

REVUE DE STATISTIQUE APPLIQUÉE

C. PARDOUX

Apport de l'analyse factorielle à l'étude d'un processus

Revue de statistique appliquée, tome 37, n° 4 (1989), p. 41-60

http://www.numdam.org/item?id=RSA_1989__37_4_41_0

© Société française de statistique, 1989, tous droits réservés.

L'accès aux archives de la revue « *Revue de statistique appliquée* » (<http://www.sfds.asso.fr/publicat/rsa.htm>) implique l'accord avec les conditions générales d'utilisation (<http://www.numdam.org/conditions>). Toute utilisation commerciale ou impression systématique est constitutive d'une infraction pénale. Toute copie ou impression de ce fichier doit contenir la présente mention de copyright.

NUMDAM

Article numérisé dans le cadre du programme
Numérisation de documents anciens mathématiques

<http://www.numdam.org/>

APPORT DE L'ANALYSE FACTORIELLE A L'ÉTUDE D'UN PROCESSUS

C. PARDOUX

Université PARIS IX

RÉSUMÉ

Pour comparer deux traitements médicaux, on dispose d'un échantillon de trajectoires d'un processus réel continu stationnaire, les trajectoires étant observées à des dates précises. L'analyse spectrale mettant en évidence deux périodicités multiples l'une de l'autre (d'ordre 4 et 8), l'étude se fera sur les séries des différences au retard 4 afin d'éliminer la périodicité commune d'ordre 4. L'analyse factorielle d'un processus permet sa décomposition en une somme de fonctions (orthogonales) certaines du temps pondérées par des variables aléatoires scalaires non corrélées. Pour la mise en oeuvre de l'analyse factorielle du processus dont on possède un échantillon, on utilise une méthode qui revient à interpoler linéairement les trajectoires entre les dates d'observation. La première composante principale différencie nettement les deux traitements et met en évidence une modification physiologique induite par l'un des traitements.

ABSTRACT

To compare two medical treatments, we have observed a sample of a real stationary continuous process; each series of this sample had been only observed at specified moments. Spectral analysis showed a daily periodicity and another of two days; to filter the common daily periodicity, we transformed the observed series with the corresponding lagged difference. Factorial analysis gives the decomposition of a process into a sum of non-stochastic orthogonal functions weighted by uncorrelated random variables. In our sample of time-series, we need to perform a linear interpolation between the dates of observation before applying a factor analysis. The first principal component differentiated clearly the two treatments and showed that one of them induced a pathological repercussion.

Mots-clés : *Processus stationnaire, Décomposition de Karhunen-Loève, Analyse spectrale, Composantes principales, Membrane de dialyse.*

Introduction

L'objectif de cet article est de présenter l'intérêt des méthodes de l'analyse factorielle à l'étude d'un processus dont on possède un échantillon de N trajectoires observées à n dates équidistantes.

Dans le cas de l'analyse factorielle d'un processus, les fonctions propres permettent de classer les individus selon leur évolution. Mais, ne disposant que de données discrètes pour étudier un processus à temps continu par nature, on sera obligé d'interpoler les trajectoires entre les dates d'observation. On commencera par exposer l'analyse factorielle d'un processus et sa mise en oeuvre sur un échantillon de trajectoires interpolées linéairement.

Après avoir posé le problème soulevé par les données médicales temporelles qui sont à l'origine de ce travail, on procédera à une première analyse des chroniques observées par la recherche de leurs périodicités à l'aide de leurs périodogrammes.

On montrera l'utilité d'élargir le champ d'application de l'analyse factorielle des processus en y incorporant les procédés de filtrage du traitement usuel des séries chronologiques.

Les résultats concernant la différenciation des deux traitements médicaux par l'analyse factorielle des trajectoires interpolées linéairement, seront présentés et commentés dans la dernière partie.

Des applications de l'analyse factorielle des processus ont déjà été menées dans le domaine démographique (DEVILLE, 1974 et 1978), dans le domaine économique (OTTER et SCHUUR, 1982), mais chaque type de problème présentait sa spécificité impliquant une utilisation appropriée de la méthode.

I. Décomposition orthogonale d'un processus du second ordre

Soit X_t un processus stochastique réel centré sur un intervalle de temps fini $[0, T]$ appartenant à $L^2(T)$, c'est-à-dire tel que :

- * pour tout t , X est de variance finie ($\sigma^2(t)$ fini pour tout t),
- * la trajectoire de tout individu ω est de carré intégrable.

On supposera de plus le processus X_t centré pour la suite de l'exposé de ce paragraphe.

Soit $C(t, s)$ sa fonction de covariance réelle symétrique :

$$C(t, s) = E(X_t X_s)$$

La fonction de covariance du processus X_t définit l'opérateur de covariance C du processus dans l'espace $L^2(T)$ par la formule :

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

Des propriétés de cet opérateur (compact, hermitien et positif) découlent le résultat fondamental de la décomposition de KARHUNEN-LOEVE (DEVILLE 1974, 1978) :

Il existe une suite λ_k de nombres réels positifs décroissant vers zéro (les valeurs propres de l'opérateur C) et une suite f_k de fonctions de $L^2(T)$ formant une famille orthonormale (les fonctions propres de l'opérateur C associées aux

valeurs propres) qui permettent de représenter le processus X_t sous forme d'une somme de processus quasi-déterministes :

$$x_t(\omega) = \sum_k \xi_k(\omega) f_k(t)$$

les ξ_k étant les composantes principales : $\xi_k(\omega) = \int_0^T x_t(\omega) f_k(t) dt$.

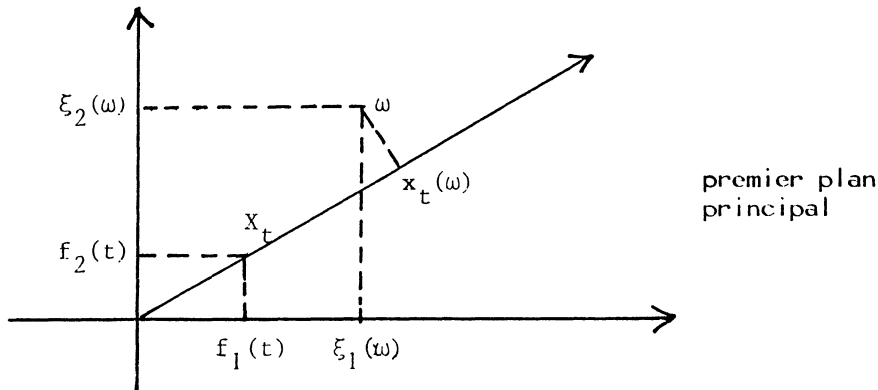
Les ξ_k sont donc des v.a. réelles non corrélées, d'espérance nulle et de variance λ_k .

Si on numérote les ξ_k selon les λ_k décroissants, le processus $\sum_{k=1}^l \xi_k f_k(t)$ est la meilleure approximation en moyenne quadratique de X_t par une somme de l processus quasi-déterministes (SAPORTA, 1981).

Tous les résultats de l'analyse factorielle se transpose donc dans le domaine des processus aléatoires.

En particulier, on peut interpréter les composantes principales à l'aide des corrélations :

$$r(\xi_k, X_t) = (\lambda_k / \text{var}(X_t))^{1/2} * f_k(t)$$

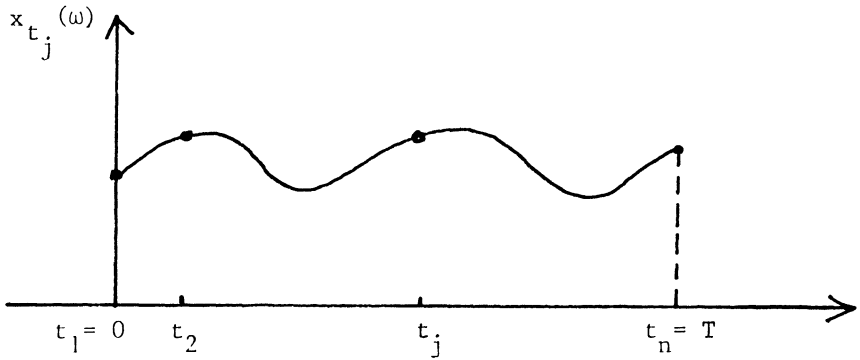


La connaissance exacte des premières valeurs propres et des fonctions propres associées, ainsi que l'identification des composantes principales correspondantes permet d'avoir une information très importante sur la structure du processus X_t .

II. Mise en oeuvre de la méthode : l'interpolation linéaire

Le processus étudié, généralement continu, n'est observé qu'à n dates précises.

Chaque individu ω est repéré par les n observations de sa trajectoire :



On peut lui associer un vecteur $x(\omega)$ de $\mathbb{R}^n$ dont les coordonnées sont $x_{t_j}(\omega)$ ($j = 1, \dots, n$).

On supposera que les n dates d'observations sont les mêmes pour tous les individus de l'échantillon.

On posera $t_1 = 0$ et $t_n = T$.

La fonction de covariance du processus peut être estimée pour les n^2 points (t_j, t_k) ($j, k = 1, \dots, n$).

On notera $\hat{C}$, la matrice symétrique de dimension n d'éléments :

$$\hat{c}_{jk} = \sum_{\omega=1}^N (x_{t_j}(\omega) - \bar{x}_{t_j}) (x_{t_k}(\omega) - \bar{x}_{t_k}) / (N - 1)$$

avec $\bar{x}_{t_j} = \sum_{\omega=1}^N x_{t_j}(\omega) / N$, moyenne empirique des x_{t_j} , et de même pour $\bar{x}_{t_k}$.

Lorsqu'on dispose d'observations en continu concernant N individus telles que les N trajectoires soient des réalisations indépendantes de X_t , on peut démontrer, sous des hypothèses assez larges, que lorsque N tend vers l'infini, l'analyse factorielle des N trajectoires converge vers celle du processus X_t (DEVILLE, 1974).

On suppose alors que les N trajectoires sont entièrement connues sur $[0, T]$, donc que l'opérateur de covariance associé à l'échantillon est calculable pour tout couple (t, s) . Or, ce n'est généralement pas le cas puisque l'échantillon de trajectoires du processus étudié n'a été observé qu'à des dates précises de l'intervalle $[0, T]$; il est néanmoins essentiel que la mesure soit théoriquement possible à tout instant.

DEVILLE (1974) montre que la diagonalisation de la matrice de covariance $\hat{C}$ pour obtenir une estimation de l'analyse factorielle du processus est légitime seulement dans le cas où les observations des trajectoires ont été faites à des dates équidistantes et si on interpole entre ces dates par des fonctions constantes par intervalles.

Mais, il est généralement beaucoup plus réaliste de faire une interpolation linéaire plutôt qu'une interpolation par des fonctions constantes par intervalles; c'est en particulier le cas pour l'exemple médical traité plus loin. On va exposer la mise en oeuvre de l'interpolation linéaire dans le cas particulier d'observations faites à des dates équidistantes, ce qui permet de rendre plus concis l'exposé de la méthode générale due à DEVILLE (1973).

II.1. Mise en place de l'interpolation linéaire

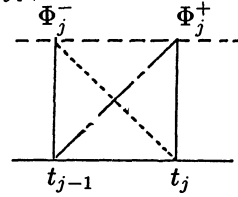
L'intervalle de temps entre deux observations successives est supposé égal à l'unité. Les couples (t_j, x_{t_j}) seront interpolés par la fonction :

$$\Phi_x(t) = x_{t_{j-1}} + (t - t_{j-1})(x_{t_j} - x_{t_{j-1}}) \quad \text{si } t \in [t_{j-1}, t_j[\quad j = 2, \dots, n$$

Introduisons pour chaque intervalle $I_j = [t_{j-1}, t_j[$ ($j = 2, \dots, n-1$), $I_n = [t_{n-1}, t_n[$, les fonctions :

$$\Phi_j^+(t) = \begin{cases} t - t_{j-1} & \text{si } t \in I_j \\ 0 & \text{sinon} \end{cases}$$

$$\Phi_j^-(t) = \begin{cases} t_j - t & \text{si } t \in I_j \\ 0 & \text{sinon} \end{cases}$$



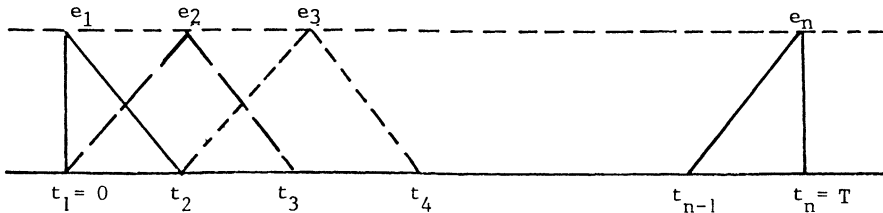
Si $(|)$ désigne le produit scalaire de $L^2(T)$, on a les relations suivantes :

$$\begin{aligned} * \quad (\Phi_j^+ | \Phi_k^+) &= (\Phi_j^- | \Phi_k^-) = (\Phi_j^+ | \Phi_k^-) = 0 & \text{si } j \neq k (j, k = 2, \dots, n) \\ * \quad (\Phi_j^+ | \Phi_j^-) &= 1/6 & \text{si } j = 2, \dots, n \\ * \quad (\Phi_j^+ | \Phi_j^+) &= (\Phi_k^+ | \Phi_k^+) = 1/3 & \text{si } j = 2, \dots, n \end{aligned}$$

Soient e_j ($j = 1, \dots, n$), les n fonctions d'interpolation simples définies ainsi :

$$\begin{cases} e_1 = \Phi_2^-, e_n = \Phi_n^+ \\ e_j = \Phi_j^+ + \Phi_{j+1}^- \end{cases} \quad \text{pour } j = 2, \dots, n-1 \quad (1)$$

Graphiquement :



Les fonctions e_j forment une base non orthogonale des fonctions continues et linéaires par morceaux sur l'intervalle $[0, T]$.

On a avantage sur le plan des calculs à se placer dans l'espace, de dimension $2(n-1)$, des fonctions linéaires sur chaque intervalle $I_j (j = 2, \dots, n)$ sans imposer de condition de continuité.

Une base *orthogonale* de cet espace est constituée par les $2(n-1)$ fonctions :

$$\begin{cases} g_j(t) = \Phi_j^+(t) + \Phi_j^-(t) \\ h_j(t) = \Phi_j^+(t) - \Phi_j^-(t) \end{cases} \quad \text{pour } j = 2, \dots, n \quad (2)$$

Dans notre cas, avec les conventions adoptées ($t_j - t_{j-1} = 1$), on a :

$$g_j(t) = \begin{cases} 1 & \text{si } t \in I_j \\ 0 & \text{sinon} \end{cases}$$

$$h_j(t) = \begin{cases} 2t - (t_{j-1} + t_j) & \text{si } t \in I_j \\ 0 & \text{sinon} \end{cases}$$

On vérifie l'orthogonalité de cette base d'après les propriétés des fonctions Φ_j^+ et Φ_j^- données précédemment; on a en effet :

$$\begin{cases} *(g_j|h_j) = 0 & \text{pour } j = 2, \dots, n \\ *(g_j|g_k) = (h_j|h_k) = (g_j|h_k) = 0 & \text{pour } j \neq k; j, k = 2, \dots, n \\ *(g_j|g_j) = 1 & \text{pour } j = 2, \dots, n \\ *(h_j|h_j) = 1/3 & \text{pour } j = 2, \dots, n \end{cases} \quad (3)$$

Les fonctions continues linéaires par morceaux forment un sous-espace de dimension n caractérisé par les $(n-2)$ équations linéaires suivantes :

$$\boxed{\begin{aligned} &\text{si } f = \sum_{j=2}^n (\xi_j g_j + \eta_j h_j) \\ f \text{ continue} &\iff \begin{cases} \text{pour } j = 2, \dots, n-1 : \\ \xi_j + \eta_j = \xi_{j+1} - \eta_{j+1} \end{cases} \end{aligned}} \quad (4)$$

En effet, si f est une fonction continue interpolée : $f = \sum_{j=1}^n x_j e_j$ et en tenant

compte des relations (1) et (2), on montre que :

$$f = \sum_{j=2}^n ((x_j + x_{j-1}) g_j + (x_j - x_{j-1}) h_j) / 2 \quad (5)$$

Et inversement, si f satisfait aux $(n - 2)$ équations linéaires (4), elle a aussi pour expression en tenant toujours compte des relations (1) et (2) :

$$f = \sum_{j=2}^n (\xi_j - \eta_j) e_{j-1} + (\xi_n + \eta_n) e_n = (\xi_2 - \eta_2) e_1 + \sum_{j=2}^n (\xi_j + \eta_j) e_j$$

Notons x , le vecteur colonne de coordonnées $x_{t_j} (j = 1, \dots, n)$, la fonction $\Phi_x(t)$ interpolée sur les couples (t_j, x_{t_j}) a pour expression d'après la relation (5) :

$$\Phi_x(t) = \psi(t)Ax$$

$\psi(t)$ étant le vecteur ligne à $2(n - 1)$ composantes :

$$\psi(t) = (g_2(t), h_2(t), \dots, g_j(t), h_j(t), \dots, g_n(t), h_n(t)) \quad (6)$$

et A , la matrice $(2(n - 1), n)$ formée de $(n - 1)$ blocs 2×2 et donnée à l'annexe I.

II.2 Interpolation linéaire du processus et recherche des éléments propres

Le processus X_T est interpolé sur les n v.a. réelles X_{t_j} selon la formule :

$$\Phi X_t = \psi(t)AX$$

X désignant le vecteur aléatoire de dimension n des X_{t_j} .

On a alors : $E(\Phi X_t) = \psi(t)AE(X)$

La fonction de covariance du processus ΦX_t est égale à :

$$C_\Phi(t, s) = \text{Cov}(\Phi X_t, \Phi X_s) = \psi(t)ACA'\psi'(s),$$

C étant la matrice -symétrique d'ordre n - des covariances du processus telle que :

$$c_{jk} = C(t_j, t_k) \quad (j, k = 1, \dots, n)$$

On notera γ_{jk} les éléments de la matrice symétrique ACA' de dimension $2(n - 1)$.

La fonction de covariance du processus interpolé apparaît donc comme une interpolation de la fonction de covariance sur $[0, T] \times [0, T]$ du processus initial à l'aide des fonctions d'interpolation linéaire, et on a bien :

$$C_\Phi(t_j, t_k) = C(t_j, t_k) \quad \text{pour tout } j, k = 1, \dots, n$$

Le problème qui se pose maintenant est le suivant :

$$\begin{aligned} \text{trouver } f(t) \text{ telle que : } \lambda f(t) &= \int_0^T C_{\Phi}(t, u) f(u) \, du \\ &= \sum_{k,l} \gamma_{kl} \psi_k(t) \int_0^T \psi_l(u) f(u) \, du \end{aligned}$$

Cette équation prouve que f est de la forme :

$$f(t) = \sum_k z_k \psi_k(t) = Z' \psi'(t) = \psi(t) Z, \quad Z \text{ étant un vecteur de } \mathbb{R}^{2(n-1)}$$

Le problème revient à chercher le vecteur Z tel que :

$$\lambda \psi(t) Z = \sum_{k,l} \gamma_{kl} \psi_k(t) \int_0^T \psi_l(u) \psi(u) Z \, du = \psi(t) A C A' D Z$$

Soit :

$$\lambda Z = A C A' D Z \quad (7)$$

où D est la matrice carrée de dimension $2(n-1)$ des produits scalaires $(\psi_k | \psi_l)$, matrice qui d'après (3) et (6) est diagonale, les éléments diagonaux valant :

$$\delta_{2j-1, 2j-1} = 1 \quad \text{et} \quad \delta_{2j, 2j} = 1/3 \quad \text{pour } j = 1, \dots, (n-1)$$

La recherche des éléments propres de C_{Φ} revient ainsi à la recherche des valeurs et vecteurs propres de la matrice $A C A' D$.

On montre que les vecteurs propres associés à des valeurs propres non nulles de la matrice $A C A' D$ appartiennent au sous-espace de dimension n des fonctions continues caractérisées par les $(n-2)$ relations linéaires (4).

Mais, cette matrice n'étant pas symétrique, on se ramènera à un problème symétrique en recherchant les éléments propres d'une matrice de rang n en prémultipliant l'équation (7) par $D^{1/2}$:

$$\lambda \left(D^{1/2} Z \right) = D^{1/2} A C A' D^{1/2} \left(D^{1/2} Z \right)$$

$D^{1/2}$ étant la matrice diagonale de dimension $2(n-1)$ dont les éléments sont les racines carrées de ceux de la matrice D .

II.3 Traitement numérique

On dispose d'un N-échantillon du processus X_t observé à n dates équidistantes qui sont les mêmes pour tous les individus de l'échantillon. On peut ainsi calculer une estimation $\hat{C}$ de la matrice C .

On détermine les premiers éléments propres de la matrice symétrique $M = D^{1/2} A C A' D^{1/2}$ de dimension $2(n-1)$ et de rang n .

Les vecteurs propres permettent de déduire des estimations des fonctions propres f de C_Φ aux n dates $t_1, \dots, t_n$; soit Z^* , un vecteur propre de M :

$Z = D^{-1/2} Z^*$ et si : $Z = (\xi_2, \eta_2, \dots, \xi_j, \eta_j, \dots, \xi_n, \eta_n)$, on aura pour fonction propre associée à Z , la fonction f telle que :

$$f(t_1) = \xi_2 - \eta_2, \quad f(t_n) = \xi_n + \eta_n$$

et pour $j = 2, \dots, n-1$,

$$f(t_j) = \xi_j + \eta_j = \xi_{j+1} - \eta_{j+1}$$

Seules nous importent les valeurs $f(t_j)$ puisque les fonctions f sont linéaires par morceaux.

III. Etude descriptive des données

III.1 Présentation des données et position du problème

Les données se présentent sous la forme d'un échantillon de 32 trajectoires d'un processus dont on possède des observations toutes les 6 heures pendant une semaine, plus une fois au début de la semaine suivante (soit 29 observations). Il s'agit de mesures de températures et les intervalles de temps séparant deux relevés sont donc toujours égaux. Le but est de rechercher si un traitement de dialyse est responsable de modifications dans le cycle thermique physiologique.

L'étude faite pour deux membranes de dialyse (notées PAN et CUP), s'attache à distinguer les modifications induites par chaque membrane et à les comparer entre elles. Pour tenir compte d'un éventuel facteur hormonal, on s'intéressera également aux sous-groupes hommes et femmes.

L'échantillon comprend 15 malades dialysés avec la membrane CUP (8 hommes et 7 femmes) et 17 malades dialysés avec la membrane PAN (10 hommes et 7 femmes).

Chaque malade est dialysé 3 fois dans la semaine : entre la 5° et la 6° prise de température, entre la 13° et 14° prise et entre la 21° et 22° prise.

L'idée est d'étudier si une périodicité d'ordre 8 liée à la dialyse, peut être mise en évidence pour les dialysés sur la membrane CUP et/ou la membrane PAN, l'évolution biologique habituelle induisant un cycle journalier, donc de période 4.

III.2 Analyse spectrale descriptive

Les séries chronologiques sont constituées chacune de 29 observations. La première analyse exploratoire consiste à étudier les périodicités des chroniques

individuelles et de chroniques agrégées à l'aide de leur "périodogramme" (ou "spectre").

Le périodogramme est défini à partir de la décomposition de Fourier; on en obtient une représentation graphique en portant la fréquence des différentes oscillations en abscisse et en ordonnée le carré de leur amplitude. La somme totale des carrés des valeurs de la chronique étant répartie additivement entre les différentes fréquences, le périodogramme permet de repérer les oscillations qui jouent un rôle important (GOURIEROUX et MONFORT, 1983; CHATFIELD, 1985).

III.2.1 Analyse des séries brutes

Pour chacun des 32 individus, l'évolution de la température pendant la semaine d'observation, ainsi que le périodogramme, ont été représentés graphiquement (logiciel STATISTICAL GRAPHICS SYSTEM, 1986).

L'examen de ces graphiques montre que ces chroniques peuvent être considérées *sans tendance* et qu'elles possèdent des périodicités. Ce même travail a été réalisé pour les chroniques des températures moyennes de chacun des deux groupes CUP et PAN.

Pour la chronique du groupe CUP, on distingue des rythmes de longueur 4 et 8, tandis que pour la chronique moyenne du groupe PAN, on ne remarque qu'un rythme de longueur 4 (Fig. 1).

III.2.2 Analyse des séries des différences au retard 4

Nous avons donc mis en évidence pour tous les individus un rythme lié à la journée de longueur 4 et pour une partie d'entre eux, un rythme lié à la dialyse de longueur 8. Ces deux composantes périodiques ne sont pas indépendantes, puisque la seconde est multiple de la première. Pour montrer l'existence d'un rythme d'une longueur multiple de 4, nous avons filtré cette périodicité 4 en étudiant les séries des différences au retard 4.

La série moyenne des différences au retard 4 du groupe CUP présente une périodicité 8 d'amplitude beaucoup plus élevée que celle du groupe PAN (Fig. 2).

Les périodogrammes des séries moyennes des différences au retard 4 selon le sexe pour les individus traités sur la membrane CUP sont quasiment les mêmes pour les hommes et les femmes et montrent l'existence d'une composante de période 8 dont le carré de l'amplitude est égale à 78,5 % de la somme totale des carrés de la chronique pour les hommes et à 80,5 % de la somme totale des carrés de la chronique pour les femmes (Fig. 3).

Pour les individus traités sur la membrane PAN, les périodogrammes des séries moyennes des différences au retard 4 n'ont pas la même allure selon le sexe. Cette série est à l'évidence un bruit blanc pour les femmes, alors que pour les hommes, le périodogramme présente deux "pics" dont les carrés des amplitudes représentent 65 % de la somme totale des carrés de la chronique pour la fréquence 0.125 (période 8), et 21 % de la somme totale des carrés de la chronique pour

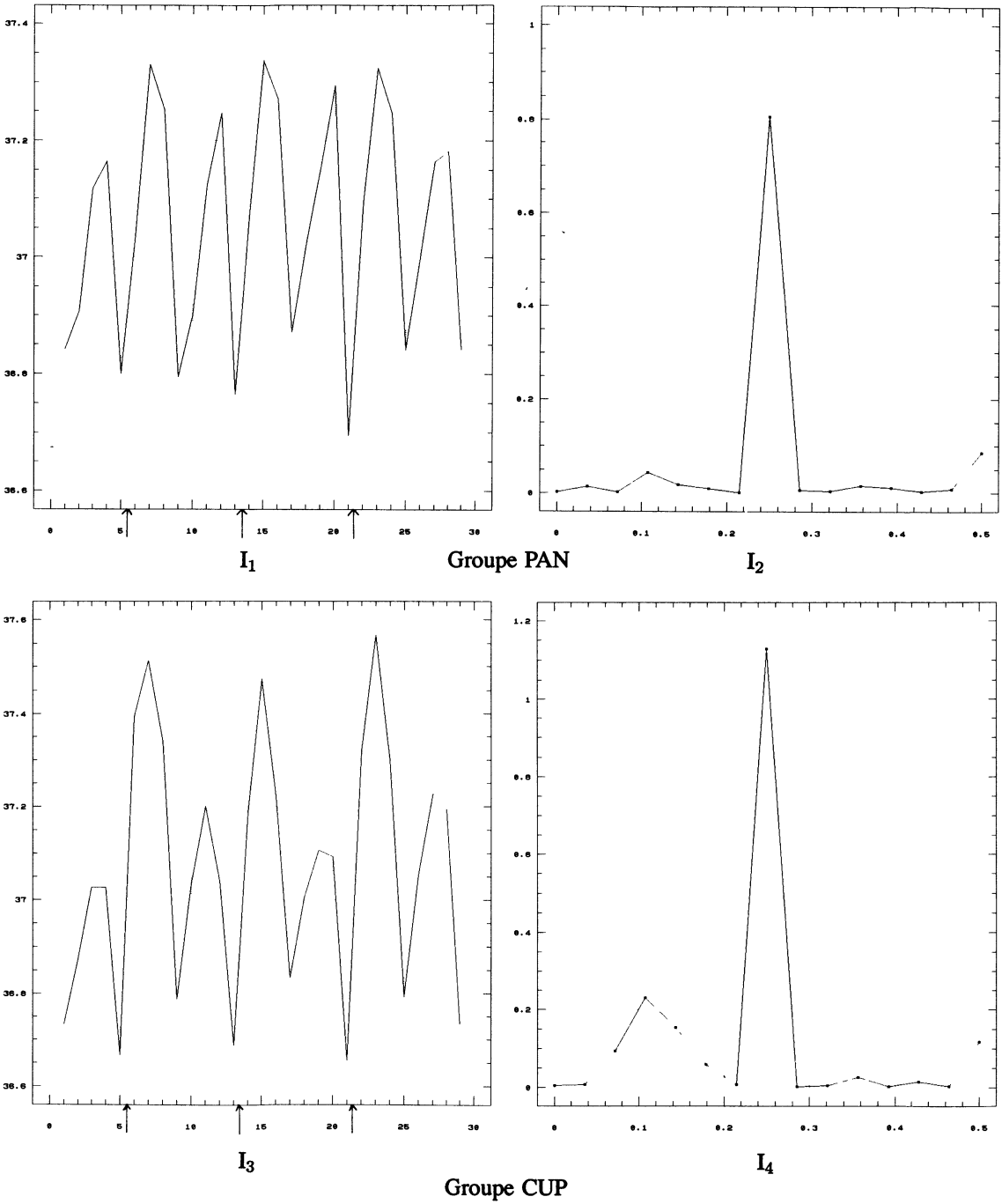
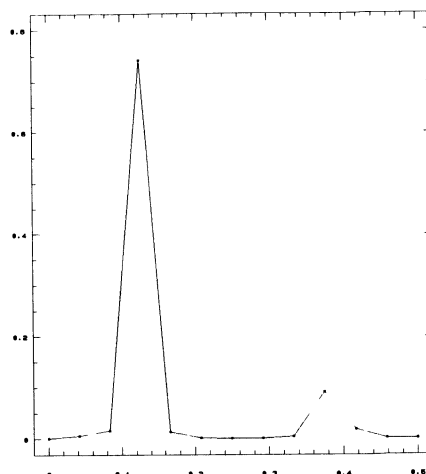
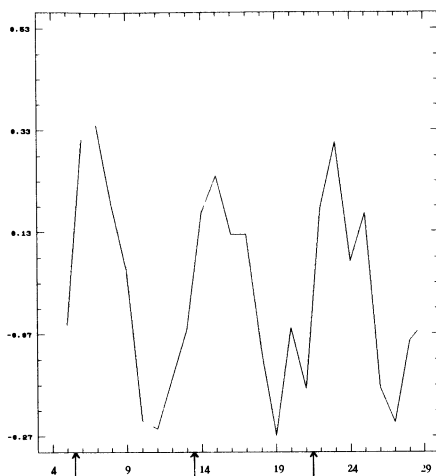
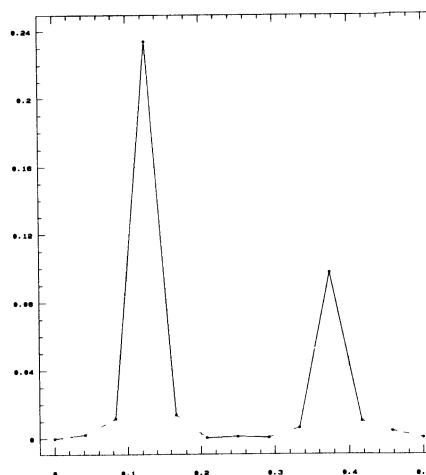
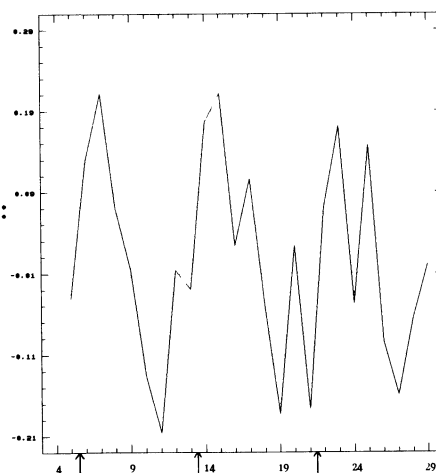


FIGURE 1
Chroniques moyennes par groupes : évolution et périodogramme

Les 32
individus



Groupe PAN :
17 individus



Groupe CUP
15 individus

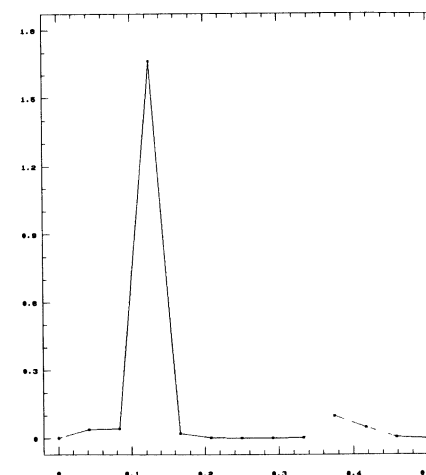
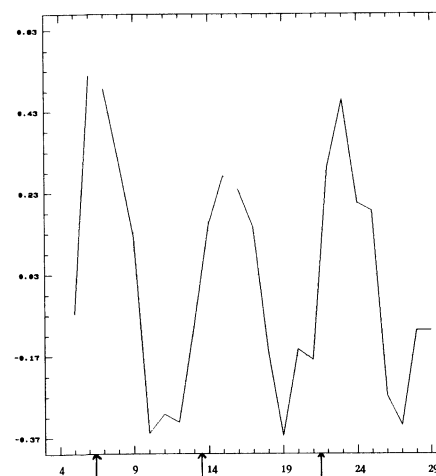
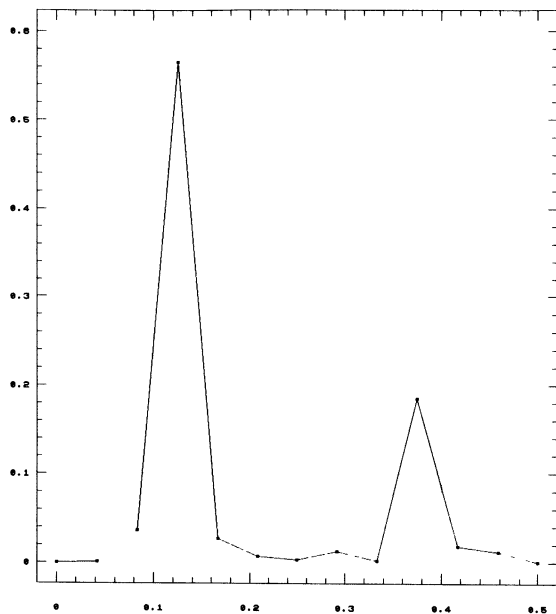
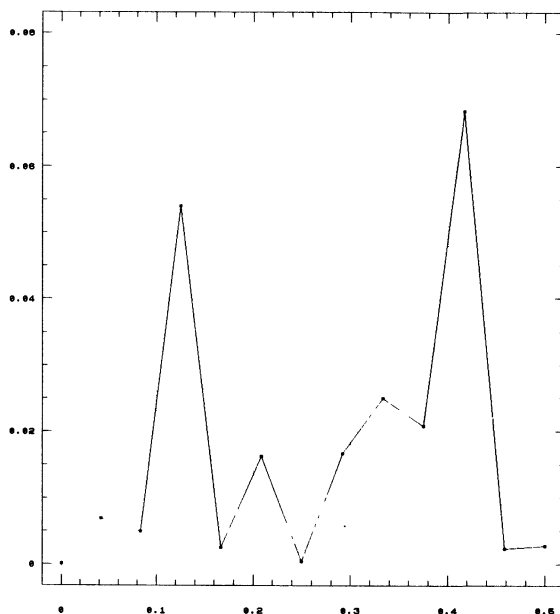


FIGURE 2
Séries moyennes des différences au retard 4 : évolution et périodogramme

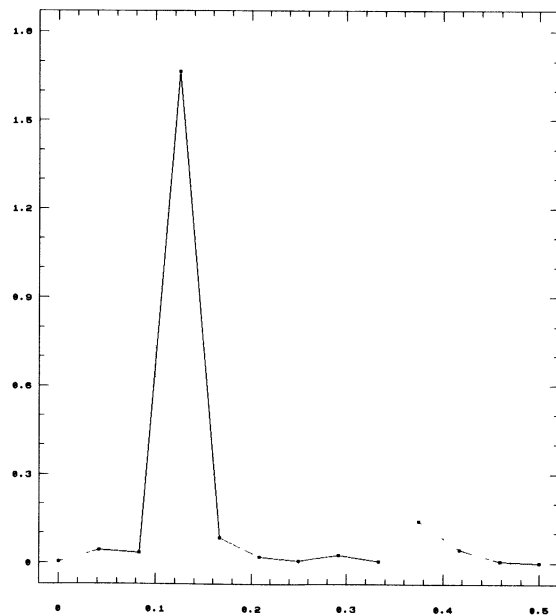
PAN Hommes



PAN Femmes



CUP Hommes



CUP Femmes

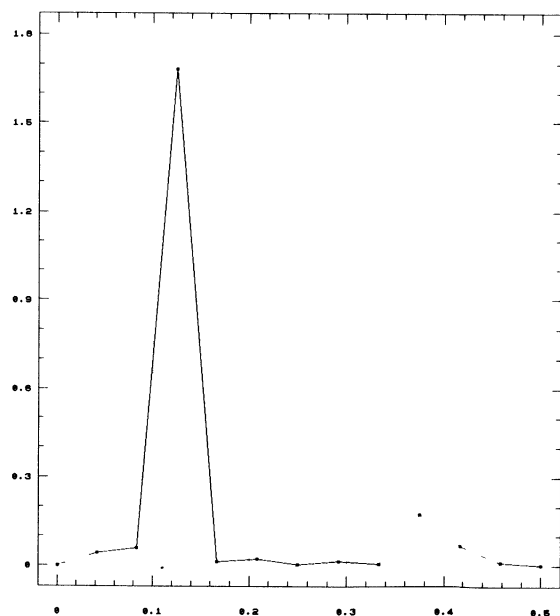


FIGURE 3

Périodogramme des séries moyennes des différences au retard 4 selon le groupe et le sexe

la fréquence 0.375 qui peut s'interpréter comme une harmonique de la première composante de fréquence 0.125.

Cette analyse descriptive permet de déceler des différences entre les deux traitements. L'analyse factorielle va permettre de situer les individus les uns par rapport aux autres.

IV. Analyse factorielle du processus des températures

IV.1 Décomposition du processus

Les 32 séries chronologiques de températures peuvent être considérées comme un échantillon d'un processus continu. Les observations ont été faites à des dates équidistantes qui sont les mêmes pour chaque malade. D'après l'étude descriptive des trajectoires, le processus peut être considéré comme sans tendance et stationnaire.

L'analyse spectrale ayant conduit à privilégier l'étude des séries des différences au retard 4, on a réalisé une analyse factorielle sur l'échantillon du processus des différences au retard 4 -en disposant donc de 25 observations par trajectoire- en appliquant la méthode exposée au paragraphe II.

On a cherché les premiers éléments propres de la matrice des covariances estimées de ce processus interpolé linéairement entre les dates d'observation. On a vu que cette recherche amène à traiter une matrice symétrique $D^{1/2} \hat{A} \hat{C} A' D^{1/2}$ de dimension 48 (cf. paragraphe II).

La trace de cette matrice est égale à 2.40 . Seules, les trois premières valeurs propres ont une valeur au moins égale à 10 % de la trace de la matrice : 39.44 % pour la première, 13.01 % pour la seconde et 11.37 % pour la troisième.

La chronique associée à la première fonction propre : $f_1(t_5), f_1(t_6), \dots, f_1(t_{29})$, est périodique de période 8, tandis que le processus dont la seconde fonction propre, orthogonale à la première, serait une réalisation, peut être considérée comme un bruit blanc (Fig. 4).

Dans ces conditions, on retient la décomposition suivante pour chacune des 32 chroniques :

$$y_{t_j}(\omega) = \bar{y}_{t_j} + \xi_1(\omega) * f_1(t_j) + \xi_{t_j}(\omega)$$

$$\text{avec : } y_{t_j}(\omega) = x_{t_j}(\omega) - x_{t_{j-4}}(\omega), \bar{y}_{t_j}(\omega) = \sum_{\omega=1}^{32} y_{t_j}(\omega) / 32 \quad \text{pour } j = 5, \dots, 29$$

Les séries chronologiques $\bar{y}_{t_j}$ et $f_1(t_j)$ -associées à la première fonction propre normée -sont données à l'annexe II, ainsi que le premier facteur ξ_1 .

Pour chaque individu, on a calculé la chronique résidu $\varepsilon_{t_j}(\omega)$ et on a utilisé la procédure de STATGRAPHICS pour la recherche d'éventuelles périodicités.

Pour chacune de ces 32 chroniques, on ne rejette pas l'hypothèse d'un bruit blanc pour le processus dont elle serait une réalisation.

A condition de modéliser les deux chroniques $\bar{y}_{t_j}$ et $f_1(t_j)$, l'analyse

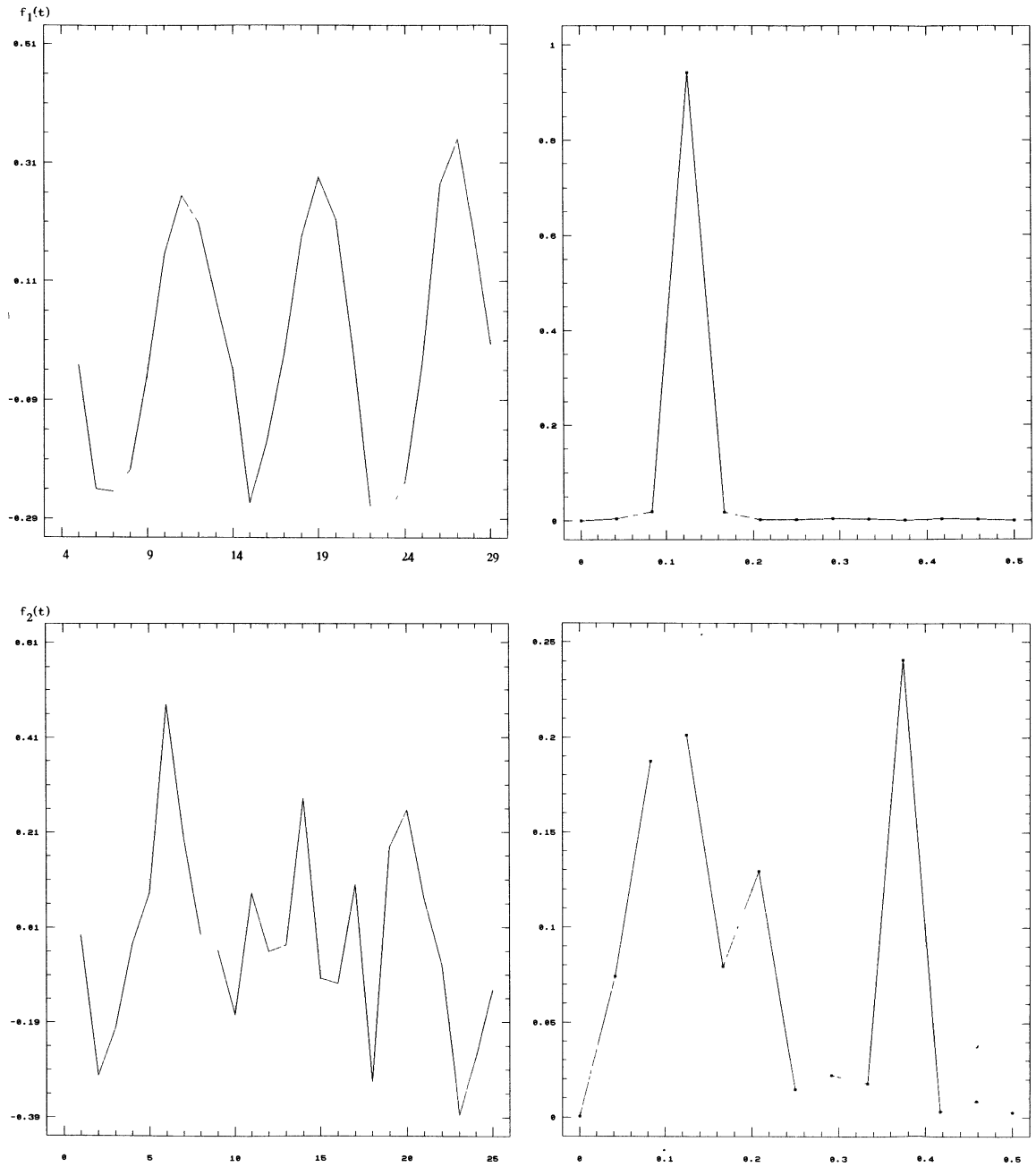


FIGURE 4
Evolutions et périodogrammes
des deux premières fonctions propres

factorielle permet le calcul d'une approximation de la décomposition du processus des températures. Mais, la recherche d'un tel modèle ne nous a pas paru présenter un intérêt puisque cette étude n'est pas menée à des fins de prévision.

IV.2 Interprétation de la première composante principale

On peut calculer les corrélations entre la première composante principale et les 25 variables Y_{t_j} ($j = 5, \dots, 29$); λ_1 désignant la première valeur propre, on a (cf. paragraphe I) :

$$r(\xi_1, Y_{t_j}) = (\lambda_1 / \text{var}(Y_{t_j}))^{1/2} * f_1(t_j)$$

D'après la représentation graphique de l'évolution de cette corrélation (Fig. 5), la première composante principale apparaît fortement liée à l'intervention de la dialyse.

L'interprétation de la première composante principale va s'enrichir par le calcul des contributions des individus et des groupes d'individus à l'inertie du premier axe factoriel (CTR_1), et par le calcul de l'inertie de chaque individu relativement à l'inertie totale du nuage (INR); pour la définition de ces indicateurs, on peut se reporter à l'ouvrage de VOLLE (1985).

Le tableau I donne les valeurs des indicateurs pour chacun des groupes.

TABLEAU 1

Groupe	PAN	CUP	PANH	PANF	CUPH	CUPF
nombre d'individus	17	15	10	7	8	7
coordonnée moyenne sur le 1 ^o axe factoriel	.390	-.442	.148	.736	-.420	-.468
$\sum_{\omega} \text{CTR}(\omega)$	47%	53%	30%	17%	39%	14%
$\sum_{\omega} \text{INR}(\omega)$	53%	47%	36%	17%	29%	18%

Le premier axe factoriel oppose les deux groupes CUP et PAN et c'est le groupe CUP qui contribue le plus à l'inertie expliquée par cet axe, bien qu'il compte *moins d'individus* que le groupe PAN. D'autre part, la première fonction propre est négative après l'intervention de la dialyse et les individus du groupe CUP ont une coordonnée moyenne négative sur le premier axe, ce qui fait apparaître une poussée de température plus accentuée pour les malades traités par la membrane CUP.

L'examen des coordonnées individuelles confirment les conclusions faites à partir des coordonnées des centres de gravité des groupes :

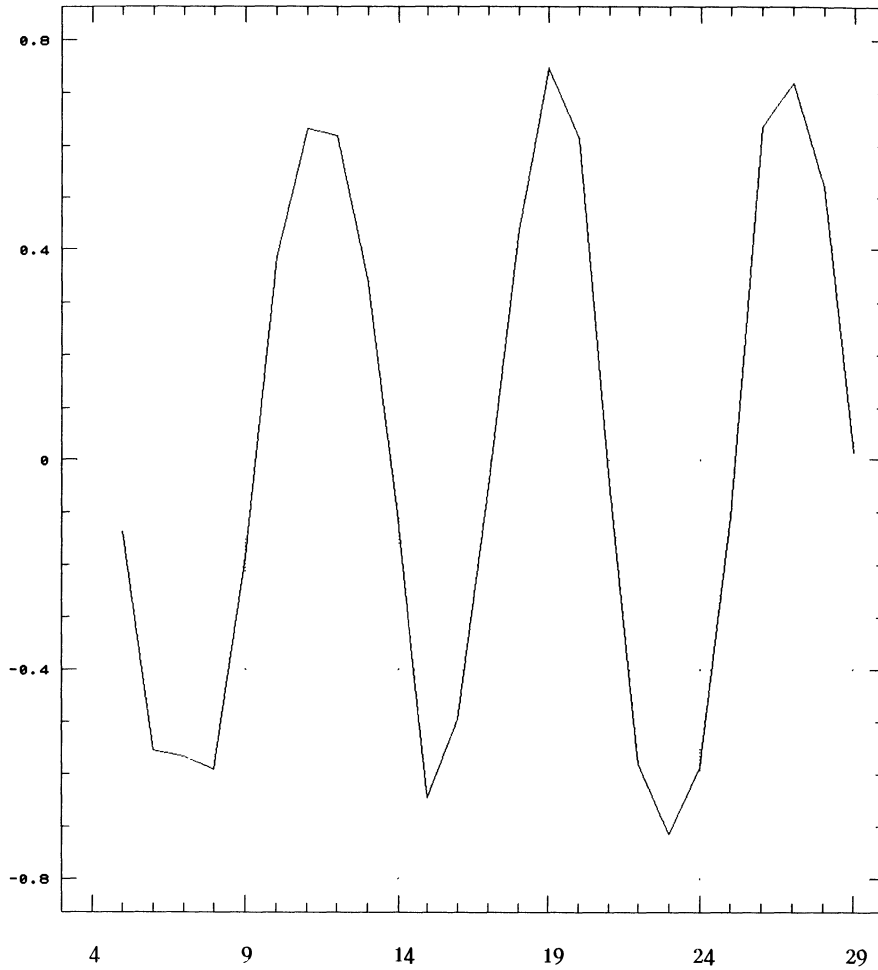


FIGURE 5
Evolution du coefficient de corrélation linéaire entre
la première composante principale et les variables Y_{tj} ($j = 5, \dots, 29$)

- * parmi les quatre individus traités par la membrane PAN qui ont des coordonnées négatives sur le premier axe factoriel, trois d'entre eux sont des hommes qui ont subi des changements de membrane au cours de leur traitement;
- * le premier axe oppose plus nettement aux malades traités par la membrane CUP, les femmes traitées par la membrane PAN que les hommes traités par

la membrane PAN; ce résultat est en accord avec la remarque précédente et les résultats de l'analyse spectrale du paragraphe III.2 .

Conclusion

L'analyse factorielle peu employée jusqu'à présent pour l'analyse d'échantillons de processus, s'est révélée très fructueuse. Elle nous a permis de différencier les deux traitements médicaux et de mettre en évidence un effet induit par l'un d'entre eux, à savoir une modification du cycle journalier de la température.

Il aurait été aussi possible d'utiliser les résultats de l'analyse factorielle pour décomposer le processus des températures en une somme de processus approximables de façon quasi-déterministe -comme l'ont fait OTTER et SCHUUR (1982) pour des séries temporelles économiques-, mais cette approche était ici sans relation avec le problème soulevé.

Au terme de ce travail, on ne peut qu'insister à nouveau sur la nécessité d'une bonne analyse descriptive des données avant de mettre en oeuvre des méthodes plus élaborées; l'examen des périodogrammes nous a permis de trouver une transformation adéquate des données avant l'application de l'analyse factorielle.

Annexe I

Ecriture de la matrice A à $2(n-1)$ lignes et à n colonnes intervenant au paragraphe II.1

Cette matrice est formée de $(n-1)$ blocs 2×2 :

$$\begin{pmatrix} 1/2 & 1/2 & 0 & 0 & \dots & 0 & 0 \\ -1/2 & 1/2 & 0 & 0 & \dots & 0 & 0 \\ 0 & 1/2 & 1/2 & 0 & \dots & 0 & 0 \\ 0 & -1/2 & 1/2 & 0 & \dots & 0 & 0 \\ 0 & 0 & 1/2 & 1/2 & \dots & 0 & 0 \\ 0 & 0 & -1/2 & 1/2 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & 0 & \dots & 1/2 & 1/2 \\ 0 & 0 & 0 & 0 & \dots & -1/2 & 1/2 \end{pmatrix}$$

On a donc :

$$\begin{cases} a_{2j-1,j} = a_{2j-1,j+1} = a_{2j,j+1} = 1/2 \\ a_{2j,j} = -1/2 \\ a_{jk} = 0 \end{cases} \quad \begin{matrix} \text{pour } j = 1, \dots, n-1 \\ \\ \text{sinon} \end{matrix}$$

Annexe II

Résultats numériques concernant le paragraphe IV

1) Vecteur des différences au retard 4 des températures moyennes

$$\bar{y}_{t_j} = \bar{x}_{t_j} - \bar{x}_{t_j-4} \text{ pour } j = 5, \dots, 29 \quad (\text{pour le graphe, cf.Fig.2}) :$$

-.05312	.31250	.34063	.18437	.05312	-.24063	-.25625	-.15937	-.0625	.16563
.24063	.125	.125	-.10313	-.26875	-.05937	-.17812	.175	.30625	.07187
.16562	-.175	-.24375	-.08437	-.05					

2) Première fonction propre normée :

$$f_1(t_j) \quad j = 5, \dots, 29 \quad (\text{pour le graphe, cf.Fig.4}) :$$

-3.240534E - 02	-.2400262	-.2447298		-.2079959	-.0454262	.154855
.2535324	.2073377	7.881921E - 02		-4.213681E - 02	-.2634761	-.1602022
-1.338402E - 02	.1848104	.2860157		.2134837	-1.286325E - 02	-.2686527
-.2839071	-.2282533	-2.883474E - 02		.2744338	.3500726	.1906056
3.935078E - 03						

3) Premier facteur $\xi_1(\omega)$, $\omega = 1, \dots, 32$

PAN

1.291848	3.759667E - 02	.8415416	.185628	.2528933	.5691953	1.84687
.8393818	.620361	.8694472	-.638887	-1.568949	-.6485511	1.80794
-.531863	.2167764	.6455514				

CUP

-.1046182	.1037288	-.26349	-2.900153	-.7599041	-.4624092	1.465264
-.8773166	-.1575101	.6883858	-1.101944	-.9529883	-.3600581	.5175525
-1.47132						

Remerciements

L'auteur exprime ses remerciements à B. BURTSCHY qui l'a aidée pour ce travail par ses conseils, et au Docteur G. LONDON (centre hospitalier F.X. MANHES) qui lui a confié les chroniques de températures relevées dans le cadre d'une étude clinique en cours.

Bibliographie

- CHATFIELD C. (1985) - The analysis of time series : Theory and practice, *Chapman and Hall*, 4^o édition.
- DEVILLE J.C. (1973) - L'analyse harmonique dans le cas de données discrètes, *note interne de l'INSEE*.
- DEVILLE J.C. (1974) - Méthodes statistiques et numériques de l'analyse harmonique, *Annales de l'INSEE*, n°15, 3-97.
- DEVILLE J.C. (1978) - Analyse et prévision des séries chronologiques multiples non stationnaires, *Statistique et Analyse des Données*, n°3.
- GOURIEROUX C., MONTFORT A. (1983) - Cours de séries temporelles, *Economica*.
- OTTER P.W., SCHUUR J.F. (1982) - Principal component analysis in multivariate forecasting of economic time-series, *Time series analysis*, O.D. Anderson (editor).
- SAPORTA G. (1981) - Méthodes exploratoires d'analyse de données temporelles, *Cahiers du B.U.R.O.* n°37-38.
- STATISTICAL GRAPHICS CORPORATION, STSC, Inc. (1986) - *Statistical graphics system*.
- VOLLE M. (1985) - Analyse des données, *Economica*, 3^o édition.