REVUE DE STATISTIQUE APPLIQUÉE

J. B. KAZMIERCZAK

Analyse logarithmique. Deux exemples d'application

Revue de statistique appliquée, tome 33, n° 1 (1985), p. 13-24

<http://www.numdam.org/item?id=RSA_1985__33_1_13_0>

ANALYSE LOGARITHMIQUE
DEUX EXEMPLES D'APPLICATION

J.B. KAZMIERCZAK

Ecole Centrale des Arts et Manufactures.
Grande Voie des Vignes. 92290 Châtenay-Malabry.

RESUME

Le principe d'équivalence distributionnelle et le principe de Yule définissent deux classes de métriques vérifiant une seule et même propriété : le principe d'équivalence distributionnelle large. Le premier principe conduit à l'analyse factorielle des correspondances, le second à l'analyse logarithmique. Ces deux méthodes possèdent quelques points de ressemblance.

Deux exemples illustrent l'intérêt de l'analyse logarithmique.

SUMMARY

The principle of distributional equivalence and Yule's principle define two sets of metrics with one common property : they all satisfy a generalized principle of distributional equivalence. The first principle leads to correspondance analysis, the second to logarithmic analysis. This two methods have some similarity.

Two examples show the interest of logarithmic analysis.

C'est au début des années soixante que J.-P. BENZECRI énonce le principe d'équivalence distributionnelle (p.e.d.) – principe que nous rappelons plus loin. Pour la première fois, le choix de la métrique dans une méthode factorielle est posé en termes d'invariance.

En 1912 l'analyse des données est encore à l'état embryonnaire. Ce sont des préoccupations plus "classiques" qui conduisent YULE, pour l'étude des tables de contingence 2 x 2, à définir un coefficient d'association satisfaisant déjà à un principe d'invariance.

Dans une optique d'analyse de données, il est naturel – et immédiat – de généraliser le principe de YULE à des tableaux de dimension quelconque et qui ne seront pas obligatoirement des tables de contingence. On introduit de cette façon l'analyse logarithmique.

Le but de cet article est double.

D'une part, éclairer les liens qui unissent ces deux principes en montrant que tous deux satisfont à un principe d'invariance très voisin du p.e.d. Sous cette lumière, on ne s'étonnera pas de constater que l'analyse logarithmique (A.L.) et l'analyse factorielle des correspondances (A.F.C.) fournissent dans certains cas des résultats très voisins.

D'autre part, présenter deux exemples d'application de l'A.L. à des tableaux de mesures.

Mot clés : **Principe d'équivalence distributionnelle, Principe de Yule, Analyse logarithmique.**

1. TROIS PRINCIPES D'INVARIANCE EN ANALYSE DES DONNEES

Dans ce paragraphe, $X = (x_{ij})$ désigne une correspondance définie sur le produit IxJ où I et J sont deux ensembles finis de cardinal respectif n et p.

On note $(x_{i.})$ et $(x_{.j})$ les lois marginales de ce tableau, i.e. les totaux en ligne et en colonne. Un profil-ligne est défini par les rapports :

$$\{x_{ij}/x_{i.} \mid j \in J\}$$

On définit de façon analogue un profil-colonne.

Rappelons maintenant :

le *p.e.d.* : on ne change pas la distance entre deux profils-lignes (resp. colonnes) d'un tableau en remplaçant deux colonnes (resp. lignes) proportionnelles par une seule, somme des deux précédentes.

Dans le cadre euclidien, si on se limite aux métriques définies par des formes quadratiques diagonales ne dépendant que des marges, ce principe conduit à la métrique du chi-deux. (cf Annexe, in fine).

On est ainsi conduit à l'A.F.C. ; il ne reste plus qu'à munir chaque profil de sa masse naturelle correspondante (élément marginal du tableau). De cette façon, on ne modifie pas les résultats de l'analyse (recherche des axes principaux d'inertie) en effectuant sur le tableau l'opération de sommation décrite plus haut, dans l'énoncé du p.e.d.

Il est trivial de constater que cette transformation (lorsqu'elle est possible, i.e. lorsqu'il existe deux colonnes – ou deux lignes – proportionnelles) diminue la taille du tableau ! Nous proposons ici un énoncé légèrement modifié du p.e.d. où il est question d'une transformation qui n'affecte pas les dimensions du tableau. Enonçons :

le *p.e.d.-large* : on ne change pas la distance entre deux profils-lignes (resp. colonnes) en remplaçant deux colonnes (resp. lignes) proportionnelles par deux autres colonnes (resp. lignes) proportionnelles aux précédentes de telle sorte que la somme des deux initiales et des deux finales reste la même.

Il est immédiat de montrer que toute métrique vérifiant le p.e.d. vérifie également le p.e.d.-large. La réciproque est fausse : on peut trouver des métriques vérifiant le p.e.d.-large ne satisfaisant pas au p.e.d. (d'où la terminologie).

Il n'est pas de notre propos d'étudier ici les analyses induites par le p.e.d.-large, ce dernier n'est là que pour servir de passerelle entre le p.e.d. et le principe de YULE que nous énonçons plus loin. Signalons toutefois que, dans le cadre euclidien, en nous limitant, comme précédemment, aux formes quadratiques diagonales ne dépendant que des éléments marginaux, on obtient une classe de métriques assez simples. (cf Annexe) Si on note D_I^{-1} la matrice de la forme quadratique associée à la métrique du chi-deux sur l'ensemble des profils-lignes, les métriques de la forme $u D_I^{-1} + v D_I^{-2}$ sont les seules satisfaisant au p.e.d.-large.

On peut encore regretter que les deux principes dont il vient d'être question exigent, pour être appliqués, une structure particulière des données (existence de lignes ou colonnes proportionnelles). Nous modifions cette fois le p.e.d.-large pour écrire :

le *principe de YULE* : on ne change pas la distance entre deux lignes ni entre deux colonnes d'un tableau en remplaçant les lignes et les colonnes de ce tableau par d'autres lignes et colonnes qui leur sont proportionnelles.

On vérifie aisément qu'une "distance de YULE" satisfait au p.e.d.-large ; la réciproque est fausse.

Ainsi, l'ensemble des distances déduites du p.e.d.-large englobe-t-il d'une part les distances vérifiant le p.e.d., d'autre part celles vérifiant le principe de YULE. Le fait que les énoncés du p.e.d. et du p.e.d.-large sont très voisins nous permet d'espérer de bonnes propriétés pour les analyses déduites du principe de YULE.

2. L'ANALYSE LOGARITHMIQUE

Imaginons la situation suivante :

x_{io} est une "valeur de référence" mesurée pour l'individu i (ou encore à l'instant i) et, pour diverses situations notées $j \in J$, on a relevé pour les mêmes individus (ou aux mêmes instants) les valeurs $\{x_{ij} \mid j \in J\}$. Ces valeurs seront supposées strictement positives. Il est habituel de définir les ratios r_{ij} par la formule :

$$r_{ij} = x_{ij}/x_{io}$$

Notons $R = \{r_{ij} \mid i \in I, j \in J\}$ le tableau (n , p) des ratios. La question est alors : quelle analyse effectuer sur R ? On pourra objecter qu'il est préférable de travailler sur les données brutes plutôt que sur les ratios ([1]). A cela nous opposerons les deux arguments suivants :

– il peut arriver que les données brutes ne nous soient pas fournies ou encore que, par principe même, seuls les ratios soient accessibles à l'observation. C'est le cas, dans l'exemple 1 étudié plus loin, où R représente les cours de différentes monnaies (change) par rapport à une monnaie de référence.

– même si nous sommes en possession des données brutes (x_{ij}) ce sont peut être des rapports (tels que r_{ij}) qui intéressent l'analyste. C'est le cas de l'exemple 2 ci-dessous : les x_{ij} étant des mensurations effectuées sur le pied d'animaux d'une même famille mais d'espèces (et de tailles !) différentes, on s'intéresse non pas à des notions de grandeur absolue mais bien plus à des notions de forme (grandeurs relatives).

Le problème étant posé, l'analyse de R devra satisfaire aux conditions suivantes :

– un changement d'unité de mesure (e.g. exprimer la monnaie des USA en dollar ou en cent, celle de la RFA en mark ou en pfennig. . .) ne modifiera pas les résultats de l'analyse.

– un changement de valeur de référence ne modifiera pas les résultats de l'analyse.

(1) Si X est un tableau de comptage et $x_{io} = x_{i\cdot}$ alors les ratios r_{ij} ne sont rien d'autre que les profils-lignes. L'objection ne vaut pas. Pour une brève discussion de ce cas cf § 5. Conclusion.

Le premier des deux principes est clair. On le rencontre dans certaines analyses classiques (e.g. ACP sur des données centrées-réduites). Le second mérite que l'on s'y arrête plus longtemps.

Reprenons les deux exemples que nous avons très brièvement évoqués.

Dans le premier, on souhaitera que l'analyse fournisse les mêmes résultats qu'elle soit effectuée en France (où les cours seront exprimés par rapport au franc), en RFA (la référence sera le mark) ou ailleurs encore.

Dans le second le problème peut sembler encore plus délicat : que choisir pour valeur de référence ? Doit-on prendre la longueur de tel ou tel os, plus ou moins arbitrairement choisi, la taille ("sous la toise") de l'individu. . . ?

Nous voyons là tout l'intérêt de cette deuxième condition.

Si on note r^*_{ij} les ratios définis par rapport à une nouvelle référence x^*_{io} :

$$r^*_{ij} = x_{ij}/x^*_{io} = (x_{ij}/x_{io}) \cdot (x_{io}/x^*_{io})$$

$$r^*_{ij} = a_i\, r_{ij} \quad \text{où} \quad a_i = x_{io}/x^*_{io}$$

D'autre part, il est clair que le changement d'unité de mesure correspond quant à lui à une transformation de la forme :

$$r^*_{ij} = r_{ij}\, b_j$$

En résumé, en notant $A = \mathrm{diag}(a_1 \ldots a_i \ldots a_n)$ et $B = \mathrm{diag}(b_1 \ldots b_j \ldots b_p)$ deux systèmes de masses strictement positives, les analyses de

$$R \quad \text{et} \quad A\,R\,B$$

devront conduire aux mêmes résultats. Nous voyons là une illustration du principe de YULE.

Au passage nous noterons qu'en choisissant A et B convenablement on peut identifier $A\,R\,B = X$, de sorte que la discussion engagée plus haut sur les préférences à donner à l'analyse de X ou à celle de R devient caduque.

Il y a certes plus d'une façon de résoudre le problème que nous venons de poser. Toutefois, l'aspect multiplicatif nous conduit naturellement à poser :

$$z_{ij} = \mathrm{Log}\, r_{ij}$$

Ainsi, à la transformation $r^*_{ij} = a_i\, r_{ij}\, b_j$ correspondra :

$$z^*_{ij} = \mathrm{Log}\, a_i + z_{ij} + \mathrm{Log}\, b_j$$

En notant a le vecteur (n , 1) des $\mathrm{Log}(a_i)$, b le vecteur (p , 1) des $\mathrm{Log}(b_j)$ et 1_r le vecteur (r , 1) dont toutes les composantes valent 1 (r = n ou p), on écrira :

$$Z^* = a\, 1'_p + Z + 1_n\, b'$$

la notation $'$ est utilisée pour la transposition.

Si C_n et C_p désignent les opérateurs de centrage définis par :

$$C_r = \mathrm{Id}_r - \frac{1}{r}\, 1_r\, 1'_r \quad (r = n\,,\, p)$$

compte tenu de la propriété $C_r\, 1_r = 0$, il est clair que :

$$C_n\, Z\, C_p = C_n\, Z^*\, C_p$$

On obtient de la sorte un tableau que nous notons Z (où les deux points rappellent que ce tableau est doublement centré) indépendant des systèmes *a* et *b*. Nous soumettrons ce tableau à une ACP simple (métrique et système de masse : identité) : c'est ce que nous appelons analyse logarithmique de X (ou de R).

Signalons qu'une telle analyse a été proposée récemment – et pour des raisons bien différentes – par Aitchison.

3. QUELQUES PROPRIETES ELEMENTAIRES DE L'ANALYSE LOGARITHMIQUE

(i) Invariance multiplicative (principe de YULE) : compte tenu de ce qui précède, il est clair que les analyses de

$$X \qquad \text{et} \qquad AXB$$

conduiront aux mêmes résultats : même liste de valeurs propres, mêmes systèmes de facteurs.

(ii) Symétrie entre lignes et colonnes : analyser

$$X \qquad \text{et} \qquad X'$$

est équivalent. Ceci provient du fait que le tableau Z des logarithmes est traité de façon identique en ligne et en colonne ("bicentrage").

(iii) Valeur propre et facteur trivial : l'A.L. de X fournit toujours la valeur propre triviale $\lambda = 0$ associée aux facteurs constants 1_n et 1_p. Ceci provient des opérations de centrage.

L'ACP de $\ddot{Z}$ conduit à diagonaliser :

soit $$V_p = (C_n Z C_p)' (C_n Z C_p) = C_p Z' C_n Z C_p$$

soit $$V_n = (C_p Z' C_n)' (C_p Z' C_n) = C_n Z C_p Z' C_n$$

et par suite :

$$V_p 1_p = 0 \quad \text{et} \quad V_n 1_n = 0$$

(iv) Les facteurs sont centrés : cela provient de la propriété précédente et de l'orthogonalité du système de facteurs. En notant φ_α et ψ_α les facteurs unitaires (sur I et sur J), on a les relations :

$$1'_n \varphi_\alpha = 0 \qquad C_n \varphi_\alpha = \varphi_\alpha$$

$$1'_p \psi_\alpha = 0 \qquad C_p \psi_\alpha = \psi_\alpha$$

On en déduit une simplification dans les formules de transition :

$$\lambda_\alpha^{1/2} \varphi_\alpha = C_n Z C_p \psi_\alpha$$
$$= C_n Z \psi_\alpha$$

de même : $$\lambda_\alpha^{1/2} \psi_\alpha = C_p Z' \varphi_\alpha$$

(v) A.L. et A.F.C. : on peut énoncer la propriété :

Au voisinage de l'indépendance, pour une correspondance dont les lois marginales sont uniformes, AL et AFC fournissent des résultats sensiblement identiques.

Avec les notations habituelles, on a :

$$\forall i,\ x_{i.} = 1/n \quad \forall j, x_{.j} = 1/p$$

L'AFC de X fournit les valeurs propres λ_α (supposées petites) et les facteurs unitaires φ_α et ψ_α. La formule de reconstitution des données à partir des facteurs s'écrit :

$$x_{ij} = x_{i.}\, x_{.j} \left(1 + \sum_\alpha \lambda_\alpha^{1/2}\, \varphi_\alpha (i)\, \psi_\alpha (j)\right)$$

La propriété d'invariance multiplicative nous permet d'affirmer que l'AL de X est équivalente à l'AL du tableau de terme général :

$$1 + \sum_\alpha \lambda_\alpha^{1/2}\, \varphi_\alpha (i)\, \psi_\alpha (j)$$

En passant aux logarithmes et en se limitant aux termes du premier ordre, il reste simplement :

$$\sum_\alpha \lambda_\alpha^{1/2}\, \varphi_\alpha (i)\, \psi_\alpha (j)$$

Ces termes sont centrés : c'est donc ce tableau que l'on soumet à l'ACP ! Les φ_α et ψ_α étant orthonormés pour la métrique identité, on identifie facilement : ce sont les facteurs de l'ACP de Z donc les facteurs de l'AL, associés aux valeurs propres λ_α.

4. DEUX EXEMPLES D'APPLICATION

4.1. Evolution du cours de quelques monnaies entre 1965 et 1980

Les données que nous soumettons à l'analyse proviennent de la documentation générale de l'INSEE. Les neuf principales puissances économiques du monde occidental y figurent : Belgique, Canada, France, Grande-Bretagne, Italie, Pays-Bas, RFA, Suisse et USA. On regrettera toutefois l'absence du Japon.

Le tableau 16 x 9 : $X = (x_{ij})$ donne pour chaque année i (de 1965 à 1980) la valeur moyenne de la monnaie du pays j. L'AL de ce tableau montre une décroissance particulièrement rapide des valeurs propres. La première représente plus de 93 % de l'inertie totale, la seconde 3.4 %, la troisième 1.5 %. La figure 1 représente le diagramme plan (1,2). On y remarque essentiellement deux époques : 1965-1973 et 1975-1980, chacune marquée par un effet Guttman.

Cette division se reflète également au niveau des pays. La valeur de la lire italienne, de la livre anglaise mais aussi du dollar (US et canadien) était en hausse très marquée au cours de la première époque. Dans la seconde période, ces mêmes

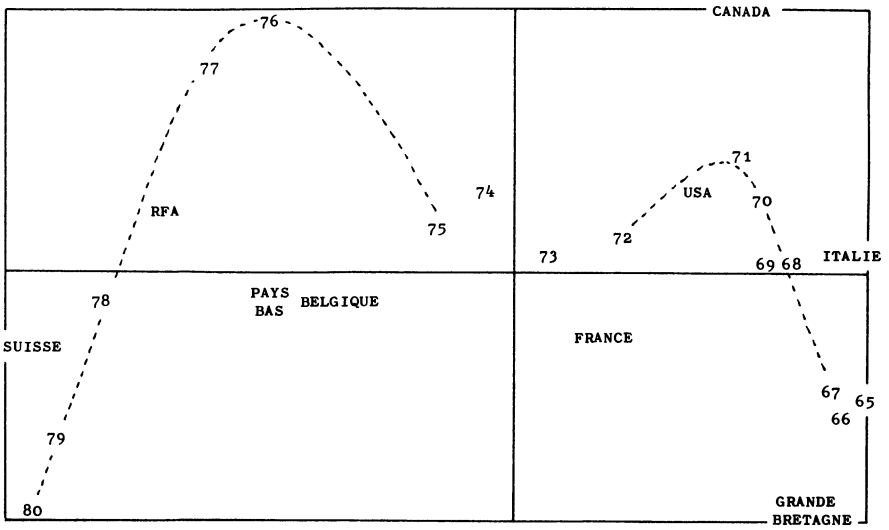

Figure 1. – Diagramme plan (1, 2) de l'analyse des cours monétaires.

monnaies vont subir une dépréciation alors que d'autres, le franc suisse et le mark allemand (que beaucoup considéraient comme sous évalué) connaîtront une croissance très forte.

Une telle dichotomie est encore mise en évidence sur la figure 2 où l'on a représenté la valeur du premier facteur en fonction du temps. On remarque que sur les périodes 1965-1971 et 1972-1979 on a une bonne approximation

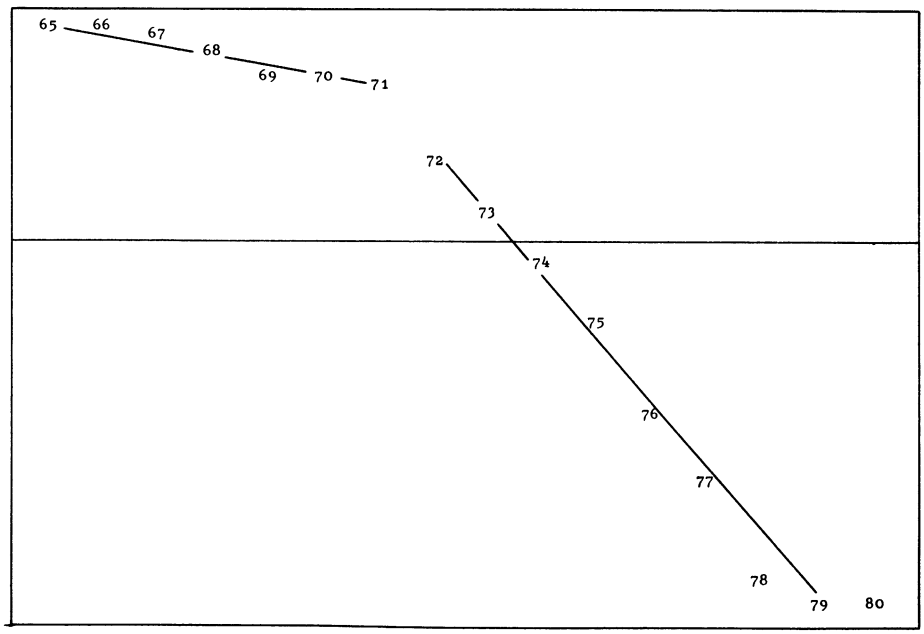

Figure 2. – Premier facteur de l'A.L. tracé en fonction du temps

du premier facteur en écrivant :

$$\varphi(t) = a\,t + b$$

où l'on a préféré, pour plus de clarté, remplacer le temps jusqu'ici indicé par i, par l'indice t.

Ainsi, en se limitant au premier facteur, on peut écrire la formule de reconstitution des données :

$$\ddot{z}_{tj} = \lambda\,\psi_j\,(a\,t + b)$$

or

$$\ddot{z}_{tj} = z_{tj} - \bar{z}_t - \bar{z}_j + \bar{\bar{z}}$$

où l'on a noté :

$$z_{tj} = \text{Log}\,x_{tj} \quad \text{encore noté Log}\,x_j\,(t)$$

$$\bar{z}_t = \Big(\sum_j z_{tj}\Big)/p \qquad \bar{z}_j = \Big(\sum_t z_{tj}\Big)/n$$

$$\bar{\bar{z}} = \Big(\sum_t \bar{z}_t\Big)/n = \Big(\sum_j \bar{z}_j\Big)/p$$

Si l'on remarque que le terme $\bar{z}_t$ reste sensiblement constant au cours du temps (variation inférieure à 1 %) on pourra admettre, pour chacune des deux époques, un modèle de la forme :

$$\text{Log}\,x_j\,(t) = \lambda\,\psi_j\,(a\,t + b) + \bar{z}_j + C$$
$$= k_j\,t + C_j$$

soit finalement :

$$x_j\,(t) = x_{oj}\,e^{k_j t}$$

Nous nous contenterons ici de signaler que la charnière entre les deux époques est constituée par les années 1971-1972 :
c'est la fin de l'ordre monétaire défini à Bretton Woods : le 15 août 1971, NIXON déclare l'inconvertibilité du dollar en or.

Enfin, l'effet de la mise en place, en juillet 1978, du système monétaire européen permettant de limiter les fluctuations des monnaies des pays de la Communauté entre elles n'est pas observable sur les diagrammes : elle est, dans notre chronologie, trop tardive. Toutefois, il n'est peut-être pas abusif de penser que l'inflexion que l'on peut deviner pour les années 78-79-80 sur la figure 2 s'explique par cette mesure.

4.2. La forme du pied chez quelques primates

Nous présentons ici quelques résultats d'une étude effectuée en collaboration avec F.K. JOUFFROY (CNRS).

Trente-huit lémuriens, primates africains nocturnes, caractérisés par une excellente aptitude au saut, appartenant à cinq espèces voisines (*Euoticus elegantulus, Galago crassicaudatus, G. alleni, G. senegalensis, G. demidovii*) sont décrits par seize variables. Qu'il nous suffise de dire ici que ces variables corres-

pondent aux mensurations du tibia, du fémur et de quelques os du pied (naviculaire, cuboïde, astragale, calcaneum).

On s'intéresse essentiellement à une description de la forme du pied et à une caractérisation de l'espèce à partir de la forme.

Il existe ici une approche classique. On peut soumettre le tableau des mensurations soit à l'ACP (normée, i.e. sur matrice de corrélation) soit à l'analyse factorielle discriminante (à but descriptif) si c'est l'aspect "caractérisation de l'espèce" que l'on privilégie. Dans les deux cas le premier axe obtenu est un facteur de taille et le plan (1, 2) permet de discriminer parfaitement les espèces. Aucun autre *plan* ne permet de le faire : la forme ne semble pas être liée de façon *simple* à l'espèce.

Pour les raisons évoquées plus haut, l'AL est bien adaptée à l'étude du problème. La pratique le confirme : dans le plan (1, 2) il nous est possible de discriminer les différentes espèces et cette fois indépendamment de la taille (invariance multiplicative ; cf § 2).

L'axe 1 (cf. Fig. 3) caractérise essentiellement la forme du naviculaire (épais et court pour Euoticus : Longueur/épaisseur = 4.5, mince et long pour *G. senegalensis* L/e = 12 et *G. demidovii* L/e = 11 ; les valeurs intermédiaires sont observées pour *G. crassicaudatus* et *G. alleni* L/e = 6). On notera qu'avec les naviculaires "trapus" on observe les cuboïdes importants (longs et larges) alors qu'avec les naviculaires "allongés" on trouve les plus grandes longueurs du calcaneum (et plus particulièrement la partie postérieure) ainsi qu'une zone articulaire calcaneo-naviculaire importante.

Cette zone caractérise l'axe 2 en s'opposant aux mensurations du naviculaire. Elle permet de séparer *G. crassicaudatus* de *G. alleni.* Cette dernière espèce (et c'est encore plus marqué pour *G. Demidovii*) possède une zone articulaire calcaneo-naviculaire faible relativement aux dimensions du calcaneum ou du naviculaire.

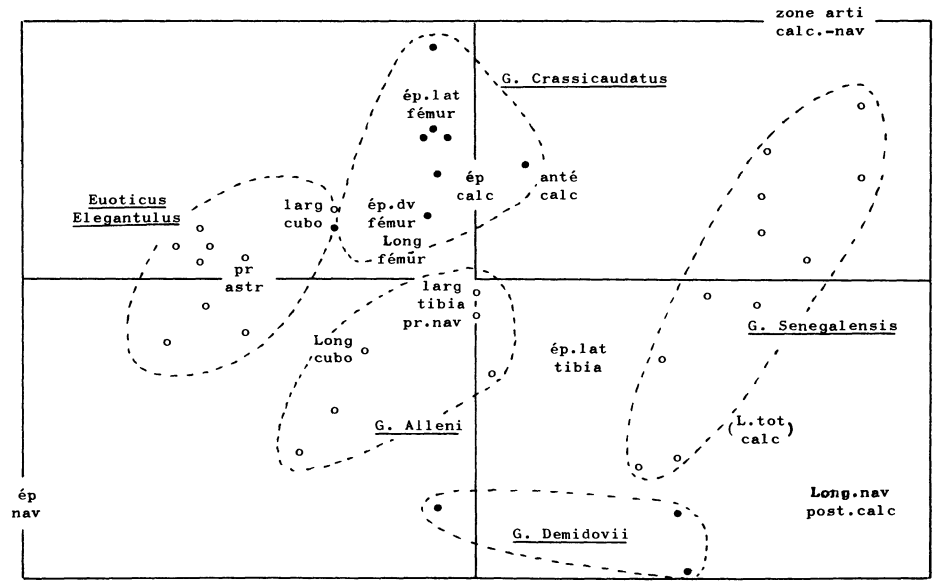

Figure 3. – Plan (1, 2) de l'A.L. On a indiqué les principales variables et symbolisé par des "ronds" les divers individus. (L. tot. calc. est un élément supplémentaire).

5. CONCLUSION

Nous venons de voir, brièvement sur ces deux exemples, que l'A.L. se justifiait pour l'analyse de certains tableaux de mesures. Ce faisant nous avons abandonné le support qui nous a permis de l'introduire plus naturellement par le principe de YULE : les tables de contingences.

Dans ce cas, l'A.F.C. reste, dans l'immense majorité des cas, l'outil privilégié et cela d'autant plus que se pose alors le problème épineux des valeurs propres nulles.

Nous réservons à plus tard la présentation de quelques études dans ce domaine mais signalons dès à présent un cadre intéressant où l'A.L. semble bien adaptée : celui de certaines enquêtes (ou autres expériences) où l'une des marges est fixée *a priori*. Ainsi, au lieu d'obtenir de véritables tables de contingence $X = (x_{ij})$ on pourra n'obtenir que des tableaux de la forme $(x_{io} \cdot x_{ij}/x_{i.})$ ou $(x_{oj} \cdot x_{ij}/x_{.j})$ sans que les masses $(x_{i.})$ ou $(x_{.j})$ soient connues, les marges (x_{io}) et (x_{oj}) étant fixées *a priori*.

Notre analyse possède alors l'avantage de fournir un résultat indépendant de cet *a priori*. En effet, l'A.L. des trois tableaux :

$$(x_{ij}) \quad (x_{io} \cdot x_{ij}/x_{i.}) \quad (x_{oj} \cdot x_{ij}/x_{.j})$$

fournira les mêmes résultats.

Reste le problème des valeurs nulles. Si celles-ci sont structurelles il est hors de question d'utiliser l'A.L. Sinon, une parade élégante nous semble être celle de la reconstitution de données manquantes : une fréquence nulle pour cause d'effectif trop petit est alors remplacée par une fréquence "estimée" strictement positive (la reconstitution pouvant s'effectuer sur le tableau des logarithmes).

BIBLIOGRAPHIE

J. AITCHISON. – Principal component analysis of compositional data. *Biometrika,* 1983, 70, 1, 57-65.

J.P. BENZECRI. – *L'analyse des données* (2 tomes) Paris, Dunod, 1973.

J.P. BENZECRI. – *Histoire et préhistoire de l'analyse des données.* Paris, Bordas-Dunod, 1982.

B. ESCOFIER. – Analyse factorielle et distances répondant au principe d'équivalence distributionnelle. *Rev. Stat. Appli.,* 1978, XXVI, 4, 29-37.

M.G. KENDALL, A. STUART. – *The advanced theory of statistics.* t.2. Inference and relationship. London, Griffin, 1973.

G.U. YULE. – On the methods of measuring association between two attributes. *J.R. Statist. Soc.,* 1912, 75, 579-642.

ANNEXE

Une classe de métriques euclidiennes satisfaisant au principe d'équivalence distributionnelle large

On suppose que le tableau $X = (x_{ij})$ possède deux lignes proportionnelles, que l'on supposera, sans nuire à la généralité, être les deux premières ($i = 1$ et $i = 2$). On pourra noter :

$$\forall j,\ x_{1j} = a_1\, x_{oj}\ ,\ x_{2j} = a_2\, x_{oj}$$

où $a_i = \sum_j x_{ij}$ de sorte que $\sum_j x_{oj} = 1$

On construit alors un tableau $X^* = (x^*_{ij})$ ne différant de X que par les deux premières lignes :

$$\forall j\ ,\ x^*_{1j} = a^*_1\, x_{oj}\ ,\ x^*_{2j} = a^*_2\, x_{oj}$$

avec $$a_1 + a_2 = a^*_1 + a^*_2$$

En nous limitant aux formes quadratiques diagonales, le carré de la distance entre deux profils colonnes de X s'écrit :

$$d^2(j\ ,j') = \sum_i f_i \left(\frac{x_{ij}}{x_{.j}} - \frac{x_{ij'}}{x_{.j'}}\right)^2$$

On décompose cette somme en trois termes : les deux premiers correspondent aux deux premières lignes ($i = 1$, $i = 2$), le troisième est une somme, notée S, étendue à toutes les autres lignes.

$$d^2(j\ ,j') = f_1\left(\frac{x_{1j}}{x_{.j}} - \frac{x_{1j'}}{x_{.j'}}\right)^2 + f_2\left(\frac{x_{2j}}{x_{.j}} - \frac{x_{2j'}}{x_{.j'}}\right)^2 + S$$

que l'on réécrit :

$$d^2(j\ ,j') = (a_1^2\, f_1 + a_2^2\, f_2)\,.\left(\frac{x_{oj}}{x_{.j}} - \frac{x_{oj'}}{x_{.j'}}\right)^2 + S$$

Le même calcul, effectué cette fois sur le tableau X^* conduit à la formule :

$$d^2(j\ ,j') = (a_1^{*2}\, f_1 + a_2^{*2}\, f_2)\,.\left(\frac{x_{oj}}{x_{.j}} - \frac{x_{oj'}}{x_{.j'}}\right)^2 + S$$

En effet, la loi marginale $(x_{.j})$ reste inchangée compte tenu de la propriété : $a_1 + a_2 = a^*_1 + a^*_2$.

Le p.e.d.-large nous conduit ainsi à écrire l'identité :

$$a_1^2\, f_1 + a_2^2\, f_2 = a_1^{*2}\, f_1 + a_2^{*2}\, f_2 \qquad (1)$$

Nous supposerons maintenant que les f_i ne dépendent que des masses :

$$f_1 = f(a_1)\,,\ f_2 = f(a_2)\,,\ \ldots$$

Il est alors naturel de poser : $g(a) = a^2\ f(a)$ pour obtenir :

$$g(a_1) + g(a_2) = g(a_1^*) + g(a_2^*)$$

$$a_1 + a_2 = a_1^* + a_2^*$$

et reconnaître une propriété caractéristique des fonctions affines. On a :

$$g(a) = u\ .\ a + v$$

d'où le résultat annoncé :

$$f(a) = u\ .\ \frac{1}{a} + v\ .\ \frac{1}{a^2}$$

Remarque : Pour satisfaire au p.e.d. (strict), il est nécessaire que l'identité (1) reste vérifiée lorsque les deux termes du membre de droite sont remplacés par un seul :

$$a_1^2\ f_1 + a_2^2\ f_2 = (a_1 + a_2)^2\ f^*$$

ce qui revient à écrire : $g(0) = 0$; $g(a)$ est linéaire et $f(a)$ proportionnelle à $(1/a)$. On obtient la métrique du chi-deux.