\right]^2$, ce qui conduit à l'A.C.1 compacte.

b) ou à utiliser, pour $i \in K$, les opérateurs $P_i \Big/ \overline{\sum_{j\neq i} H_j}$, ce qui constitue l'A.C.2.L. . On peut d'ailleurs dans certains cas proposer une A.C. différente, appelée A.C.2.N.L.

2 - Soit en généralisant d'une part les problèmes de maximisation (ou de minimisation) "pas à pas", étudiés pour deux sous-espaces, et d'autre part les contraintes d'orthogonalité .

On peut envisager la maximisation sur $\prod_{i=1}^{k} H_i$ de :

a) $\Phi : u = (u_1,\ldots,u_k) \rightarrow \sum_{\substack{(i,j)\in K^2 \\ i<j}} \langle u_i,u_j\rangle$

b) $\Psi : u = (u_1,\ldots,u_k) \rightarrow \sum_{\substack{(i,j)\in K^2 \\ i<j}} \langle u_i,u_j\rangle^2$.

ou la minimisation, toujours sur $\prod_{i=1}^{k} H_i$, de :

c) $G : u = (u_1,\ldots,u_k) \rightarrow G(u_1,u_2,\ldots,u_k) = \det(\langle u_i,u_j\rangle)$

(déterminant de Gram de $(u_1,\ldots,u_k)$).

Par ailleurs on peut considérer deux types de contraintes généralisant le cas de l'A.C. de 2 sous-espaces :

a) Les contraintes "faibles" : on opère alors sous condition d'orthonormalité dans la somme hilbertienne H', c.à.d., si u est la solution à une étape et v celle à l'étape suivante :

$\| u \|_{H'} = \| v \|_{H'} = 1$, et $\langle u,v\rangle_{H'} = \sum_{i=1}^{k} \langle u_i,v_i\rangle = 0$

b) Les contraintes "fortes" : c'est-à-dire l'orthonormalité dans chacun des H_i :

$\forall i \in K$, $\| u_i \| = \| v_i \| = 1$ et $\langle u_i,v_i\rangle = 0$

A priori, on peut choisir comme extension de maximiser Φ , Ψ ou minimiser G, et itération, sous contraintes "faibles" ou sous contraintes "fortes".

On peut noter que la condition (iii) ne peut pas être généralisée sauf dans des cas très particuliers.

La définition et l'étude des diverses k-analyses évoquées ci-dessus font l'objet de [5.6.] ; l'A.C.1. y fait l'objet de la 1ère partie ; l'A.C.2.L. et l'A.C.2.N.L. constituent la seconde partie.

On y étudie d'autre part l'analyse "pas à pas" sous contraintes "faibles" correspondant à $\underline{\Phi}$ et on montre qu'elle coïncide avec l'A.C.1. .

Les analyses "pas à pas" sous contraintes "fortes" obtenues à partir de $\underline{\Phi}$ (resp. Ψ ; resp. G) y font l'objet de la 3ème partie.

Elles correspondent à des critères proposés par Kettenring et Horst dans un cadre linéaire. Ils sont traités dans un cadre plus général englobant le cas non-linéaire. Certains résultats récents (cf. [8.2] , et [12]) sont ainsi resitués et complétés.

4 - ECHANTILLONNAGE ET DISCRETISATION

L'étude de la convergence des analyses factorielles obtenues par échantillonnage ou discrétisation vers l'analyse correspondante de la v.a. ou de la fonction aléatoire considérée est essentielle pour donner un sens à l'interprétation qui en est souvent faite. On va indiquer ici comment on peut conduire cette étude sur deux exemples : l'A.C.P. linéaire en ce qui concerne l'échantillonnage, et l'A.C. non linéaire pour ce qui est de la discrétisation.

4.1. Analyse en composante principale linéaire, et échantillonnage

Soit $(\Omega, \mathcal{A}, P)$ et $(T, \mathcal{E}, \mu)$ deux espaces probabilisés , et $X = (X_t)_{t \in T}$ une fonction aléatoire mesurable à valeurs dans $(H, \mathcal{B}_H)$ (espace de Hilbert réel séparable muni de sa tribu borélienne). On suppose X de norme carrée $P \otimes \mu$-intégrable, et centrée. L'A.C.P. linéaire de X se ramène (cf.[5.5] et [5.7.]) à l'analyse spectrale de l'opérateur $V = E[X \otimes X]$.

X peut être considérée comme une v.a. définie sur $(\Omega, \mathcal{A}, P)$ et à valeurs dans $E = L^2_H(\mu)$ muni de sa tribu borélienne $\mathcal{B}_E$. Soit $X^1, \ldots, X^n$ une suite de v.a. indépendantes, chacune de même loi que X (c'est-à-dire un échantillon). L'étude du cas fini conduit à estimer V par l'opérateur $V_n = \frac{1}{n} \sum_{i=1}^{n} X^i \otimes X^i$, v.a. intégrable définie sur $(\Omega, \mathcal{A}, P)$ et à valeurs dans l'espace de Banach séparable $\mathcal{G}_1(E)$ muni de sa tribu borélienne *. D'après la loi forte des grands nombres pour des v.a. à valeurs dans un espace de Banach séparable, la suite $(V_n(\omega))_{n \in \mathbb{N}}$ converge uniformément, pour P-presque tout ω de Ω, vers V. On en déduit la convergence uniforme (presque sûre) de l'A.C.P. linéaire obtenue à partir de l'échantillon vers l'A.C.P. linéaire de X.

La pratique courante conduit à poser le problème en termes de convergence presque sûre. On peut s'intéresser néanmoins à d'autres modes de convergence : la convergence en moyenne quadratique est établie en [6] sous l'hypothèse supplémentaire d'existence d'un moment d'ordre 4.

4.2. Analyse canonique non linéaire, et discrétisation .

Soit (X,Y) un couple de v.a. réelles définies sur $(\Omega, \mathcal{A}, P)$, $\mathcal{D}_0$ (resp. $\mathcal{E}_0$) la tribu sur $\mathbb{R}$ engendrée par une partition finie $\{A_i\}_{i=1,\ldots,k}$ (resp. $\{B_j\}_{j=1,\ldots,\ell}$) de $\mathbb{R}$ en intervalles semi-ouverts, et $\{\mathcal{D}_n\}_{n \in \mathbb{N}}$ (resp. $\{\mathcal{E}_n\}_{n \in \mathbb{N}}$) la suite croissante de tribus sur $\mathbb{R}$ définie par la récurrence suivante : $\mathcal{D}_{n+1}$ (resp. $\mathcal{E}_{n+1}$) est engendrée par la partition obtenue en partageant en deux intervalles semi-ouverts de même amplitude chaque intervalle borné de la partition engendrant $\mathcal{D}_n$ (resp. $\mathcal{E}_n$), et en remplaçant tout intervalle de la forme $]-\infty, a[$ par $\{]-\infty, a-\alpha[, [a-\alpha, a[\}$ et tout intervalle de la forme $[b, +\infty[$ par $\{[b, b+\beta[, [\beta+\beta, +\infty[\}$, où α et β sont des réels strictement positifs, arbitraires. On note $\mathcal{B}_n$ la sous-tribu complétée de $\mathcal{A}$ engendrée par $X^{-1}(\mathcal{D}_n)$, et $\mathcal{C}_n$ celle engendrée par $Y^{-1}(\mathcal{E}_n)$. On montre alors que

(*) $\mathcal{G}_1(E)$ espace de Banach des opérateurs nucléaires de E dans E.

$(\mathcal{B}_n)_{n\in\mathbb{N}}$ (resp. $(\mathcal{C}_n)_{n\in\mathbb{N}}$) converge fortement vers $\mathcal{B}$ (resp. $\mathcal{C}$), sous-tribu de $\mathcal{A}$ engendrée par X (resp. Y); cela permet, puisque l'A.C. non linéaire de (X,Y) [resp. de $(\mathcal{B}_n, \mathcal{C}_n)$] est obtenue par l'analyse spectrale de $E^{\mathcal{B}} \circ E^{\mathcal{C}}$ (resp. $E^{\mathcal{B}_n} \circ E^{\mathcal{C}_n}$), de démontrer la convergence uniforme de l'A.C. de $(\mathcal{B}_n, \mathcal{C}_n)$ vers celle de (X,Y), lorsqu'elle est compacte.

Comme $\mathcal{B}_n$ et $\mathcal{C}_n$ sont finies, on peut remarquer que leur analyse canonique est une analyse des correspondances.

A partir d'une partition finie de $\mathbb{R}$, il est clair qu'il existe une infinité de façons de construire une suite croissante de tribus conduisant à la convergence . On s'est limité ici à la plus simple.

On pourra trouver les démonstrations, dans les deux cas étudiés ci-dessus, et l'étude des convergences des autres types d'analyses factorielles dans [5.7.].

BIBLIOGRAPHIE :

[1] T.W. ANDERSON : "Introduction to multivariate statistical analysis" Wiley (1958)

[2] A. BACCINI : "Aspect synthétique de la segmentation et traitement de variables qualitatives à modalités ordonnées" Thèse de 3ème cycle, Toulouse (1975)

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

[4] J.P. BENZECRI : "Analyse des données" - Dunod (1973)

[5] J. DAUXOIS et A. POUSSE :

[5.1.] "Sur l'analyse canonique de deux tribus" C.R.A.S. Paris T. 278 série A (février 1974)

[5.2.] "Analyse canonique de deux tribus". Publication du Laboratoire de Statistique N° 01-74 Université Paul Sabatier Toulouse (1974)

[5.3.] "Image d'une analyse canonique" Publication du Laboratoire de Statistique N° 02-74 Université Paul Sabatier, Toulouse (1974)

[5.4.] "Une extension de l'analyse canonique ; quelques applications" Annales de l'Institut Henri Poincaré n° 11 Vol. 4 (1975)

[5.5.] "Analysès factorielles et dualité : une approche générale" Publication du Laboratoire de Statistique N°02-75 Université Paul Sabatier, Toulouse (1975)

[5.6.] "k-analyses canoniques ; applications statistiques" Publication du Laboratoire de Statistique N° 03-75 Université Paul Sabatier, Toulouse (1975)

[5.7.] "Discrétisation et échantillonnage en analyses factorielles. Etude des problèmes de convergence". Publication du Laboratoire de Statistique Université Paul Sabatier, Toulouse (à paraître)

[6] J.C. DEVILLE : "Méthodes statistiques et numériques de l'analyse harmonique".
Annales de l'I.N.S.E.E. N° 15 (1974)

[7] Y. ESCOUFIER : "Echantillonnage dans une population de variables aléatoires réelles"
Thèse de Doctorat d'Etat - Université de Montpellier (1970)

[8] M. MASSON :

[8.1.] "Analyse non linéaire de données"
C.R.A.S. Paris t. 278 série A (mars 1974)

[8.2.] "Processus linéaires et analyses des données non linéaires". Thèse de doctorat d'Etat. Université de Paris VI (1974)

[9] J.P. PAGES et collaborateurs : "Analyse des données multidimensionnelles". C.E.E.E. (1971)

[10] J.P. PAGES : "A propos des opérateurs d'Y. Escoufier" . Note du C.E.A. Fontenay (1975)

[11] A. POUSSE :

[11.1.] "Sur l'analyse canonique considérée comme une analyse en composantes principales". C.R.A.S. Paris t. 276 Série A (Janvier 1973)

[11.2.] "Une remarque sur la segmentation dans le cas d'indépendance conditionnelle".
Publication du Laboratoire de Statistique N° 01-75
Université Paul Sabatier,Toulouse (1975)

[12] G. SAPORTA : "Liaisons entre plusieurs ensembles de variables et codage de données qualitatives". Thèse de 3ème cycle Université de Paris VI (1975)