Statistique et analyse des données

Jean Le Nouvel

Analyse morphologique de courbes et application à un problème médical

Statistique et analyse des données, tome 7, n° 3 (1982), p. 26-54
<http://www.numdam.org/item?id=SAD_1982__7_3_26_0>

Statistiques et Analyse de données
1982 - Vol. 7, n° 3 pp 26-54

ANALYSE MORPHOLOGIQUE DE COURBES ET APPLICATION A UN PROBLEME MEDICAL

Jean LE NOUVEL

Département Statistique
Institut Universitaire de Technologie
Université de Haute Bretagne
Rue Montaigne - 56008 VANNES

Laboratoire de Statistique
Unité 49 de l'INSERM
CHR Pontchaillou
35000 RENNES

Résumé : *Nous proposons ici une application de l'analyse en composantes principales au cas de variables définies sur un continuum d'individus (intervalle (a,b)).*

Le calcul étant pratiquement impossible, nous avons recours à une approximation par des fonctions en escalier définies sur une partition en n intervalles de (a,b).

Nous montrons, sans utiliser les processus aléatoires, que l'ACP des fonctions en escalier converge vers celle des fonctions continues ($n \to \infty$), qu'elle est identique à celle des fonctions discrétisées en n points, et qu'elle rend ainsi possible l'emploi des programmes classiques d'ACP.

Les composantes principales étant de même nature que les fonctions initiales, elles peuvent donc être représentées graphiquement, et servir de base de décomposition.

Le centrage et la réduction d'échelle n'affectant pas les formes, cette méthode permet une analyse morphologique de la famille de courbes.

Un exemple commenté en montre le mode d'utilisation.

Abstract: *We propose here an application of the Principal Component Analysis in case of the variables defined on a continuum of the individuals (in the interval (a,b)).*

The calculation being practically impossible, we need to use the approximation by step functions defined on the partition of the interval (a,b) in n subdivisions.

It is shown, without using the random processes, that the Principal Component Analysis of the step functions converges to that of the continuous variables as $n \to \infty$ *, and is identical to the Principal Component Analysis of the functions "descretized" to n points. This property allows the use of the classical computer program of the Principal Component Analysis.*

Since the Principal Components are of the same nature as the initial functions, they can be represented graphically and be used as the basis of decomposition.

Standardization of the scale having no effect on the forms, this method enables a morphological analysis of a family of curves.

A commented example shows how to use the method.

Mots clés : *Variables continues, Approximation, Fonctions en escalier, Analyse en Composantes Principales, Convergence, Analyse morphologique.*

1 - INTRODUCTION.

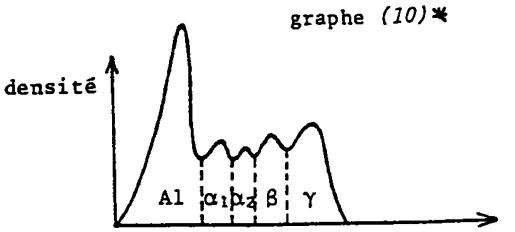
Le problème méthodologique traité dans cet article a eu pour origine l'application médicale suivante:

L'électrophorèse des protéines de sérum humain est un examen discriminant dans certains protocoles d'aide au diagnostic, et donc très utilisé par les médecins. (200 examens en moyenne, par semaine, au C.H.R. de Pontchaillou à Rennes).

La séparation des protéines résulte de leurs différences de vitesse de migration, dans une solution tampon soumise à un champ électrique.

Le spectre ainsi obtenu est lu par un densitomètre qui le transforme en une courbe du type suivant :

graphe *(10)**



électrophorèse de protéines sériques

* Les graphes sont numérotés en italique de graphe *(10)* à graphe *(38)*.

Et c'est par intégration qu'est estimé le pourcentage de la surface sous la courbe relative à chaque fraction protéique.

Les médecins ont pour habitude d'exploiter l'information contenue dans l'électrophorégramme, en tenant compte :

- de la morphologie de la courbe,
- des poids relatifs des 5 fractions protéiques.

L'interprétation des profils et leur liaison avec des pathologies relèvent de l'expérience acquise au cours des observations médicales.

Le traitement statistique des résultats classiques des électrophorégrammes est, en général, fondé sur les poids des 5 caractéristiques classiques : albumine, alpha 1, alpha 2, bêta et gamma globulines, à l'exclusion de l'allure générale de la courbe qui est appréciée qualitativement par le médecin. Des typologies de courbes associées à des classes de pathologie ont été ainsi obtenues par des méthodes d'analyse factorielle. (7)

Au Laboratoire de Statistique de l'INSERM, U 49, à Rennes, il nous a semblé intéressant d'affiner les typologies, en prenant en compte, directement, la morphologie des courbes

En concertation avec les médecins, nous avons donc choisi un protocole de saisie automatique des courbes sous une forme numérique très voisine du relevé effectué par le densitomètre.

Hypothèse de continuité. La justesse de la mesure alliée à la qualité de la numérisation engendrent des courbes pouvant être considérées comme continues eu égard à la variation du phénomène.

Il apparaît ainsi naturel d'assimiler un tel corpus de courbes à une famille finie de fonctions continues sur un intervalle commun T de $\mathbb{R}$.

Cette phase délicate (6) nous a permis d'obtenir un échantillon de 628 courbes discrétisées en 99 points et couvrant un large champ pathologique.

Objectifs. Afin d'analyser cette famille de courbes, nous souhaitons proposer une méthode permettant de :

- déceler les courbes qui se ressemblent ou se distinguent du point de vue de leurs formes;
- mettre en évidence ce qui les différencie ou les rapproche;
- pouvoir situer de nouvelles courbes;
- relier la typologie ainsi réalisée sur la famille de courbes au phénomène qu'elles représentent (classes pathologiques).

Méthode. Dans un espace euclidien, considérons une famille finie de courbes discrétisées en n points d'un intervalle T.

L'application de l'analyse en composantes principales à leurs ordonnées en chaque point permet de dégager une base orthonormée de courbes ajustée à la famille.

L'étude des éléments de cette base et des représentations des courbes dans le sous-espace ajusté rend possible l'analyse de la structure de cette famille.

En analyse en composantes principales, cet ajustement peut être réalisé soit dans l'espace des individus, soit dans celui des variables.Le choix entre ces deux modèles sera effectué en fonction de leur aptitude à respecter la forme des courbes.

Dans le traitement classique (ACP centrée ou normée), les courbes sont considérées comme des individus, et les instants de discrétisation jouent le rôle de variables.

Le centrage et, éventuellement, la réduction en chaque point de discrétisation sont des transformations non indépendantes affectant la forme des courbes.

En effet, la moyenne et l'écart-type varient avec les instants et dépendent de toutes les courbes.

Ce modèle dont la généralisation en dimension infinie sur T a donné l'analyse harmonique (!) est donc mal adapté à l'étude morphologique des courbes.

En revanche, la seconde présentation, dans laquelle les courbes jouent le rôle de variables d'une ACP normée, substitue à l'ACP des courbes initiales celle des courbes centrées réduites. Le centrage et la réduction effectués sur chacune des courbes sont alors indépendants.

Les fonctions f et af + b (où a et b sont des scalaires) sont considérées comme identiques du point de vue de l'analyse qui ne prend donc en compte, essentiellement,que la forme des courbes.

Dans l'hypothèse de continuité avancée précédemment, une semblable démarche suppose la généralisation de l'ACP à des variables continues sur un intervalle T de $\mathbb{R}$.

Nous devons donc situer cette analyse en dimension infinie, et recourir à des approximations permettant les calculs et dont on devra vérifier la convergence et la stabilité.

Afin de distinguer clairement cette méthode de l'Analyse Harmonique, nous emploierons le terme d'Analyse Morphologique pour désigner l'ACP normée de variables continues sur un intervalle T de $\mathbb{R}$.

2 - GENERALISATION DE L'ACP A DES VARIABLES CONTINUES DEFINIES SUR UN INTERVALLE T DE $\mathbb{R}$.

2.1 - Propriétés des fonctions.

Les variables soumises à l'ACP sont en réalité des courbes, représentations graphiques de fonctions possédant les propriétés suivantes:

- définies sur un intervalle T = (a,b) fermé;
- à valeurs dans $\mathbb{R}$;
- continues sur (a,b).

Elles constituent une famille F finie de p fonctions notées:

$$(f_j)_{(j=1,\ldots p)}$$

2.2 - L'espace $L^2(T)$.

Considérons l'intervalle T = (a,b) muni de ses boréliens et de sa mesure de Lebesgue normalisée à 1 comme espace de probabilité.

Les courbes $f_j(t)$ définissent des variables aléatoires sur cet espace. L'ACP classique se généralise à des variables définies sur un intervalle T (continuum d'individus) en se plaçant dans l'espace de Hilbert $L^2(T)$.

Les fonctions f_j présentées précédemment, étant continues sur T, appartiennent bien à $L^2(T)$.

Le produit scalaire de deux fonctions f_j et f_k se définissant par :

$$< f_j, f_k > = \frac{1}{b-a} \int_a^b f_j(t).f_k(t)dt$$

- si les fonctions sont centrées, nous obtenons la covariance;
- si elles sont centrées réduites, nous obtenons le coefficient de corrélation.

2.3 - Définition des composantes principales.

Elles se définissent comme dans le cas fini.

La première composante principale C_1 de p variables f_j est la combinaison linéaire de ces dernières de variance maximale. C'est un élément de $L^2(T)$.

La seconde possède les mêmes propriétés sous contrainte d'orthogonalité à la première.

Les suivantes se définissent par itération du procédé, sous contrainte d'orthogonalité aux composantes précédentes.

2.4 - Propriétés des composantes principales.

Fonctions de même nature que les fonctions initiales.

Les composantes principales sont des combinaisons linéaires des variables initiales $f_j \in F$; les coefficients s'obtiennent en diagonalisant la matrice des

- . produits scalaires en ACP non centrée,
- . covariances en ACP centrée,
- . corrélations en ACP normée (analyse morphologique).

Le vecteur associé à la $k_{ième}$ valeur propre prise dans l'ordre décroissant donne les coefficients u_{jk} de la $k_{ième}$ composante principale sur la $j_{ième}$ courbe.

$$C_k(t) = \sum_{j=1}^{p} u_{jk} . f_j(t)$$

Combinaisons linéaires de fonctions continues, les composantes principales sont continues sur T. Elles appartiennent également à $L^2(T)$ et sont donc de même nature que les fonctions initiales.

Elles constituent un système orthonormé S de $L^2(T)$.

Les composantes principales normées constituent un système orthonormé S de $L^2(T)$ de dimension $l \leq p$ (l étant le nombre de valeurs principales non nulles). En ACP centrée ou centrée normée, elles sont de moyenne nulle, de variance unité, et sans corrélation.

Décomposition orthogonale des fonctions initiales.

Une fonction quelconque f_j se décomposera comme suit :

$$f_j(t) = \sum_{j=1}^{l} \lambda_k^{1/2} . a_{kj} . C_k(t)$$

en ACP non centrée	:	$a_{kj} = \langle f_j, C_k \rangle$	produit scalaire
en ACP centrée	:	$a_{kj} = \mathrm{cov}(f_j, C_k)$	covariance
en ACP normée (analyse morphologique)	:	$a_{kj} = r(f_j, C_k)$	coefficient de corrélation .

Représentation graphique des composantes principales.

Etant de même nature que les fonctions initiales, les composantes principales sont des fonctions réelles, continues sur T.
Elles peuvent être représentées par une courbe. L'analyse des courbes représentant les principales composantes sera essentielle pour mettre en évidence la structure de la famille F.

Sous-espace ajusté à la famille F des courbes.

Le système orthonormé S des composantes principales engendre le sous-espace le plus proche de la famille F, au sens des moindres carrés. En effet, les composantes principales maximisent la variance du nuage représentant F, sous contrainte d'orthogonalité aux composantes de rangs inférieurs.

2.5 - ACP et décomposition de Fourier.

La décomposition des courbes sur une base orthonormée peut être comparée à la décomposition de Fourier. Dans celle-ci, les éléments de la base sont fixés; ce sont des fonctions sinwt et coswt, alors que dans l'ACP, les éléments de la base sont calculés à partir de la famille F, de manière à s'ajuster le mieux possible à cette dernière.

2.6 - Influence du centrage et de la réduction.

Ces deux transformations indépendantes conduisent à substituer à l'ACP des courbes initiales, celle des courbes centrées réduites.

Notons que les fonctions f et af + b ont même fonction centrée réduite $\frac{f - \bar{f}}{\sigma_f}$, et sont donc considérées comme identiques du point de vue de l'analyse. Celle-ci prend en compte essentiellement la forme des courbes.

Ces transformations ont pour effet de fixer, pour chaque courbe, une origine ne dépendant que d'elle même et d'en normaliser la dispersion, et donc de s'affranchir d'une échelle.

Conséquences en ACP normée (analyse morphologique).

Les courbes exceptionnelles, voire aberrantes, interviendront donc, comme les autres, dans la détermination des composantes principales, mais elles n'auront pas d'influence sur les transformations effectuées sur les ordonnées des autres courbes.

2.7 - Choix de la variante d'ACP (non centrée, centrée, normée).

Si l'on désire comparer les courbes du point de vue de leurs formes, l'emploi de l'ACP normée (analyse morphologique) s'impose, car elle permet de comparer des courbes produites à des instants ou endroits différents, sans contrainte d'étalonnage du niveau des appareils.

Si l'on veut prendre en compte le niveau des courbes, on doit disposer d'une origine et d'une échelle communes à toutes. L'emploi de l'ACP non centrée s'impose alors. Elle conduira à des résultats pouvant différer sensiblement de l'ACP normée, puisque les contributions des courbes à la détermination des composantes seront inégales. L'emploi complémentaire des deux analyses peut, cependant, être très fructueuse.

L'emploi de l'ACP centrée revient à conférer aux courbes des poids différents dans leur détermination, ce qui n'est pas forcément justifié.

2.8 - Représentation de l'intervalle T = (a,b).

L'intervalle (a,b) joue le rôle des individus en ACP finie. Le choix de la mesure de Lebesgue implique que l'on considère tous les instants individus comme équivalents.

Dans le cas fini, l'ensemble des individus est représenté dans $\mathbb{R}^p$ par un nuage de points. Les composantes sont les projections de ce nuage sur les axes principaux d'inertie. On a donc une représentation simultanée des variables et des individus, les positions des uns expliquant les positions des autres.

Les distances entre individus, aussi bien représentées que possible par les composantes principales, sont les distances induites par l'ensemble des variables.

Que se passe-t-il, lorsque les individus sont représentés par un continuum (a,b) ?

Dans $\mathbb{R}^p$, il sera représenté par un arc de courbe continu, puisque les fonctions f_j sont continues. Il est paramétré par $t \in (a,b)$. Deux points t et t' de (a,b) sont représentés par des points de cet arc de courbes, dont la distance est :

$$d^2(t,t') = \sum_{j=1}^{p} (f_j(t) - f_j(t'))^2$$

Les deux points seront proches (ou même confondus), si les p courbes (centrées normées) ont des valeurs proches (ou égales) en ces deux points.

Cet arc est le support de l'image par $F = (f_1 \ldots f_j \ldots f_p)$ de la mesure de Lebesgue sur (a,b).

Les premières composantes principales seront la projection de cet arc de courbe sur les axes principaux d'inertie. La distance entre les points sera conservée au mieux de par la qualité de l'ACP.

La projection sur un plan principal donnera un arc paramétré continu. (On notera les valeurs du paramètre de manière quelconque). Dans la réalité, on notera un certain nombre de points de (a,b). La proximité entre ces points permettra de repérer la structure d'ordre du continuum sur l'arc projeté.

On décèlera aussi les intervalles de (a,b) où les courbes ont des valeurs semblables, et éventuellement, des phénomènes périodiques.

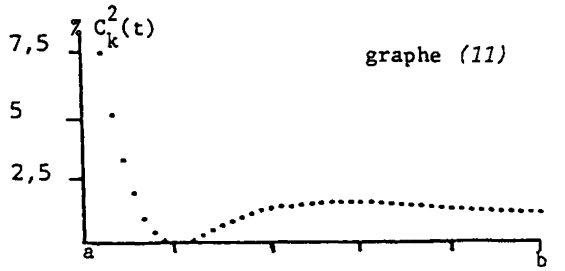
2.9 - Contributions des sous-intervalles de (a,b) à la détermination des composantes principales.

Considérons la $k^{ième}$ composante principale, qui, nous l'avons vu, est une fonction continue sur T.

Soit $C_k(t)$, l'ordonnée de cette composante pour la valeur t du paramètre, la contribution relative d'un intervalle quelconque (c,d) est alors :

$$\int_c^d C_k^2(t)dt$$

Si nous subdivisons (a,b) à pas constant, la visualisation de la courbe $C_k^2(t)$ permettra de déceler immédiatement les zones de forte contribution.



graphe *(11)*

Contributions relatives des individus (intervalles)

2.10 - Reconstitution approchée d'une fonction; importance des premières composantes principales.

Qualité de la représentation.

La reconstitution approchée donne : $f_j(t) = \sum_{k=1}^{h} \lambda_k^{1/2} a_{kj} \mathcal{C}_k(t)$ $\quad (h<n)$

Les premières composantes principales permettent fréquemment de reconstituer, très correctement, les fonctions initiales. On devine l'intérêt qu'elles représentent, pour analyser la famille de fonctions F à laquelle elles sont ajustées.

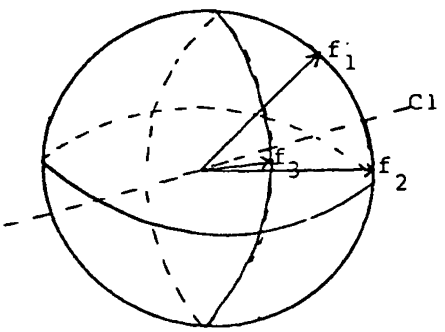
Importance des premières composantes principales.

Première composante.

La première composante principale, se trouvant sur la direction maximisant la variance du nuage des fonctions initiales $f_j \in F$, devrait constituer une excellente synthèse de la population, si celle-ci est assez homogène.

En Analyse Morphologique notamment, elle devrait rendre compte de l'allure générale des formes des fonctions de F. Remarquons qu'il ne s'agit pas, alors, de la moyenne des courbes, mais de la courbe la plus proche au sens des moindres carrés.

graphe *(12)*



Si les courbes ont des formes assez semblables, elles seront proches sur l'hypersphère, et très corrélées avec la première composante principale.

Si la famille n'est pas homogène, la première composante principale sera un résumé assez "bâtard" des courbes de la famille. Ceci devrait se traduire, dans les résultats, par des corrélations faibles avec certaines courbes. On pourra alors séparer la famille en sous-familles homogènes, dont on fera l'ACP séparément.

Le pourcentage d'inertie associé rendra compte de son importance dans la décomposition; si la famille est homogène, on peut s'attendre à une forte contribution de la première composante. (effet de forme).

Autres composantes principales.

Les composantes suivantes devraient rendre compte des différences entre fonctions ou des écarts à la première composante.

L'étude des plans principaux et des corrélations entre fonctions devrait permettre de créer une typologie sur les fonctions, à la condition qu'elles y soient bien représentées.

Représentation de fonctions supplémentaires.

Toute nouvelle fonction, dont on connaîtra les p coefficients de corrélation avec les composantes principales, pourra être représentée et projetée sur les plans principaux.

3 - L'APPROXIMATION. Résumé cf.(6)

La généralisation de l'ACP à des variables continues, définies sur l'infinité d'individus de (a,b), étant pratiquement impossible, nous aurons recours à une méthode d'approximation dont nous étudierons la convergence et la stabilité.

Notre démarche consiste à :

. effectuer un partage de I = (a,b) en une partition finie de n intervalles ;

. approximer les f_j par des fonctions en escalier $f_{(n)j}$

soit (e_i) ; $0 \leqslant i \leqslant n$ une subdivision de l'intervalle (a,b)

avec $e_0 = a$ et $e_n = b$, n réels (t_i) tels que $t_i \in]e_{i-1}, e_i[$

$$f_{(n)j}(t) = f_j(t_i) \quad \text{pour tout } t \in [e_{i-1}, e_i[$$

L'intégrale de la fonction en escalier $f_{(n)}(t)$ précédemment définie, converge au sens de Rieman vers l'intégrale de la fonction $f(t)$ quand $n \to \infty$.

Approximation et convergence du produit scalaire.

$f_{(n)j} \cdot f_{(n)k}$ produit de deux fonctions en escalier, est encore une fonction en escalier sur (a,b).

La fonction continue approximée correspondante est le produit $f_j \cdot f_k$

La convergence de l'intégrale des fonctions en escalier sur (a,b) entraîne donc celle des produits scalaires.

Le produit scalaire des fonctions en escalier converge donc vers celui des fonctions continues sur (a,b).

4 - CONVERGENCE DE L'ACP DES FONCTIONS EN ESCALIER. Résumé cf.(6).

Les composantes principales de l'ACP des fonctions continues sont obtenues en diagonalisant la matrice (C) des produits scalaires. Nous noterons C l'opérateur de $\mathbb{R}^p$ correspondant.

Les composantes principales des fonctions en escalier sont obtenues en diagonalisant la matrice (C_n), l'opérateur associé est noté C_n. On dira que l'analyse des fonctions en escalier converge vers l'analyse des fonctions continues, si C_n converge vers C .

Dauxois et Pousse (3) ont montré que la convergence des opérateurs entraîne celle des sous-espaces propres, des valeurs propres, des axes principaux d'inertie et des composantes principales.

La convergence des produits scalaires ayant été démontrée, le terme général de $(C_n) \xrightarrow[n \to \infty]{}$ terme général de (C).

Ces opérateurs de dimensions finies (p x p) convergeant terme à terme, convergent en norme.

L'ACP des fonctions en escalier converge donc vers l'ACP des fonctions continues, quand n croît indéfiniment.

5 - MISE EN OEUVRE DE L'ACP DES VARIABLES CONTINUES SUR T.

Pour une subdivision (e_i) de (a,b), l'ACP approchée des fonctions continues s'obtient en diagonalisant la matrice des produits scalaires de terme général: intégrale de Rieman.

$$C_{(n)jk} = \frac{1}{b-a} \int_a^b f_{(n)j}(t) \cdot f_{(n)k}(t)dt$$

$$= \sum_{i=1}^{n} f_j(t_i) \cdot f_k(t_i) \cdot p_i \quad \text{avec} \quad p_i = \frac{e_{i-1} - e_i}{b-a}$$

5.1 - Définition de la fonction discrétisée associée à une fonction continue.

On appelle fonction discrétisée associée à la fonction continue f_j , une fonction $\tilde{f}_{(n)j}$ définie aux n points t_i des intervalles (e_{i-1}, e_i) et égale à f_j en ces points.

Le terme général $C_{(n)jk}$, calculé plus haut, est identique au terme général qui serait obtenu en calculant la matrice des produits scalaires des fonctions discrétisées aux points $t_i \in T$, à la condition d'accorder à ces derniers un poids p_i proportionnel à la longueur de l'intervalle qu'ils représentent.

5.2 - Le traitement et l'interprétation des résultats.

Le grand intérêt de cette approximation réside dans le fait que l'on peut utiliser les programmes classiques d'analyse en composantes principales sur le tableau des valeurs des fonctions aux n points t_i , en donnant à ce point le poids p_i.

Puisque $\tilde{f}_{(n)j}$ est la fonction discrétisée correspondant à la fonction en escalier $f_{(n)j}$, il est évident que la moyenne et la variance (pondérées par p_i) de $\tilde{f}_{(n)j}$ sont égales à leurs homologues de la fonction en escalier. De même, la corrélation des fonctions en escalier coïncide avec celles des fonctions discrétisées.

Les composantes principales que l'on obtient sont des combinaisons linéaires des fonctions discrétisées $\tilde{f}_{(n)j}$ définies sur les n points t_i .

On les considérera comme des approximations des composantes principales de la famille de fonctions continues. On tracera leur graphe en considérant qu'il s'agit d'une fonction continue, dont on connaît la valeur aux points t_i . On les interprétera comme on l'a indiqué au paragraphe précédent.

6 - COMPARAISON AVEC D'AUTRES METHODES.

L'analyse des courbes peut se rattacher à la généralisation de l'ACP à des infinités de variables et d'individus établie par Pousse et Dauxois, et à ses développements (3),(4),(5), ainsi qu'à l'analyse harmonique de Deville (1),(2).

Ces auteurs recourent à un modèle aléatoire.

6.1 - Analyse harmonique.

"L'analyse harmonique est une extension de l'analyse factorielle en composantes principales au cas où chaque individu est caractérisé par une courbe temporelle, une fonction, au lieu d'être caractérisé par un vecteur de dimension finie (un ensemble de variables)". (2)

L'utilisation de l'analyse harmonique centrée ou normée change considérablement la forme des courbes puisque moyenne et écart-type varient à chaque instant et dépendent de toutes les courbes.

Plus l'hétérogénéité des formes sera grande, plus grave sera la distorsion induite par l'analyse harmonique.

Approximation et convergence. résumé cf.(6)

L'analyse harmonique des courbes nécessite une extension de l'ACP dont la démonstration est aussi simple que dans le cas de l'analyse morphologique.

Nous disposons d'un nombre fini p d'individus (courbes) qui appartiennent, comme précédemment, à l'espace $L^2(T)$.

Les fonctions $f_j(t)$ ne sont plus centrées pour le produit scalaire canonique du cas précédent, mais centrées en chaque point t.

m(t) : courbe moyenne des $f_j(t)$.

Remarque. Cela revient à munir $L^2(T)$ du produit scalaire :

$$< f_j , f_k > = \frac{1}{b-a} \int_a^b (f_j(t) - m(t)).(f_k(t) - m(t))dt$$

6.2 - Comparaison des analyses harmonique et morphologique.

En analyse harmonique (centrée), on analyse la fonction $X_j(t) = f_j(t)-m(t)$ qui reflète l'écart en chaque point t de la courbe initiale à la courbe moyenne sur T de la famille F.

La décomposition orthogonale de $X_j(t)$ concerne donc les écarts à la tendance, et non plus comme précédemment, les courbes elles-mêmes.

Les harmoniques forment la base orthogonale de décomposition de ces écarts à la tendance. Il ne s'agit donc plus de représenter au mieux la famille, mais de dégager la structure différentielle des courbes à la tendance moyenne.

Compte tenu de cette intéressante propriété, l'analyse harmonique est utilisée pour étudier des phénomènes temporels.

L'analyse morphologique fournit une base orthogonale de fonctions de même nature, en faisant jouer un même rôle à toutes les courbes.

La première composante principale est, en général, très proche de la fonction moyenne, si la population est très homogène. Dans ce cas, la composante de rang 2, reflétant les écarts à la première composante principale, a une forme très voisine de la première harmonique obtenue en analyse harmonique centrée. Par contre, si la population est hétérogène, cette coïncidence devrait disparaître.

En conclusion, il s'agit de deux méthodes analysant, différemment, l'information contenue dans un corpus de courbes.

La première est une analyse chronologique des écarts à la tendance, tandis que la seconde est une analyse morphologique de la famille F des courbes.

7 - APPLICATION MEDICALE.

Nous avons soumis l'échantillon des 628 courbes d'électrophorèse discrétisées en 99 points au traitement par l'analyse morphologique :

. les variables soumises à l'ACP sont les 628 fonctions discrétisées à pas constant;

. les 99 points de discrétisation jouent le rôle d'individus.

7.1 - Caractéristiques globales de l'analyse morphologique.

Composantes	1	2	3	4	5	6	7	8
Valeurs propres	0,859	0,0428	0,0246	0,0199	0,0118	0,0106	0,006	0,005
% d'inertie	85,95	4,28	2,46	1,99	1,18	1,06	0,62	0,47
% cumulés	85,95	90,23	92,69	94,68	95,87	96,93	97,56	98,02

L'importance du pourcentage d'inertie de la première composante caractérise, dans cet exemple, l'homogénéité des formes des courbes, ce qui n'est pas surprenant, puisqu'elles reflètent un même phénomène. Les différences de formes seront illustrées par les autres composantes.

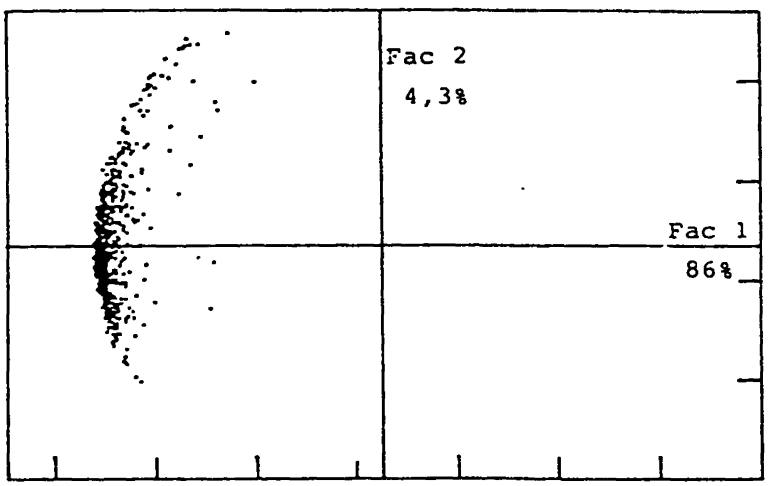
Afin de mieux faire apparaître leurs importances relatives, nous avons calculé, pour chaque composante de rang supérieur à 1, les pourcentages qu'elles prennent dans l'inertie non expliquée par la première composante :
(% d'inertie non expliquée = 100 - % d'inertie de la première composante).

Composantes	2	3	4	5	6	7	8
% d'inertie complémentaire	30,5	17,5	14,2	8,4	7,5	4,4	3,3
% cumulés	30,5	48	62,2	70,6	78,1	82,5	85,8

Nous emploierons le terme d'inertie complémentaire pour désigner ce pourcentage d'une composante.

7.2 - Plan principal (1,2) ; 90% de l'inertie totale.

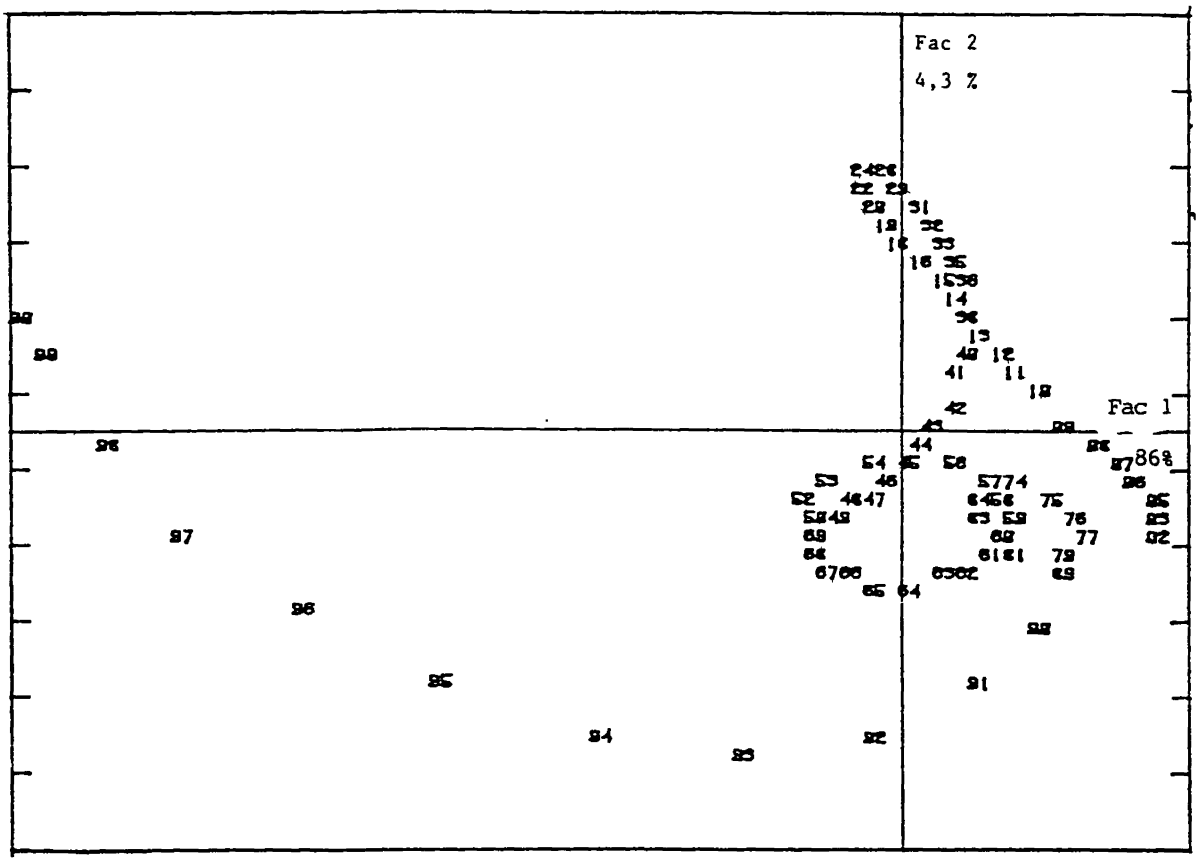
Position des courbes (variables). Graphe *(13)*.



La localisation du nuage des points variables sur l'arc de cercle voisin du premier axe reflète la forte corrélation de la plupart des courbes avec la première composante principale ainsi que leur bonne représentation sur ce plan.

Les courbes qui s'écartent du nuage sont évidemment les plus intéressantes à étudier, du point de vue de leurs formes qui devraient se différencier de celle de la première composante C1.

Arc paramétré des points de discrétisation (individus). Graphe (14).



L'arc paramétré reflète bien l'effet de chaîne dû à la structure d'ordre existant sur les points de discrétisation.

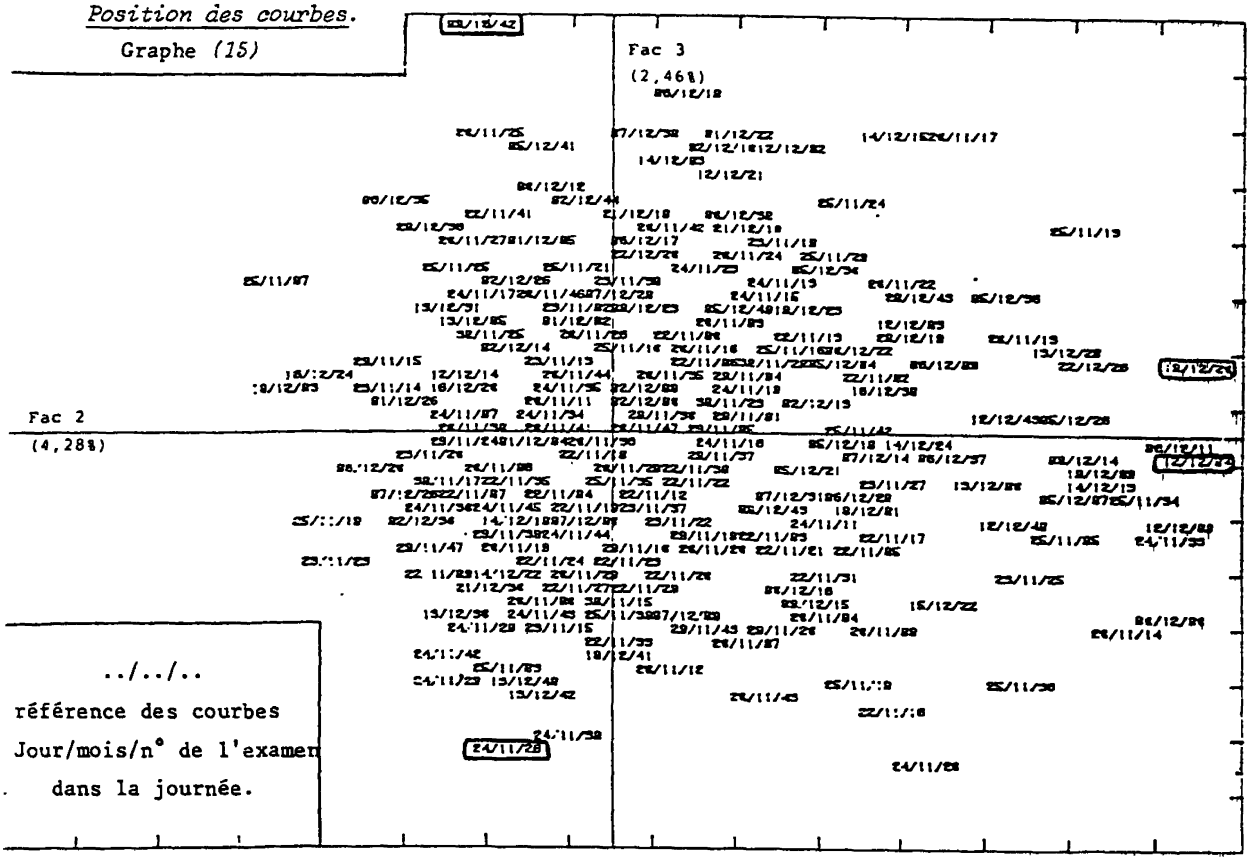
Le premier facteur oppose les derniers points (albumine) au reste, et notamment, aux dix premiers (gamma-globulines).

L'axe 2 oppose la fraction d'arc allant du 10 au 40e point à celle allant du 45 au 90e, c'est-à-dire les gamma-globulines aux autres fractions protéiques (alpha 1, alpha 2, beta et l'albumine).

7.3 - Plan principal (2,3).

(6,74% de l'inertie totale ; 47,9% de l'inertie complémentaire de la première composante).

Position des courbes.
Graphe *(15)*



C'est sur ce plan que les courbes sont les mieux séparées, l'effet de forme ne jouant plus.

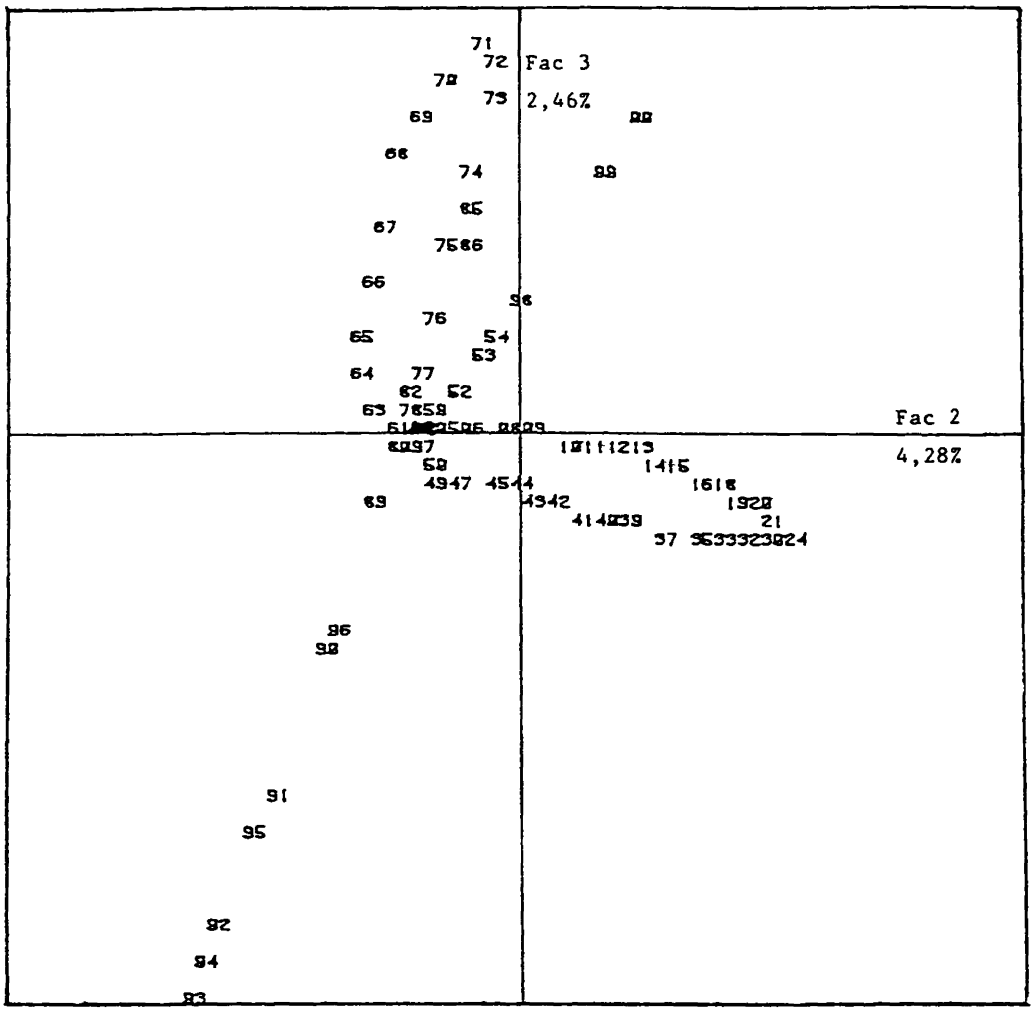
Les courbes bien représentées se trouvent à la périphérie du nuage de points, près du cercle de corrélation.

On peut ainsi déceler les courbes fortement liées aux composantes principales. Celles dont les références sont encadrées sont représentées graphiquement avec les numéros suivants : graphes n°*(28),(29),(30),(31)*.

La courbe , graphe *(30)*, présente une fraction alpha 2 plus importante que la normale associée à une fraction albuminique faible. Elle s'oppose à la courbe graphe *(31)* qui possède une fraction alpha 2 aussi forte mais associée à une albumine forte.

Arc paramétré des points de discrétisation (plan principal (2,3)).

Graphe *(16)*



La 2ème composante principale oppose, nous l'avons vu, les gamma-globulines aux autres fractions protéiques.

La 3ème oppose la partie d'arc allant du 61ème point au 77ème (correspondant, sur la courbe moyenne, aux alpha 2 globulines) aux premiers points représentant l'albumine (89 à 97).

Toutes ces remarques concernant les composantes principales (1,2,3) sont encore plus nettement décelables sur leurs représentations graphiques *(17), (18), (19)*.

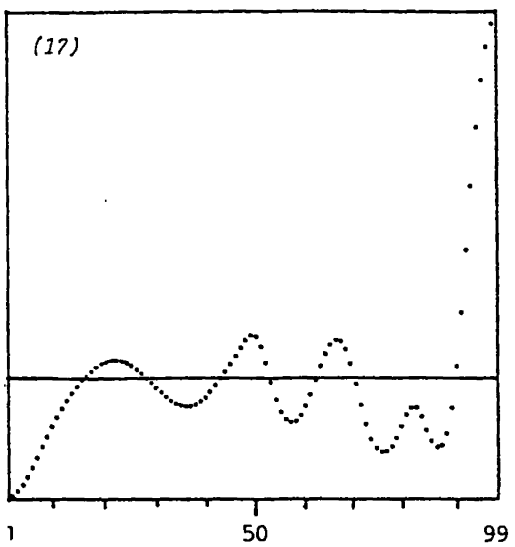
7.4 - Composantes principales.

Nous avons représenté, ci-dessous, les 4 premières composantes principales. (Graphes *(17)* à *(20)*)

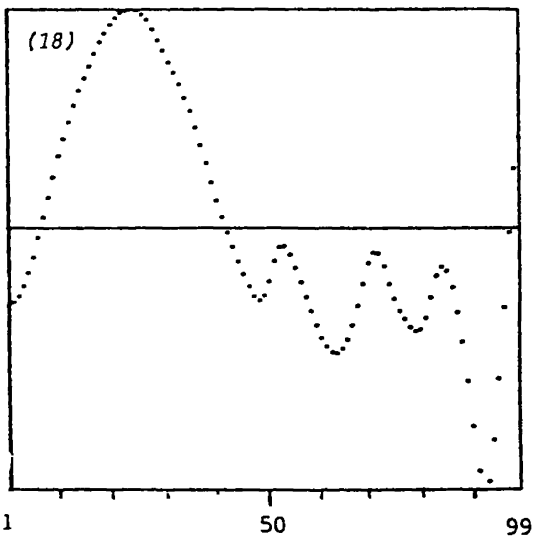
A la page suivante, figurent les graphiques des contributions des points de discrétisation à leur détermination. (Graphes *(21)* à *(24)*)

Pour rechercher leur signification, nous présentons ensuite des courbes qui leur sont fortement corrélées. (Graphes *(25)* à *(35)*)

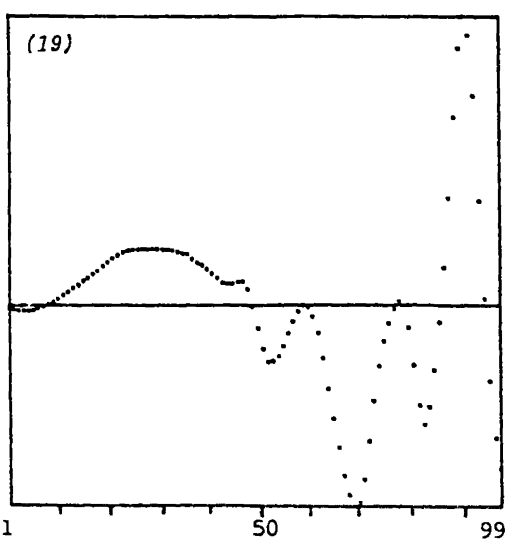
Composante principale C1



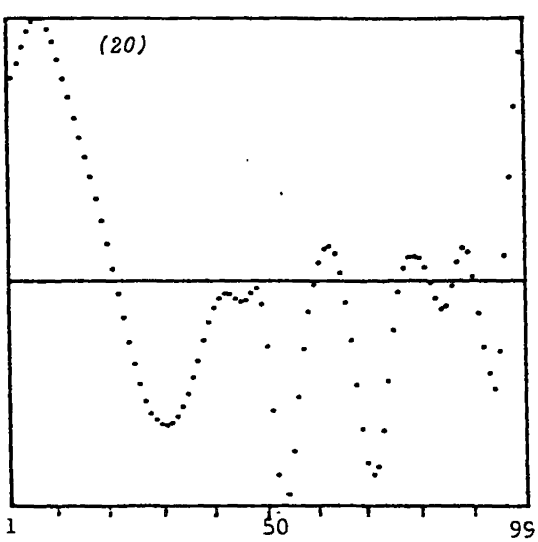
Composante principale C2



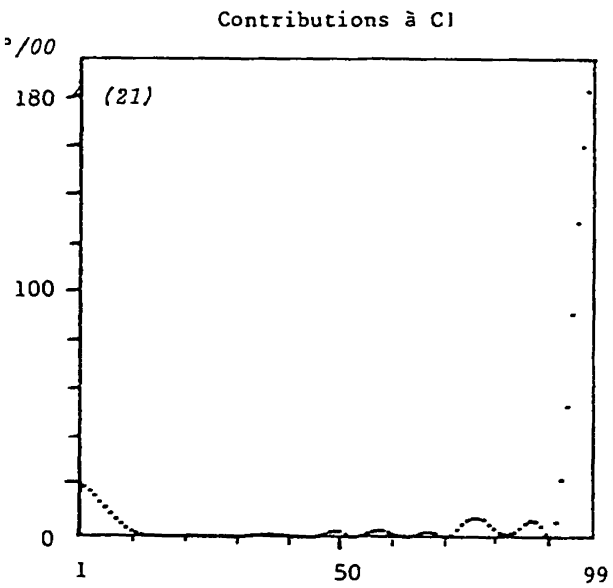
Composante principale C3

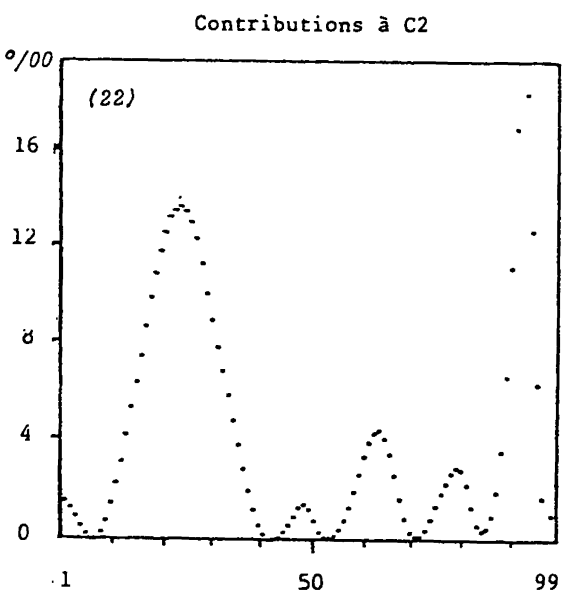


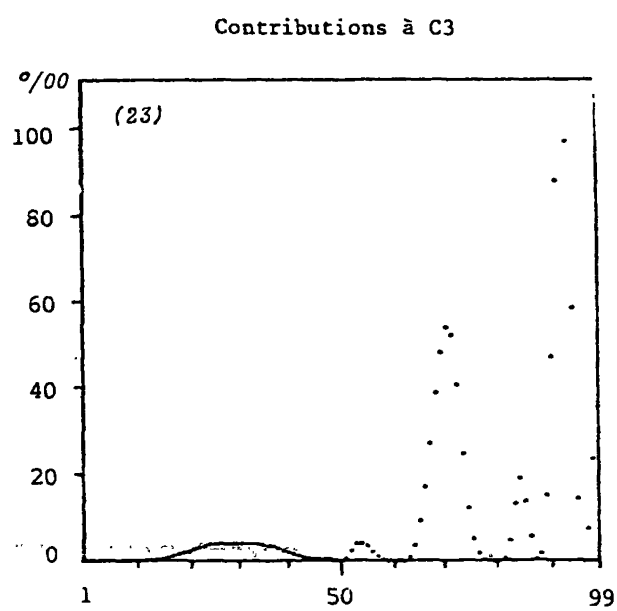
Composante principale C4

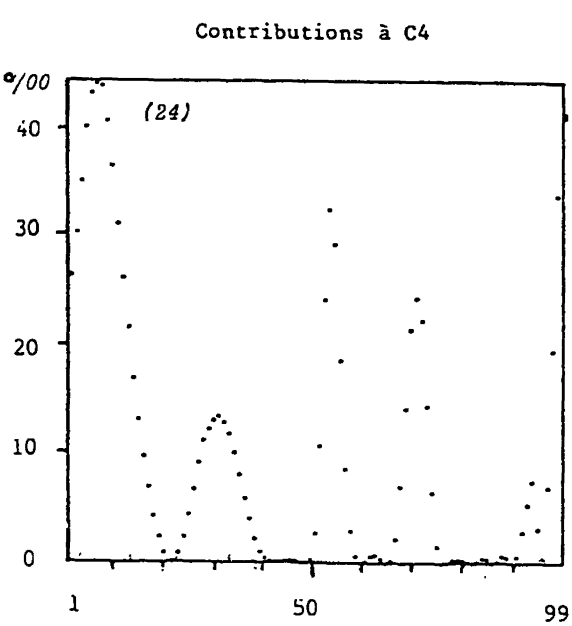


Contributions des différents points de discrétisation à la détermination des composantes.







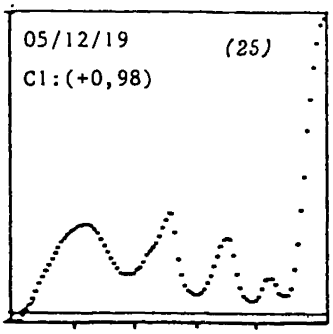


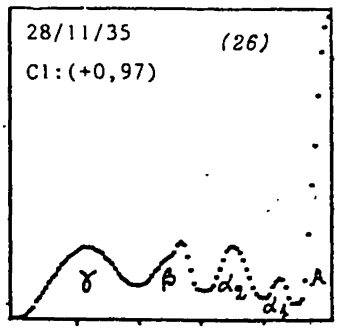
../../.. référence de la courbe.

C.:(.....)coefficient de corrélation avec la composante concernée.

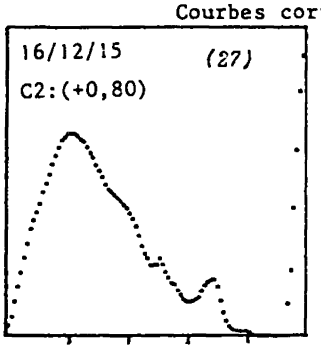
( ) n° du graphe

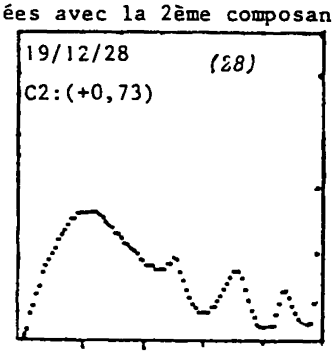
Courbes corrélées avec la 1ère composante principale

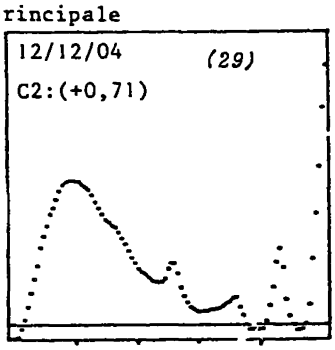




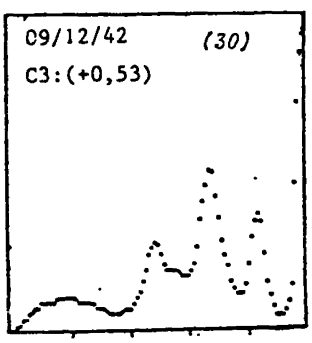
Courbes corrélées avec la 2ème composante principale

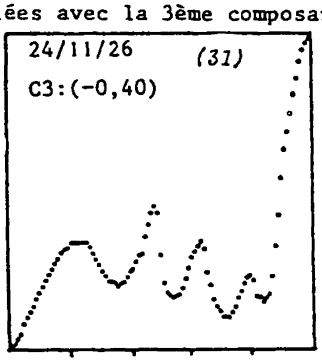


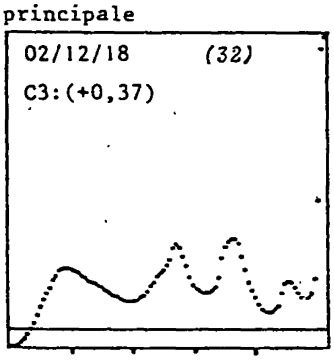




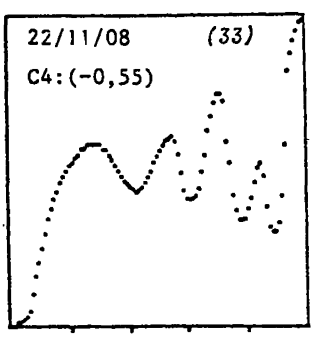
Courbes corrélées avec la 3ème composante principale

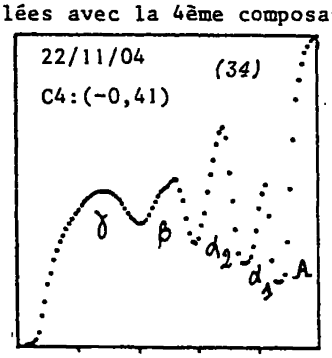


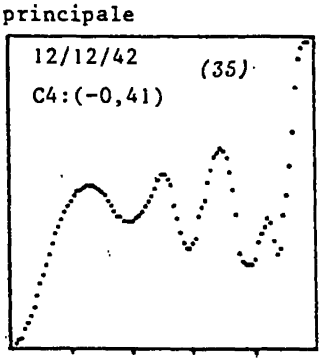




Courbes corrélées avec la 4ème composante principale







Décomposition des courbes sur la base des composantes principales.

Les composantes principales forment une base orthonormée de courbes discrétisées (en 99 points), ajustée au sens des moindres carrés à la famille des 628 courbes d'électrophorèses.

La décomposition sur la base des composantes principales s'effectue comme suit :

$$f_j(t) = \sum_{j=1}^{h} \lambda_k^{1/2} \cdot a_{kj} \cdot C_k(t)$$

k : $k^{ième}$ valeur principale

$C_k(t)$: $k^{ième}$ composante principale normée

a_{kj} : coefficient de corrélation de $f_j(t)$ avec $C_k(t)$

Première composante principale. (86% de l'inertie) C1 (graphe *(17)*).

La première composante principale reflète l'allure générale de la famille d'électrophorèses. Sa forme correspond à des courbes de sujets normaux, qui dominent dans le corpus.

L'importance du taux d'inertie (86%) , expliquée par cette composante, rend également compte de la parenté de forme existant entre les courbes. L'examen comparatif des formes des diverses composantes montre qu'elle est la seule à pouvoir être assimilée à une courbe d'électrophorèses.

Les formes des autres composantes principales devront donc s'interpréter par référence à l'allure générale dégagée par la première. Le calcul pour chacune d'entre elles, de la part d'inertie non expliquée par la première, prend alors tout son sens.

L'importance de l'effet de forme, classique en ACP, se traduit par une très forte corrélation de cette première composante avec les courbes du sous-groupe dominant (+0,90), ainsi que par des coefficients plus faibles mais non négligeables avec les courbes corrélées aux autres composantes.

Les composantes principales de rang supérieur, rangées par pourcentages d'inertie complémentaire décroissants, permettront de dégager des sous-familles, et de mettre en valeur les formes qui les caractérisent.

Contributions des points de discrétisation à la détermination de la première composante.

L'importance des 6 derniers points constituant le pic de l'albumine apparaît nettement sur le graphe *(21)*.

Seconde composante principale. (4,28% de l'inertie totale, 30,6% de l'inertie complémentaire. C2 (graphe *(18)*).

La deuxième composante n'est pas une courbe d'électrophorèse et doit s'interpréter par référence à la première. Sa forme est néanmoins plurimodale à cinq pics, dont les importances peuvent refléter les variations de forme par rapport à l'allure générale.

Elle est nettement marquée par la prééminence du premier pic caractérisant les gamma-globulines, et par le relatif effacement du dernier correspondant à l'albumine.

La sous-famille de courbes fortement corrélées à la seconde composante devrait présenter une variation de forme par rapport à la première liée au signe et à l'intensité du coefficient de corrélation.

On distinguera, en réalité, deux types de courbes selon le signe de la variation.

L'homogénéité des formes des 3 courbes présentées s'explique par un fort coefficient de corrélation de même signe positif. Graphes *(27),(28),(29)*.

Cette forme de courbes d'électrophorèses est fort connue des médecins pour son association avec les cirrhoses, qui représentent l'un des plus importants sous-groupes (109 individus; bloc beta-gamma dominant).

Contribution des points de discrétisation à la détermination de la deuxième composante.

L'examen du graphe des contributions *(22)* traduit bien l'importance des 40 premiers points représentatifs des gamma-globulines. Ils totalisent presque 60% de l'inertie totale, l'albumine ne représentant plus que 25%.

Troisième composante principale. (2,46% de l'inertie totale, 17,5% de l'inertie complémentaire. C3 : graphe *(19).*

L'examen des deux courbes (graphes *(30)* et *(32)*) corrélées positivement avec C3 distingue un profil associé par les médecins aux syndromes inflammatoires, qui représentent 134 individus dans l'échantillon.

La courbe (graphe *(31)*), corrélée négativement avec C3, a bien une forme différente des deux autres. Le pic des gamma est plus élevé et moins étendu , celui des albumines plus important, alors qu'il était presque inexistant dans les deux courbes précédentes,

Quatrième composante principale. (1,99% de l'inertie totale, 14,2% de l'inertie complémentaire. C4 : graphe *(20)*

L'examen des graphes *(33),(34),(35)* de cette composante et des courbes lui étant corrélées permet de résumer, schématiquement, un type de cette sous-famille (corrélation négative).

Les gamma-globulines (1er pic) sont nettement au-dessus de l'allure générale (C1), tandis que beta, alpha 1 et 2 sont légèrement au-dessus. Notons la faible incidence de l'albumine.

L'examen des contributions des points (graphe *(24)*) confirme cette analyse et, notamment, la faiblesse de l'apport albuminique (12% seulement).

Les 3 exemples de courbes présentent des coefficients de corrélation avec C4, de l'ordre de 0,4 à 0,5 et des formes très proches. (Graphes *(33),(34),(35).*

De l'avis des spécialistes, le niveau élevé des courbes pourrait être dû à un décalage artificiel du zéro.

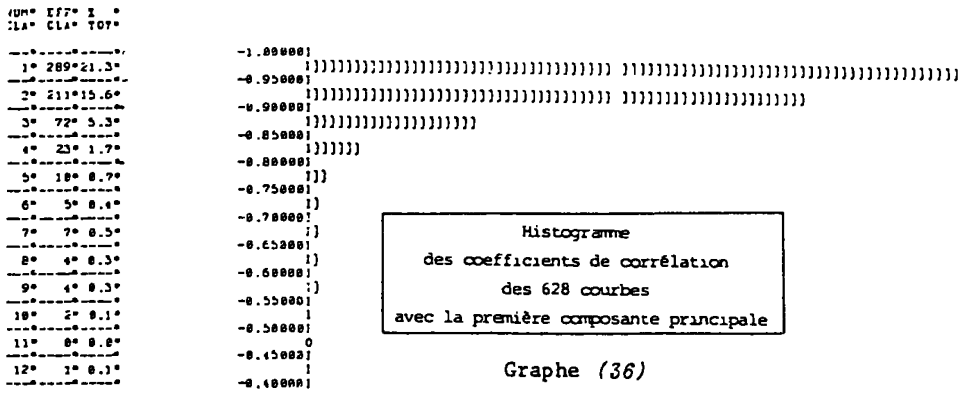
Du point de vue statistique, le fait de regrouper des courbes du même type est, au contraire, à inscrire au crédit de l'analyse morphologique.

7.5 - Détermination des sous-populations composantes.

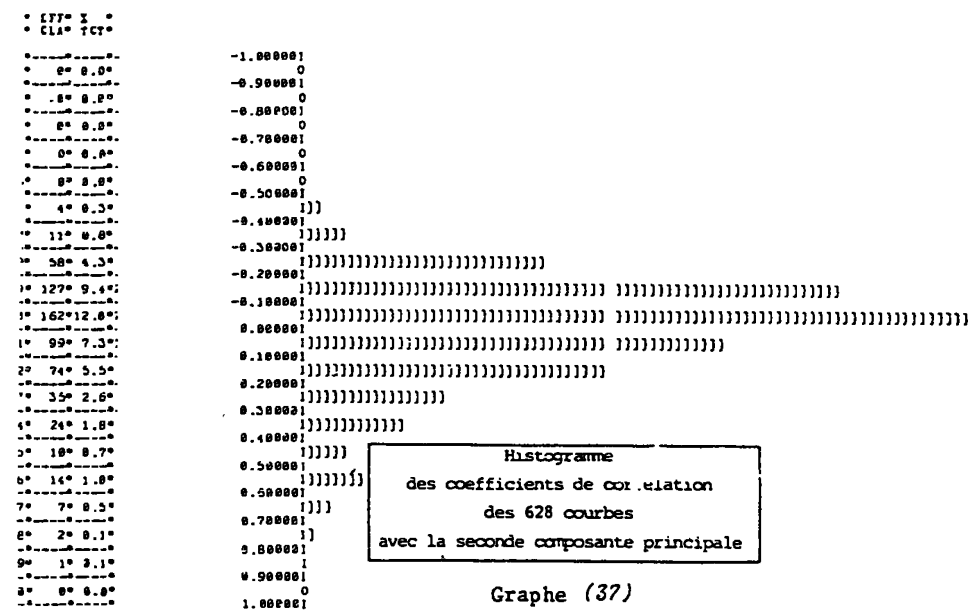
La recherche de la typologie des courbes peut se mener à partir des analyses des histogrammes :

. des coefficients de corrélation des courbes avec les composantes (graphes *(36),(37)*).

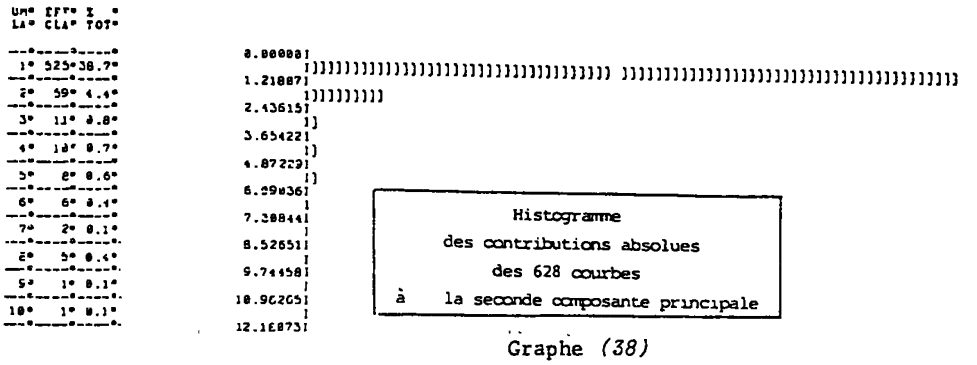
. des contributions absolues des courbes. (graphe *(38)*).



Graphe *(36)*



Graphe *(37)*



Graphe *(38)*

L'information contenue dans ces histogrammes se recoupe, et permet de de déceler les groupes d'individus liés positivement ou négativement aux composantes.

Le niveau élevé de la corrélation de forme existant entre les courbes et la première composante principale apparaît nettement sur l'histogramme des coefficients (graphe *(36)*).

On constate aussi que 38 courbes ont des coefficients de corrélation supérieurs à 0,4 (en valeur absolue) avec C2. (graphe*(37)*).

De même, on retrouve 44 courbes présentant des contributions supérieures à 3 fois la contribution normale (0,8) avec C2. (graphe *(38)*).

La recherche des sous-familles peut également être entreprise à partir des plans principaux, en relevant les groupes d'individus qui y sont bien représentés. *

7.6 - Perspectives.

Les sous-familles, mises en évidence par l'analyse morphologique, correspondent aux liaisons profils/pathologies rencontrées dans la pratique, notamment en ce qui concerne les cirrhoses et les syndromes inflammatoires.

Des profils associés à des pathologies plus rares apparaissent sur les composantes de rang supérieur, ou sur des plans principaux particuliers.

L'analyse morphologique ayant prouvé sa capacité à construire une typologie sur les formes de courbes d'électrophorèses, le prolongement médical de ces travaux consistera à établir des référentiels précis et éprouvés associant pathologies, composantes et plans principaux.

Le recueil systématique d'électrophorèses et leur numérisation à l'aide d'un convertisseur analogique digital permettrait de fournir des courbes directement analysables.

* *Dans cet article, nous nous sommes limités à l'étude des 4 premières composantes; pour plus de détails, cf.(6).*

B I B L I O G R A P H I E

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

2. DEVILLE J.C. - *"Analyse harmonique du calendrier de constitution des familles en France"*.
INED, Population n°1, 1977. (p.18 à 20).

3. DAUXOIS J. et POUSSE A.
- *"Les analyses factorielles en calcul des probabilités et en statistique. Essai d'étude synthétique"*.(p.257 à 271).
Thèse, Toulouse, 1976.

4. BESSE P. - *"Etude descriptive d'un processus"*. (p.51).
Thèse, Toulouse, 1979.

5. MONDOT A.M. - *"Problèmes numériques et statistiques relatifs à l'analyse en composantes principales des variables à valeurs dans un espace de Hilbert"*.
Thèse, Toulouse, 1979.

6. LE NOUVEL J. - *"Etude d'une famille de courbes par des méthodes d'analyse des données. Application à l'analyse morphologique de courbes provenant de données médicales"*.(p.77à91,20à22,23,133à159)
Thèse, Rennes, 1981.

7. KERBAOL M., LENOIR P., SIMON M. et BOUREL M.
- *"Systématisation d'un procédé de traitement des données cliniques"*.
Journées d'informatique médicale, Toulouse, 1970.(p.231 à 240).