Statistique et analyse des données

Philippe Besse

Deux exemples d'analyse en composantes principales filtrantes

Statistique et analyse des données, tome 5, n° 3 (1980), p. 5-15.

<http://www.numdam.org/item?id=SAD_1980__5_3_5_0>

Statistique et Analyse des Données
3 - 1980 pp. 5-15

DEUX EXEMPLES D'ANALYSE EN COMPOSANTES PRINCIPALES FILTRANTES

Philippe BESSE

Laboratoire de Statistique et Probabilités
E.R.A. - C.N.R.S. n° 591
Université Paul Sabatier - 31077 TOULOUSE Cedex

INTRODUCTION

X étant une fonction aléatoire sur un espace probabilisé $(\Omega, \mathcal{A}, P)$ et connue aux instants $t_1,\ldots,t_p$ de discrétisation, on considère les p variables aléatoires $\{X_{t_i} ; i=1,\ldots,p\}$. Les individus éléments de Ω sont alors considérés comme des vecteurs de $\mathbb{R}^p$ muni de la métrique euclidienne classique. La pratique de l'Analyse en Composantes Principales (A.C.P.) qui en découle pose alors quelques problèmes :

La stabilité est assurée en montrant que l'analyse est une approximation convergente de l'A.C.P. théorique, dans l'espace de Hilbert $L^2(T)$, de la fonction aléatoire X ([4],[3]). Mais la métrique alors considérée, attribuant un même rôle à chaque variable, rend l'analyse insensible à l'aspect évolutif du phénomène.

Cette remarque conduit à l'emploi d'autres structures hilbertiennes (espaces de Sobolev) adaptables à l'exemple étudié en fonction des hypothèses de régularité admissibles sur X. Les problèmes d'interpolation des trajectoires de X et d'approximation de l'analyse sont résolus, tandis que l'interprétation des plans factoriels, rendue délicate par l'utilisation d'une métrique non classique, est facilitée par la notion d'A.C.P. filtrante.

Le texte, tout en résumant la démarche de [2], propose de généraliser la notion d'A.C.P. d'une variable aléatoire par celle d'A.C.P. d'une fonctionnelle aléatoire linéaire dans certains cas où X ne vérifie pas des hypothèses suffisantes.

1 - A.C.P. EN GEOMETRIE HILBERTIENNE.

1.1 - A.C.P. d'une variable aléatoire.

Φ désigne un espace de Hilbert supposé séparable et Φ' son dual topologique ($(.,.)_\Phi$ et $<.,.>_{\Phi,\Phi'}$ désignent respectivement le produit scalaire et la dualité sur Φ). Soit $(\Omega, \mathcal{A}, P)$ un espace probabilisé sur lequel est définie une variable aléatoire (v.a) X du second ordre, centrée et à valeurs dans Φ:

$$E(X) = 0 \qquad \text{et} \qquad X \in L^2(\Omega, \mathcal{A}, P ; \Phi, \mathcal{B}_\Phi) \quad (\text{c-à-d. : } E(\|X\|_\Phi^2) < \infty)$$

On considère la fonctionnelle aléatoire linéaire (f.a.ℓ.) notée U^* (cf [1]; p.89), application de Φ' dans $\text{Mes}[\Omega, \mathcal{A}; \mathbb{R}, \mathcal{B}_\mathbb{R}]$ qui à tout élément u de Φ' associe la v.a.r. :

$$U^*u = < X, u >.$$

On montre alors que U^* est un opérateur linéaire de Hilbert-Schmidt (H.S.) de Φ' dans $L^2(\Omega, \mathcal{A}, P)$ et son adjoint U (également de H.S.) est l'opérateur de $L^2(\Omega, \mathcal{A}, P)$ dans Φ défini par :

$$\forall\, f \in L^2(\Omega, \mathcal{A}, P), \; Uf = E(Xf) \; .$$

Le produit $V = U \circ U^*$ est alors un opérateur nucléaire, auto-adjoint et positif : c'est l'opérateur de covariance de X. On a le schéma de dualité :

$$\begin{array}{ccc} \Phi & \xleftarrow{\quad U \quad} & L^2(\Omega, \mathcal{A}, P) \\ {\scriptstyle A}\downarrow\uparrow{\scriptstyle V} & & \uparrow{\scriptstyle I} \\ \Phi' & \xrightarrow{\quad U^* \quad} & L^2(\Omega, \mathcal{A}, P) \end{array}$$

(A désigne l'isomorphisme canonique entre Φ et Φ').

L'A.C.P. de X, qui se définit comme étant la recherche de l'élément normé de Φ' maximisant $\|U^*u\|_{L^2(\Omega, \mathcal{A}, P)}$, est obtenue par l'analyse spectrale de l'opérateur auto-adjoint, positif et nucléaire $A \circ V$.

1.2 - A.C.P. d'une f.a.ℓ.

D'une façon plus générale, on considère une f.a.ℓ. U^* de Φ' dans $L^2(\Omega, \mathcal{A}, P)$ telle que l'opérateur U^* soit borné mais non nécessairement compact. On définit alors l'A.C.P. de la f.a.ℓ. U^* comme étant l'A.C.P. de l'opérateur linéaire borné U, adjoint de U^* (cf. [3], p. 148 et suivantes). Au moins dans le cas où U^* est de Hilbert-Schmidt, la f.a.ℓ. admettant un opérateur de covariance $V = U \circ U^*$ nucléaire, se réduit à une v.a. X par la formule (cf. [1], p.101) :

$$\forall\, u \in \Phi' \;, \qquad U^*u(\omega) = < u, X(\omega) >_{\Phi',\Phi} \qquad P\text{-p.s.}$$

et l'A.C.P. de U^* est équivalente à l'A.C.P. de la v.a. X. Mais, comme dans l'exemple du § 4, si la f.a.ℓ. ne se réduit pas à une v.a., l'analyse en composantes principales d'une fonctionnelle aléatoire linéaire est une généralisation de l'analyse d'une variable aléatoire.

1.3 - A.C.P. filtrante.

Soit F un élément de $\mathcal{L}(\Phi,\psi)$ où ψ est un autre espace hilbertien supposé séparable. Si X est une v.a. à valeurs dans Φ, centrée et du second ordre, FX est une v.a. à valeurs dans ψ, également centrée et du second ordre ; l'A.C.P. de FX dans ψ est alors fournie par l'analyse spectrale de l'opérateur auto-adjoint, positif, nucléaire $B \circ F \circ V \circ F^*$ où $F \circ V \circ F^*$ désigne l'opérateur de covariance de FX et B l'isométrie canonique entre ψ et ψ'. Dans le cas où F est une isométrie entre Φ et ψ, l'A.C.P. de X dans Φ et l'A.C.P. de FX dans ψ sont dites semblables car elles conduisent à des représentations identiques dans les plans factoriels.

En pratique, ψ est l'espace $L^2(T)$ (où T est un ouvert de $\mathbb{R}^n$) identifié à son dual topologique (B=I) et pour interpréter l'A.C.P. d'une v.a. X dans un espace Φ (un espace de Sobolev) on cherche l'application F qui est une isométrie entre Φ et $L^2(T)$. Inversement, F étant fixé, on détermine l'espace Φ qui fait de F une isométrie et il est plus simple, pour connaître l'A.C.P. de FX dans $L^2(T)$, de calculer directement l'A.C.P. (appelée filtrante) de X dans Φ.

Exemple : $\Phi = H^s(\mathbb{R}^n)$, $\psi = L^2(\mathbb{R}^n)$ ($s \in \mathbb{R}$, $\mathbb{R}^n$ est muni de la mesure de Lebesgue).

Par construction des espaces $H^s(\mathbb{R}^n)$, (cf.[7])

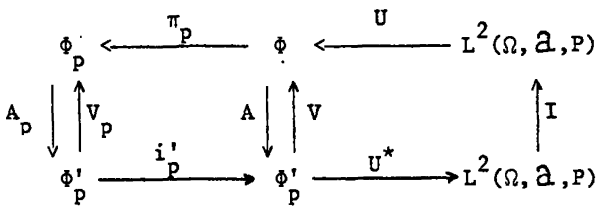
$$F : u \in H^s(\mathbb{R}^n) \to Fu = \mathcal{F}^{-1}\left[\mathcal{F} u \, (1 + |\lambda|^2)^{s/2}\right]$$

est une isométrie entre $H^s(\mathbb{R}^n)$ et $L^2(\mathbb{R}^n)$. Pour s = -2 on reconnait en F le filtre de Wiener utilisé pour débruiter un signal.

2 - APPROXIMATION

2.1 - Définition

Soit Φ_p (resp. Φ'_p) un sous-espace de Φ (resp. de Φ') (dim Φ_p = dim Φ'_p = p) muni de la structure hilbertienne induite ; π_p, i_p (resp. π'_p, i'_p) désignent le projecteur orthogonal et l'injection canonique entre Φ et Φ_p (resp. Φ' et Φ'_p). L'approximation de l'A.C.P. de X dans Φ est la recherche d'un élément de norme 1 de Φ'_p tel que $U^* \circ i_p u$ soit de variance maximale et itérations sous contraintes d'orthonormalité dans Φ'_p ; c'est encore l'analyse de la v.a. $\pi_p \circ X$ dans Φ_p qui conduit au schéma de dualité :



$A_p = \pi'_p \circ A \circ i_p$

$V_p = \pi_p \circ V \circ i'_p$

L'approximation est fournie par l'analyse spectrale de l'opérateur auto-adjoint, positif et de rang fini $A_p \circ V_p$.

2.2 - Convergence

Afin de s'assurer de la stabilité de cette méthode, on doit montrer la convergence uniforme des facteurs et valeurs principales qui est conséquence de la convergence uniforme de l'opérateur $i'_p \circ A_p \circ V_p \circ \pi'_p$ vers l'opérateur $A \circ V$ dans Φ'_p. On montre alors le

Théorème : *Si $\{\Phi_p\}_{p\in\mathbb{N}}$ est une suite croissante de sous-espaces de Φ munis de la structure hilbertienne induite et telle que :*

$$\dim \Phi_p = p, \qquad \bigcup_{p\in\mathbb{N}} \Phi_p \text{ dense dans } \Phi ,$$

alors l'analyse en composantes principales approchée de X dans Φ_p converge uniformément vers l'analyse de X dans Φ.

Remarque : lorsque l'approximation ci-dessus est associée à une approximation par échantillonnage statistique de Ω (cf. [3], p.266), on obtient une double approximation de l'A.C.P. de X dans Φ (cf. [2], p. 31) qui converge également uniformément.

2.3 - Cas des espaces de Sobolev

Soit $T =]a,b[$ un ouvert de $\mathbb{R}$; on considère le cas où Φ est un espace de Sobolev sur T (resp. le dual d'un espace de Sobolev sur T). Φ (resp. Φ') est un sous-espace hilbertien de $\mathbb{R}^T$ et admet donc un noyau reproduisant noté $g(t,t')$. Soit $\Delta_p = \{t_1,\ldots,t_p\}$ une subdivision d'éléments distincts de T ; on construit une suite $\{\Phi_p\}_{p\in\mathbb{N}}$ (resp. $\{\Phi'_p\}_{p\in\mathbb{N}}$) de sous-espaces d'approximation de la façon suivante :

$$\Phi_p (\text{resp.}\Phi'_p) = S_p = \text{Vect.}\{ g(t_i,.) \ ; \ i=1,\ldots,p \}$$

et la subdivision Δ_p est affinée de sorte que

$$\lim_{p\to\infty} \sup(t_1 - a,\ b - t_p\ ,\ t_{i+1} - t_i \ ; \ i=1,\ldots,p-1) = 0 .$$

S_p (cf.[6]) est l'espace des fonctions "spline" d'interpolation aux noeuds $t_1,\ldots,t_p$ c'est-à-dire que le projecteur orthogonal π_p associe à tout f de Φ (resp. Φ') l'élément $\pi_p f$ de Φ (resp Φ') qui prend les mêmes valeurs que f aux instants $t_1,\ldots,t_p$, et est le plus lisse en un certain sens. De plus, la suite $\{S_p\}_{p\in\mathbb{N}}$ vérifie les hypothèses du théorème de convergence.

On montre alors que l'approximation cherchée est fournie par l'analyse spectrale de la matrice $\Lambda_p^{-1} \circ V_p$ (resp. $\Lambda_p \circ V_p$ dans le cas où Φ est le dual d'un espace de Sobolev) avec

$$\Lambda_p = [g(t_i,t_j)]_{1\le i,j\le p} \quad , \quad V_p = [K(t_i,t_j)]_{1\le i,j\le p} \quad ,$$

où g est le noyau reproduisant de l'espace de Sobolev considéré et K la fonction de covariance de X. La réalisation pratique est immédiate à partir des programmes connus d'A.C.P..Deux exemples simples sont présentés dans les paragraphes suivants afin d'illustrer la démarche adoptée.

3 - COURBES DE TEMPERATURE

3.1 - Les données.

On considère une population de 32 villes françaises dont on connait les moyennes des températures mensuelles, c'est-à-dire les valeurs prises par 12 variables notées $T_1,\ldots T_9$,TO,TN,TD qui se succèdent à des intervalles de temps considérés égaux dans T = [0,1] (il s'agit en fait d'un processus "moyenne mobile" : $\overline{X}_t$ = moyenne de X_t sur un intervalle de trente jours centrés en t).

3.2 - A.C.P. dans $L^2(0,1)$.

Après centrage des données, l'approximation de l'A.C.P. dans $L^2(0,1)$ conduit (cf. [2] , p.25) à l'emploi de la métrique euclidienne classique. On retient les trois premiers vecteurs propres auxquels sont associés les "variances expliquées" :

	i=1	i=2	i=3
$100 \times \frac{\lambda_i}{\sum_{j=1}^{12} \lambda_j}$	87,0	12,1	0,378

Le premier plan factoriel (planche 1), expliquant 99% de la variance, est une très bonne représentation du nuage tandis que le troisième axe (0,4%) présente peu d'intérêt. Le premier axe discrimine fortement et sans nuances les villes froides des villes chaudes (axe Nord-Sud) alors que le second axe discrimine faiblement les villes tempérées des villes à climat plus continental (axe Ouest-Est). Mais, ce qui apparaît avec le plus de précision, c'est une "corrélation circulaire" des variables car chaque variable est fortement corrélée à celle du mois précédent et à celle du mois suivant. Il s'agit alors, par un choix judicieux de métrique sur l'espace des individus, de tenir compte des proximités dans le temps des variables et ainsi de supprimer ces corrélations triviales afin d'obtenir une représentation plus fine des données.

3.3 - A.C.P. dans $H^1(0,1)$.

Afin de comparer les deux analyses, les variables sont maintenant considérées comme les résultats de mesures mensuelles à des instants distincts. On suppose alors que les courbes obtenues sont des éléments de l'espace de Sobolev $H^1(0,1)$ (on suppose les trajectoires absolument continues) qui, lorsqu'il est muni du produit scalaire :

$$(u,v)_1 = \int_0^1 u'(t)\, v'(t)dt + u(0)\, v(0) ,$$

admet pour noyau reproduisant :

$$g(t,t') = 1 + \min(t,t') .$$

L'emploi de cette métrique, qui est représentée dans la base $\{g(t_i,.)\ ;\ i=1,\ldots,p\}$ de Φ_p par la matrice inverse de la matrice de terme général $g(t_i,t_j)$, est équivalent à une transformation des données. Les variables représentent alors (sauf aux bornes) les pentes des segments de droite qui interpolent les trajectoires (cf. [2], p.77). Cette métrique a un effet dispersif sur les "variances expliquées" qui deviennent :

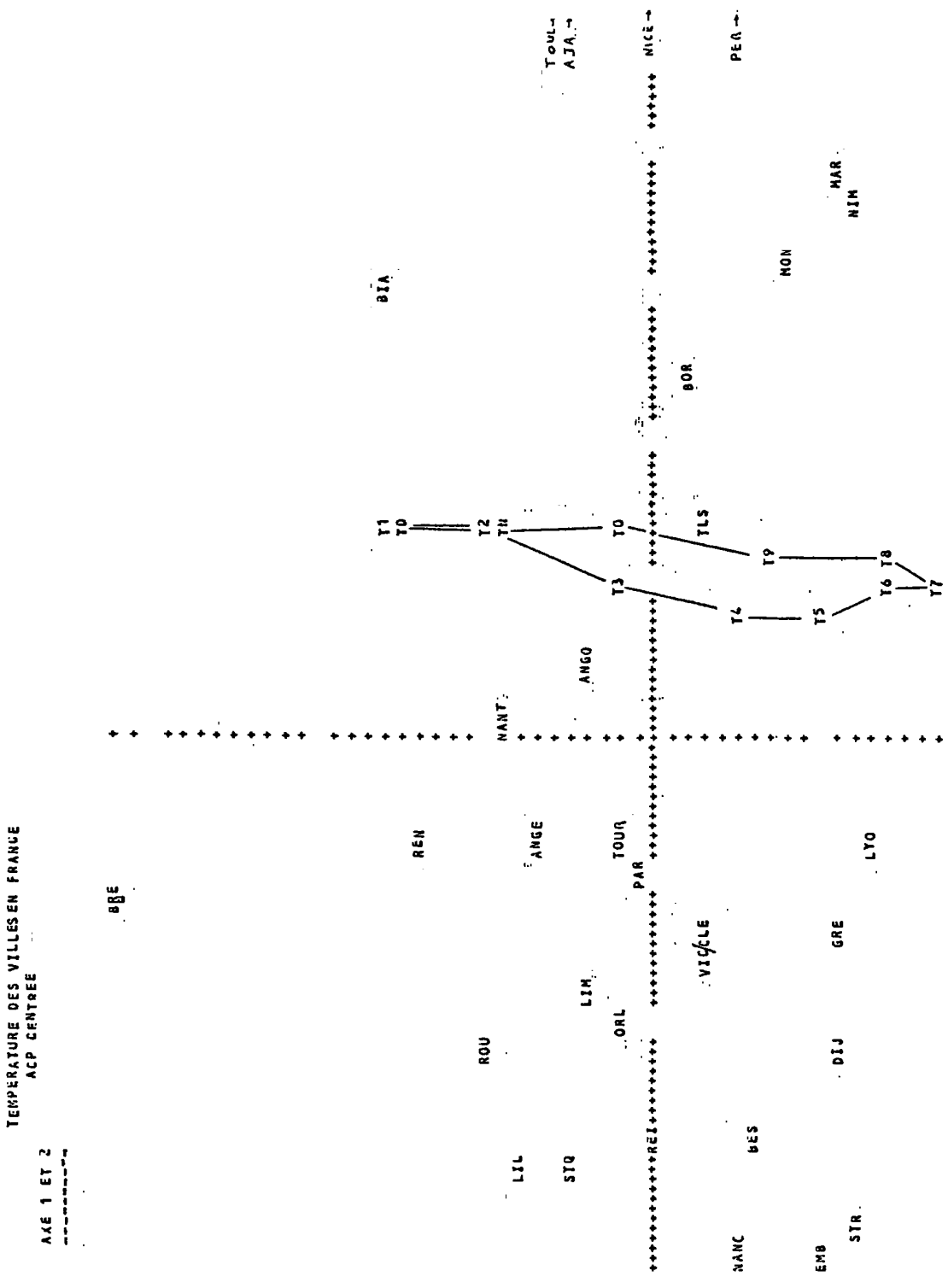

TEMPERATURE DES VILLES EN FRANCE
ACP CENTREE
AXE 1 ET 2
BRE
REN
NANT
ANGE
ANGO
TOUR
PAR
REI
ROU
LIL
STQ
LIM
ORL
VICHE
BES
NANC
EMB
DIJ
GRE
LYO
STR
BIA
BOR
TLS
NON
MAR
NIM
NICE
PER
T0
T1
T2
T3
T4
T5
T6
T7
T8
T9
TH

Planche 1

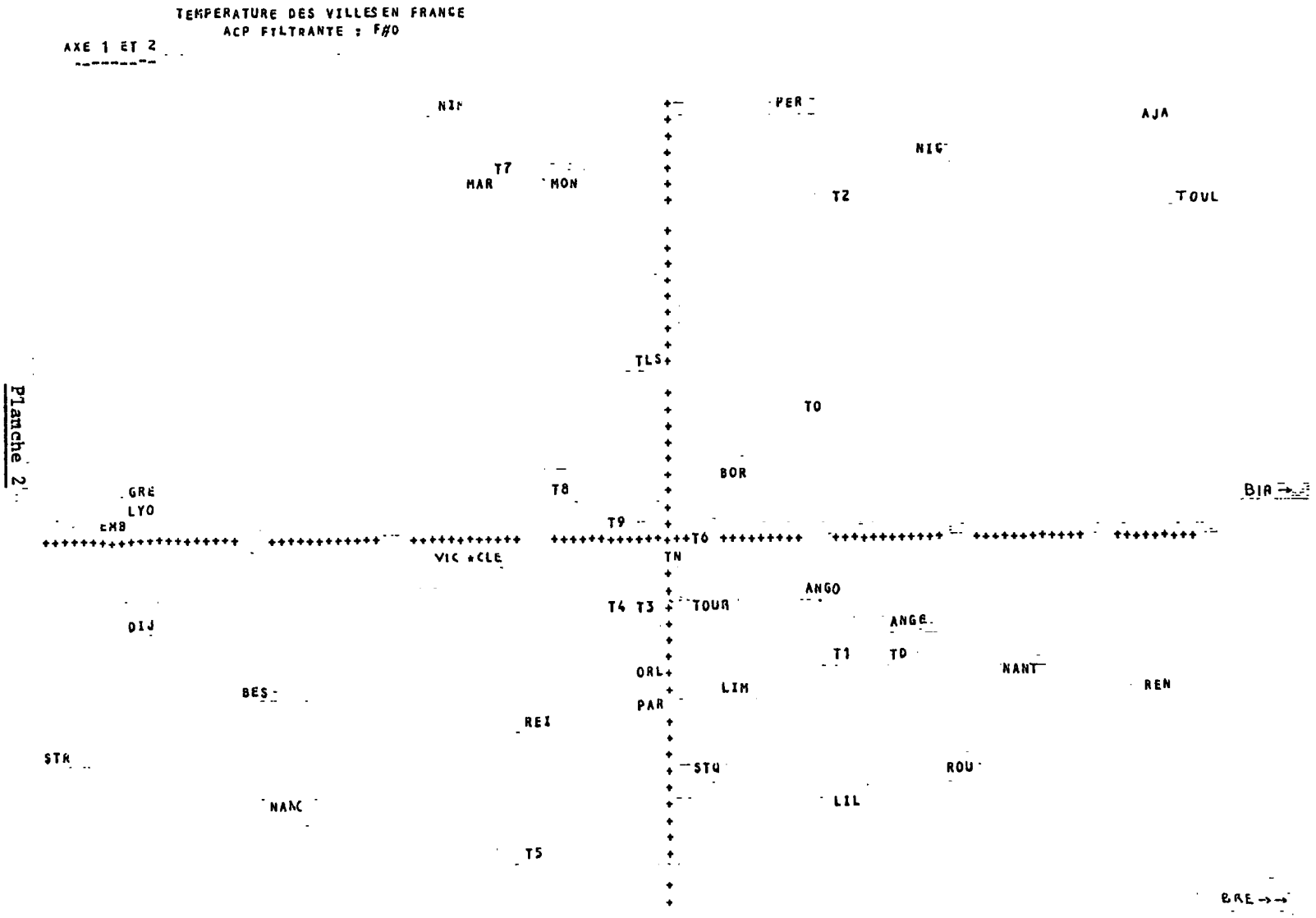

TEMPERATURE DES VILLES EN FRANCE
ACP FILTRANTE : F#0
AXE 1 ET 2
NIM
PER
AJA
NIC
T7
MAR
MON
T2
TOUL
TLS
T0
BOR
T8
BIA
GRE
LYO
EMB
T9
T6
VIC
CLE
TN
ANGO
T4
T3
TOUR
ANGE
DIJ
T1
TD
NANT
ORL
LIM
REN
BES
PAR
REI
STR
STQ
ROU
NANC
LIL
T5
BRE

Planche 2

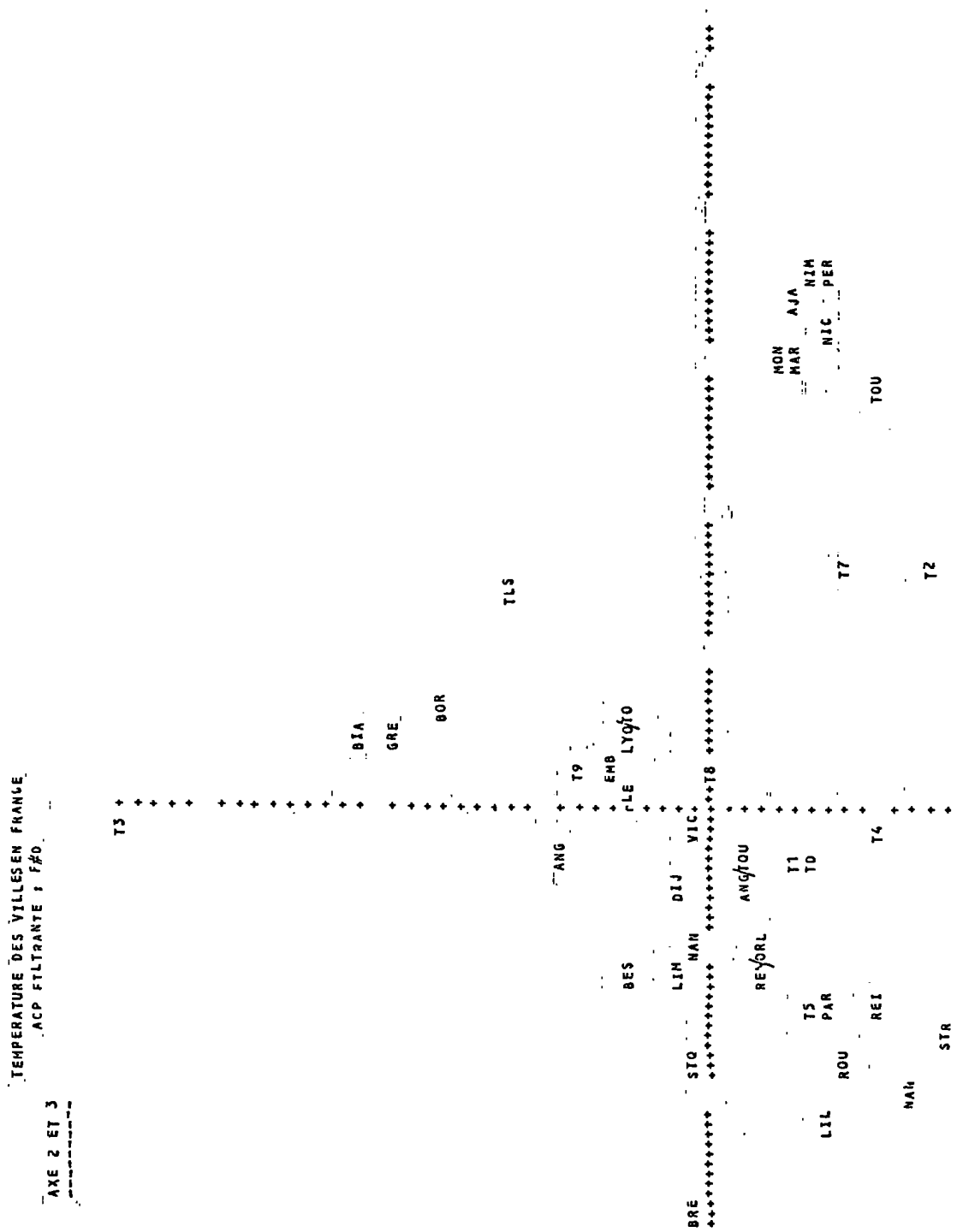

TEMPERATURE DES VILLES EN FRANCE
ACP FILTRANTE : F#0
AXE 2 ET 3
T3
BIA
GRE
BOR
TLS
ANG
T9
EMB
LE
LYOTO
T8
VIC.
DIJ
NAN
LIM
BES
STQ
BRE
REYORL
ANGTOU
T1
TO
T5
PAR
LIL
ROU
REI
T4
MAH
STR
T7
T2
MON
MAR
AJA
NIM
NIC
PER
TOU

Planche 3

	i=1	i=2	i=3
$100 \times \frac{\lambda_i}{\sum_{j=1}^{12} \lambda_j}$	68,7	17,5	4,82

Il faut cette fois considérer trois axes (planches 3 et 4) pour obtenir 91% de la variance et les deux premiers axes ont subi une permutation : l'axe 1 est devenu l'axe Est-Ouest, l'axe 2 est l'axe Nord-Sud. Le premier plan factoriel donne alors une représentation assez fidèle de la carte de la France tandis que le troisième axe complète la caractérisation des villes intermédiaires dont l'accroissement de la température se fait sentir dès le mois de mars.

Ainsi, en éliminant les corrélations triviales, l'A.C.P. dans $H^1(0,1)$ présente un compromis entre les données multidimensionnelles initiales (p=12) et la représentation trop linéaire de l'A.C.P. dans $L^2(0,1)$ et, présentée comme une A.C.P. filtrante, l'interprétation des plans factoriels reste simple.

4 - CALENDRIER DE CONSTITUTION DES FAMILLES

On propose dans ce paragraphe une relecture d'une étude présentée en [4] et en [5] à l'aide des notions décrites précédemment.

4.1 - Le modèle

La formalisation du problème proposée en [5] commence de la façon suivante :

" Sur chacune des n familles observées pendant 20 années de mariage, on connait un vecteur Z de dimension 20 où la $i^{ème}$ coordonnée représente le nombre d'enfants nés dans la famille entre les anniversaires de mariage (i-1) et i ; Z est alors constitué des valeurs prises par la mesure X, définie ci-dessous, sur des intervalles](i-1)/20, i/20] du segment T = [0,1].

A chaque famille ω de Ω , on associe un entier $N(\omega)$ qui est le nombre total d'enfants qu'a cette famille et la suite $\{t_i(\omega) ; i=1,\ldots,N(\omega)\}$ des dates de naissance des enfants. En associant à chaque naissance une masse de Dirac, on construit une mesure aléatoire :

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

qui vérifie pour toute fonction continue sur T :

$$< X(\omega), u > = \sum_{i=1}^{N(\omega)} u(t_i) \quad . \quad "$$

X est ainsi une v.a. à valeurs dans l'espace de Banach des mesures de Radon sur T, mais comme l'A.C.P. utilise les propriétés d'orthogonalité, l'étude doit nécessairement se situer dans un cadre hilbertien. La démarche suivante est donc adoptée :

- définir l'A.C.P. de X comme étant celle d'une v.a. à valeurs dans un espace de Hilbert Φ .

- comme X n'est pas une v.a. à valeurs dans $L^2(T)$, l'A.C.P. de X par rapport à la structure de $L^2(T)$ est définie comme étant celle d'une fonctionnelle aléatoire linéaire.

4.2 - A.C.P. dans $(H^1(0,1))'$

L'A.C.P. de X est donc définie dans un espace de Hilbert qui doit être, pour contenir les éléments $X(\omega)$, le dual d'un espace de fonctions continues. L'espace hilbertien convenable et le plus pratique est le dual Φ_1' de l'espace :

$$\Phi_1 = \left\{ u \in H^1(0,1) \ ; \ u(v) = 0 \right\}$$

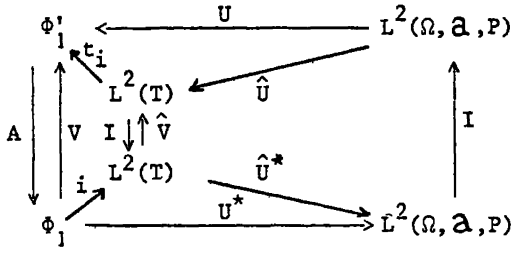
de produit scalaire :

$$(u,v)_2 = \int_0^1 u'(t)\, v'(t)dt \ .$$

On montre alors ([2] p.86) que X est bien du second ordre et admet donc un opérateur de covariance nucléaire et, après centrage, on peut en faire l'A.C.P. dans $(H'(0,1))'$ qui revient donc à calculer l'analyse de FX dans $L^2(T)$ (F est l'opérateur d'intégration).

4.3 - A.C.P. dans $L^2(T)$

Le schéma de dualité de l'A.C.P. ci-dessus peut se compléter de la façon suivante :



$V = U \circ U^*$

$\hat{V} = \hat{U} \circ \hat{U}^*$

L'injection canonique de Φ_1 dans $L^2(T)$ étant une application continue et à image dense, l'opérateur U^* de noyau X se prolonge de manière unique en un opérateur continu $\hat{U}^*$ de $L^2(T)$ dans $L^2(\Omega, \mathcal{A}, P)$ dont l'adjoint est un opérateur linéaire continu noté $\hat{U}$ de $L^2(\Omega, \mathcal{A}, P)$ dans $L^2(T)$. La f.a.ℓ. ainsi construite, dont on considère l'A.C.P., permet donc de donner un sens à la notion "d'A.C.P. de la v.a. X" relativement à la structure hilbertienne de $L^2(T)$. L'opérateur $\hat{U}^*$ n'ayant pas de raison d'être de Hilbert-Schmidt et l'opérateur de covariance $\hat{V}$ de la f.a.ℓ. $\hat{U}^*$ n'étant pas compact (cf.[5]), une telle analyse ne peut donc pas être limite uniforme d'une suite d'analyses en dimension finie.

En [5], deux analyses utilisant la métrique euclidienne classique sont comparées. La première est calculée sur des variables représentant le nombre d'enfants nés entre les dates i-1 et i ; elle apparait donc comme étant une approximation (non convergente) de l'A.C.P. de la f.a.ℓ. définie ci-dessus et fournit des résultats instables (ils dépendent du découpage de T et de l'échantillon). La deuxième analyse considère le nombre d'enfants nés à la date i ; c'est donc une approximation (convergente) de l'A.C.P. de X dans l'espace Φ_1'.

Contrairement à l'exemple précédent, la métrique de l'espace $\Phi_1^!$ employée concentre la "variance expliquée" sur le premier axe (16 à 93%) (cf.[5]). Cette concentration amenant une représentation très linéaire des données, il serait sans doute intéressant de calculer l'analyse dans un espace intermédiaire.

CONCLUSION :

Pour s'assurer de la stabilité de l'analyse en composantes principales d'un processus, il a fallu en étudier (cf.[3]) la convergence vers l'analyse théorique calculée dans une structure hilbertienne commode ($L^2(T)$). Ce travail étend ce résultat à d'autres espaces de Hilbert et les exemples présentés montrent bien que le choix des variables décrivant le phénomène, équivalant à un choix de métrique, est fondamental quant à la qualité des représentations obtenues. L'étape suivante conduira sans doute à plonger ces structures hilbertiennes dans l'espace vectoriel topologique de référence (implicitement l'espace des distributions) et à rechercher des critères permettant de sélectionner les métriques optimales, selon les exemples traités, tout en conservant (par l'expression du filtre) les possibilités d'interprétation.

BIBLIOGRAPHIE

[1] BENSOUSSAN A. - Filtrage optimal des systèmes linéaires. Dunod, Paris, 1971.

[2] BESSE P. - Etude descriptive d'un processus. Approximation et interpolation. Thèse de 3ème cycle, Toulouse, novembre 1979.

[3] DAUXOIS J., POUSSE A. - Les analyses factorielles en calcul des Probabilités et en Statistique : essai d'étude synthétique. Thèse, Toulouse, 1976.

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

[5] DEVILLE J.C. - Un exemple catastrophique d'analyse factorielle et son explication. Colloque IRIA. Analyse des données et informatique T2, septembre 1977.

[6] DUC-JACQUET M. - Approximation des fonctionnelles linéaires sur des espaces hilbertiens auto-reproduisants. Thèse, Grenoble, 1973.

[7] VO-KHAC-KOAN - Distributions, analyse de Fourier, opérateurs aux dérivées partielles, T.2. Vuibert, Paris, 1972.