L'analyse harmonique qualitative, une synthèse de la théorie

Gilbert Saporta

Recoleccion y Analisis de Datos Longitudinales,Bogota, 9-13 Décembre 1996

L'ANALYSE HARMONIQUE QUALITATIVE, UNE SYNTHESE DE LA THEORIE

SAPORTA Gilbert[1]

Résumé: L'analyse harmonique qualitative (AHQ) est une variante de l'analyse des correspondances destinée à traiter des données de calendrier décrivant les passages par différents états d'une variable d'un ensemble d'individus.

1 INTRODUCTION

L'analyse harmonique qualitative a été proposée et étudiée par Deville et Saporta dans différentes publications et également reprise par De Leeuw et Van der Heijden. Son objectif est de structurer des informations du type suivant: chaque individu d'un échantillon est décrit par la chronologie d'une succession de changements d'états parmi *m* états possibles : par exemple, changements de province, d'activité professionnelle, de statut matrimonial etc.

Il s'agit donc d'analyser des histoires individuelles (en anglais « event-history data »). Typiquement les données se présentent sous la forme d'une suite de couples (t_j, x_{t_j}) où t_j est la date de changement d'é tat , x_{t_j} le nouvel état pris à cette date.. Le nombre de changements d'état est variable d'un individu à l'autre . Nous supposerons ici que tous les individus sont observés sur une même période notée $[0, T]$ et que l'on connait les états initiaux et finaux .

D'un point de vue probabiliste, les données ne sont autres que n trajectoires d'un processus qualitatif X_t à temps continu et à nombre fini d'états, mais il ne s'agit pas ici de modéliser ce processus comme on le ferait pour un processus de Markov où on chercherait à estimer la matrice des probabilités de transition.

[1] Conservatoire National des Arts et Métiers, 292 rue Saint Martin, F 75141 Paris Cedex 03
Courrier électronique: Saporta@cnam.fr

L'AHQ est une méthode exploraroire multidimensionnelle proche de l'analyse des correspondances et qui , comme telle, peut être présentée de divers points de vue.
Partant du cas simple où la période $[0,T]$ peut être découpée en p sous périodes d'amplitudes égales, nous introduirons les diverses approches de l'AHQ jusqu'à sa forme la plus abstraite avec des opérateurs hilbertiens.

On notera X_t la variable X (aléatoire ou statistique) observée à l'instant *t, x* désignera un des m états. $\mathbf{N}_{ts} = \mathbf{X}_t^{'}\mathbf{X}_s$ la matrice à n lignes et m colonnes de ses indicatrices à l'instant *t.* $\mathbf{N_{ts}}$ la matrice (m,m) contenant les distributions conjointes $n_{ts}(x,y)$ d'individus présents dans l'état x à l'instant t et dans l'état y à l'instant s. $\mathbf{N_{tt}}$ est donc la matrice des effectifs marginaux n_t^x des états à l'instant t. Il n'est pas nécessaire que $t < s$, si c'est le cas, alors $\mathbf{N}_{tt}^{-1}\mathbf{N}_{ts}$ est la matrice des probabilités empiriques de transition de t vers s.
$\mathbf{1}_t^x$ sera l'indicatrice de l'état x à l'instant t. $\mathbf{N}_{ts} = \mathbf{X}_t^{'}\mathbf{X}_s$ est le projecteur sur l'espace engendré par les indicatrices .

2 ANALYSE DES CORRESPONDANCES D'UNE SERIE CHRONOLOGIQUE QUALITATIVE

Considérons le cas où la période d'étude $[0,T]$ est subdivisée en intervalles T_1, T_2 , ...tels que chaque individu reste dans le même état pendant durant ces intervalles. En d'autres termes, les changements d'état ne peuvent se produire qu 'aux dates de fin d'intervalles , (on supposera de plus qu'il n'y a pas de changement d'état en T.

2.1 Cas de périodes égales.

On supposera donc $T_j=T/p$, ce qui serait le cas pour des événements ne pouvant survenir qu'en fin de mois comme des promotions.
Si X_1, X_2 , ..., X_p désignent les observations successives de la variable X, on se trouve confronté à un problème de description de n individus par p variables qualitatives ayant toutes le même nombre de modalités.

L'analyse des correspondances multiples va alors donner une description efficace des trajectoires des n individus . rappelons que l'ACM consiste en une AFC du tableau disjonctif complet **X** obtenu en concaténant les tableaux d'indicatrices $\mathbf{X_t}$:

$$\mathbf{X}=(\mathbf{X_1}, \mathbf{X_2}, ..., \mathbf{X_p}).$$

L'ACM permet de représenter dans un systeme d'axes orthogonaux les individus et les états. Le vecteur **a** contenant les coordonnées des modalités est solution de l'équation:

$$\frac{1}{p}\mathbf{D}^{-1}\mathbf{Ba} = \lambda\mathbf{a} \qquad (1)$$

où **B** est le tableau de Burt contenant les p^2 croisements des X_t et **D** la matrice diagonale des effectifs marginaux.

$$\mathbf{B} = \begin{pmatrix} \mathbf{N_{11}} & \mathbf{N_{12}} & \cdots & \mathbf{N_{1p}} \\ \mathbf{N_{12}} & \mathbf{N_{22}} & \cdots & \mathbf{N_{2p}} \\ \cdots & \cdots & \cdots & \cdots \\ \mathbf{N_{p1}} & \cdots & \cdots & \mathbf{N_{pp}} \end{pmatrix} \quad \text{et} \quad \mathbf{D} = \begin{pmatrix} \mathbf{N_{11}} & 0 & \cdots & 0 \\ 0 & \mathbf{N_{22}} & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & \cdots & \mathbf{N_{pp}} \end{pmatrix}$$

On notera a_t^x la coordonnée de l'état x à l'instant *t*. Les vecteurs z contenant les coordonnées des individus sont alors les solutions de:

$$\frac{1}{p}\sum_{t=1}^{p} \mathbf{X_t}(\mathbf{X_t'X_t})^{-1}\mathbf{X_t z} = \lambda \mathbf{z} \qquad (2)$$

$$\frac{1}{p}\sum_{t=1}^{p} \mathbf{A_t z} = \lambda \mathbf{z} \qquad (3)$$

Les formules de transition expriment:

i) que la coordonnée d'un individu i est proportionnelle à la moyenne arithmétique simple des coordonnées des états qu'il a occupé pendant la période $[0, T]$:

$$\mathbf{z} = \frac{1}{\sqrt{\lambda}}\frac{1}{p}\mathbf{Xa} = \frac{1}{\sqrt{\lambda}}\frac{1}{p}\sum_{t=1}^{p}\mathbf{X_t a_t} \qquad (4)$$

$$z_i = \frac{1}{\sqrt{\lambda}}\frac{1}{p}\sum_t\sum_x a_t^x 1_t^x(i) \qquad (5)$$

ii) que la coordonnée de l'état x à l'instant *t* est proportionnelle à la moyenne des coordonnées des individus qui sont dans cet état à cet instant:

$$\mathbf{a} = \frac{1}{\sqrt{\lambda}}\mathbf{D^{-1}X'z} \qquad (6)$$

$$\mathbf{a_t} = \frac{1}{\sqrt{\lambda}}\mathbf{N_{tt}^{-1}X_t'z} \qquad (7)$$

$$\left(a_t^x\right)^{(k)} = \frac{1}{\sqrt{\lambda}}\frac{1}{n_t^x}\sum_i z_i 1_t^x(i) \qquad (8)$$

On peut donc repérer sur les graphiques de l'analyse factorielle :
d'une part les proximités entre individus (ou groupe d'individus par leur barycentre) ayant des trajectoires similaires; d'autre part les évolutions des états au cours du temps.

2.2 Cas de périodes de durées variables

Les propriétés précédentes s'étendent aisément en pondérant les sous tableaux du tableau disjonctif complet par la durée t_i de chaque sous-période.
L'équation (1) devient:

$$\frac{1}{p}\mathbf{D}^{-1}\mathbf{BQa} = \lambda\mathbf{a}$$

où $\mathbf{Q}$= Diag ($t_1\mathbf{I_1}$, $t_2\mathbf{I_2}$,..,$t_p\mathbf{I_p}$) où les $\mathbf{I}$ sont les matrices unités d'ordre m.

L'équation (3) devient:

$$\frac{1}{p}\sum_{t=1}^{p}\frac{t_i}{T}\mathbf{A_t z} = \lambda\mathbf{z}$$

Ceci revient à effectuer l'analyse des correspondances du tableau disjonctif pondéré par les durées des sous-périodes:

$$\mathbf{Z}=(t_1\mathbf{X_1},t_2\ \mathbf{X_2},\ \ldots,t_p\ \mathbf{X_p}).$$

Dans les formules barycentriques, il suffit alors de remplacer les moyennes sur t par des moyennes pondérées.

3 CAS GENERAL

Les changements d'état peuvent intervenir à n'importe quel instant et à des dates différentes d'un individu à l'autre.
On est donc amené à analyser un ensemble infini (non dénombrable en théorie) de variables aléatoires qualitatives X_t avec t parcourant $[0,T]$.
L'AHQ est donc une généralisation de l'ACM à une infinité de variables ayant toutes le même nombre de modalités.
Au lieu de reprendre le cheminement historique du développement de l'AHQ, nous présenterons tout d'abord les équations à partir d'une généralisation des formules barycentriques (4) à (8) . On verra que techniquement tout se passe comme si on remplaçait les sommes temporelles par des intégrales.

3.1 Le principe barycentrique en temps continu.

Exprimons que la cordonnée d'un individu est proportionnelle à la moyenne temporelle des coordonnées des états pris au cours de $[0,T]$.

$$\mathbf{z} = \alpha\frac{1}{T}\int_0^T \mathbf{X}_t\mathbf{a}_t dt$$

soit pour chaque individu i:

$$z_i = \alpha\frac{1}{T}\int_0^T \sum_x a_t^x 1_t^{x(i)} dt$$

et que les coordonnées des états à l'instant *t* sont proportionnelles aux moyennes des coordonnées des individus qui y séjournent:

$$\mathbf{a}_t = \alpha \mathbf{N}_{tt}^{-1} \mathbf{X}_t' \mathbf{z}$$

cette derniére équation ne dépend pas du fait que le temps soit discret ou continu.
Il vient alors par substitution:

$$\lambda \mathbf{a}_t = \frac{1}{T} \mathbf{N}_{tt}^{-1} \mathbf{X}_t' \int_0^T \mathbf{X}_s \mathbf{a}_s ds = \frac{1}{T} \int_0^T \mathbf{N}_{tt}^{-1} \mathbf{X}_t' \mathbf{X}_s \mathbf{a}_s ds$$

d'où:

$$\frac{1}{T} \int_0^T \mathbf{N}_{tt}^{-1} \mathbf{N}_{ts} \mathbf{a}_s ds = \lambda \mathbf{a}_t \qquad (9)$$

et:

$$\frac{1}{T} \int_0^T \mathbf{X}_t \mathbf{N}_{tt}^{-1} \mathbf{X}_t' \mathbf{z} dt = \lambda \mathbf{z}$$

soit:

$$\frac{1}{T} \left[\int_0^T \mathbf{A}_t \, dt \right] \mathbf{z} = \lambda \mathbf{z} \qquad (10)$$

L'équation (9) est une équation intégrale où la fonction inconnue est une fonction vectorielle du temps à *m* composantes, tandis que l'équation (10) est une équation matricielle en dimension *n* (si *n* est fini), où la matrice à diagonaliser est la moyenne temporelle des matrices de projection $\mathbf{A_t}$.

3.2 Le principe d'association maximale.

La quasi totalité des méthodes exploratoires d'analyse des données peut se formuler en termes de recherche de variables inconnues Y (facteurs ou partitions) maximisant:

$$\sum_{j=1}^{p} \Phi(Y; X_j)$$

où les Xj sont les variables observées et Φ une mesure d'association appropriée à la nature des variables.(Saporta 1988)

Ainsi l'ACP normée recherche des composantes principales c rendant maximale :

$\sum_{j=1}^{p} r^2(c; x_j)$ avec des Xj numériques , tandis que l'ACM recherche des composantes z rendant maximale $\sum_{j=1}^{p} \eta^2(z; x_j)$ où η^2 est le carré du rapport de corrélation de *z* en X_j, c'est à dire la proportion de variance de *z* expliquée par X_j.

On sait que $\eta^2(z; x_j) = R^2(\mathbf{z}; \mathbf{X_j})$ où $\mathbf{X_j}$ est l'ensemble des indicatrices des modalités de X_j d'où :

$$\eta^2(\mathbf{z};\mathbf{x}_t) = \frac{\mathbf{z}'\mathbf{A}_t\mathbf{z}}{\mathbf{z}'\mathbf{z}}$$

L'extension au cas continu de ce critère est immédiate et conduit alors à l'AHQ.
En effet, en remplaçant la somme finie par une intégrale il vient:

$$\max \int_0^T \eta^2(\mathbf{z};\mathbf{x}_t)dt = \max \mathbf{z}'\left(\int_0^T \mathbf{A}_t dt \right)\mathbf{z} / \mathbf{z}'\mathbf{z}$$

et donc **z** est vecteur propre de $\int_0^T \mathbf{A}_t dt$

3.3 Une distance entre trajectoires liée à un indice de présence-rareté.

L'AHQ peut être présentée également comme une méthode de « multidimensional scaling » basée sur un indice de similarité particulier.
On sait en effet que si on dispose d'une matrice de similarité euclidienne inter-individuelle **S** , c'est à dire semi-définie positive, cette matrice peut être considérée comme une matrice de produits scalaires. Les axes principaux du nuage et les coordonnées sur ces axes s'obtiennent directement à l'aide des vecteurs propres de **S** .
L'opérateur de projection $\mathbf{A_t}$ est une telle matrice de similarité: il est facile de voir que $A_t(i,j)=0$ si *i* et *j* sont dans des états différents à l'instant *t*, et vaut $1/n_t^x$ si *i* et *j* sont dans le même état *x* à l'instant *t*. Leur similarité est d'autant plus grande qu'ils sont dans un état peu fréquent.
L'opérateur $\int_0^T \mathbf{A}_t dt$ est alors une matrice de similarité intégrée, telle que deux individus seront d'autant plus semblables qu'ils seront restés longtemps ensemble dans les mêmes états, et d'autant plus que ces états seront rares.

3.4 Présentation abstraite

Nous reprenons ici la formulation de Deville-Saporta,1979. Les X_t et les indicatrices $\mathbf{1}_t$ sont ici des variables aléatoires (et non plus des vraiables statistiques définies sur un ensemble fini) définies sur un espace probabilisé Ω.
On notera $\mathbb{B}_t$ la tribu engendrée par X_t et $L^2(t)$ l'ensemble des variables aléatoires $\mathbb{B}_t$ - mesurables qui ne sont autres que les combinaisons linéaires $\sum_x a_t^x 1_t^x$.
L'opérateur matriciel $\mathbf{A_t}$ est alors remplacé par E_t , opérateur de projection orthogonal sur $L^2(t)$, en d'autres termes l'opérateur d'espérance conditionnelle à $\mathbb{B}_t$.
Pour un processus scalaire, on définit l'opérateur de covariance *C* tel que *Cf=G* par $g(t) = \int_0^T C(t,s)f(s)ds$. *C* transforme une fonction du temps en une autre fonctiondu temps.
L'analyse harmonique scalaire (Deville 1974) consiste en l'analyse spectrale de *C* et est donc l'extension de l'ACP à un processus à temps continu.
Pour un processus qualitatif, on définit un opérateur K de $L^2(\Omega xT)$ dans lui-même, qui transforme donc un processus scalaire en un autre par:

$K\xi=\psi$ c'est à dire $\psi_t = \int_0^T E_t \circ E_s(\xi_s)ds$. Le noyau de K est donc lui-même un opérateur et non une fonction.
K résume toutes les dépendances temporelles de X_t. Ses fonctions prpores sont alors les processus scalaires codant successivement X_t. de manière optimale.

$$\lambda\xi_t = \int_0^T E_t \circ E_s(\xi_s)ds$$

Les processus propres ξt s'obtiennent aisément à partir de variables aléatoires indépendantes du temps, appelées génératrices et notées *z*.

$$\lambda\xi_t = E_t \circ \int_0^T E_s(\xi_s)ds = E_t(z) \text{ avec } z = \int_0^T E_s(\xi_s)ds$$

Il vient en intégrant:

$$\lambda\xi_t = E_t(z)$$

$$\int_0^T E_t(z)dt = \lambda\int_0^T \xi_t dt = \lambda z$$

On retrouve l'équation (10). Il suffit ensuite de projeter *z* sur $L^2(t)$ pour obtenir les ξt, de la même façon qu'en analyse canonique généralisée de *p* groupes de variables on obtient les variable canoniques en projetant les variables auxiliaires sur les sous-espaces associés à chaque groupe (Carroll 1968).

4 INTERPRETATION ET RESOLUTION NUMERIQUE

4.1. Graphiques et contributions

En supposant résolues les équations (5) et (6) on obtient des axes orthogonaux correspondant à des variables aléatoires non corrélées et décrivant au mieux l'évolution du processus qualitatif selon des variables aléatoires indépendantes du temps (les *z*) et des fonctions non aléatoires du temps (les coordonnées des états).
Donc, dans un plan factoriel, un individu est un point fixe dont les coordonnées résument sa trajectoire , et les catégories *x* sont des courbes paramétrées par le temps
Chaque valeur propre λ s'écrit:

$$\lambda = \int_0^T \sum_x (a_t^x)^2 \frac{n_t^x}{n} dt .$$

On peut alors calculer différents types de contributions (non normalisées) qui seront, selon le cas, des fonctions de *t* ou non:

contribution de l'état *x* à l'instant *t* : $(a_t^x)^2 \frac{n_t^x}{n}$;

contribution de l'état *x* au cours du temps: $\int_0^T (a_t^x)^2 \frac{n_t^x}{n} dt$;

contribution de l'instant t : $\sum_x (a_t^x)^2 \frac{n_t^x}{n}$.

On peut également calculer des contributions de sous-périodes Θ en intégrant sur Θ.

4.2. Résolution numérique des équations

4.2.1. Résolution exacte

Pour un nombre fini d'individus, le nombre total de dates de changement d'état sera fini. Il suffit donc d'ordonner toutes ces dates: entre deux de ces dates, il n'y a aucun changement d'état et on se trouve alors dans la situation décrite en 2.2.
On a donc la solution exacte du problème en effectuant l'analyse des correspondances du tableau disjonctif pondéré. Les fonctions a_t^x sont alors constantes par morceaux (pour obtenir une représentation « esthètique » on pourra lisser ces courbes en reliant par exemple les milieux des segments par des droites.
Cette solution ne peut être raisonnablement employée que si n n'est pas trop élevé: 3 changements d'états.en moyenne d'une variable à 5 modalités pour 600 individus conduit à un tableau disjonctif pondéré de 600 lignes et 9000 colonnes!

4.2.2. Solution approchée

Cette méthode, valable également dans le cas infini consiste à rechercher des solutions a_t^x constantes par morceaux sur un découpage de T en p sous-périodes, de façon à n'effectuer une analyse des correspondances sur un tableau à n lignes et mp colonnes. La meilleure approximation au sens des moindres carrés est alors donnée par le procédé suivant:
on se donne un découpage en p périodes à partir des instants: $0, t_1, t_2,\ldots,t_{p-1}, T$.
A la période $[t_{j-1}, t_j]$ correspond un tableau $\mathbf{Z}_j$ à n lignes et m colonnes tel que $Z_j(i\ ,\ x)$ soit égal au temps total passé dans l'état x par l'individu i entre les instants t_{j-1} et t_j.
En d'autres termes:

$$\mathbf{Z}_j = \int_{t_{j-1}}^{t_j} \mathbf{X}_t dt$$

L'approximation de l'AHQ est alors donnée par l'analyse des correspondances du tableau :

$$\mathbf{Z} = (\mathbf{Z}_1, \mathbf{Z}_2,\ldots,\mathbf{Z}_p).$$

En cumulant les temps de séjour dans chaque période, on perd une partie de l'information (ordre de passage dans les états, retours éventuels..), mais d'autant moins que le découpage est plus fin.
Il est recommandé de découper $[0,T]$ en périodes courtes lorsque le processus change souvent d'états et de relacher le découpage lorsque les changements d'états deviennent rares.

5. CONCLUSION ET PERSPECTIVES

L'AHQ en tant qu'extension des méthodes d'analyse des données permet de décrire, sans utiliser des modèles restrictifs, une situation complexe.
Comme toute analyse des correspondances, la faculté d'utiliser des variables supplémentaires, fixes ou variables avec le temps, enrichit les interprétations.
Elle ne nécessite pour être utilisée qu'une interface d'entrée-sortie avec un logiciel standard d'analyse des correspondances.
Cependant des problèmes théoriques et pratiques nécessitent encore des développements: citons entre autres l'analyse de certains processus à modèles connus (Markov, semi-Markov..), la prise en compte de données manquantes ou censurées car l'hypothèse d'une information complète pour tous les individus entre $[0, T]$ est souvent irréaliste, et enfin le traitement simultané de plusieurs chroniques qualitatives.

BIBLIOGRAPHIE

CARROLL J.D., 1968, « A generalization of canonical correlation analysis to three or more sets of variables », *Proceedings of the 76th convention of the American Psychological Association,*227-228

DEVILLE J.C., 1974, « Méthodes statistiques et numériques de l'analyse harmonique », *Annales de l'INSEE* 15, 3-101

DEVILLE J.C., SAPORTA G., 1979, « Analyse harmonique qualitative », *Data Analysis and Informatics,* E. Diday eds., North-Holland, 375-389

DEVILLE J.C., SAPORTA G., 1983, « Correspondence analysis, with an extension towards nominal time-series », *Journal of Econometrics* 22, 169-189

HEIJDEN PGM van der.,1987, *Correspondence analysis of longitudinal categorical data*, DSWO Press, Leiden

LEEUW J. de, HEIJDEN PGM van der, KREFT I.,1985, « Homogeneity analysis of event-history data », *Methods of Operations Research*, 50, 299-316

SAPORTA G., 1981, « Méthodes exploratoires d'analyse de données temporelles »,Cahiers du Buro n°37-38

SAPORTA G., 1985, « Data analysis for numerical and categorical individual time-series », *Applied Stochastic Models and Data Analysis* vol.1., n°2,, 109-119

SAPORTA G,1988, «About maximal association criteria in linear analysis and in cluster analysis »*Classification and related Methods of Data Analysis*, H.H.Bock ed., North-Holland, 541-550