Statistique et analyse des données

Philippe Besse

Choix de la métrique pour l' A.C.P. de séries d'événements discrets

Statistique et analyse des données, tome 12, n° 3 (1987), p. 1-16

<http://www.numdam.org/item?id=SAD_1987__12_3_1_0>

Statistique et Analyse des Données

1987 - vol. 12 n° 3 p. 1-16

CHOIX DE LA METRIQUE POUR L'A.C.P. DE SERIES D'EVENEMENTS DISCRETS

Philippe BESSE

Laboratoire de Statistique et Probabilités
U.A.-C.N.R.S. 745 - Université Paul Sabatier
31062 Toulouse Cedex.

Résumé: *Cet article considère des données constituées de séquences d'évènements discrets ou, plus généralement, des fonctions du temps présentant des discontinuités. L'analyse en composantes principales (a.c.p.) de ce type de données nécessite un lissage ou, c'est équivalent, l'utilisation d'une métrique appropriée, parmi une famille parametrée, afin d'obtenir des résultats suffisamment stables donc fiables. Ceci entraine une discussion sur le choix du parametre puis, les problèmes de stabilité soulevés par la discrétisation temporelle sont résolus en termes de convergence.*

Abstract: *This paper deals with data that are discrete event dates or, more generally, noisy time functions. Thus the problem is to smooth the raw data in order to obtain p.c.a. results that are stable enough to be reliable. This is achieved by defining a parametric family of subject space metrics equivalent to a parametric smoothing of the data. Then, the choice of the parameter value is discussed and, at last, the stability problems raised by the time discretisation are solved.*

Mots clés: *Analyse en composantes principales, données évolutives, choix de métrique.*
Indices de classification STMA: 06-070.

Manuscrit reçu le 15 novembre 1986
Révisé le 8 septembre 1987

1 - INTRODUCTION

Cet article s'intéresse à des données du type de celles étudiées par DEVILLE 1974, 1977 et constituées de séries de dates d'évènements discrets ou encore plus généralement, de fonctions bruitées ou présentant des discontinuités. Dans l'exemple considéré, on connait les dates de mariage et de naissances des enfants de n=250 familles. Les deux approches précédentes, qui ont abordé ce problème, opèrent des analyses en composante principale (a.c.p.) classiques d'un tableau (n×p) dont chaque ligne i est associée à une famille et chaque colonne j est définie par :

1. le nombre d'enfants nés jusqu'à la $j^{ième}$ année de mariage (j=1,...,P=20) (DEVILLE 1974).
2. le nombre d'enfants nés pendant la $j^{ième}$ année de mariage. (DEVILLE 1977).

La première analyse fournit des résultats intéressants mais beaucoup trop marqués par le nombre total d'enfants dans une famille qui détermine à lui seul le premier axe. La deuxième approche est jugée catastrophique par l'auteur car les résultats dépendent étroitement de la localisation et du nombre de points de discrétisation ; ils sont inexploitables.

L'objet de cet article est d'offrir une démarche générale posant le problème en termes de lissage paramétré des données brutes. Le choix du paramètre permet alors à l'utilisateur de déterminer, parmi toutes les analyses intermédiaires aux situations limites précédentes, celle qui convient à ses objectifs.

Ce travail comprend deux parties, la première, appliquée, décrit tous les outils nécessaires à une mise en oeuvre concrète : a.c.p. d'un tableau n×p, changement de métrique, famille paramétrée de lissages et de métriques. La deuxième partie développe une approche théorique pour aborder les problèmes de stabilité : approximation, analyse limite et convergence.

2 - APPROCHE CONCRETE

2.1. a. c. p., rappels et notations.

a. *a. c. p. relative au triplet* (X, Q, D).

Soit X une matrice de n lignes et p colonnes contenant les valeurs prises pour n individus par p variables réelles. L'espace des individus noté E et isomorphe à $\mathbb{R}^p$, est muni de la base canonique et de la métrique euclidienne associée à la matrice Q. De manière identique, l'espace des variables, noté F et isomorphe à $\mathbb{R}^n$, est muni de la base canonique et de la métrique associée à la matrice diagonale des poids.

On note m le vecteur colonne contenant les moyennes (pondérées par D) des colonnes X^j, X' la matrice transposée et V la matrice de covariance associée à X : V=X'DX-mm'. L'a. c. p. de X relativement aux métriques Q et D est obtenue par l'analyse spectrale de la matrice Q-symétrique définie et positive VQ dont les vecteurs propres Q-normés $\{\varphi_\alpha;\ \alpha=1,\ldots,p\}$, associés à la suite pleine décroissante $\{\lambda_\alpha;\alpha=1,\ldots,p\}$ des valeurs propres, fournissent une base de représentation dans l'espace des individus.

Lorsque le contexte évite toute ambiguïté, un même identificateur désigne un endomorphisme et sa représentation matricielle dans les bases canoniques.

b. *changement de métrique.*

Soit F un endomorphisme de $\mathbb{R}^p$. Le tableau XF' admet pour moyennes des colonnes le vecteur Fm et pour matrice de covariance :

$$(XF')'DXF' - (Fm)(Fm)' = FVF'.$$

Si, de plus, F est une isométrie de $(\mathbb{R}^p, Q)$ dans $(\mathbb{R}^p, N)$, c'est-à-dire si F'NF=Q, alors l'a. c. p. de XF' relativement à N conduit à l'analyse spectrale de FVF'N qui admet les mêmes valeurs propres $\{\lambda_\alpha;\alpha=1,\ldots,p\}$ que VQ associées aux vecteurs propres $\{F\varphi_\alpha;\alpha=1,\ldots,p\}$. Aussi, les deux a. c. p. (X, Q, D) et (XF', N, D) sont dites semblables car elles conduisent aux mêmes représentations graphiques des individus. En particulier, si F vérifie F'F=Q, les a. c. p. (X, Q, D) et (XF', I, D), où I est la matrice identité, sont semblables. Ainsi, dans ce cadre, une transformation linéaire des données est équivalente à un changement de métrique.

2.2. Illustration.

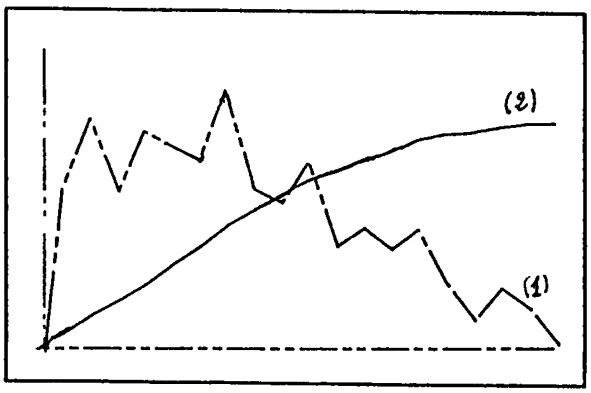
On considère le tableau X où le terme général X_i^j représente le nombre d'enfants pris dans la $i^{ème}$ famille pendant la $j^{ème}$ année de mariage. A l'analyse brute de ce tableau (DEVILLE 1977) qui fournit des résultats inexploitables, est préférée celle du tableau Y (DEVILLE 1974) contenant le nombre d'enfants nés jusqu'à la $j^{ème}$ année de mariage.

Notons F_0 l'endomorphisme représenté par la matrice triangulaire :

$$F_0 = \begin{bmatrix} 1 & \dots & 0 \\ \dots & 1 & \dots \\ 1 & \dots & 1 \end{bmatrix}$$

alors $Y=XF_0'$. D'après ce qui précède, l'a.c.p. de (Y, I, D) est obtenue par celle qui lui est semblable de (X, Q_0, D) avec $Q_0 = F_0'F_0$.

Les variables de ces analyses sont des mesures ordonnées dans le temps; aussi, à chaque individu-famille, est associé un vecteur de $\mathbb{R}^p$ assimilable à une trajectoire. Les vecteurs principaux des analyses éléments de $\mathbb{R}^p$ peuvent également être assimilés à des trajectoires appelées "harmoniques" (DEVILLE 1974) et représentées sur un graphe :



Première harmonique de l'a.c.p. de

1) (X, I, D) variance expliquées : 21,6%
2) (X, Q_0, D) variance expliquée : 90,4%

Pour la première analyse, la grande instabilité de (1) et la faiblesse de la variance expliquée souligne l'inintérêt des résultats obtenus.

A l'opposé, pour la deuxième, le premier axe de l'a.c.p. de (X, Q_0, D) est très fiable; malheureusement il est trivial: il représente avant tout la taille des familles et l'importance prépondérante qu'il prend masque l'objectif principal qui est l'étude de la façon dont se constituent les familles. Cette information est disponible sur les axes suivants mais la faible valeur des variances expliquées (4,7; 1,4;...) les rend trop peu fiables.

2.3. Famille paramétrée de métriques.

En d'autres termes, la transformation F_0 ou, ce qui est équivalent, l'utilisation de la métrique Q_0 permet de stabiliser l'a.c.p. de (X, I, D) par un lissage des données. L'a.c.p. de (X, Q_0, D) est alors appelée a.c.p. lissante (ou filtrante). Mais, dans l'exemple présenté, le lissage est trop fort car il détermine à lui seul la forme de la première harmonique. L'objectif de ce paragraphe est donc de définir une approche permettant de le pondérer ; on se propose de construire les analyses intermédiaires à celles déjà décrites afin de pouvoir limiter la prépondérance du facteur "taille de la famille" tout en conservant une stabilité suffisante des premiers axes principaux.

a - construction.

On considère la matrice F_a de dimension p'xp :

$$[F_a]_i^j = \begin{cases} e^{-a(i-j)} & \text{si} \quad i \geq j \quad ; \quad a \in \mathbb{R}^{+*} \\ 0 & \text{sinon} \end{cases}$$

pour i variant de 1 à p' et j de 1 à p. Le choix de p=20 est imposé a priori, implicitement il signifie qu'il n'y a pas de naissance après 20 ans de mariage. Limiter p' à la même valeur 20 serait erroné, car pénalisant pour les enfants nés tardivement, puisque le "poids" ou l'"importance" des enfants deviendrait nul après 20 ans de mariage. Aussi, il semble plus

rigoureux de considérer une matrice F_a de dimension p' infinie. En pratique, une telle matrice n'est pas manipulable mais, comme on s'intéresse à l'a.c.p., il suffit d'exhiber la métrique équivalente à cette transformation pour se ramener à un problème en dimension finie :

$$Q_a = F_a^{\cdot} F_a$$

est la métrique à utiliser. Le terme général de la matrice associée est donné par :

$$\sum_{k=\sup(i,j)}^{\infty} e^{-a(k-i)}\, e^{-a(k-j)} = e^{-a|j-i|} \sum_{k=0}^{\infty} e^{-2ak}$$

$$= \frac{1}{1-e^{-2a}}\, e^{-a|j-i|}.$$

Comme la constante $1/(1-e^{-2a})$ ne joue aucun rôle significatif dans l'a.c.p., on cherchera donc l'a.c.p. de (X, Q_a, D) semblable à celle de $(XF_a^{\cdot}, I, D)$ où,

$$Q_a = [e^{-a|j-i|}]_{1\leqslant i,\, j\leqslant 20}.$$

Remarques :

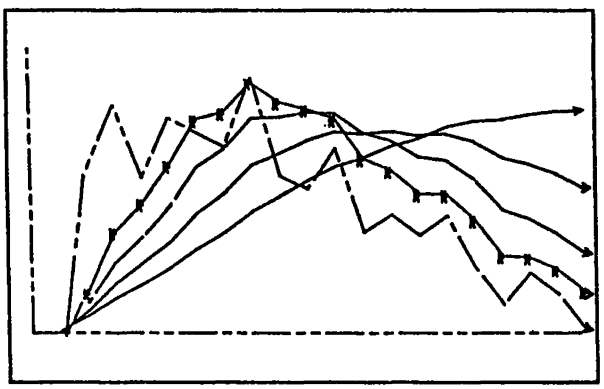
Pour a=0 cette métrique n'est pas définie mais, concrètement, on trouve que les résultats de l'a.c.p. obtenus pour de petites valeurs de a sont tout à fait semblables à ceux du §2.2. obtenus avec la métrique Q_0 aussi gardera-t-on cette notation abusive.

Lorsque a croît, Q_a tend vers la matrice identité (au sens de la convergence uniforme des opérateurs) et donc, l'a.c.p. de (X, Q_a, D) tend vers celle de (X, I, D) (au même sens).

Ainsi, on a obtenu une famille paramétrée d'a.c.p. intermédiaires à celles proposées par DEVILLE. Dans un cas (Q_0), la naissance d'un enfant est associée à un saut de la trajectoire figuré par la fonction indicatrice d'un intervalle: $(\mathbb{I}_{[t_j, 20[})$, dans l'autre cas extrême (I) elle est associée à un pic $(\mathbb{I}_{[t_{j-1}, t_j[})$; enfin l'utilisation de Q_a revient à lui associer un saut amorti $(\mathbb{I}_{[t_j, \infty[} e^{-a|t_j - t|})$. Le choix de la forme de l'amortissement est tout à fait arbitraire, c'est l'affaire d'un spécialiste du domaine concerné. Dans l'exemple présenté, le critère déterminant a été la simplicité des calculs du paragraphe 3 mais tout autre solution pourrait être envisagée.

b - *application.*

Le graphique ci-dessous donne, pour différentes valeurs de a, la forme de la première harmonique de l'a.c.p. de (X, Q_a, D) ainsi que les pourcentages de variance expliquée des trois premiers axes.



a	1	2	3
0.00	90.4	4.7	1.4
0.10	75.9	11.4	4.0
0.25	54.7	15.1	7.6
0.50	36.3	14.0	8.5
1.00	21.6	11.4	8.5

Première harmonique pour différentes valeurs de a
et pourcentages d'inertie expliquée par les premiers axes.

Remarques :

les harmoniques associées aux axes suivants n'ont été, dans cet exemple, que peu modifiées aussi ne sont-elles pas représentées. Il serait évidemment intéressant de disposer d'un critère objectif dont l'optimisation fournirait une "meilleur" valeur pour le choix de a. Ce n'est malheureusement pas ou pas encore le cas aussi faut-il se contenter d'appréciations empiriques :

- le lissage doit être suffisamment marqué, c'est-à-dire la première harmonique suffisamment lisse, pour être assuré d'une certaine stabilité des plans factoriels vis à vis de la situation et du nombre des instants de discrétisation.

- l'interprétation des résulats (DEVILLE 1974) tient compte évidemment du premier axe mais nécessite également l'information apportée par les q suivants. Le choix de q n'est pas discuté ici mais celui-ci dépend directement de la répartition des valeurs propres ou des variances expliquées. Une "meilleur" répartition,

c'est-à-dire une répartition correspondant à des valeurs propres bien distinctes pour les premiers axes, est obtenue pour des valeurs de a "avoisinant" 0,25.

3. APPROCHE THEORIQUE.

Le paragraphe précédent est suffisant pour une utilisation concrète des métriques proposées mais il laisse dans l'ombre l'étude de la stabilité des axes factoriels vis à vis de la répartition ou du nombre des dates t_i qui constituent une discrétisation de la période concernée. Ceci nous conduit à étudier la convergence de l'a.c.p. lorsque le nombre p de dates t_i croît vers l'infini.

Pour l'a.c.p. de (X, I, D), DEVILLE 1977 puis BESSE 1980 ont montré qu'il n'y avait pas convergence car l'opérateur de covariance, dont on tente de définir une approximation, n'est pas compact. Ceci explique la sensibilité des harmoniques à la situation des dates de discrétisation.

L'objet de ce paragraphe est donc de montrer que l'a.c.p. de (X, Q_g, D) définie précédemment est une approximation convergente d'une analyse en dimension infinie.

3.1 - Modélisation.

Soit Ω la population des familles, $\mathcal{A}$ la tribu des parties et μ une probabilité sur cet espace. On considère des variables aléatoires définies sur $(\Omega, \mathcal{A}, \mu)$:

$N(\omega)$ est le nombre total d'enfants nés dans la famille ω.

$\{t_i(\omega) \ ; \ i=1,\ldots,N(\omega)\}$ la suite des dates de naissance de ces enfants.

Soit X la mesure aléatoire :

$$X = \sum_{i=1}^{N} \delta_{t_i}$$

qui vérifie pour toute fonction continue u :

$$\langle X(\omega), u \rangle = \sum_{i=1}^{N(\omega)} u(t_i).$$

X est une variable aléatoire (v.a.) à valeurs dans l'espace de

Banach des mesures de Radon (pour lequel $\langle , \rangle$ désigne de façon classique la dualité) mais, comme l'a.c.p. nécessite une structure hilbertienne pour etre définie, elle sera considérée comme une v.a. à valeurs dans un espace de Hilbert H_a' construit ci-dessous. Les calculs des paragraphes suivants ont été réalisés en collaboration avec ZAAMOUN 1985.

3.2 - Construction de H_a', $a \neq 0$.

On note H_a l'espace de Sobolev sur $T = [0, +\infty[$:

$$H_a = \{f \in L^2(T) \mid f' \in L^2(T)\}$$

muni du produit scalaire :

$$\forall (f,h) \in H_a^2, \quad (f,h)_a = af(0)h(0) + a^2 \int_0^\infty f(t)h(t)dt + \int_0^\infty f'(t)h'(t)dt.$$

Proposition 1 : *L'espace* H_a *admet pour noyau reproduisant (ARONSZAJN 1950) la fonction* $g(s,t) = (2a)^{-1} e^{-a|s-t|}$.

En effet, pour tout s de T, $g(s,.)$ est un élément de H_a et, d'autre part, pour tout f de H_a :

$$\begin{aligned}(f, g(s,.))_a &= a^2 \int_0^\infty f(t)g(s,t)dt + \int_0^\infty f'(t)g_t'(s,t)dt + af(0)g(s,0) \\ &= \frac{a}{2}\int_0^s f(t)e^{-a(s-t)}dt + \frac{a}{2}\int_s^\infty f(t)e^{-a(t-s)}dt + \frac{1}{2}f(0)e^{-as} \\ &\quad + \frac{1}{2}\int_0^s f'(t)e^{-a(s-t)}dt - \frac{1}{2}\int_s^\infty f'(t)e^{-a(t-s)}dt\end{aligned}$$

il vient :

$$(f, g(s,.))_a = \frac{1}{2}[f(t)e^{-a(s-t)}]_0^s - \frac{1}{2}[f(t)e^{-a(t-s)}]_s^\infty + \frac{1}{2}f(0)\,e^{-as}$$

d'où,

$$(f, g(s,.))_a = f(s).$$

De plus, si on note H_a' le dual topologique de H_a et G_a l'opérateur défini par le crochet de dualité :

$$\forall\, u \in H_a',\quad G_a u(s) = \langle u, g(s,.)\rangle_{H_a', H_a},$$

G_a est l'isométrie canonique de H_a' dans H_a; c'est plus précisément le noyau de SCHWARTZ (1964) de l'espace H_a.

Proposition 2 : *L'opérateur* F_a *défini sur* H_a' *par* $\forall\, s \in T,\ \forall\, u \in H_a',\ F_a u(s) = \langle u, \mathbb{I}_{[0,.]}(s)\, e^{-a|s-.|}\rangle_{H_a', H_a}$ *est une isométrie de* H_a' *sur* $L^2(T)$.

On montre d'abord que $G_a = {}^tF_a \circ F_a$ (${}^tF_a$ est l'opérateur transposé) :

$$\forall\, (s,t) \in T^2,\ \langle \delta_s, {}^tF_a \circ F_a\, \delta_t\rangle_{H_a', H_a} = \langle F_a\delta_s, F_a\delta_t\rangle_{L^2(T)}$$

$$= \langle \mathbb{I}_{[0,.]}(s)\, e^{-a|s-.|}, \mathbb{I}_{[0,.]}(t)\, e^{-a|t-.|}\rangle_{L^2(T)}$$

$$= \int_0^\infty \mathbb{I}_{[s,\infty[}(x)\, e^{-a|s-x|}\, \mathbb{I}_{[t,\infty[}(x)\, e^{-a|t-x|}\, dx.$$

L'intégrale est bien définie et les égalités ci-dessus supposent, comme c'est l'usage, que $H_a(T)$ soit identifié à une partie de $L^2(T)$ et $L^2(T)$ lui-même identifié à son dual topologique ainsi qu'à une partie de $H_a'(T)$. Alors,

$$\langle \delta_s, {}^tF_a \circ F_a\, \delta_t\rangle = \int_{\sup(s,t)}^\infty e^{a(s+t)}\, e^{-2ax}\, dx = \frac{1}{2a}\, e^{-a|s-t|}.$$

Comme la famille $\{\delta_t;\ t \in T\}$ est dense dans H_a', $G_a = {}^tF_a \circ F_a$ et ainsi :

$$\forall\, (u,v) \in H_a'^{\,2},\ (F_a u, F_a v)_{L^2(T)} = \langle u, {}^tF_a \circ F_a\, v\rangle_{H_a', H_a} = \langle u, G_a v\rangle_{H_a', H_a}$$

$$= (u,v)_{H_a'}.$$

3.3 - a.c.p. de X dans H_a'.

Les derniers paragraphes sont une application de la définition et de l'étude par BESSE 1979 de l'a.c.p. d'une variable aléatoire du second ordre et à valeurs dans un espace de Hilbert séparable. Cette a.c.p. conduit à l'analyse spectrale de l'opérateur compact $V \circ G$ où V désigne l'opérateur de

covariance de la v.a. et G l'isométrie entre l'espace de Hilbert de référence et son dual topologique. Pour appliquer ces résultats à la variable X définie au paragraphe 3.1. et à valeurs dans H_a', il suffit de montrer qu'elle est bien du second ordre :

$$E(\|X\|^2_{H_a'}) = E(\sum_{i=1}^{N} \delta_{t_i}, \sum_{j=1}^{N} \delta_{t_j})_{H_a'}$$

$$= E(\sum_{i,j=1}^{N} (G\delta_{t_i}, \delta_{t_j})_a)$$

$$= \frac{1}{2a} E(\sum_{i,j=1}^{N} e^{-a|t_i - t_j|}) \leqslant \frac{1}{2a} \sup_{\omega\in\Omega} N(\omega)^2.$$

qui est supposé fini, dans l'exemple étudié, lorsque $a \neq 0$.

Ainsi, la mesure aléatoire X est considérée comme une v.a. à valeurs dans un espace de Hilbert de distributions car c'est l'espace dual d'un espace de fonction continues. L'a.c.p. de X dans H_a' est encore appelée a.c.p. lissante (ou filtrante) car elle est semblable à celle de $F_a \circ X$ dans $L^2(T)$. En effet, le filtre F_a, qui est dans ce cas un opérateur de convolution, est bien une isométrie de H_a' dans $L^2(T)$ identifié à son dual. On obtient donc, en dimension infinie, une situation identique à celle décrite dans le paragraphe 2 en dimension finie. Il reste à montrer que l'une est bien l'approximation convergente de l'autre.

3.4 - Approximation.

On montre aisément (ZAAMOUN 1985) que la famille des fonctions $\left\{\mathbb{1}_{[T,\infty[}(1)\, e^{-a|t-\cdot|}\,;\, t\in T\right\}$ est dense dans $L^2(T)$; on peut en extraire une base dénombrable de $L^2(T)$. Soit $\{0=t_0, t_1, \ldots, t_p\}$ une suite d'instants de discrétisation qui sont, dans l'exemple cité, les dates anniversaires de mariage.

La suite des fonctions:

$$\left\{ e_i = \mathbb{I}_{[t_i, \infty[}(.)\, e^{-a|t_i - .|}, \quad i = 1, \ldots, p \right\}$$

engendre un sous-espace E_p de dimension p de $L^2(T)$ et l'opérateur défini par :

$$F_a^p = \sum_{i=1}^{p} \mathbb{I}_{[t_{i-1}, t_i[} \otimes e_i$$

est une application de $H_a^{\cdot}$ dans E_p qui à toute réalisation de la mesure aléatoire $X(\omega)$ de $H_a^{\cdot}$ associe le vecteur ["nombre d'enfants nés entre les dates t_{i-1} et t_i" ; $i = 1, \ldots, p$] exprimé dans la base $\{e_i\}$ de E_p :

$$\left[F_a^p\, X(\omega) \right]_i = \left\langle \mathbb{I}_{[t_{i-1}, t_i[}, \sum_{j=1}^{N(\omega)} \delta_{i_j}(\omega) \right\rangle .$$

Ainsi, le choix de la forme du lissage ou de "l'amortissement" est aussi le choix de la base permettant de construire le sous-espace d'approximation de $L^2(T)$.

> <u>Définition</u> : *L'approximation de l'a.c.p. de* X *dans* $H_a^{\cdot}$ *est donnée par l'a.c.p. de* $F_a^p \circ X$ *dans le sous-espace* E_p *de* $L^2(T)$.

Elle est obtenue par l'analyse spectrale d'un opérateur V^p relativement à la métrique de E_p (opérateur identité). La matrice associée à V^p s'exprime comme au paragraphe 2 tandis que, dans la base $\{e_i\}$ de E_p, celle de la métrique est donnée par Q_a^p de terme général :

$$(e_i, e_j)_{L^2(T)} = g(t_i, t_j) = \frac{1}{2a}\, e^{-a|t_i - t_j|} .$$

On retrouve bien, à une constante non significative près, la situation du paragraphe 2 qui est donc considérée comme une approximation de l'a.c.p. limite de X dans $H_a^{\cdot}$.

3.5 - <u>Convergence</u>.

a - *par discrétisation.*

Proposition 3 : *L'a.c.p. de* $F_a^p \circ X$ *dans* E_p *converge uniformément vers l'a.c.p. de* $F_a \circ X$ *dans* $L^2(T)$ *semblable à celle de X dans* H_a' *lorsque la suite* $\bigcup_{i=1}^{p} E_i$ *a une limite dense dans* $L^2(T)$.

C'est une conséquence directe d'un théorème de BESSE 1979 où la convergence uniforme est celle de la suite des opérateurs $V^p \circ Q_a^p$. Celà assure la convergence uniforme des vecteurs et valeurs propres et donc la stabilité de l'analyse en dimension finie.

La condition de densité de la limite des sous-espaces emboités E_p est assurée par une hypothèse raisonnable sur la suite $\{t_i ; i=1,\ldots,p\}$ des instants de discrétisation ; par exemple :

$$\lim_{p\to\infty} \sup_{i=1,\ldots,p} \left\{ |t_{i-1} - t_i|, \frac{1}{p} \right\} = 0.$$

b. *par discrétisation et échantillonnage.*

La structure probabiliste $(\Omega, \mathcal{A}, \mu)$ n'est utile que si l'on considère les n familles étudiées comme un échantillon pris dans une population plus vaste. C'est rarement explicitement le cas dans l'usage de "l'analyse des données" même si il est fréquent que l'interprétation des plans factoriels présente un caractère de généralité implicitement plus large que celui des n individus étudiés. La résolution des problèmes de convergence de l'a.c.p. par échantillonnage (DAUXOIS-POUSSE 1976) a ouvert la voie à des études asymptotiques (DAUXOIS et al. 1982) et BESSE 1979 a montré la convergence de la double approximation par discrétisation et échantillonnage.

3.6. Le cas a=0.

Les situations présentées dans les paragraphes 3.2 à 3.5 ne s'appliquent pas lorsque a est nul, c'est-à-dire pour l'a.c.p. notée (X, Q_0, D) précédemment. Dans ce cas, il suffit de suivre la même démarche en choisissant pour intervalle $T=[0, b]$ et, pour

espace de référence H_0' le dual topologique de

$$H_0 = \{f \in L^2(T) \ ; \ f' \in L^2(T) \text{ et } f(b) = 0\}$$

muni du produit scalaire :

$$\forall (f,h) \in H_0^2 \qquad (f,h)_0 = \int_0^b f'(x) h'(x) dx.$$

Il admet pour noyau reproduisant :

$$g(s,t) = \langle \mathbb{I}_{[0,.]}(s), \ \mathbb{I}_{[0,.]}(t) \rangle = \int_0^b \mathbb{I}_{[s,b]}(x) \ \mathbb{I}_{[t,b]}(x) dx$$
$$= b - \sup(s,t)$$

et est isométrique de $L^2(T)$ par l'opérateur F_0 :

$$\forall s \in T, \ \forall u \in H_0' \qquad F_0 u(s) = \langle u, \ \mathbb{I}_{[0,.]}(s) \rangle = \langle u, \ \mathbb{I}_{[s,b]}(.) \rangle$$

On retrouve les mêmes résultats d'approximation et de convergence en utilisant le sous-espace de $L^2(T)$ engendré par la base des indicatrices :

$$\left\{ e_i = \mathbb{I}_{[t_{i-1},b]}(.) \ ; \ i=1,\ldots,p \right\}.$$

4 - CONCLUSION.

L'objectif principal de cet article, comme celui sans doute d'autres articles de ce recueil, est de montrer qu'au delà de l'utilisation de l'a.c.p. classique par l'exécution systématique de logiciels standards, l'utilisateur a la possibilité d'adapter cette technique à la nature de ses données ou à ses objectifs. Cette flexibilité de la méthode peut être obtenue en réfléchissant explicitement au rôle de la métrique mesurant les distances entre individus car l'usage systématique, et souvent implicite, de la métrique identité n'est en rien justifié a priori.

Il semble que, dans le cadre de l'étude de données évolutives, les métriques induites par un espace de Sobolev (BESSE, RAMSAY 1986) ou encore, comme dans ce papier, par le dual d'un espace de Sobolev offrent une étendue et une souplesse de choix qui méritent une investigation plus approfondie. Cette étude a été présentée dans le contexte limité de l'a.c.p. mais l'usage des métriques proposées peut être transposé sans

difficulté à toute technique pouvant utiliser une structure euclidienne et, en particulier, à certaines méthodes classiques de classification.

Finalement c'est la grande variété des choix possibles qui pose un problème ou, plutôt, c'est l'absence de critère permettant d'objectiver ce choix. Il est clair qu'un premier groupe de critères concerne essentiellement l'utilisateur c'est-à-dire la connaissance qu'il a a priori de ses données, de ses objectifs (ici, par exemple, le choix de la forme de "l'amortissement"). Mais, d'autres critères doivent être nécessairement liés à la qualité, la fiabilité des résultats (ici le choix de a) et sont donc spécifiques à la méthode. Ce dernier point a été déjà largement évoqué dans la littérature (test de sphéricité, rééchantillonnage) à propos du choix de la dimension de "l'espace latent" à partir de la répartition des valeurs propres. Une ouverture possible consiste à réunir ces deux questions (dimension et métrique), dans l'unique problématique d'un choix de modèle (CAUSSINUS 1985, BESSE et al. 1986).

REFERENCES.

ARONSZAJN, N., 1950 "Theory of reproducing kernel". Trans. Amer. Math. soc. , Vol.68, 337-404.

BESSE PH., 1979 "Etude descriptive d'un processus ; approximation, interpolation". Thèse de 3ème cycle, Toulouse III.

BESSE Ph., 1980 "Deux exemples d'a.c.p. filtrante". Stat. et Analyse des données n°3, 1980, 5-15.

BESSE Ph., CAUSSINUS H., FERRE L., FINE J., 1986 "Sur l'utilisation optimale de l'Analyse en Composantes Principales". Note C.R.A.S. t.304, Série I, n°15, 459-462.

BESSE PH., RAMSAY J., 1986 "Principal Components Analysis of sampled functions", Psychometrika, 51, 285-311.

CAUSSINUS H., 1985 "Quelques réflexions sur la part des modèles probabilistes en analyse des données", 4ème Journées

internationales, Anal. des données et inform., Versailles.

DAUXOIS J., POUSSE A., 1976 "Les analyses factorielles en calcul des probabilités et en statistique: essai d'étude synthétique", thèse, TOULOUSE III.

DAUXOIS J., POUSSE A., ROMAIN Y., 1982 "Asymptotic theory for the p.c.a. of a vector random function : some applications to statistical inferences". Journal of Multivariate Analysis, 12, 136-154.

DEVILLE J.C., 1974 "Méthodes statistiques et numériques de l'analyse harmonique". Annales de l'INSEE, n°15 janvier-avril.

DEVILLE J.C., 1977 "Un exemple catastrophique d'analyse factorielle et son explication". Colloque INRA, Analyse des données et inform.

SCHWARTZ L., 1964 "Sous-espaces hilbertiens d'espaces vectoriels topologiques et noyaux reproduisants", J., Anal., Math., Vol. 13.

ZAAMOUN S., 1985 "A.c.p. de mesures aléatoires". Mémoire de DEA, Toulouse III.