L'analyse harmonique qualitative

Jean-Claude Deville, Gilbert Saporta

Data Analysis and Informatics
E. Diday et al. (eds.)

ANALYSE HARMONIQUE QUALITATIVE

Jean-Claude DEVILLE
Institut National de la Statistique et des Etudes Economiques

Gilbert SAPORTA
Université de Paris V - Institut Universitaire de Technologie

SUMMARY :

This paper deals with nominal time-series. The spectral analysis of an operator which sums up time dependences provides a decomposition of the nominal process into a sequence a time-invariant uncorrelated random variables and a set of non-random time functions which give changing scores to the states of the process. These results extend scalar harmonic analysis to nominal processes, multiple correspondence analysis to a continuous set of nominal variables and generalized canonical analysis to multidimensional stochastic processes. Numerical computation comes down to a particular kind of correspondence analysis.

RESUME :

On s'intéresse à des données qualitatives évoluant dans le temps. L'analyse spectrale d'un opérateur résumant les dépendances temporelles fournit une décomposition du processus qualitatif en une suite de variables aléatoires réelles non corrélées indépendantes du temps, et une famille de fonctions certaines du temps donnant des codages évolutifs des états du processus. Ces résultats étendent l'analyse harmonique scalaire au cas des processus qualitatifs, l'analyse des correspondances multiples au cas d'une famille continue de variables quantitatives et l'analyse canonique généralisée aux processus vectoriels. La résolution numérique se ramène à une analyse des correspondances d'un type particulier.

I - POSITION DU PROBLEME ET MOTIVATIONS

De nombreuses données d'enquêtes permettent de retracer l'histoire d'individus pendant une certaine période de temps : évolution de l'activité professionnelle, de la situation matrimoniale, de la résidence. Le but de cet article est de montrer qu'on peut analyser de telles données de façon analogue à ce que l'on ferait pour un processus scalaire (analyse harmonique [4]) et pour un ensemble fini de variables qualitatives non temporelles (analyse canonique généralisée ou analyse des correspondances multiples [9]).

Formellement, on considérera comme donnés les éléments suivants :

- Un intervalle de "temps" réel compact $T = [0,T]$ muni de la mesure de Lebesgue dt (la généralisation à une mesure finie arbitraire ne pose aucun problème).
- Un ensemble fini $\mathfrak{X}$ comportant m états.
- Un espace probabilisé $(\Omega, \mathcal{A}, P)$ et un processus X_t à valeur dans $\mathfrak{X}$. Pour tout t, X_t est donc une variable qualitative à m modalités. On supposera, hypothèse qui n'a rien de minimale, que X_t est continu en probabilité, c'est-à-dire

$$P(X_{t+h} \neq X_t) \to 0 \text{ quand } h \to 0.$$

On notera 1_t^x la variable indicatrice de l'évènement $X_t = x$, $\underline{1}_t$ (ou $\vec{1}_t$) le vecteur dont les m composantes sont les 1_t^x, $p_x(t) = P(X_t = x) = E(1_t^x)$ et $p_{x,y}(t,s) = P(X_t = x \cap X_s = y) = E(1_t^x 1_s^y)$.

Un codage (réel) du processus X_t est une fonction $f_t(x)$ de $T \times \mathfrak{X}$ dans R. Notons $\vec{f}_t$ (ou $\underline{f}_t$) le vecteur dont les m composantes sont les $f_t(x)$ $(x \in \mathfrak{X})$; on obtient un processus scalaire Y_t par :

$$Y_t = f_t(X_t) = \langle \vec{f}_t, \vec{1}_t \rangle = \sum_x f_t(x)\, 1_t^x$$

Une démarche naturelle pourrait être de faire l'analyse harmonique du processus Y_t pour un codage bien choisi. On pourrait aussi choisir une famille finie de temps $0 < t_1 < t_2 < t_n < T$ et faire l'analyse canonique des variables qualitatives $X_{t_1}, X_{t_2} \ldots X_{t_n}$. En fait ou peut arriver assez facilement à un traitement qui synthétise et dans une certaine mesure justifie ces deux idées.

II - DEPENDANCES TEMPORELLES

Soit B_t la tribu engendrée par X_t, $L^2(t)$ l'espace des variables réelles B_t mesurables (c'est-à-dire de la forme $\sum a_x 1_t^x$), et E^t l'opérateur d'espérance conditionnelle à B_t c'est-à-dire, dans $L^2(\Omega, \mathcal{A}, P)$ la projection orthogonale sur $L^2(t)$.

Les opérateurs E^t sont donc idempotents, hermitiens, de rang m (au plus), et transforment une v.a.r. positive en une v.a.r. positive. Inversement, tout opérateur ayant ces propriétés est une espérance conditionnelle relative à une tribu engendrée par une partition de taille m au plus [7]. Il y a donc équivalence entre la donnée d'un opérateur E^t et d'une variable qualitative (définie à une P-équivalence près). Toutes les dépendances entre variables qualitatives pourront donc s'exprimer en terme d'opérateurs, ou si on veut, on pourra se contenter, pour étudier les liaisons entre variables, d'étudier les "rapports" entre opérateurs correspondants.

Les dépendances deux à deux, au second ordre en quelque sorte, seront donc entièrement prises en compte par la "fonction de covariance " $K_{t,s} = E^t E^s$. Notons que $K_{t,t} = E^t$ et que $K_{s,t} = K_{t,s}^*$. La fonction K est à valeurs <u>opérateurs</u> (et non pas scalaire) et transforme donc une v.a.r. en une autre. Il est facile de vérifier qu'elle est continue pour la norme habituelle des opérateurs.

Si ξ_t est un processus scalaire sur T de variance totale finie (continu m.q par exemple), la fonction $K_{t,s}$ définit un opérateur intégral K

opérant sur les processus de variance finie de la façon suivante :

$$(K\xi)_t = \eta_t = \int_T K_{t,s}\ \xi_s\ ds$$

Soit $\mathcal{H}$ l'espace de Hilbert des (classes d'équivalence des) processus du second ordre de variance totale finie muni du produit scalaire classique suivant (covariance en moyenne) :

$$(\xi|\eta)_{\mathcal{H}} = \int_T E(\xi_t\ \eta_t)\ dt$$

K est un opérateur dans $\mathcal{H}$; il est immédiat, en vertu des hypothèses, de vérifier que K est hermitien, positif, compact, de trace finie. On va donc pouvoir utiliser sa décomposition spectrale :

$$K = \sum_{i=1}^{\infty} \lambda_i\ \xi^{(i)} \otimes \xi^{(i)}$$

où les $\xi^{(i)}$ sont des processus de variance totale unité, orthogonaux dans $\mathcal{H}$, vecteurs propres associés à λ_i de l'opérateur K c'est-à-dire vérifiant :

$$\lambda_i\ \xi_t^{(i)} = \int_T K_{t,s}\ \xi_s^{(i)}\ ds \qquad (1)$$

Remarque 1 : L'opérateur $\xi \otimes \eta$ transforme le processus ψ_t en le processus :

$$E(\eta_t \otimes \psi_t)\ \xi_t.$$

Remarque 2 : Les processus $\xi^{(i)}$ forment une base hilbertienne de l'image fermée de K dans $\mathcal{H}$. Autrement dit on a, pour tout processus de $\mathcal{H}$, soit η, en notant Π la projection orthogonale sur $\overline{\mathrm{Im}K}$:

$$\Pi\eta = \sum_{i=1}^{\infty} (\eta\,|\,\xi^{(i)})_{\mathcal{H}}\ \xi^{(i)}$$

III - GENERATRICES, UN AUTRE ASPECT DE LA DECOMPOSITION

L'équation (1) s'écrit aussi :

$$\xi_t = E^t \int_T E^s(\xi_s)\ ds = E^t \int_T \xi_s\ ds \text{ puisque } E^s(\xi_s) = \xi_s$$

Les processus propres sont donc nécessairement tels que ξ_t soit B_t-mesurable.

Notons alors $Q = \int_T E^s ds$. C'est un opérateur hermitien autoadjoint compact sur $L^2(\Omega,\mathcal{A},P,)$ (et c'est aussi, en quelque sorte, la trace de K).

Définissons une variable aléatoire z par :

$$z = \int_T \xi_s ds \text{ d'où } \lambda\xi_t = E^t(z) \qquad (2)$$

On appellera z la génératrice du processus propre ξ. Par intégration de (2) sur T et par définition de Q il vient :

$$\lambda z = Qz \qquad (3)$$

Les génératrices sont donc vecteurs propres de Q. Inversement, soit z un vecteur propre de Q et posons $\xi_t = E^t z$. En prenant l'espérance à B_t des deux membres de (3) on voit que ξ_t est processus propre de K et que la recherche des éléments propres de l'équation (1) est équivalente à celle de l'équation (3).

On en déduit que :

- Les génératrices sont sans corrélation .
- $z^{(\circ)} = C^{te}$ est génératrice propre de valeur propre T. Comme $\|Q\| < 1$ dès que E^t n'est pas constant (dt-presque partout) toutes les valeurs propres sont plus petites que T.
- $z^{(i)}$ (i=1...) est de moyenne nulle
- ξ_t° est constante
- $\xi_t^{(i)}$ est de moyenne nulle tout tout t et tout $i \geq 1$

On trouve de plus aisément que les $z^{(i)}$ sont de variance λ_i.

IV - LES CODAGES PROPRES ET LEURS CALCULS

Si ξ_t est un processus propre, on peut lui associer un codage car ξ_t étant B_t-mesurable on a :

$$\xi_t = \sum_x a_t^x\ 1_t^x$$

L'équation (1) s'écrit alors :

$$\lambda \sum_x a_t^x\ 1_t^x = \int_T \sum_y a_s^y\ E^t(1_s^y)\ ds$$

comme $E^t(1_s^y) = \sum_x 1_x^t\ E\ (1_s^y/1_x^t = 1)$, on trouve que :

$$\lambda \sum_x a_t^x\ 1_t^x = \sum_x 1_t^x \int_T \sum_y a_s^y\ \frac{p_{x,y}(t,s)}{p_x(t)}\ ds \text{ si } p_x(t) \neq 0$$

d'où par identification :

$$\lambda\ a_t^x = \int_T ds \sum_y \frac{p_{x,y}(t,s)}{p_x(t)} a_s^y \qquad (4)$$

ou enfin, vectoriellement, en notant $P_{t,s}$ la matrice dont les éléments sont les $p_{x,y}\ (t,s)$:

$$\lambda \underline{a}_t = \int_T P_{t,t}^{-1}\ P_{t,s}\ \underline{a}_s\ ds. \qquad (5)$$

Les génératrices $z^{(i)}$ et les codages $\underline{a}_t^{(i)}$ fournissent donc une décomposition du processus en une suite orthogonale de variables aléatoires indépendantes du temps et de codages (non aléatoires) fonctions du temps. Leur inter-

prétation pratique peut se faire comme celle des composantes et des harmoniques en analyse harmonique : graphes des codages de chaque état en fonction du temps, corrélation des génératrices avec certaines variables connues (cercle des corrélations).

Cela dit, comme nous allons le voir, l'analyse harmonique qualitative est une généralisation de méthodes familières d'analyse des données et s'y ramène d'ailleurs sur le plan numérique. Toutes les techniques d'interprétations habituelles dans ces méthodes sont donc encore valables dans notre cas.

V - LIEN AVEC L'ANALYSE DES CORRESPONDANCES MULTIPLES

Supposons qu'il existe p+1 instants $0=t_0 < t_1 \ldots < t_p = T$, tels que toutes les trajectoires du processus soient constantes sur les intervalles (t_{j-1}, t_j). Autrement dit les changements d'états ne pourront se produire qu'aux instants t_j $(j=1,\ldots,p-1)$ (*)

L'étude d'un tel processus est semblable à la temporalité près, à celle d'un ensemble de p variables qualitatives à valeurs dans le même ensemble

Notons ξ_j la valeur de ξ_t sur (t_{j-1}, t_j), E^j l'espérance conditionnelle à ξ_j et $l_j = t_j - t_{j-1}$; les équations (1) (2) et (3) prennent maintenant la forme suivante :

$$\lambda \xi_k = \sum_{j=1}^{p} l_j E^k E^j \xi_j \qquad (1')$$

$$z = \sum_{j=1}^{p} l_j \xi_j \qquad (2')$$

$$\lambda z = \left(\sum_j l_j E^j\right) z \qquad (3')$$

Il est clair, enfin, que les codages associés aux variables ξ_k seront des fonctions constantes sur les intervalles du découpage et qu'ils seront donnés par la modification suivante de l'équation (4) : (on note $p_{x,y}(j,k)$ la valeur constante de $p_{x,y}(t,s)$ sur le rectangle $(t_{j-1},t_j) \times (t_{k-1},t_k)$; $Px(j)$ est défini de façon analogue) :

$$\lambda a_k^x = \sum_{j=1}^{p} \sum_y l_j \frac{p_{x,y}(k,j)}{p_x(k)} a_j^y \qquad (4')$$

(*) Il ne s'agit évidemment pas des hypothèses de continuité faites au paragraphe I, nous avons déjà dit qu'elles n'avaient rien de minimum. Il est clair que, dans le contexte de ce paragraphe, tout "marche" encore bien.

L'équation (3') n'est autre que celle de l'analyse canonique généralisée de J.D.CARROLL [2] appliquée aux p groupes des m indicatrices des états. Dans le cas où Ω est fini et comporte n individus on retombe sur un des avatars de l'analyse des correspondances : les codages et les génératrices sont les "facteurs" de l'analyse de correspondances d'un tableau disjonctif modifié à n lignes et mp colonnes :

$$\left(l_1 X_1 \mid l_2 X_2 \mid \ldots \mid l_p X_p \right)$$

où la $j^{ème}$ "question" serait pondérée par la durée l_j . Les valeurs propres de cette AFC sont alors égales à λ_i / T .

VI - LIEN AVEC L'ANALYSE HARMONIQUE VECTORIELLE

Une variable qualitative est équivalente à l'ensemble des indicatrices de ses modalités ; un processus qualitatif à m états est donc équivalent au processus vectoriel à valeurs dans $\mathbb{R}^m$; $\vec{1}_t = \left\{ 1_t^x \quad x \in \mathcal{X} \right\}$

L'analyse des correspondances multiples étant un cas particulier de l'analyse canonique généralisée (ACG), on s'attend à ce que l'analyse harmonique qualitative se déduise de l'analyse harmonique d'un processus vectoriel. Nous devrons pour cela modifier la définition de l'analyse harmonique vectorielle donnée dans [4].

Soit $\vec{X}_t$ un processus aléatoire m-dimensionnel $\vec{X}_t = (X_t^1, X_t^2, \ldots X_t^m)$ tel que $\int_T E(X_t^i)^2 \, dt < \infty$ $i=1,2,\ldots m$

Soit C l'opérateur de covariance du processus $\vec{X}_t$ tel que $C_{t,s}$ est la matrice d'éléments $C_{t,s}^{i,j} = E\left[X_t^i X_s^j\right]$ si X_t est centré. Nous supposerons $C_{t,t}$ inversible pour tout t.

L'analyse harmonique vectorielle consiste dans l'analyse spectrale de l'opérateur R déduit de C par $R_{t,s} = C_{tt}^{-1} C_{ts}$. Autrement dit on cherche les solutions $\vec{f}_t$, fonctions vectorielles certaines du temps vérifiant :

$$\lambda \vec{f}_t = \int_T C_{tt}^{-1} C_{ts} \vec{f}_s \, ds \text{ avec } \int_T \left\| \vec{f}_t \right\|^2_{C_{t,t}} dt = 1 \qquad (6)$$

En posant z, variable aléatoire scalaire indépendante du temps :

$$z = \int_T \langle \vec{X}_t, \vec{f}_t \rangle \, dt \qquad E(z^2) = \lambda$$

on a alors :

$$\vec{X}_t = \sum_{i=1}^{\infty} z^{(i)} C_{t,t}^{-1} \vec{f}_t^{(i)}$$

Par rapport à la définition de [4] où l'on effectuait l'analyse

spectrale de C, ceci revient à munir les espaces F_t des "individus" à l'instant t ($\equiv \mathbb{R}^m$) d'une métrique M_t variable avec t, ici $M_t = C_{t,t}^{-1}$.

Le choix de cette métrique s'impose pour diverses raisons :

* elle seule permet l'identification de l'analyse harmonique vectorielle aux cas particuliers de l'analyse harmonique qualitative et de l'analyse canonique généralisée (si le processus est constant par morceaux).

** elle traite les dépendances temporelles differemment des dépendances "spatiales" et respecte donc la temporalité du phénomène. En effet travailler sur l'opérateur C au lieu de R, c'est donner la même importance à la liaison entre X_t^k et X_t^l qu'à celle entre X_t^l et X_s^l ce qui n'est pas souhaitable. On retrouve ici la même différence que celle qui existe entre une ACP ordinaire sur le tableau $X = (X_1 | X_2 | \ldots | X_p)$ avec la métrique I qui ne tient pas compte du fait que les variables sont réparties en groupes, et l'analyse canonique généralisée (ou ACP réduite généralisée [8]) qui traite chaque bloc comme l'ACP normée traite une variable. On démontre alors pour l'analyse harmonique vectorielle les propriétés suivantes qui généralisant celle de l'ACG à un continuum de vecteurs aléatoires.

Propriété 1 :

Les variables z sont fonctions propres de l'opérateur $Q = \int_T A_t \, dt$ où A_t désigne le projecteur orthogonal sur W_t sous-espace vectoriel (de"variables") engendré par les composantes de $\vec{X}_t$ à l'instant t

Démonstration :

En multipliant scalairement par $\vec{X}_t$ chaque membre de l'équation (6) et en intégrant sur t on trouve :

$$\lambda \int_T < \vec{X}_t ; \vec{f}_t > dt = \int_T < \vec{X}_t ; \int_T C_{tt}^{-1} C_{ts} \vec{f}_s \, ds > dt$$

soit $$\lambda z = \int_T < \vec{X}_t ; C_{tt}^{-1} \int_T C_{ts} \vec{f}_s \, ds > dt$$

et en explicitant C_{ts} comme matrice de covariance :

$$\lambda z = \int_T < \vec{X}_t ; C_{tt}^{-1} E (\vec{X}_t \int_T <\vec{X}_s ; \vec{f}_s > ds) > dt$$

d'où : $$\lambda z = \int_T <\vec{X}_t ; C_{tt}^{-1} E (\vec{X}_t z) > dt$$

or l'opérateur A_t défini par $A_t(x) = < \vec{X}_t ; C_{tt}^{-1} E(\vec{X}_t x) >$ n'est autre que l'opérateur de régression linéaire de x sur $\vec{X}_t$ c'est-à-dire le projecteur orthogonal sur W_t.

Remarque : A_t est l'opérateur d'Escoufier associé à $\vec{X}_t$ mais avec la métrique C_{tt}^{-1} [5]

L'analyse harmonique vectorielle est équivalente au problème suivant : trouver le processus scalaire ξ_t, combinaison linéaire à tout instant des composantes de $\vec{X}_t$ $\xi_t = <\vec{X}_t ; \vec{f}_t>$ tel que son analyse harmonique réduite (i.e. sur opérateur de corrélation) fournisse une première valeur propre maximale. On alors $\xi_t = A_t z$.

La démonstration de cette propriété est identique à celle faite par exemple en [9] pour l'analyse canonique généralisée; nous ne la reproduirons donc pas.

VII - L'ANALYSE HARMONIQUE QUALITATIVE COMME METHODE DE CODAGE

Si on remplace $\vec{X}_t$ par le processus $\vec{1}_t$ des indicatrices d'un processus qualitatif l'équation (6) s'identifie à l'équation (5) puisqu'alors $C_{t,s} = P_{t,s}$

La propriété 2 énoncée ci-dessus s'interprète alors en termes de recherche de codage, puisque toute combinaison linéaire de variables indicatrices définit un codage. L'analyse harmonique qualitative revient alors à chercher à coder un processus qualitatif de façon à obtenir une analyse harmonique réduite optimale (au sens de la première valeur propre) du processus scalaire ainsi crée.

Evidemment poser à priori un tel critère, n'est pas sans arbitraire et on aurait très bien pu imaginer d'autres critères de codage optimal conduisant à d'autres types d'analyses (p.ex. maximiser la norme de l'opérateur de corrélation...) mais certainement moins fécondes et moins cohérentes que celle proposée ici.

VIII - TRAITEMENT NUMERIQUE DES DONNEES : L'APPROXIMATION

Le calcul pratique des codages propres nécessite le recours à des approximations. Si on suppose qu'on s'intéresse à des variables statistiques, la fonction $P_{x,y}$ (t,s) est certes calculable entièrement, elle est même constante sur des (petits!) rectangles $[t,t'[\times [s,s'[$ où t et t' sont deux instants successifs d'une liste comprenant toutes les dates, rangées dans l'ordre, des changements d'états de tous les individus. Même pour quelques centaines d'individus subissant chacun en moyenne deux ou trois changements d'états on aurait un problème aux valeurs propres de taille colossale ! Si d'autre part, on considère qu'on a affaire à un échantillon tiré d'une population assimilée à un espace probabilisé, on cherche alors à déterminer une famille de fonctions

de T dans $\mathbb{R}$; le problème ressort alors de l'analyse numérique et sa solution est, en fait, une solution approchée.

Nous proposerons donc ici deux méthodes d'approximation assez simples, qui, toutes deux, se ramènent à une analyse des correspondances multiples.

Première méthode : échantillonnage des temps

Reprenons l'équation (3) :

$$\lambda z = \int_T (E^s ds) z.$$

L'idée est d'approximer l'intégrale par une somme de Riemann. Supposons que l'on ait un découpage de T en p intervalles $[t_0,t_1[\ \ldots \ [t_{p-1},t_p[$ avec $t_0 = 0$, $t_p = T$ et $l_j = t_j - t_{j-1}$. Supposons en outre que l'on ait :

$$\sup_{j=1,..p} \ \sup_{t_{j-1} \leq s,t \leq t_j} \| E^s - E^t \| \leq \varepsilon/T$$

En pratique, il suffira de vérifier que $E^{t_j} - E^{t_{j-1}} \leq \varepsilon/T$ pour $j=1...p$, ce qui se traduit par diverses conditions sur les $P_{x,y}(t,s)$ que nous n'expliciterons pas ici. Soit alors $\theta_j \in [t_{j-1},t_j[$ et $\hat{Q} = \sum_{j=1} l_j E^{\theta_j}$ (somme de Riemann)

On obtient alors que $\|Q - \hat{Q}\| < \varepsilon$. D'après [4] (p.93.94), la décomposition spectrale de $\hat{Q}$ fournit des approximations des premières valeurs propres et des premiers vecteurs propres de $\hat{Q}$ satisfaisantes. Or la décomposition spectrale de $\hat{Q}$ n'est autre que l'analyse des correspondances multiples des variables qualitatives X_{θ_j} pondérées, il est vrai par des poids l_j. Si le découpage du temps est fait en intervalles de longueurs égales on se ramènera donc très exactement à ce type d'analyse.

On remarquera qu'on n'a pas tenu compte, dans cette méthode, de toute l'information disponible. Par exemple les variables X_0 et X_T (les deux extrémités) ne sont pas prise en compte. Pour le faire on pourrait, dans le cas où les l_j sont égaux (et égaux à T/p) prendre $\hat{Q} = \frac{T}{p} \frac{1}{2} (E^0 + E^T) + \sum_{i=1}^{p-1} E^{\frac{iT}{p}}$.

Dans le cas d'intervalles de longueurs inégales on pourrait utiliser :

$$\hat{Q} = \frac{1}{2} l_1 E^0 + \sum_{i=1}^{p-1} \frac{1}{2} (l_i + l_{i+1}) E^{t_i} + \frac{1}{2} l_p E^T.$$

Néanmoins on peut trouver une méthode qui prend mieux en compte l'information disponible, et qui semble à tous égards plus recommandable.

Seconde méthode : "moyennage des temps"

Numériquement cette méthode sera à peu près aussi simple que la précédente mais sans doute plus efficace car elle tient compte des durées passées

dans chaque état entre deux instants t_{j-1} et t_j du découpage et pas seulement des états en t_{j-1} et t_j. Sa compréhension nécessite quelques éléments théoriques préalables dont voici le résumé :

Soit H l'espace de Hilbert des variables de carré intégrables mesurables par rapport au processus X_t, c'est à dire engendré par les indicatrices 1_t^x ($t \in T$; $x \in \mathbb{X}$).

Soit $f_t(x)$ un codage continu par morceaux tel que $\int_T E[f_t(x_t)]^2 dt < \infty$, et $z = \int_T f_t(x_t)\, dt$ la variable de H associée à ce codage.

L'ensemble des variables z ainsi définies est un espace vectoriel. Sous des hypothèses assez larges sur le processus (par exemple $p_{x,y}(t,s) > 0$), ce sous-espace de H est partout dense. On en déduit que les z^i, vecteurs propres de Q (avec $\|z^i\|^2 = \lambda_i > 0$) forment un système orthonormé maximal de H (et donc les $z^i/\sqrt{\lambda_i}$ une base hilbertienne).

Soit maintenant H_p le sous-espace vectoriel de H engendré par les codages $f_x(t)$ tels que $f_x(t)$ soit constant sur $[t_{j-1},t_j[$; il est facile de voir que si on dispose d'une suite de découpages de plus en plus fins de T les espaces H_p correspondant forment au sens de [4] une suite approximante de H. Notons P le projecteur orthogonal dans H sur H_p ; on sait que l'analyse spectrale de l'opérateur PQP fournira une approximation de celle de Q.

La résolution de l'équation (4) s'avère équivalente au problème suivant :

Trouver les extrémums de :

$$\iint_{s,t} \sum_{x,y} a_s(y)\, a_t(x)\, p_{t,s}(x,y)\, ds\, dt$$

sous la contrainte : $\sum_x \int_t (a_t(x))^2 p_t(x)\, dt = 1$.

Résoudre ce problème dans H_p revient à ajouter les contraintes supplémentaires :

$$a_t(x) = a_j^x \quad \text{si } t \in [t_{j-1},t_j[$$

Notons $p_{i,j}(x,y) = \iint p_{x,y}(t,s)\, dt\, ds$ où la somme est faite sur le rectangle $[t_{i-1},t_i[\times [t_{j-1},t_j[$ et $p_j(x) \int_{t_{j-1}}^{t_j} p_x(t)\, dt$.

Le problème revient maintenant à trouver les extrémums de :

$$\sum_{i,j,x,y} a_i(x)\, a_j(y)\, p_{i,j}(x,y)$$

sous la contrainte : $\sum_{i,x} a_i(x)^2 p_i(x) = 1$.

On arrive alors immédiatement au problème de valeurs propres suivant:

$$\sum_{j,y} \frac{p_{i,j}(x,y)}{p_i(x)} a_j(y) = \lambda a_i(x) \qquad (7)$$

Définissons maintenant les variables aléatoires t_j^x suivantes :

$$t_j^x = \int_{t_{j-1}}^{t_j} 1^x(t)dt$$

t_j^x est donc le temps passé dans l'état x au cours de la $j^{ème}$ période. Il est clair que :

$$E(t_j^x) = P_x(j)$$

$$E(t_j^x\, t_k^y) = P_{x,y}(t,k).$$

La variable z associée à un codage constant par intervalles aura l'expression simple :

$$z = \sum_{j,x} a_x(j)\, t_j^x$$

Associons maintenant à chaque individu de l'échantillon les mp valeurs de t_j^x; si les données de base sont les dates de changement d'état successifs et la liste de ces états ce calcul est très facile à mettre en oeuvre automatiquement .

C'est d'ailleurs unesimple extension de ce qu'on obtient par codage disjonctif des variables en analyse des correspondances multiples.
Il est maintenant facile de voir que le calcul proposé à l'équation (7) résulte exactement de l'analyse des correspondances appliquée au tableau N de nombres positifs dont les lignes sont les individus échantillonés et les colonnes les mp variables t_j^x :

$$\begin{array}{c} \\ 1 \\ 2 \\ \vdots \\ n \end{array} \begin{array}{c} \begin{array}{cccccccccccc} t_1^1 & t_1^2 & \ldots & t_1^m & \ldots & t_j^1 & t_j^2 & \ldots & t_j^m & \ldots\ldots & t_p^1 & t_p^2 \ldots t_p^m \end{array} \\ \left(\begin{array}{cccccccccccc} & & & & & & & & & & & \\ & & & & & & & & & & & \\ & N_1 & & & \ldots & & N_j & & & \ldots\ldots & & N_p \\ & & & & & & & & & & & \\ & & & & & & & & & & & \end{array}\right) \end{array}$$

Cette pratique est voisine de l'analyse des correspondances mixtes de M. MASSON [6]

Le calcul des codages propres de l'analyse harmonique qualitative pourra donc se faire comme celui des facteurs de l'analyse des correspondances multiples sur tableau disjonctif complet. Il suffira, par un traitement préalable des données de base, de calculer le temps passé par chaque individu dans chaque état au cours de chaque intervalle de temps d'un découpage de $[o,T]$ choisi de façon raisonnable, afin de ne pas avoir un nombre de variables exagérément grand.

BIBLIOGRAPHIE

[1] BOUROCHE J.M. ; SAPORTA G. ; TENENHAUS M. (1975)
Généralized canonical analysis of qualitative data
U.S. JAPAN Seminar on multidimensional scaling San Diego

[2] CARROLL J.D.(1968) A generalization of canonical correlation analysis to three or more sets of variables . Proceedings 76th convention. Am. Psych. Ass. p 227-228

[3] DAUXOIS J. ; POUSSE A. (1976) Analyses factorielles étude synthétique Thèse d'Etat Toulouse.

[4] DEVILLE J.C. (1974) Methodes statistiques et numériques de l'analyse harmonique . Annales de l'INSEE n°15 p 3-101

[5] ESCOUFIER Y. (1970) Echantillonnage dans une population de variables aléatoires réelles. Thèse d'Etat.Montpellier

[6] MASSON M. (1978) Essai de synthèse de géométrie d'approximation sur $L^2(\Omega)$. Communication Journée de Statistique. Nice

[7] NEVEU J. (1975) Martingales à temps discrets. Masson

[8] PAGES J.P. ; CAILLIEZ F. ; ESCOUFIER Y.(1979)Analyse factorielle : un peu d'histoire et de géométrie . Revue de Statistique Appliquée XXVII n°1 p5-28.

[9] SAPORTA G. (1975) Liaisons entre plusieurs ensembles de variables et codage de données qualitatives. Thèse 3e cycle Paris VI.

ANNEXE : Quelques résultats concernant l'analyse canonique généralisée (ACG)

Soient X_i i=1,2,...,p des tableaux de données associés a p groupes de m_i variables centrées, mesurées sur n individus. La définition de p-uples de variables canoniques $\underline{\xi}_i = X_i \underline{a}_i$ n'étant pas unique et dépendant du critère choisi (*), J.D. CARROLL [2] avait suggéré de rechercher plutôt des variables auxiliaires $\underline{z}$ (nos génératrices) combinaisons linéaires des Σm_i variables et maximisant la somme de leurs carrés de corrélation multiple avec les X_i :

$$\max_{\underline{z}} \sum_{i=1}^{p} R^2(\underline{z} ; X_i)$$

On trouve immédiatement que les $\underline{z}$ sont vecteurs propres de la somme des projecteurs A_i sur les sous-espaces W_i engendrés par les X_i. Les projections $\underline{\xi}_i = A_i \underline{z}$ constituent alors un système de variables canoniques, $\underline{z}^1$ est à la fois la somme et la première composante principale des $\underline{\xi}_i$ (ACP normée). Les $\underline{z}$ sont aussi les composantes principales de l'ACP du tableau $X = (X_1 | X_2 | \ldots | X_p)$ avec pour métrique $M = \text{diag } V_{ii}^{-1}$. Cette technique apparait donc aussi comme une généralisation de l'ACP réduite à des groupes de variables : chaque groupe est traité comme une variable réduite et on ne tient compte que des corrélations intergroupes et non intragroupes.

Lorsque les X_i sont des tableaux d'indicatrices la méthode est identique à l'analyse des correspondances multiples de X. [1] [3] [9]

Les résultats obtenus au paragraphe V se transposent aisément à l'ACG en remplaçant les opérateurs d'espérance conditionnelle E^t par les projecteurs A_i Avec $l_j = 1$ l'équation (1') devient :

$$\sum_{j=1}^{p} A_i A_j \underline{\xi}_j = \lambda \underline{\xi}_i \qquad i = 1,2,\ldots p$$

Le vecteur $\underline{\xi}$ à np composantes obtenu en empilant les $\underline{\xi}_i$: $\underline{\xi}' = (\underline{\xi}'_1 \ \underline{\xi}'_2 \ .. \ \underline{\xi}'_p)$ est alors vecteur propre de la matrice K de taille np dont les blocs sont les produits $A_i A_j$:

$$\begin{pmatrix} A_1 & A_1 A_2 & \ldots & A_1 A_p \\ A_2 A_1 & A_2 & \ldots & A_2 A_p \\ \vdots & & & \\ \vdots & \ldots\ldots\ldots & & A_p \end{pmatrix} \begin{pmatrix} \underline{\xi}_1 \\ \underline{\xi}_2 \\ \vdots \\ \underline{\xi}_p \end{pmatrix} = \lambda \begin{pmatrix} \underline{\xi}_1 \\ \underline{\xi}_2 \\ \vdots \\ \underline{\xi}_p \end{pmatrix}$$

Cette propriété généralise de manière naturelle la propriété suivante (inédite à notre connaissance) de l'analyse canonique ordinaire de deux tableaux X_1 et X_2 :

(*) Il y a autant de généralisations de l'analyse canonique que de façons de définir la liaison globale entre p variables.

les variables canoniques $\underline{\xi}$ et $\underline{\eta}$ vérifient la relation :

$$\begin{pmatrix} A_1 & A_1 A_2 \\ A_2 A_1 & A_2 \end{pmatrix} \begin{pmatrix} \underline{\xi} \\ \underline{\eta} \end{pmatrix} = (1+r) \begin{pmatrix} \underline{\xi} \\ \underline{\eta} \end{pmatrix}$$

où r est le coefficient de corrélation canonique. Les vecteurs propres de K étant deux à deux orthogonaux on obtient alors immédiatement la propriété d'"orthogonalité faible" [3] des $\underline{\xi}_i$:

$$\sum_{i=1}^{p} r(\underline{\xi}_i^k ; \underline{\xi}_i^l) = 0 \text{ si } k \neq l$$

qui complète la propriété d'orthogonalité des $\underline{z}$: $r(\Sigma \underline{\xi}_i^k ; \Sigma \underline{\xi}_i^l) = \delta_{kl}$