

Fouille de données fonctionnelles

Gilbert Saporta

Chaire de Statistique Appliquée

Conservatoire National des Arts et Métiers

292 Rue Saint Martin

75141 Paris Cedex 03

gilbert.saporta@cnam.fr

INTRODUCTION

● Premiers travaux:

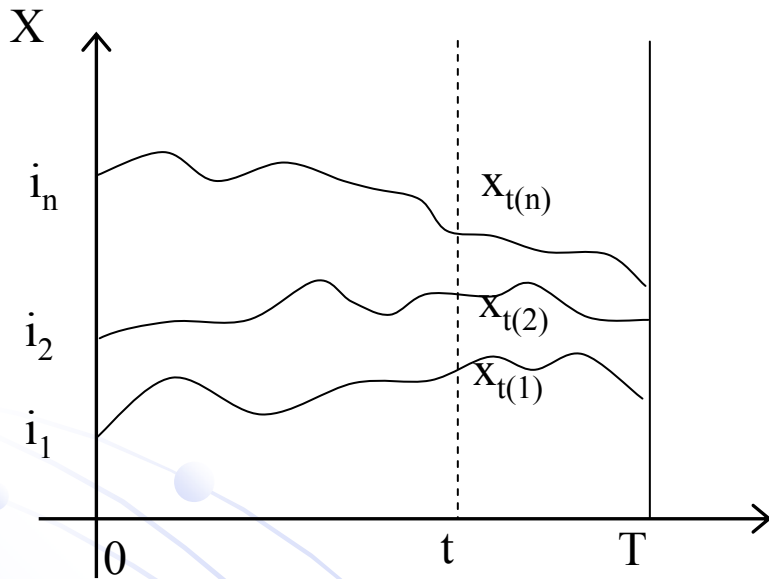
- J. C. Deville – 1974
- P. Besse – 1979
- G. Saporta – 1981

• Ensuite...

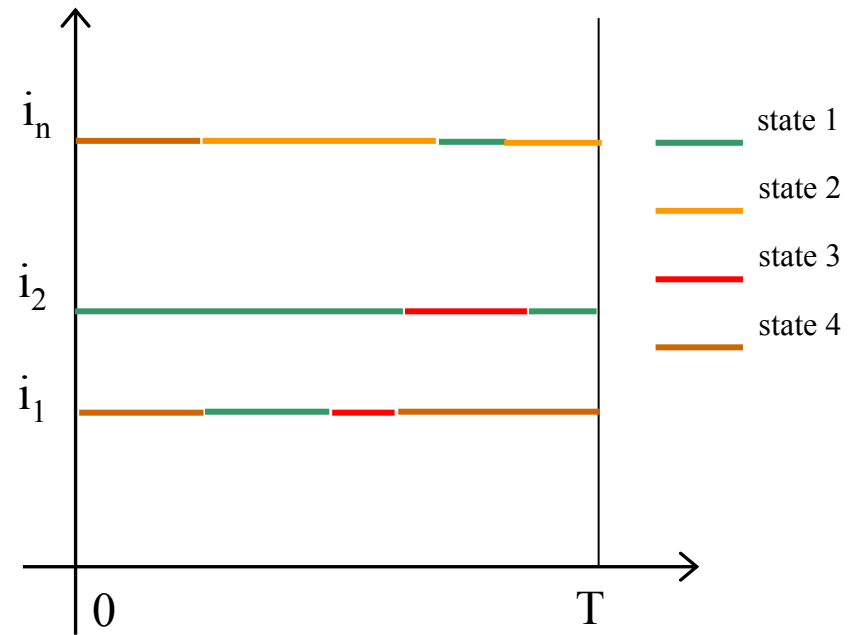
- Aguilera, Valderrama – 1993, 1995, 1998
- Ramsay, Silverman – 1995, 1997
- Van der Heijden – 1997
- Preda, Cohen – 1999
- Cardot, Ferraty, Vieu - 1999, 2005

Données fonctionnelles

NUMERIQUES



QUALITATIVES



« event history data »

PROCESSUS NUMERIQUE

Exemples:

- Taille d'une famille t années après le mariage
- Valeur boursière

Pour chaque t

variable numérique:

$$x_t = \begin{pmatrix} x_t(1) \\ \cdot \\ \cdot \\ \cdot \\ x_t(n) \end{pmatrix}$$

Infinité non dénombrable de variables si $t \in [0;T]$

PROCESSUS QUALITATIF

Exemples:

- Phases du sommeil
- Statut social
- Statut matrimonial

A chaque instant t
variable nominale x_t à m catégories.

I ACP FONCTIONNELLE

X_t centré $X_t \in L^2(\Omega \times T)$

➤ fonction de covariance: $C(t, s) = E(X_t X_s)$

➤ Opérateur de covariance C

$$f(s) \rightarrow \int_0^T C(t, s) f(s) ds$$

➤ Combinaison linéaire

$$\xi = \int_0^T f(t) X_t dt$$

Opérateurs linéaires associés à X_t

$$L^2(T) \rightarrow L^2(\Omega)$$

➤ U

$$f(t) \rightarrow \int_0^T X_t f(t) dt$$

➤ U^*

$$L^2(\Omega) \rightarrow L^2(T)$$
$$Y \rightarrow E(YX_t)$$

➤ $U^* \circ U = C$

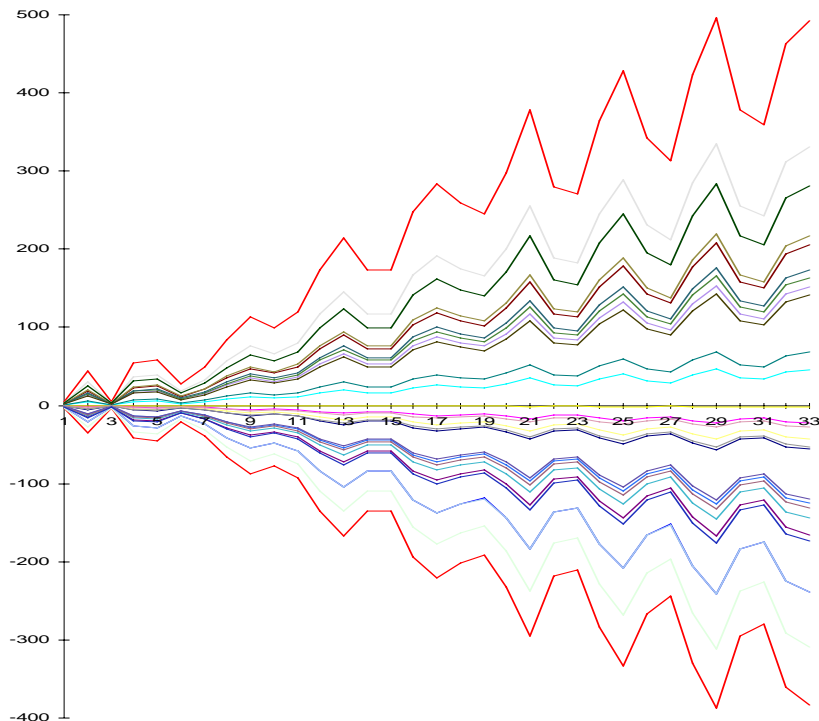
➤ $U \circ U^* = W$

$$L^2(\Omega) \rightarrow L^2(\Omega) \quad Y \rightarrow \int_0^T X_t E(X_t Y) dt$$

I.1 Processus quasi – déterministe

$$X_t(i) = \xi_i f(t)$$

Même forme à une constante près ξ_i relative à l'individu i



Tout processus peut être approché par une somme de processus quasi déterministes.

$$X_t \approx \sum_k \xi^k f^k(t)$$

• $\xi_i(k)$ = coordonnée sur l'axe k

Choix de la base $f^k(t)$

Fonctions orthogonales de $L^2(T)$

$$\int_0^T f^k(t) f^l(t) dt = \begin{cases} 1 & \text{si } k = l \\ 0 & \text{si } k \neq l \end{cases}$$

Fourier par exemple:

$$f^k(t) = \cos \frac{2k\pi t}{T} \text{ ou } \sin \frac{2k\pi t}{T} \Rightarrow \xi_k = \int_0^T X_t f^k(t) dt$$

MAIS les ξ_k sont corrélés.

I.2 Décomposition de Karhunen – Loeve

Décomposition unique

$$X_t = \sum_{k=1}^{\infty} \xi_k f_k(t)$$

f_k = ensemble orthonormé de fonctions de $L^2(T)$

ξ_k = ensemble orthogonal (non-corrélation) de variables de $L^2(\Omega)$

f_k fonctions propres de C

$$\lambda_k f_k(t) = \int_0^T C(t,s) f_k(s) ds$$

$$V(\xi_k) = \lambda_k$$

ξ_k fonctions propres de W

$$\xi_k = \int_0^T X_t f_k(t) dt$$

Preuve

1. Multiplier des deux côtés par $f_k(t)$ et intégrer sur t

$$X_t f_k(t) = \sum_{j=1}^{\infty} \xi_j f_j(t) f_k(t)$$

$$\int_0^T X_t f_k(t) dt = \sum_{j=1}^{\infty} \xi_j \int_0^T f_k(t) f_j(t) dt = \xi_k$$

2 Multiplier des deux côtés par ξ_k et prendre l'espérance

$$X_t \xi_k = \sum_{j=1}^{\infty} \xi_j \xi_k f_j(t)$$

$$E(X_t \xi_k) = E\left(\sum_{j=1}^{\infty} \xi_j \xi_k f_j(t)\right) = E(\xi_k^2) f_k(t)$$

$$\begin{aligned} E\left(X_t \int_0^T X_s f_k(s) ds\right) &= \int_0^T E(X_t X_s) f_k(s) ds \\ &= \int_0^T C(t, s) f_k(s) ds = \lambda_k f_k(t) \end{aligned}$$

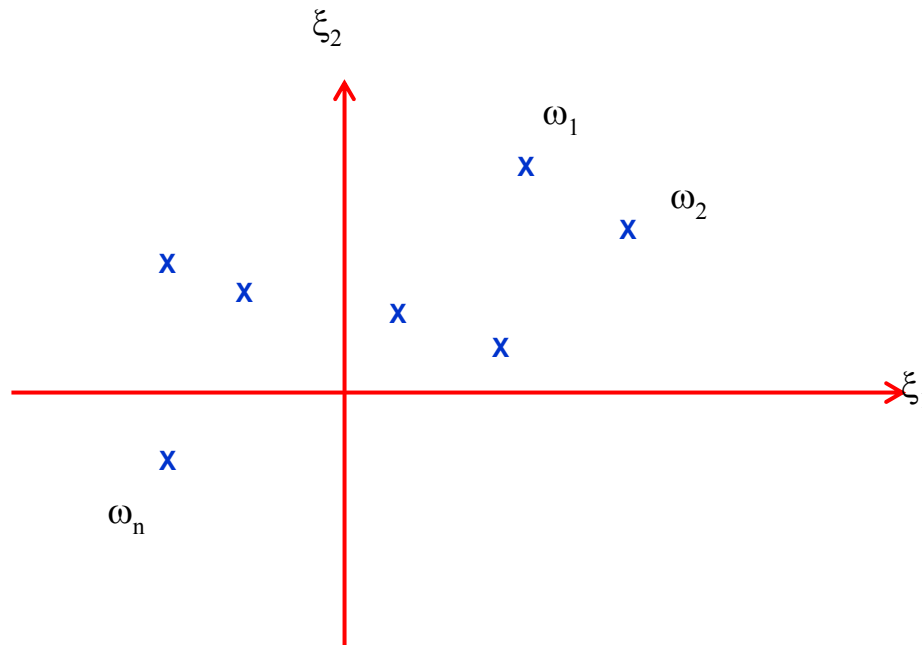
$$\begin{aligned}W \xi &= \int_0^T X_t E(X_t \xi) dt \\&= \int_0^T X_t E\left(X_t \int_0^T X_s f(s) ds\right) dt \\&= \int_0^T X_t \left(\int_0^T E(X_t X_s) f(s) ds\right) dt \\&= \int_0^T X_t \lambda f(t) dt = \lambda \xi\end{aligned}$$

I.3 décomposition de Karhunen– Loeve et ACP

f_k facteurs principaux

ξ_k composantes principales
(coordonnées sur l'axe k)

KARHUNEN – LOEVE \equiv SVD (SINGULAR
VALUE DECOMPOSITION)



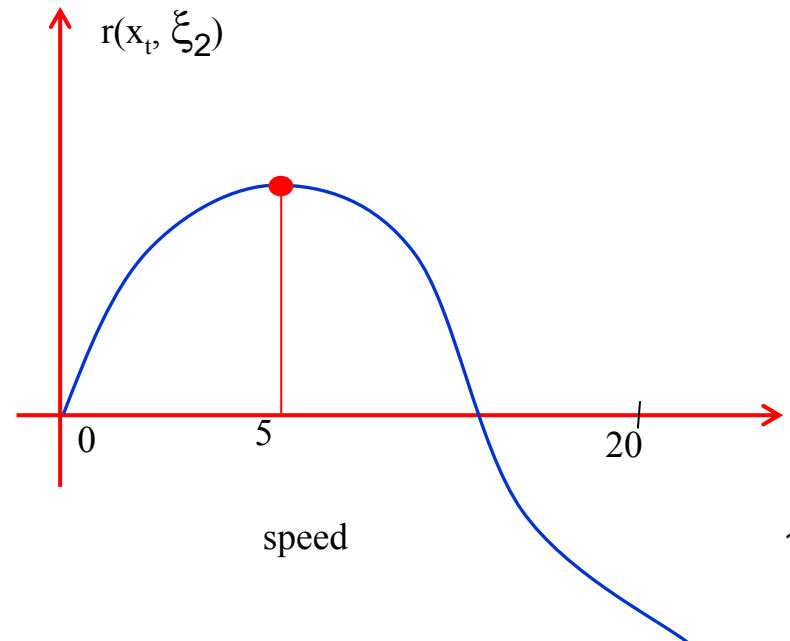
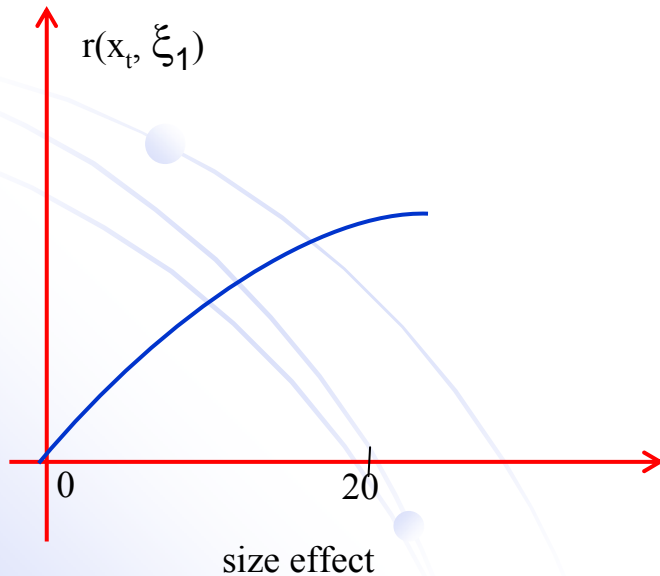
Coordonnées indépendantes de t; facteurs
fonctions de t

Interprétation:

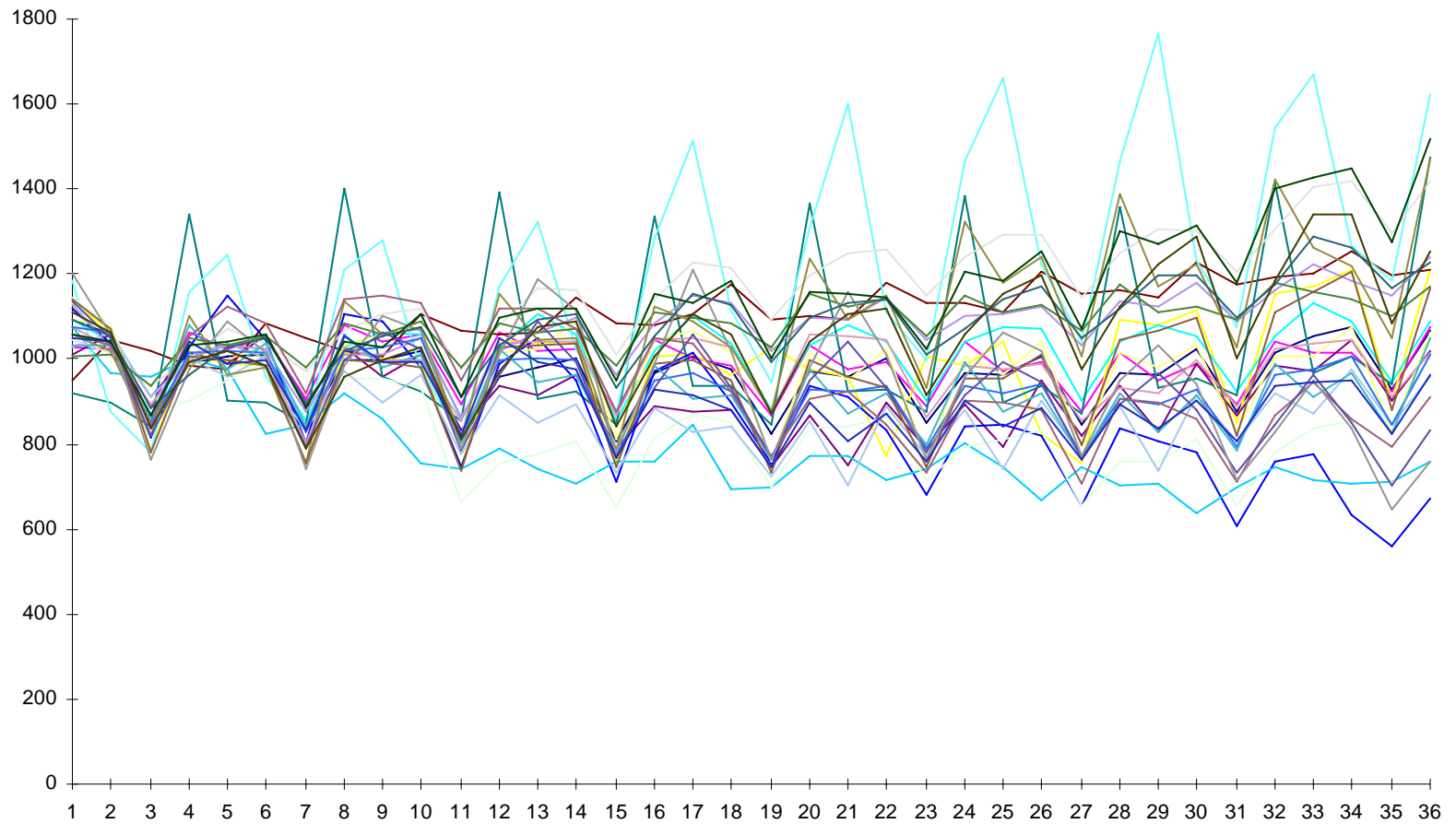
Comme en ACP

$$r(X_t; \xi_k) = \frac{\sqrt{\lambda_k} f_k(t)}{\sigma(t)}$$

Exemple: taille des familles



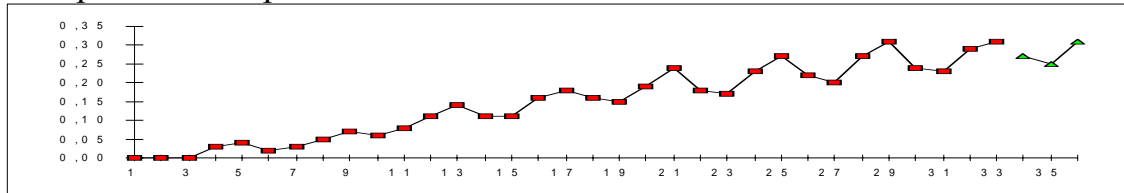
LA PRODUCTION INDUSTRIELLE EN FRANCE DE 1980 A 1988



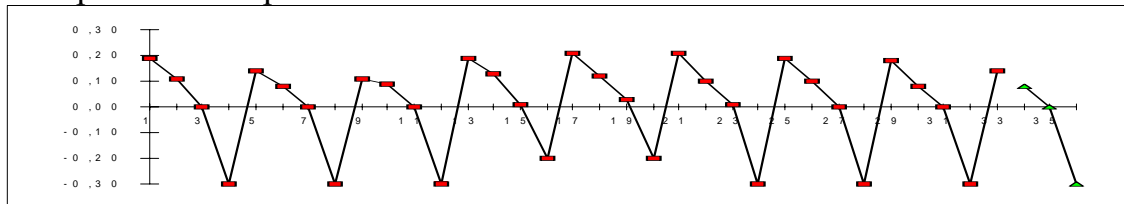
thèse G.Cohen, 1999

LES SIX PREMIÈRES COMPOSANTES TEMPORELLES ET LEURS PRÉVISIONS

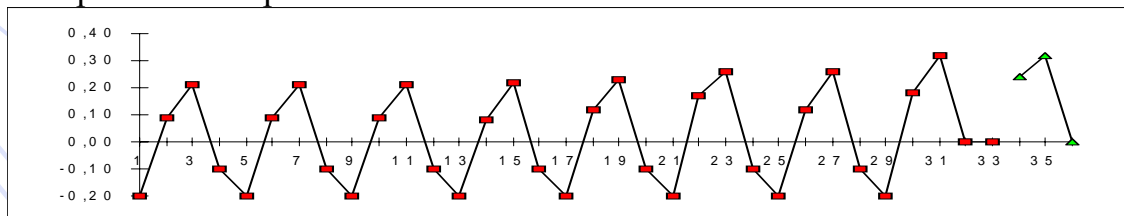
Composantes temporelles N°1



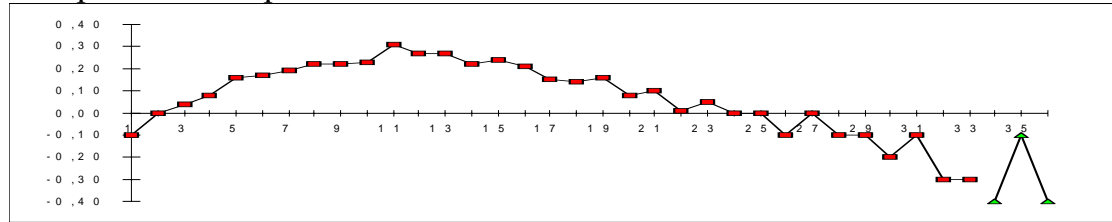
Composantes temporelles N°2



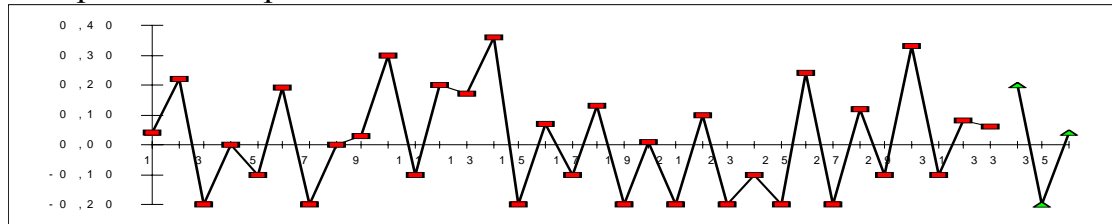
Composantes temporelles N°3



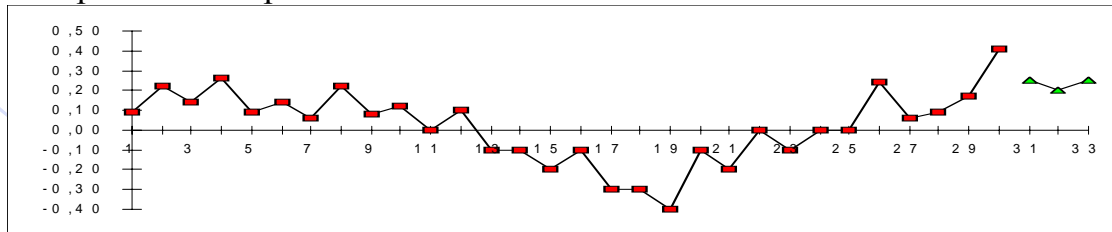
Composantes temporelles N°4



Composantes temporelles N°5



Composantes temporelles N°6



- Remarques:

- Si le processus X_t est périodique en moyenne quadratique de période sous multiple de T : $C(t, s+T/k) = C(t, s)$, alors la décomposition de Karhunen-Loeve est la décomposition de Fourier avec valeurs propres doubles associées à $\sin(2k\pi t/T)$ et $\cos(2k\pi t/T)$
- L'ordre des valeurs propres n'est pas celui des fréquences

I.4 Résolution numérique

Résoudre: $\int_0^T C(t,s) f(s) ds = \lambda f(t)$ ou $\frac{1}{n} W \xi = \lambda \xi$

Solutions analytiques pour quelques processus bien connus (eg mouvement et pont brownien motion)

Pur n trajectoires:

W matrice $n \times n$ (multidimensional scaling);

w_{ij} se calcule facilement pour des processus de sauts \Rightarrow solutions exactes \Rightarrow

$$w_{ij} = \int_0^T x_i(t) x_j(t) dt$$

$$f(t) = \frac{1}{n} \frac{1}{\lambda} \sum_{i=1}^n \xi_i X_i(t)$$

Difficile si n est grand

Discrétisation t_0, t_1, \dots, t_p
Intégration numérique de

$$\sum_{j=0}^{p-1} C(t, t_j) f(t_j) a_j = \lambda f(t)$$

$$\int_0^T C(t, s) f(s) ds = \lambda f(t)$$

Matrice diagonale de poids $CAf = \lambda f(t)$

✓ méthode des rectangles: $a_j = t_{j+1} - t_j$

✓ méthode des trapèzes:

$$a_0 = \frac{t_1 - t_0}{2} \quad a_j = \frac{t_{j+1} - t_{j-1}}{2} \quad \dots \quad a_p = \frac{t_p - t_{p-1}}{2}$$

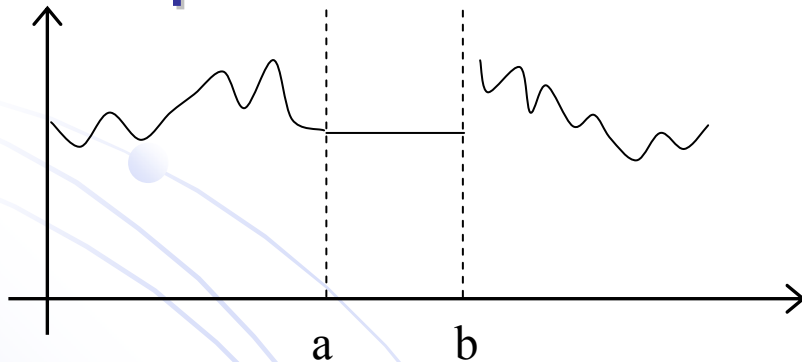
✓ Simpson ...

✓ Projection on a subspace E_p

$$X_t \rightarrow PX_t; C \rightarrow PCP$$

✓ Convergence is proved if $E_p \rightarrow L^2$

✓ Step functions



$$Y_t = X_t \text{ if } t \notin [a, b]$$
$$= \int_a^b X_s f(s) ds \text{ if } t \in [a; b]$$

$$V(Y_t) \leq V(X_t); \quad \max V(Y_t) \Rightarrow \int_a^b C(t, s) f(s) ds = \lambda f(t)$$

$$f(t) = \frac{1}{b-a}$$

$$V(X_t) - V(Y_t) = \int_a^b E(X_s - M_{ab}) ds$$

$$\|C - C^\Delta\| \leq \sqrt{\varepsilon^2 + 2TM\varepsilon} \text{ with } \|C\| = \sup_f \|Cf\|$$

$$M > E(X_t^2)$$

$$|\lambda_j - \lambda_{j^1}^\Delta| < \sqrt{\varepsilon^2 + 2TM\varepsilon}$$

$$\|f_j - f_{j^1}^\Delta\| < \frac{2\sqrt{2}}{\delta} \sqrt{\varepsilon^2 + 2TM\varepsilon}$$

$$\text{where } \sqrt{\varepsilon^2 + 2TM\varepsilon} < \frac{\delta}{2}$$

The “best” discretization problem:

Min ε

If X_t brownian motion: $N(0; \sigma^2(t))$

Best Δ : fixed length intervals

$$t_i - t_{i-1} = l = \frac{T}{p}$$

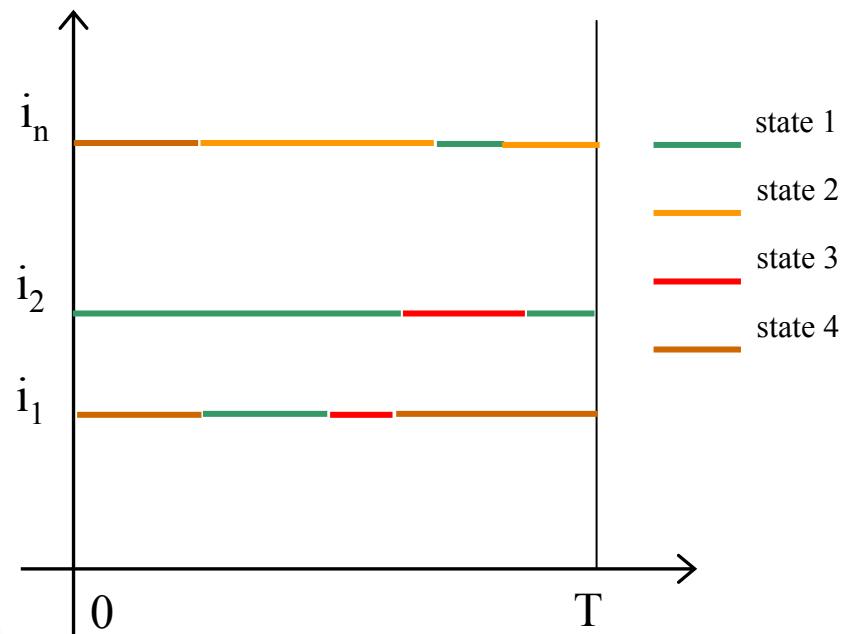
for $\varepsilon = \sum_{i=0}^{p-1} \frac{(t_{i+1} - t_i)^2}{6}$ $|\lambda_i - \lambda_i^\Delta| \leq \frac{l^2}{6} \sqrt{p(3p-1)}$

Generalized to process with independent and stationary increments

II. L'analyse harmonique qualitative

(Deville, Saporta 1979)

- Évolution d'une variable qualitative, trajectoires d'un processus qualitatif



- Cas général

$$X_t \quad t \in [0; T]$$

Principe barycentrique en temps continu:

$$\left\{ \begin{array}{l} z = \alpha \frac{1}{T} \int_0^T X_t a_t dt \\ z_i = \alpha \frac{1}{T} \int_0^T \sum_x a_t^x 1_t^x(i) dt \end{array} \right.$$

$$a_t = \alpha N_{tt}^{-1} X_t' z \quad a_t = (a_t^1, \dots, a_t^n)$$

$$\frac{1}{T} \int_0^T N_{tt}^{-1} N_{ts} a_s ds = \lambda a_t \text{ équation intégrale}$$

$$\frac{1}{T} \left[\int_0^T A_t dt \right] z = \lambda z \text{ équation matricielle}$$

- « Multidimensional scaling » avec un indice de présence – rareté

$$A_t = X_t (X_t' X_t)^{-1} X_t' = i \begin{pmatrix} j \dots \\ \vdots \\ \dots \end{pmatrix}$$

The diagram shows a dot placed at the intersection of the second row and second column of the matrix structure. Two arrows originate from this dot: one points upwards and to the right towards the value '0', and the other points downwards and to the right towards the value '1/n_t^x'.

A_t – matrice de similarité

$\int A_t dt$ matrice intégrée, positive définie

~ produits scalaires



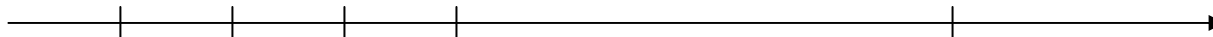
représentation euclidienne

- Résolution numérique

- ✓ Exacte

- Rassemblement de toutes les dates de changement d'état

AFC



- Approchée

- ✓ Décomposer T en p périodes

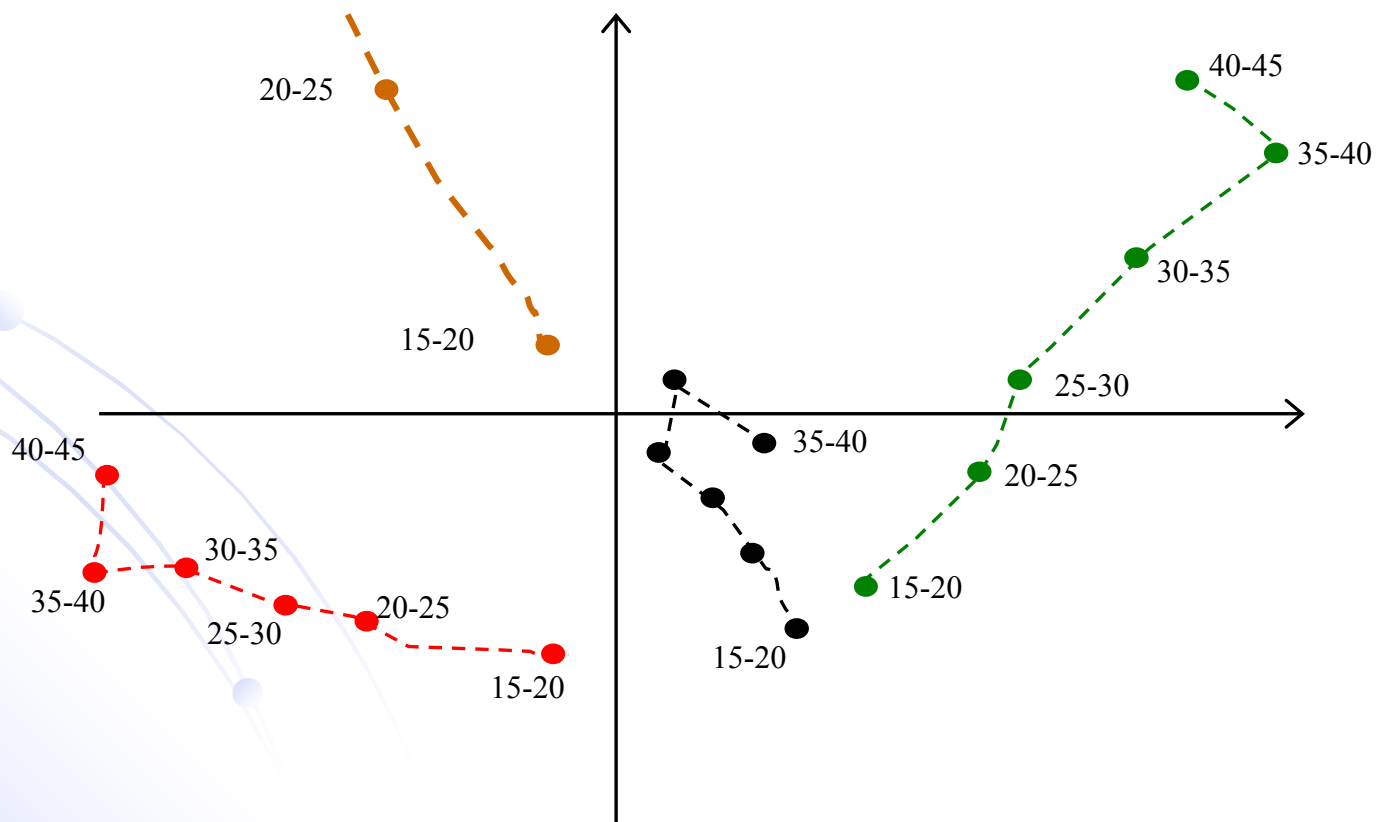
- ✓ Approximation par des fonctions constantes par intervalles

● Exemple (Deville 82):

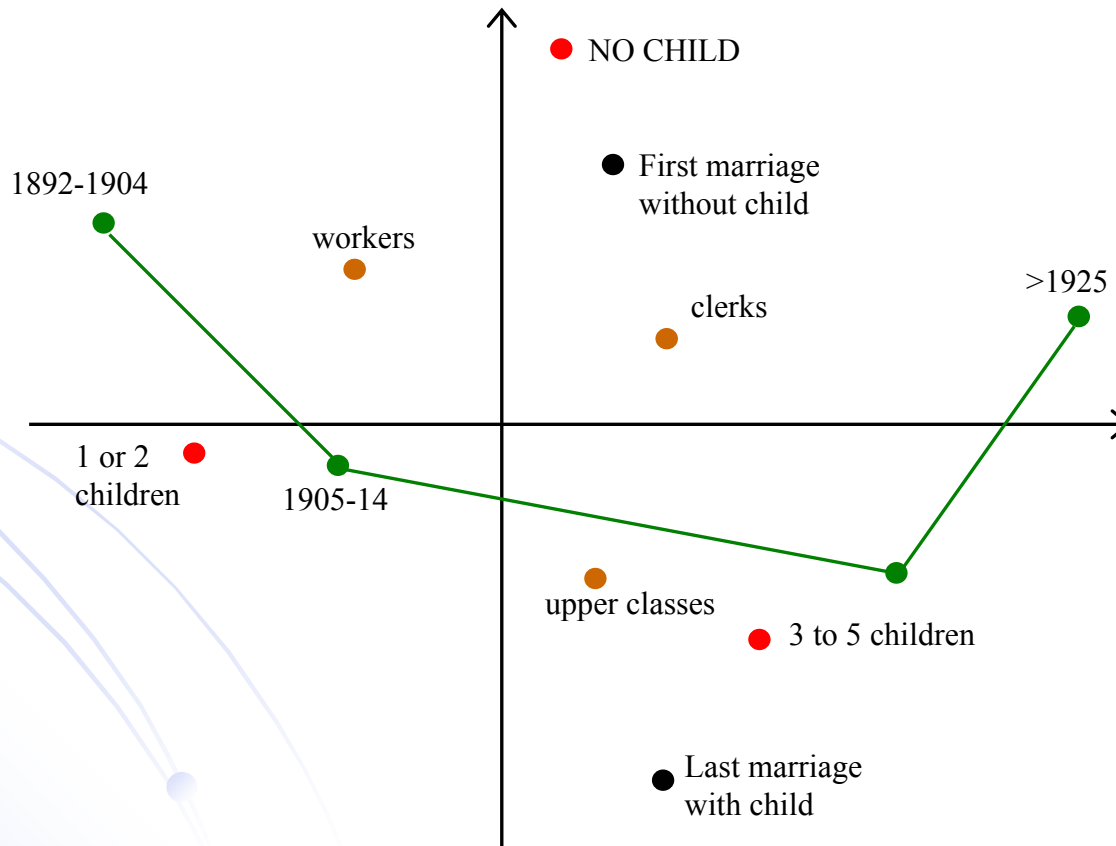
French women married more than 3 times

n=423

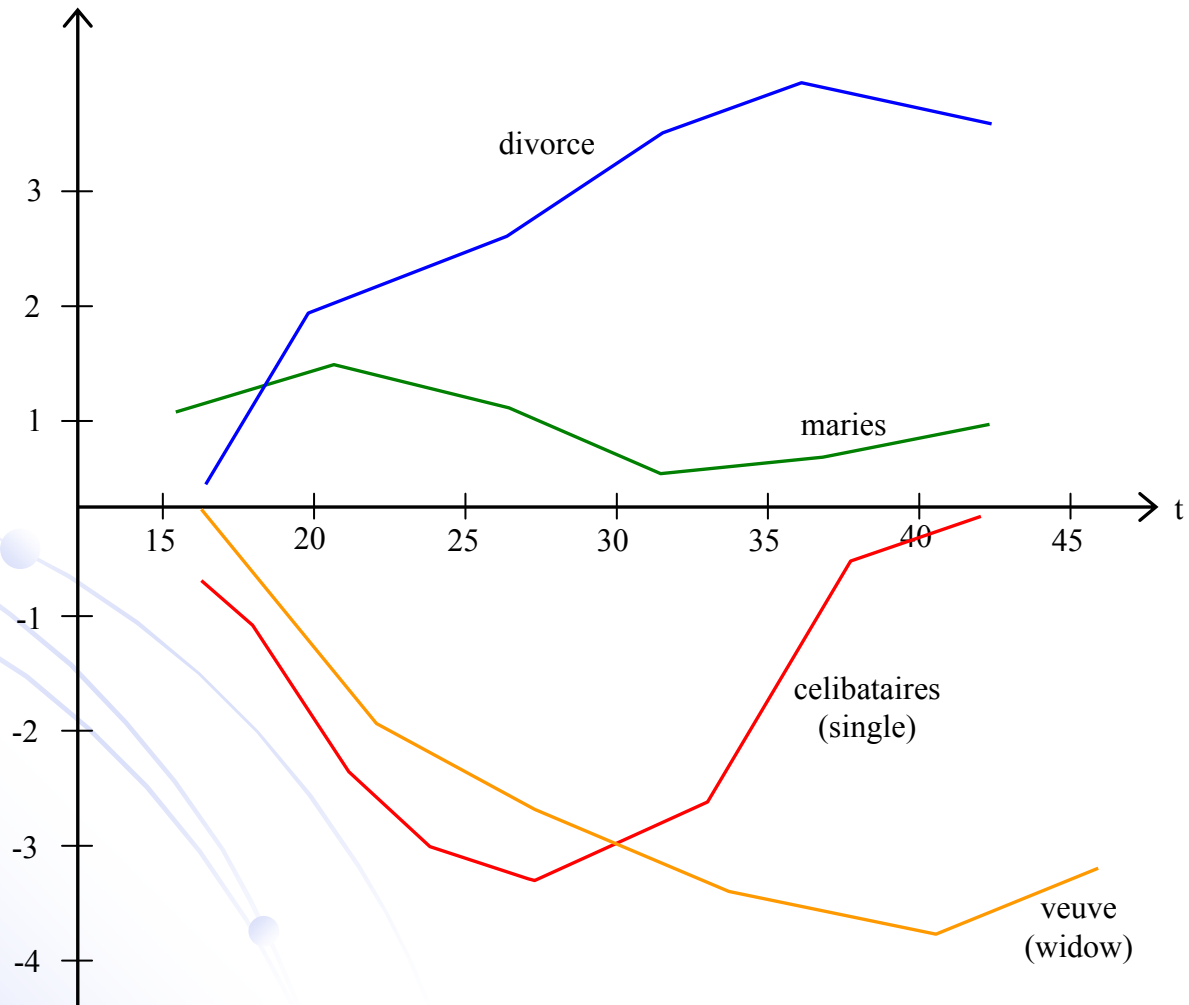
m=4 single ——— t∈[15; 45]
married ———
divorced ———
widowed ———



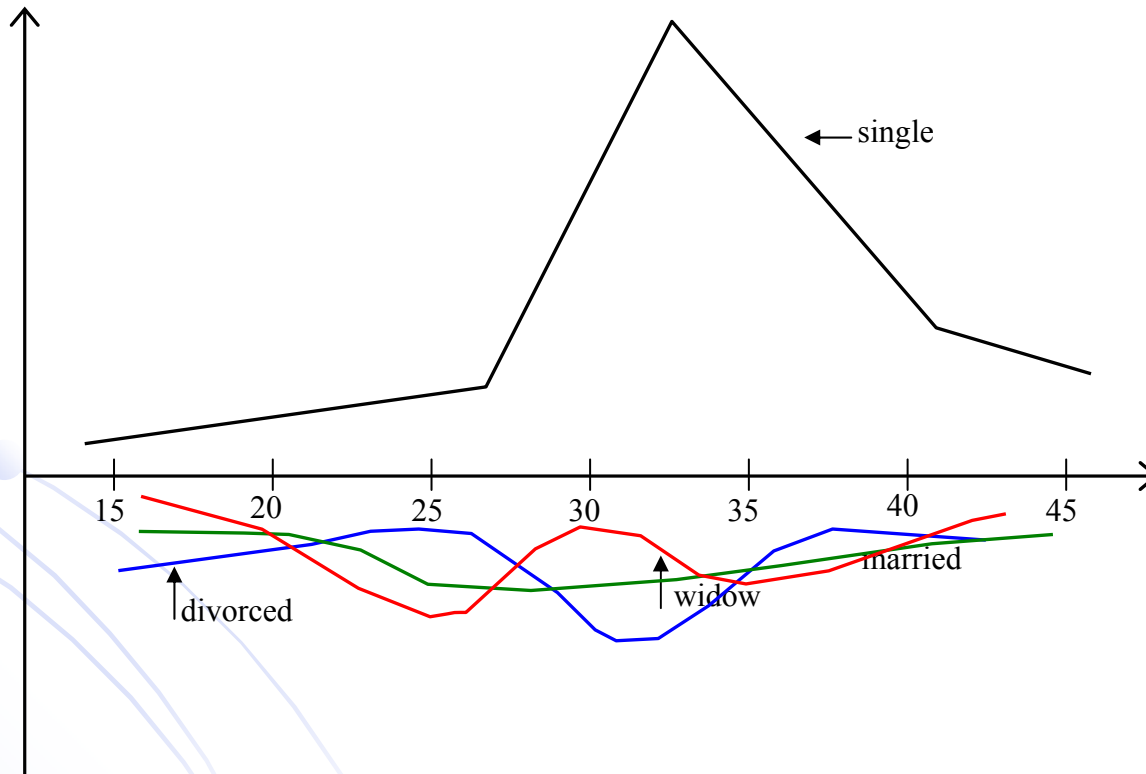
● Variables supplémentaires



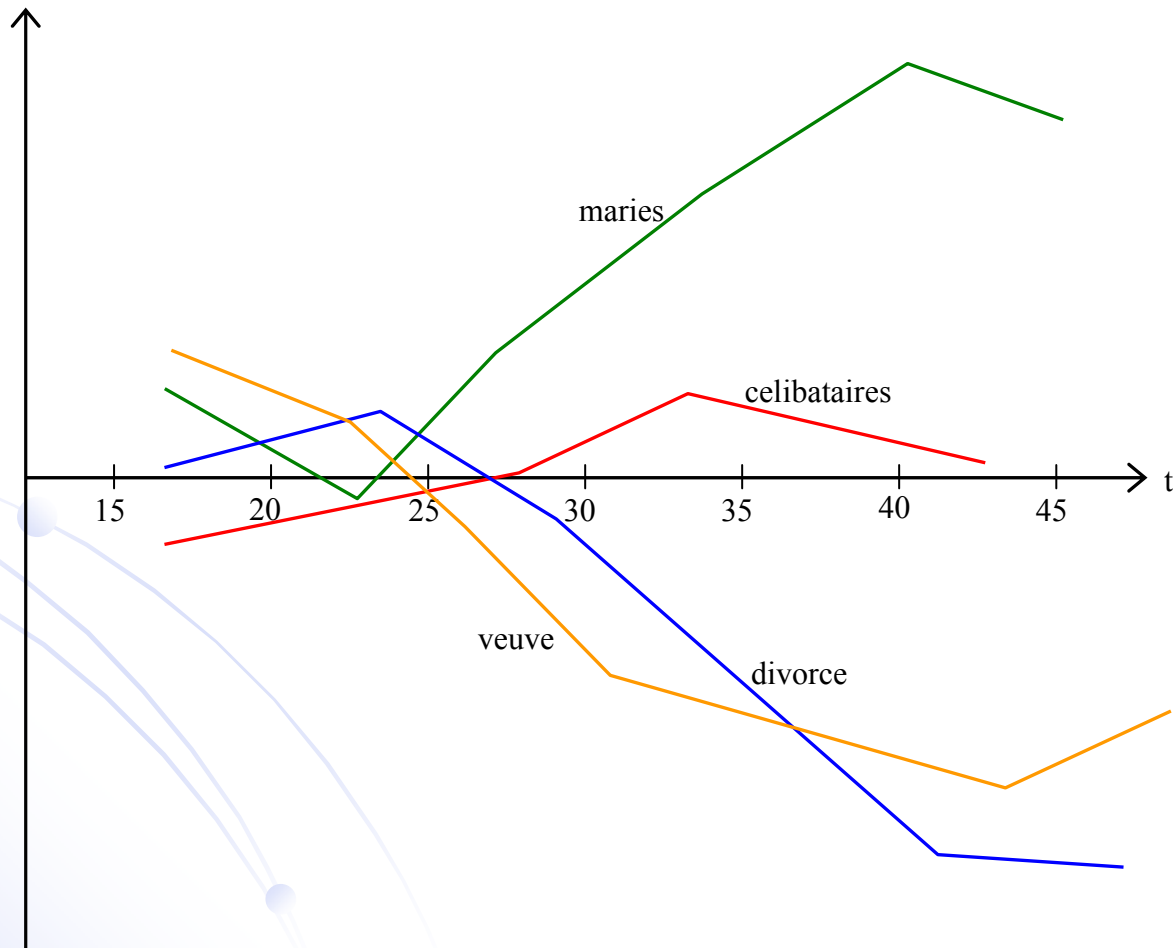
●Premier facteur



● Deuxième facteur



● Troisième facteur



III Classification (non supervisée)

- k-means, Ward etc. applicables sans difficulté à des courbes dès que l'on a une distance.
- Processus numériques

$$d^2(i; j) = \int_0^T (x_i(t) - x_j(t))^2 dt$$

- D'autres distances sont possibles

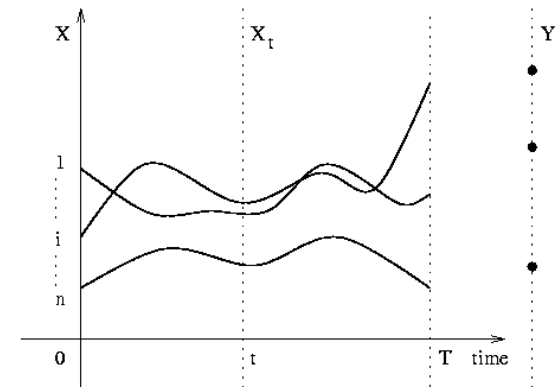
- Espaces de Sobolev

$$\langle x_i(t), x_j(t) \rangle = \int_0^T \left(x_i(t)x_j(t) + \frac{dx_i(t)}{dt} \frac{dx_j(t)}{dt} \right) dt$$

- Conséquences inattendues d'un lissage ou d'une interpolation lissée quand il manque des observations (Besse, 1979): utiliser des splines de lissage revient à changer de métrique et donc d'espace car c'est aussi faire des hypothèses de régularité (dérivabilité..) sur le processus.
- Interpoler revient à faire une transformation des données donc implicitement des produits scalaires;
- L'analyse n'est pas meilleure, elle est différente.

Développements et perspectives

- Utiliser les composantes de l'ACP fonctionnelle pour faire du supervisé
 - Mais moins bon que PLS



- Plusieurs fonctions : fonctionnel vectoriel extension des méthodes 3-way

Références

- COHEN G. , 1999 *Contribution à la prévision des processus aléatoires par l'analyse harmonique*, Ph.D. CNAM
- Dabo-Niang S., Ferraty F. (2008): *Functional and Operatorial Statistics*, Springer-Verlag
- 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
- PREDA C. , 1999, *Analyse factorielle d'un processus*, Ph.D. Université Lille 1
- RAMSAY, J.O. and SILVERMAN, B.W. ,2005: *Functional Data Analysis. 2nd ed.* Springer
- RAMSAY, J.O. and SILVERMAN, B.W., 2002: *Applied Functional Data Analysis. Methods and Case Studies.* Springer
- SAPORTA G., 1985, « Data analysis for numerical and categorical individual time-series », *Applied Stochastic Models and Data Analysis* vol.1., n°2, 109-119