MATHÉMATIQUES ET SCIENCES HUMAINES

H. LE BRAS

Vingt analyses multivariées d'une structure connue

Mathématiques et sciences humaines, tome 47 (1974), p. 37-55

<http://www.numdam.org/item?id=MSH_1974__47__37_0>

Math. Sci. hum. (12e année, n° 47, 1974, p. 37-55)

VINGT ANALYSES MULTIVARIÉES D'UNE STRUCTURE CONNUE

par
H. LE BRAS [1]

RÉSUMÉ

L'usage croissant des analyses multivariées modifie la méthode de plusieurs sciences sociales. Le débat sur leur validité est toutefois assez confus car des questions de pure mathématique se mélangent à des problèmes perceptifs (représentation) et logiques (modélisation).

Nous n'avons actuellement ni les moyens ni l'ambition d'aborder théoriquement leur statut, mais nous pensons fournir un document utile au débat en appliquant la majorité des méthodes à un même objet déjà connu et en comparant leur résultat à la réalité de cet objet.

SUMMARY

Multivariate analyses being used more and more frequently, their applications in different social sciences is changing. However, the argumentation on their validity is rather confuse since questions of pure mathematics are mixed to problems of perception (representation) and logics (modelization).

Presently, we have neither the means nor the ambition to approach their status from the theoretical viewpoint, but we hope to present a useful document to the debate in applying the majority of these methods to one same and already known object by comparing their results to the reality of this object.

1. Maître de conférences à l'Ecole Polytechnique.

Les analyses multivariées se répandent toujours plus dans les sciences sociales et modifient leur épistémologie[1].

Comme ce phénomène est diffus et ne s'appuie pas sur une démarche théorique, il a ses partisans, mais aussi ses critiques.

Les partisans soulignent la généralité des représentations obtenues, leur facilité à illustrer les ensembles de données dont on connaît la structure. Ils pensent aussi que les méthodes multivariées servent à dialoguer entre des données quantitatives et des hypothèses qualitatives encore incertaines, parfois contradictoires jusqu'à assurer leur cohérence d'ensemble.

Mais quelques excès suscitent les critiques :

— ces modèles ne peuvent faire apparaître que certaines structures ; il se peut que la structure cherchée ne soit pas de leur catégorie ;
— le délicat processus de formalisation, particulier à chaque discipline, est escamoté : on constate d'ailleurs que ces analyses multivariées pénètrent plus facilement dans des sciences qui se montrèrent un peu rétives à la formalisation ;
— ces méthodes semblent souvent être le biais pour introduire des mathématiques ; leur principe est d'ailleurs presque toujours exposé avec un formalisme pompeux qui décourage vite l'honnête homme.

Le débat entre partisans et adversaires achoppe sur la preuve de l'efficacité des méthodes : dans l'immense majorité des études une ou deux analyses sont produites. Leur réussite est jugée sur les données dont on connaissait mal la structure. On ne peut donc ni faire la preuve qu'une autre analyse eût été plus pertinente, ni que la structure « découverte » est la structure réelle. Parfois l'argument « on retrouve ce que l'on savait déjà » est utilisé. Il est alors curieux que l'on ait négligé de formaliser ce savoir *a priori*.

Pour dépasser ces arguments, il nous a paru utile d'effectuer un certain nombre d'analyses *différentes* sur un même ensemble de données dont la structure est *connue* puisque nous l'avions construite.

En d'autres termes, à partir d'une structure réelle, nous collecterons des données, puis nous oublierons cette structure et tenterons de la retrouver par la seule analyse des données. Entre la réalité (la structure) et l'observation (les données), un écran plus au moins opaque existera : il figurera ce que nous pouvons savoir de la structure en dehors des données. En faisant varier l'opacité de l'écran, c'est-à-dire en introduisant ou en retirant certaines informations sur la réalité, nous serons conduits à faire divers types d'analyse (toutes les méthodes de représentation euclidienne, plusieurs méthodes hiérarchiques et une analyse en facteurs communs et spécifiques) et nous jugerons de quelle manière elles ont pu reconstituer la réalité. Les analyses que nous utiliserons sont toutes connues et fréquemment employées. Notre critique ne portera aucunement sur les méthodes elles-mêmes qui sont presque toujours techniquement irréprochables, mais sur leur usage. Il paraîtra que nombre de ces critiques étaient connues, mais elles sont délibérément ignorées dans la pratique. Pour mieux conserver l'analogie entre notre étude et des travaux concrets, pour simuler leur démarche, nous partirons d'un phénomène réel. Il va cependant de soi que nous ne nous intéresserons pas aux propriétés profondes de ce phénomène, mais à l'adéquation entre les résultats des analyses et cette structure.

1. La réalité et l'observation

Un problème de neuropsychologie sera l'anecdote à partir de laquelle nous allons introduire une structure : dans l'étude de l'hémisphère droit du cerveau, on suppose que certaines zones correspondent à certaines performances. Lorsque des zones sont détruites (tumeur, hémorragie, traumatisme), les performances leur correspondant doivent alors disparaître. Pour savoir si une performance a disparu, on utilise un test psychologique et désormais nous omettrons la performance et admettrons que le lobe frontal est un ensemble de petites cases telles qu'à chacune corresponde un test et un seul. Dans l'observation, certains malades passeront les tests, mais on ne sait pas quelles cases sont atteintes, on saura seulement à quels tests ils ont échoué et à quels tests ils ont réussi. A partir des résultats des tests d'un certain nombre de malades, il faudra donc reconstituer la géométrie des petites cases du lobe frontal ; à partir de l'observation, donc, retrouver la réalité.

Pratiquement, nous supposerons que la réalité a une géométrie très forte : l'hémisphère droit sera constitué de 25 petites cases égales disposées en carré. Il leur correspondra 25 tests dans l'observation, et le problème deviendra un problème de codage : assigner les 25 tests à chacune des 25 cases. Les malades seront

1. Nous remercions le Laboratoire de Pathologie du Langage EPHE (ERA au CNRS) et la Grant Foundation qui nous ont permis de réaliser les calculs ainsi que l'Institut de l'Environnement qui a publié une première version de ce travail et qui a eu l'obligeance de nous fournir les figures, « Comparaison d'analyses multivariates d'une structure connue », *Notes méthodologiques en architecture*, n° 1, janvier 1973, pp. 29 à 53.

simulés : nous aurons une lésion pour chacun, c'est-à-dire un certain nombre de cases détruites, donc d'échecs aux tests correspondants. Si les cases détruites étaient choisies au hasard, nous n'aurions aucun lien pour retrouver, à partir de l'observation, la réalité. Il faut donc donner une certaine structure à l'ensemble des cases atteintes : nous exigerons, ce qui est tout à fait légitime, pour les lésions tumorales et les traumatismes, que l'ensemble des cases atteintes soit connexe.

Concrètement, pour simuler l'hémisphère droit d'un malade, nous tirerons au hasard une des 25 cases du carré, ce sera l'origine de la lésion, puis nous tirerons au hasard l'une des cases adjacentes à cette première case ; de proche en proche, on tirera au hasard une case parmi les cases qui sont adjacentes à l'ensemble des cases déjà tirées. Ce modèle est un modèle correct de croissance cellulaire donc de croissance tumorale. Nous arrêterons les tirages de deux manières : soit en se fixant une taille, toujours la même, pour la lésion (15 cases dans le modèle des pages suivantes), soit en se fixant le nombre total de cases menacées à un moment ou à un autre (modèle des dernières pages). Ici aussi ces deux types d'arrêt sont conformes à des croissances tumorales. Nous obtiendrons alors une série de malades dont les lésions (ce que nous avons appelé la réalité) se présentent comme sur la Figure 1 : en noir les cases atteintes, en blanc les autres. Dans l'observation, nous n'avons que les résultats de chaque malade aux 25 tests : nous avons codé 1 l'échec et 0 la réussite, la Figure 2 représente donc les données telles qu'elles seraient obtenues par collecte.

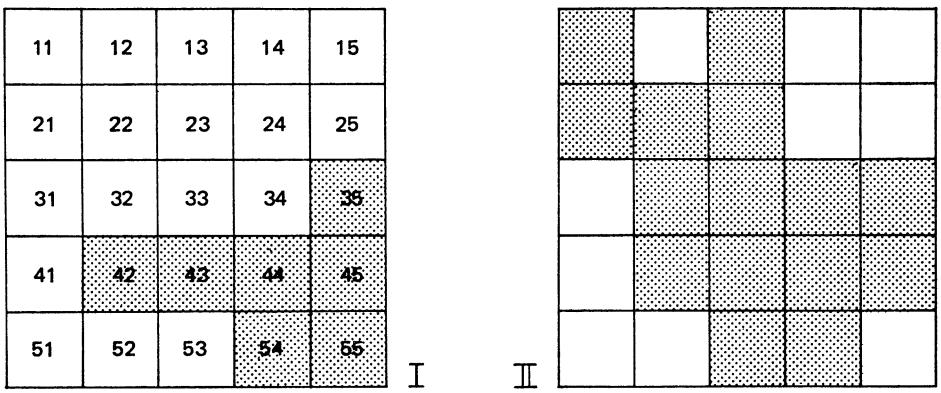


Figure 1. — *Hémisphère droit de deux malades simulés (on a porté, sur le premier, le code des tests correspondant aux cases)*

Entre cette réalité et ces données, qui doivent permettre de la reconstituer, se situe l'écran, qui figure ce que nous savons de la structure en plus des données : cette structure est bi-dimensionnelle, ou bien les lésions sont connexes, ou bien absolument rien (écran noir). Au fur et à mesure de l'analyse, nous augmenterons l'opacité de l'écran : nous restreindrons ce que nous connaissons en plus des données.

Nous n'examinerons que le modèle décrit ici dans ses deux variantes sur la « taille » des lésions, mais nous voyons que de nombreuses extensions sont possibles : ici la bijection entre cases et tests peut être remplacée par une application univoque ou même multivoque, le carré peut être remplacé par une forme géométrique quelconque à trois dimensions [1]. Nous noterons que les résultats se présentent sous forme de variables binaires, excluant ainsi certaines méthodes. Mais ces variables sont intéressantes car elles sont au confluent de nombreuses méthodes et par conséquent revendiquées par les tenants de ces méthodes séparément.

2. Résultat des analyses multivariates dans le cas d'un grand nombre de malades

Nous avons d'abord simulé 1 000 malades, pour réduire les causes de variations imputables au faible nombre d'observations et discuter ainsi de la validité des analyses avec une bonne sécurité. Dans le chapitre suivant, nous reprendrons ces analyses avec un faible nombre de malades (100), ce qui est fréquemment la taille des échantillons étudiés.

1. Bien que le modèle ait des fondements sérieux, il ne s'agit pas ici de l'étudier en tant que tel (nous aurions prêté plus d'attention à sa construction), mais d'utiliser sa simplicité et sa géométrie pour « tester » les analyses de données.

INDIVIDU	VARIABLES																								
NO 1	1	1	1	1	1	1	1	0	1	0	1	1	0	0	0	1	1	1	1	0	1	0	0	0	0
NO 2	0	0	0	0	1	0	0	0	1	1	0	1	1	1	1	0	0	1	1	1	1	1	1	1	1
NO 3	0	0	0	0	0	0	1	1	1	0	1	1	1	1	1	0	0	1	1	0	1	1	1	1	1
NO 4	0	1	1	1	0	0	1	1	1	1	0	1	1	1	1	0	0	1	1	0	0	0	1	1	1
NO 5	0	1	1	1	1	0	0	1	1	1	0	0	1	1	1	0	1	1	1	1	0	1	0	0	0
NO 6	0	0	1	0	0	0	1	1	0	0	1	1	1	0	0	0	1	1	1	1	1	1	1	1	1
NO 7	0	1	1	1	0	0	0	0	1	1	0	1	0	1	1	0	1	1	1	1	0	0	1	1	1
NO 8	1	1	1	1	1	0	1	1	1	1	0	1	0	1	1	0	0	1	1	1	0	0	0	0	0
NO 9	0	0	0	1	1	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0
NO 10	1	1	0	0	0	1	1	1	1	0	1	1	0	0	0	1	1	1	1	1	1	1	0	0	0
NO 11	0	0	0	0	0	1	1	0	1	1	1	1	1	1	1	1	1	0	1	1	1	0	0	0	1
NO 12	1	1	1	1	0	0	1	1	1	0	1	1	0	0	0	1	1	0	0	1	1	1	1	1	0
NO 13	0	0	1	0	0	0	0	1	0	1	1	1	1	1	1	0	1	1	1	1	0	0	1	1	1
NO 14	1	1	1	1	0	1	1	1	1	0	1	1	1	1	1	0	0	1	0	0	0	0	1	0	0
NO 15	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	0	0	1	1	0	0	0	0	0	0
NO 16	0	1	0	1	1	0	1	1	1	1	0	0	1	1	1	0	0	1	1	0	0	1	1	1	0
NO 17	0	1	1	1	1	1	1	1	1	1	0	0	1	1	1	0	0	0	1	1	0	0	0	0	1
NO 18	1	1	1	1	0	1	1	1	1	0	1	1	0	0	0	0	1	0	0	0	1	1	1	1	0
NO 19	1	0	0	1	1	1	1	0	1	1	0	1	1	1	1	0	0	1	0	1	0	0	0	1	1
NO 20	0	0	0	1	1	0	0	0	1	1	0	0	1	1	1	0	1	1	1	1	0	1	1	1	1
NO 21	0	0	1	1	1	0	0	0	1	0	0	0	1	1	1	0	0	1	1	1	1	1	1	1	1
NO 22	0	0	0	0	0	1	1	0	1	1	1	1	0	1	1	0	1	1	1	1	1	1	0	1	0
NO 23	0	1	1	1	1	0	1	1	1	1	1	1	1	0	0	0	1	1	1	0	0	1	0	0	0
NO 24	0	0	1	0	0	0	0	1	1	0	0	1	1	1	1	1	1	1	1	1	0	1	0	1	1
NO 25	0	0	0	0	0	1	0	0	0	0	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1
NO 26	1	0	0	1	0	1	1	1	1	0	1	1	1	1	0	1	1	1	1	0	0	0	1	0	0
NO 27	0	0	0	0	0	0	0	1	0	0	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1
NO 28	1	1	1	1	1	0	0	1	1	1	0	1	1	1	1	0	0	0	1	1	0	0	0	1	1
NO 29	0	0	0	1	0	0	1	1	1	0	0	0	1	1	1	0	1	1	1	1	1	1	1	1	0
NO 30	0	1	1	1	0	0	0	1	0	0	1	1	1	1	0	0	1	1	1	1	0	0	1	1	1
NO 31	1	1	1	1	1	0	0	1	1	1	1	0	1	1	1	0	1	1	0	1	0	0	0	0	0
NO 32	0	1	1	0	0	1	1	1	1	1	0	1	1	1	1	0	0	1	1	1	0	0	1	0	0
NO 33	1	1	0	0	0	1	1	1	1	0	1	1	1	1	0	1	1	1	0	0	1	1	0	0	0
NO 34	1	1	1	1	1	1	1	1	1	0	1	1	0	1	1	1	0	0	0	0	1	0	0	0	0
NO 35	0	0	0	1	1	0	0	1	1	1	1	1	1	0	1	1	0	1	0	1	1	1	0	1	1
NO 36	0	1	1	1	1	0	1	1	1	1	0	0	1	1	1	0	0	0	1	1	0	0	0	1	1
NO 37	0	1	0	0	0	0	1	1	1	0	1	1	1	1	0	0	1	0	1	0	1	1	1	1	1
NO 38	0	0	1	1	0	0	0	1	1	1	0	1	1	1	0	0	0	1	1	1	0	1	1	1	1
NO 39	0	1	1	0	0	1	1	1	1	1	1	1	1	1	1	1	1	0	1	0	0	0	0	0	0
NO 40	1	1	1	0	0	1	1	1	0	0	1	1	1	1	1	0	1	1	0	1	0	0	0	0	1
NO 41	0	0	0	0	0	0	1	0	0	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	0
NO 42	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	1	0	0	0	0	0	0
NO 43	0	0	1	1	0	1	1	1	1	1	0	1	1	1	1	1	1	1	0	0	0	1	0	0	0
NO 44	0	0	1	1	1	0	0	0	1	1	0	0	1	1	1	0	1	1	1	1	0	1	1	1	0
NO 45	0	0	1	1	1	0	1	1	1	1	0	0	0	1	1	0	0	1	1	1	0	0	1	1	1
NO 46	1	0	1	0	0	1	1	1	1	1	1	0	1	1	1	1	0	0	1	0	0	0	0	1	1
NO 47	1	1	0	0	0	1	1	1	0	0	1	1	1	1	0	1	1	1	1	0	0	1	1	0	0
NO 48	1	0	0	0	0	1	0	0	1	0	1	1	1	1	0	1	1	1	1	1	0	1	1	1	0
NO 49	1	1	1	1	1	1	1	1	1	0	1	1	1	1	0	1	1	0	0	0	0	0	0	0	0
NO 50	0	0	0	0	1	0	0	0	1	1	0	1	1	1	1	0	1	1	1	1	0	1	1	1	1
NO 51	0	0	1	1	1	0	0	1	1	1	0	1	1	1	1	0	0	1	1	0	0	1	1	1	1
NO 52	0	1	0	0	0	0	1	0	1	0	1	1	0	1	0	1	1	1	1	1	0	1	1	1	1
NO 53	1	1	1	0	0	1	1	1	1	1	0	1	1	1	1	0	0	1	1	0	0	0	1	0	0
NO 54	1	1	1	0	0	1	1	1	1	1	1	1	1	1	1	0	1	0	0	0	0	1	1	0	0
NO 55	0	0	0	0	0	0	0	1	0	1	1	0	1	1	1	1	1	1	1	1	0	1	1	1	1
NO 56	0	0	0	1	1	0	0	1	1	1	1	0	1	1	1	1	0	0	1	1	1	0	0	1	1
NO 57	1	1	1	1	0	1	1	1	1	1	1	1	1	1	0	0	1	1	1	0	0	0	0	0	0
NO 58	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	0	1	1	0	0	0	0	0	0	0
NO 59	0	1	1	1	0	0	1	1	1	1	0	0	1	1	1	0	0	1	1	0	0	0	1	1	1
NO 60	1	1	1	1	0	0	1	1	1	1	0	1	1	1	1	0	1	0	1	1	0	0	0	0	0
NO 61	1	1	1	1	0	1	1	1	1	1	1	1	0	1	1	0	0	0	1	1	0	0	0	0	0
NO 62	1	1	1	1	0	0	1	1	1	1	0	1	1	1	0	0	1	1	1	0	0	1	1	0	0
NO 63	1	1	1	1	1	1	1	1	1	1	0	1	1	0	1	0	0	0	1	1	0	0	0	0	0
NO 64	1	1	0	1	0	1	1	1	1	1	0	1	0	1	1	0	0	0	0	1	1	0	0	1	1
NO 65	1	1	1	1	1	1	0	1	1	0	1	1	0	1	0	0	1	1	1	0	0	1	0	0	0
NO 66	0	0	1	0	0	0	1	1	0	1	0	1	1	0	1	1	1	1	1	1	0	1	1	1	0
NO 67	0	1	1	1	1	0	1	1	1	0	1	0	1	1	0	0	0	0	1	1	0	0	1	1	1
NO 68	1	1	1	1	1	1	0	1	1	0	1	0	1	1	1	0	0	1	1	0	0	0	0	1	0
NO 69	1	1	1	0	0	1	1	1	1	0	0	0	1	1	0	0	0	0	1	1	0	1	1	1	1
NO 70	1	1	1	0	0	1	1	1	0	0	1	1	1	1	0	1	1	1	0	0	1	1	0	0	0

Figure 2. — *Tableau des résultats simulés (réussite = 0, échec = 1) pour 70 malades (chaque test ou colonne correspond à une case de la structure réelle)*

2.1. *Analyse des correspondances*

Supposons que l'écran soit peu opaque ; nous savons que l'échec signifie que la case est atteinte par la lésion, nous savons que cette lésion est connexe et que le lobe frontal est assimilable à un plan. Une analyse des correspondances où nous coderons 1 l'échec et 0 la réussite est alors légitime puisque la distance euclidienne qu'elle représente dans les espaces de plus faible dimension que 25 tient bien compte du phénomène : elle revalorise les cases frontières qui sont moins fréquemment impliquées dans les lésions, mais elle en tient peu compte dans la recherche successive des axes, puisque ces cases ont des poids faibles [1].

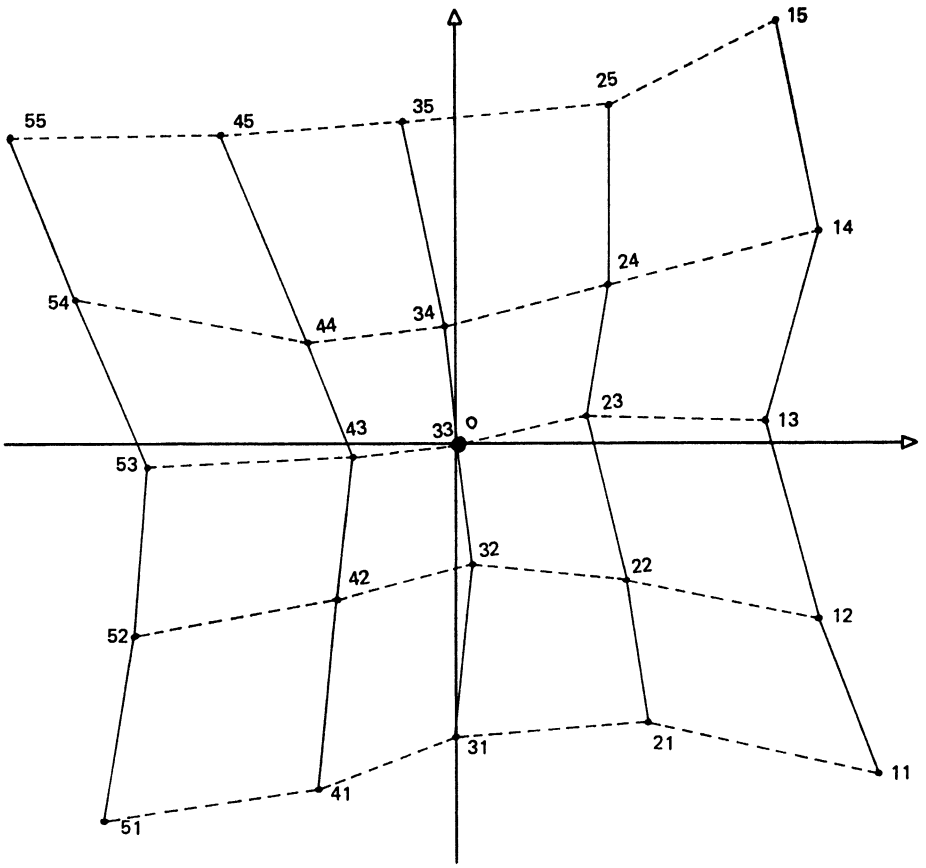


Figure 3. — *Premier plan d'analyse factorielle des correspondances*

Le résultat est conforme aux prévisions (Fig. 3) et sur ce résultat nous avons recopié la réalité en joignant entre eux les tests qui correspondent à des cases adjacentes (le nom des tests est à deux indices, ainsi 32 signifie test correspondant à la case de la troisième ligne, deuxième colonne). La structure n'est pas complètement régulière car les fluctuations du hasard jouent encore (heureusement car de nombreuses valeurs propres seraient multiples).

---

1. Posons Pij = résultat au $i^{ème}$ test du $j^{ème}$ individu et $Pi. = \sum_j Pij$.

Expliquons simplement cette propriété = l'analyse des correspondances peut être envisagée comme l'analyse en composantes principales des 25 objets Xi dont chacune des 1 000 coordonnées est $\frac{Pij}{Pi.}$ (parce que ici tous les individus ont 15 cases atteintes), ces objets étant munis des poids Pj.

Ainsi, lorsque $Pi.$ est faible (test peu fréquemment à la valeur 1) le point correspondant est éloigné du centre de gravité, mais comme son poids est faible, cet éloignement est compensé dans le calcul de l'inertie (qui a la dimension : poids × carré des distances).

Autrement dit, l'éloignement « visible » est plus important que son effet sur la recherche des axes, car on lit toujours un nuage de points en attribuant visiblement le même poids à chacun des points.

Pour simplifier les descriptions ultérieures nous dirons que l'aspect planaire est conservé si les arêtes qui joignent les cases adjacentes, dans la réalité, ne se recoupent pas dans la représentation. Nous parlerons d'aspect géométrique pour indiquer que les petits carrés reconstitués sont assez réguliers et assez carrés.

Ici l'aspect planaire est parfaitement conservé et, fait paradoxal, l'aspect géométrique aussi, à un étirement près des coins, étirement expliqué par leur faible poids. Les axes au-delà du second n'apportent plus aucune information intéressante sur la « réalité » bien qu'ils représentent 60 % de l'inertie totale.

2.2. *L'analyse en facteurs communs et spécifiques* [1]

Gardons à l'écran sa faible opacité : puisque l'on suppose que le phénomène est à deux dimensions, autant le postuler et adopter l'analyse en facteurs communs et spécifiques. Bien que les variables ne soient pas multinormales, l'estimation par le maximum de vraisemblance est recommandée [2].

Le résultat est porté sur la Figure 4 ; la réalité (cases adjacentes) est figurée comme dans l'analyse des correspondances. Le résultat est tout aussi bon du point de vue planaire et presque meilleur du point de vue géométrique à l'arrondissement près des angles du carré. Si au lieu des deux dimensions, nous en avions supposé une seule, le résultat serait celui de la Figure 5 : c'est une diagonale du carré qui aurait été saisie. Ce résultat est meilleur que le premier axe de l'analyse des correspondances, mais il n'entre pas en ligne de compte puisque dans les deux cas nous avons retenu l'hypothèse d'un phénomène plan.

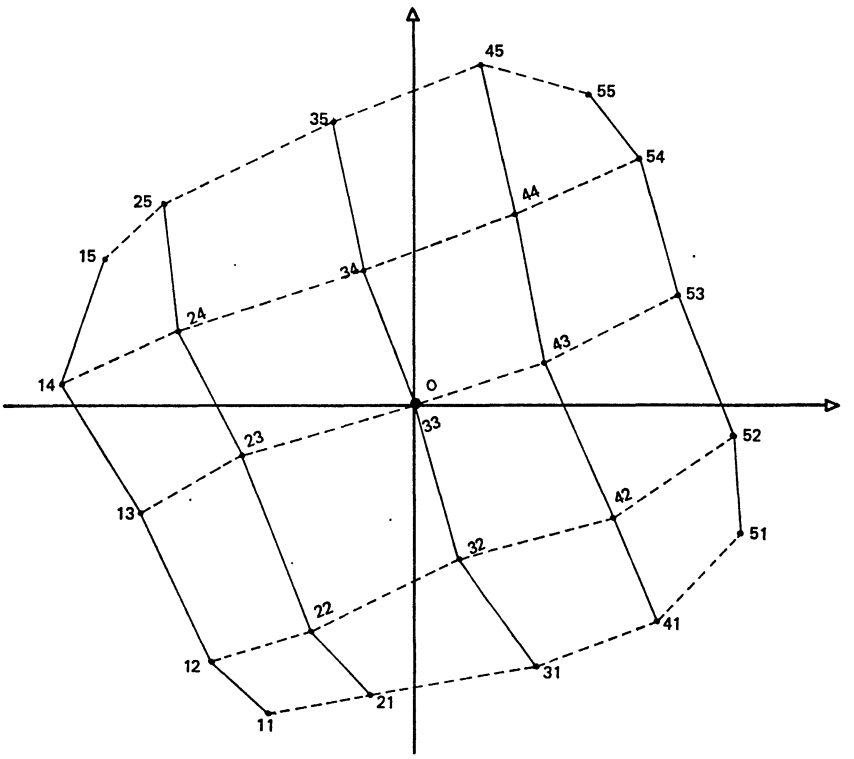


Figure 4. — *Plan des deux facteurs de l'analyse en facteurs communs et spécifiques*

2.3. *L'analyse en composantes principales*

Puisqu'il s'agit de représenter des « variables » et non des individus, nous envisagerons cette analyse sous l'angle des représentations euclidiennes : ce sont les meilleures représentations euclidiennes successives dans des sous-espaces de dimension croissante où les points variables sont à la distance $d^2_{ij} = \sigma_{ii} + \sigma_{jj} - \sigma_{ij}$,

1. Une excellente description de cette méthode se trouve dans *Factor analysis as a statistical method*, Lawley et Maxwell, Londres, Butterworth, 1963.

2. Morrison D.F., *Multivariate statistical methods*, New York, McGraw-Hill, 1967, pp. 286-289.

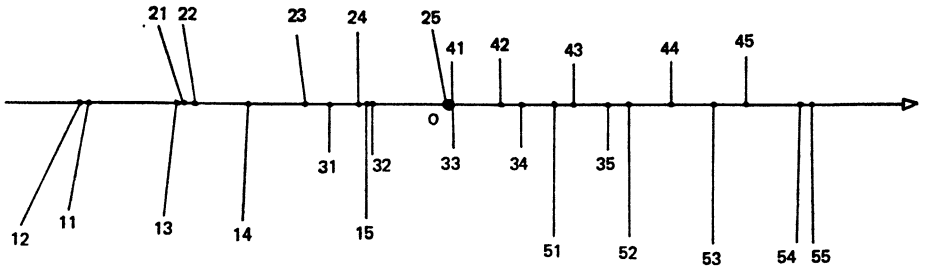


Figure 5. — *Position des tests sur l'axe factoriel de l'analyse à un seul facteur*

Remarquons en passant une belle propriété qui nous permettra de baptiser cette analyse « l'analyse d'interaction » : faire comme nous l'avons décrit revient à chercher la position en composantes principales, mais de variables transformées par rapport aux variables initiales par $\xi_{ij} = \chi_{ij} + \chi_{oo} - \chi_{io} - \chi_{oj}$

où χ_{io} = moyenne de la variable i

χ_{oj} = moyenne de l'individu j

χ_{oo} = moyenne générale

L'inertie totale $\sum_i \sum_j \xi_{ij}^2$ est la somme des carrés utilisée dans l'analyse de variances à deux facteurs pour tester l'interaction. Les variables, comme les individus, sont centrés dans cette analyse.

La représentation obtenue (Fig. 6) est très voisine de la représentation de l'analyse en facteurs communs et spécifiques ; tout au plus l'aspect géométrique est moins bon (disproportion des cases centrales).

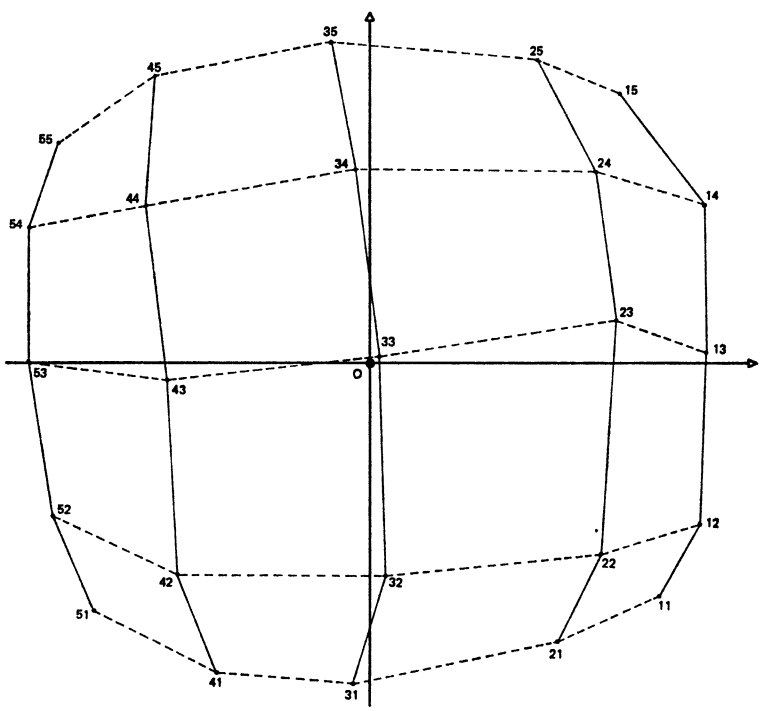


Figure 6. — *Premier plan principal de l'analyse d'interaction*

L'inertie des 22 axes ultérieurs est de 60 % comme dans le cas de l'analyse des correspondances et ces axes n'apprennent plus rien sur la configuration du carré.

Jusqu'ici tout marche à merveille et l'on serait porté à partager l'opinion de certains auteurs : s'il y a une structure, elle est facilement détectée. Nous pensons au contraire que c'est la connaissance que nous

avions de la structure qui nous a permis de la retrouver (déjà, sans aller plus loin, si nous ignorions que le phénomène était bi-dimensionnel, la configuration du troisième et quatrième axe nous aurait plongé dans un grand embarras). Pour mieux le prouver, supprimons donc certaines connaissances de la « réalité » ; opacifions l'écran en ignorant désormais la connexité des lésions [1] : dès lors, il n'y a plus de raisons de coder l'échec 1 et la réussite 0, il est parfaitement légitime de prendre le codage opposé : échec = 0, réussite = 1 ; adoptons-le et voyons si la structure apparaît avec la même facilité que dans les analyses précédentes.

2.4. *L'analyse des correspondances avec le second codage*

Ici, la surprise est forte : sur le plan des deux premiers axes (Fig. 7), le carré n'existe plus ; son aspect planaire n'est pas trop profondément perturbé (6 recoupements d'arêtes), mais le carré a bel et bien disparu. On peut naturellement l'imaginer déployé sur une sphère mais la prise en compte du troisième et du quatrième axe infirme totalement cette imagination. En outre, l'inertie des deux premiers axes n'est plus la même que pour la première analyse, bien qu'assez voisine. Certes, le spécialiste ne s'étonnera pas car il sait que l'analyse des correspondances n'est invariante ni par homothétie ni par translation de variables ou des individus. Mais le profane a de quoi se faire une philosophie : très souvent le choix d'échec = 1 et de réussite = 0 est arbitraire et cet arbitraire suffit pour faire apparaître deux résultats radicalement différents. Puisque ici nous avions créé la réalité, nous pouvons très facilement expliquer ce terrible changement : les poids faibles deviennent les poids forts, ainsi les coins du carré sont attirés vers le centre ; surtout : la lésion était connexe, mais le complémentaire de la lésion ne l'était pas. Par contre, l'analyse en facteurs communs et spécifiques et l'analyse d'interaction demeurent invariantes dans ce changement de codage.

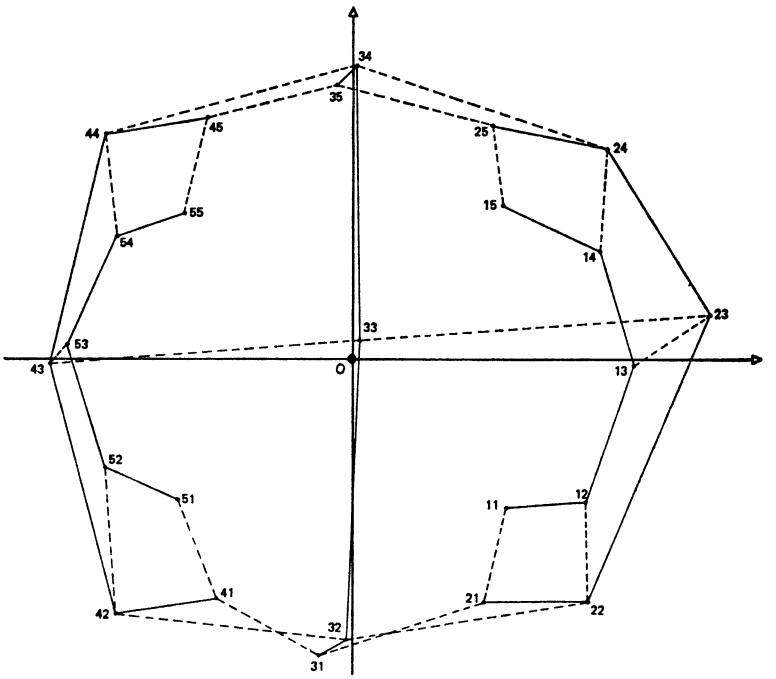


Figure 7. — *Premier plan de l'analyse des correspondances en codage inverse*

Plus généralement, l'analyse factorielle permet d'appliquer toutes transformations linéaires des variables à leurs coordonnées sur les axes. L'analyse d'interaction ne résiste qu'à des changements globaux d'origine et d'échelle. L'analyse en composantes principales normées (chaque variable est transformée linéairement en variable de moyenne nulle et de variance unité) résiste en outre aux changements locaux positifs d'échelle des variables.

1. Si la lésion est connexe par construction, son complémentaire peut être formé de plusieurs parties disjointes par exemple si la lésion était constituée par les cinq carrés de la 3e colonne.

La seule distance qui serait invariante pour toute transformation serait la distance de Mahalanobis [1], malheureusement elle n'est ici d'aucun secours : d'une part nous travaillons sur les variables, d'autre part cette distance tue la représentation puisqu'elle redonne à l'ellipsoïde des données une forme hypersphérique ! Cependant les spécialistes de l'analyse des correspondances peuvent faire remarquer qu'en cas de codage arbitraire des données il fallait « dichotomiser » les variables, autrement dit, créer une variable échec et une variable réussite, amalgamer ainsi les 25 variables du codage direct et les 25 du codage inverse. Le premier plan de cette analyse dichotomique est porté sur la Figure 8. Il consiste à peu de choses près à superposer les deux premiers plans des deux précédentes analyses des correspondances. A notre avis, un tel plan est illisible en termes euclidiens car le même objet, ou test, y est représenté par deux points : sa réussite ou son échec. Or ici, précisément, la réussite est l'inverse de l'échec. Ce n'est cependant pas toujours le cas : ainsi, en génétique, l'analyse des fréquences géniques est effectuée en double codage puisque l'un des génotypes ne signifie pas le contraire de l'autre. Mais, en général, pour des données binaires, réussite et échec sont en dualité. Donc l'analyse des correspondances perd pied la première mais les autres vont rapidement suivre quand on opacifie l'écran.

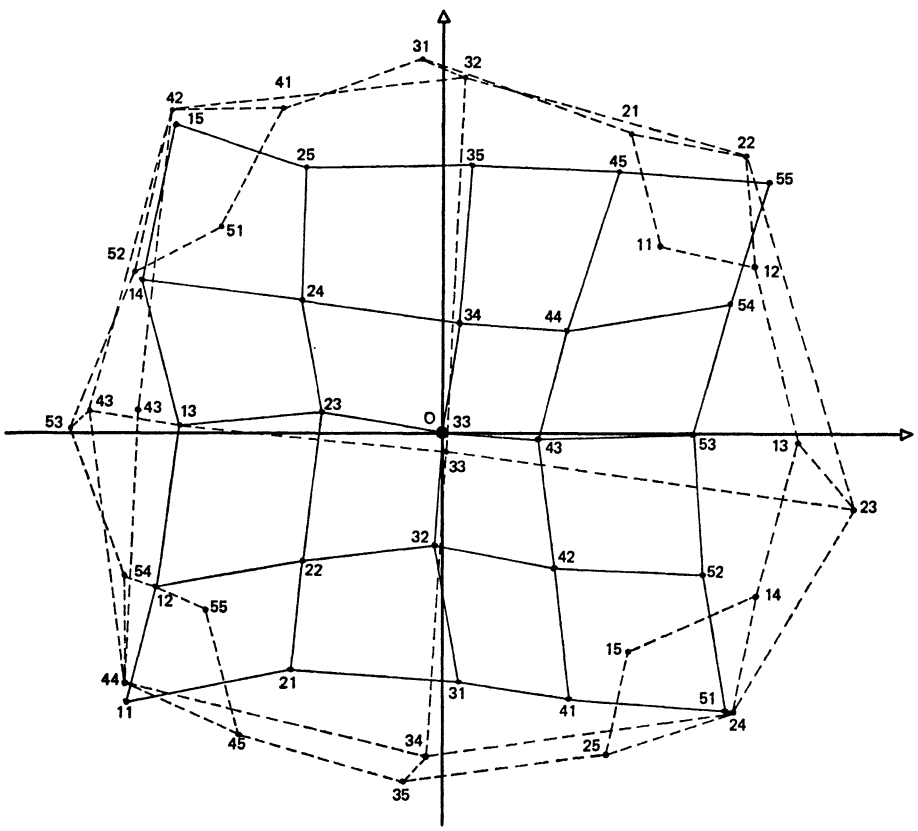


Figure 8. — *Premier plan de l'analyse des correspondances des 25 variables dichotomisées* (50 *variables*)

2.5. *Inversions locales du codage des variables*

Supposons maintenant que nous ne sachions même plus, pour un test donné, ce qui est échec et ce qui est réussite — c'est le cas de nombreuses variables en sociologie : attribuer 1 à catholique, 0 à non catholique, puis 1 à bachelier, 0 à non bachelier ; l'arbitraire porte alors sur chaque variable séparément.

Le premier plan principal d'une analyse des correspondances faite après inversion locale des codages est figuré sur la Figure 9. On a cerclé les variables qui avaient gardé le codage initial échec = 1, réussite = 0 et on les a jointes comme précédemment. Mais si l'on joint toutes les variables correspondant à des cases

1. Cette distance s'obtient en utilisant l'inverse de la matrice des covariances. Elle peut ainsi s'introduire par la construction des « fonctions linéaires discriminantes » de Fisher. On en trouve un bon exposé dans Morrison, *op. cit.*

adjacentes, on obtient la Figure 10 où cette fois toute structure du carré a bien sûr disparu, aspects géométrique et planaire confondus. Le même phénomène se produit avec les composantes principales et avec l'analyse en facteurs communs et spécifiques. On remarque cependant sur la Figure 9 que, séparément, les variables inversées et non inversées correspondent à leur position dans le carré inversé et non inversé. Au fond, la structure interne de chacun de ces deux groupes a résisté ; il va en être de même si seulement certaines variables étaient analysées.

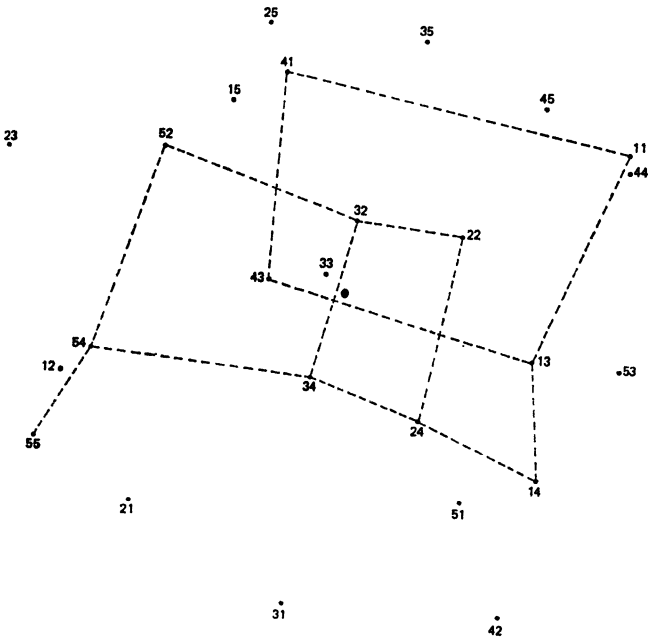


Figure 9. — *Premier plan de l'analyse des correspondances des variables localement inversées (sont jointes les variables non inversées)*

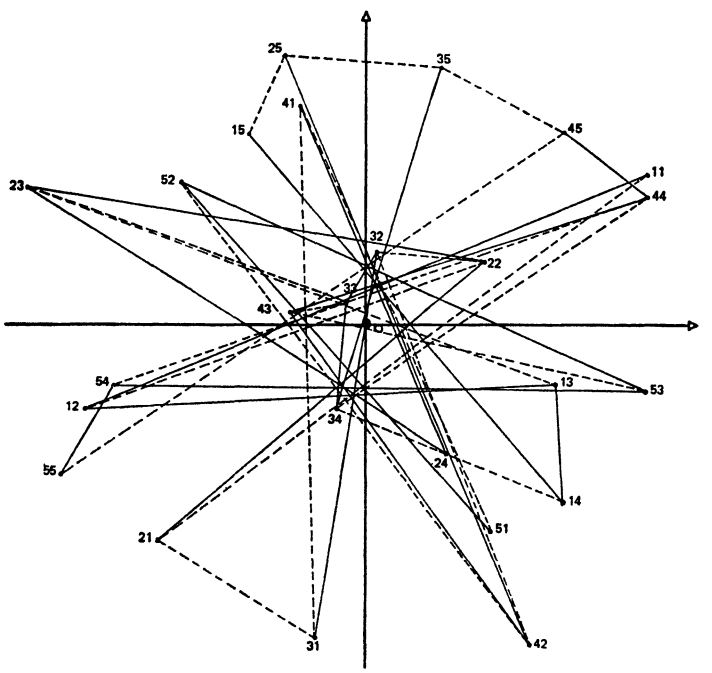


Figure 10. — *Premier plan de l'analyse des correspondances avec inversion locale des codages (même Figure que 9 mais en joignant les tests qui correspondent à des codes adjacents)*

2.6. *Analyse de certaines des variables*

On a supposé une bijection entre tests et cases de l'hémisphère droit, il serait légitime de considérer qu'il y a plutôt une injection des tests dans les cases : certaines cases existent qui correspondent à des compétences qu'aucun test ne saisit.

Supposons donc que seulement 13 variables soient retenues. Le premier plan principal de l'analyse des correspondances (Fig. 11) restitue bien la structure et ressemble de très près au premier plan des 25 variables dont on aurait gommé les 11 variables absentes. Cette propriété se retrouve pour toutes les autres analyses, puisque les distances réciproques des variables sont ou peu ou pas modifiées par les analyses. Mais imaginons que, parmi les 13 variables retenues, certaines soient en codage direct, d'autres en codage inverse ; alors le carré a définitivement disparu. Hélas, ce dernier cas est peut-être le plus proche de ce que nous rencontrons sur des données réelles. Faut-il chercher une structure euclidienne que l'on n'a aucune chance de mettre en évidence ?

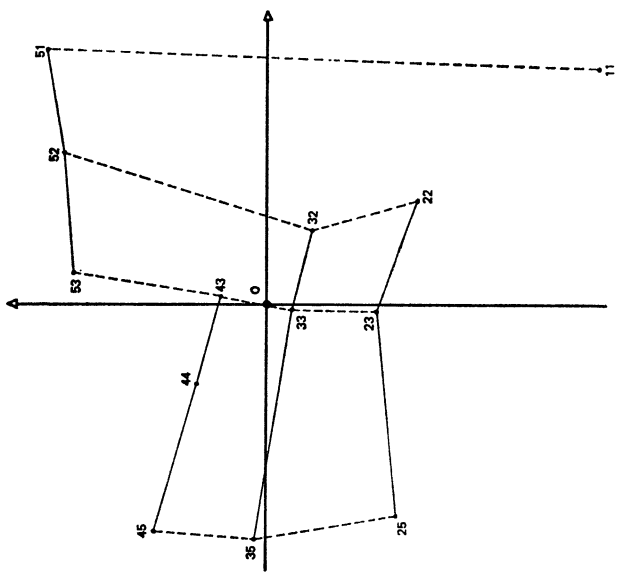


Figure 11. — *Premier plan de l'analyse des correspondances de 13 tests sur les 25*

Aussi beaucoup d'analystes préfèrent-ils s'orienter vers les structures d'arbres, l'analyse euclidienne permettant seulement de rectifier ce que les hiérarchies obtenues auraient de trop sommaire (n'oublions pas que N points dans un plan nécessitent l'estimation de $2N - 3$ paramètres, tandis que N points dans un arbre n'en exigent que $N - 1$). L'écran donc devenu plus opaque, nous ne savions même plus que la structure était plane et nous pouvions, comme les neuropsychologues autrefois, supposer des rapports hiérarchiques entre tests. Appliquons donc plusieurs méthodes hiérarchiques.

2.7. *L'analyse hiérarchique en codage direct*

Cette méthode utilise la distance du χ^2 et procède par regroupements successifs avec le critère d'inertie intragroupe minimale. La Figure 12 montre l'arbre obtenu (l'orientation des branches de l'arbre est systématique à l'aide d'un critère de post-optimisation). Cet arbre ne nous apprend que peu de choses : tout au plus distingue-t-on cinq groupes qui, grosso modo, correspondent au centre du carré (groupe très volumineux) et aux quatre coins. Comme la structure est régulière, l'un des inconvénients majeurs des arbres apparaît facilement : un regroupement en interdit d'autres et de petites dissymétries finissent par jouer un rôle important. Pour rapprocher cet arbre de la structure réelle que nous nous étions donnée, nous avons, sur le carré initial, illustré les regroupements successifs par des courbes de niveau (Fig. 13). On voit mieux ce qui est arrivé et l'on comprend combien les conclusions que l'on peut tirer de cet arbre sont ou sommaires ou précaires. Inversons maintenant le codage.

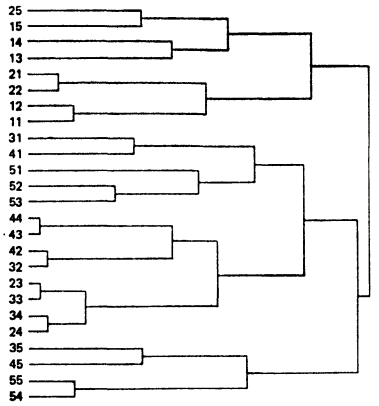


Figure 12. — *Analyse hiérarchique en codage direct*

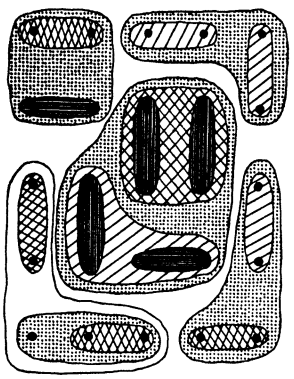


Figure 13. — *Résultat de l'analyse hiérarchique représentée sur la structure réelle*

2.8. *Analyse hiérarchique en codage inverse et en codage dichotomique*

Pas plus que l'analyse des correspondances, l'analyse hiérarchique ne résiste à l'inversion de codage. L'arbre obtenu est porté sur la Figure 14 et les courbes de niveau utilisant la structure réelle, sur la Figure 15. Les conclusions déjà sommaires de la précédente hiérarchie sont bousculées : le groupe volumineux du centre s'est volatilisé, seuls restent les quatre groupes des coins.

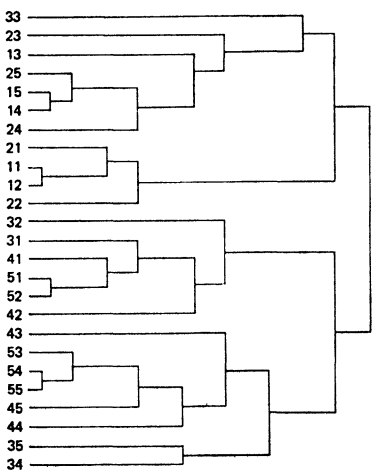


Figure 14. — *Analyse hiérarchique en codage inverse*

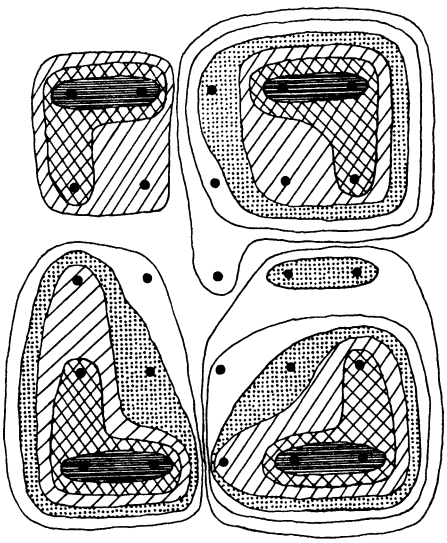


Figure 15. — *Résultat de l'analyse hiérarchique en codage inverse, représenté sur la structure réelle*

Enfin, cette analyse hiérarchique a été effectuée sur les données dichotomiques et, comme pour l'analyse des correspondances, le résultat est embrouillé (Fig. 16). Nous avons utilisé des pointillés chaque fois que des variables inversées et non inversées coexistaient dans le même groupe. On voit que l'amorce des coins subsiste, mais ils s'agglomèrent vite avec les coins diamétralement opposés dans le codage contraire. Seul le groupe du centre du codage direct subsiste.

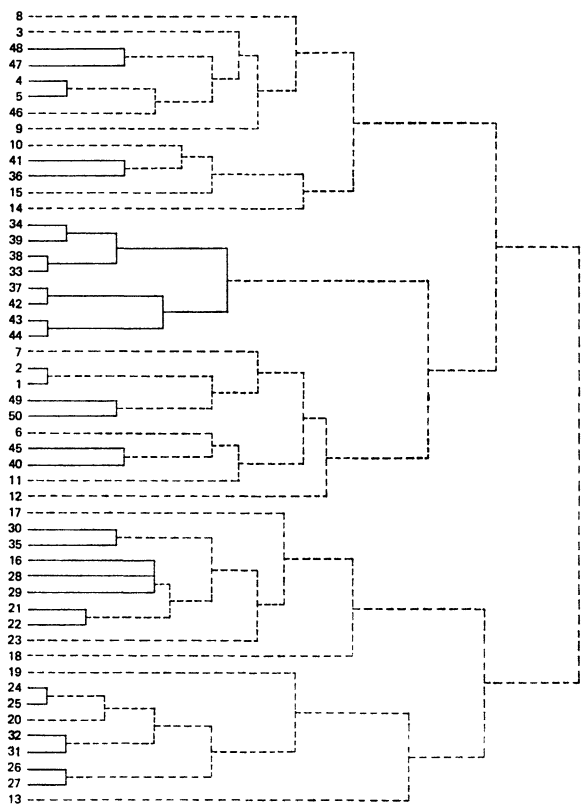


Figure 16. — *Analyse hiérarchique des variables dichotomisées*
(les tests sont numérotés ligne par ligne : de 1 *à* 25 *en codage direct,* de 26 *à* 50 *en codage inverse)*

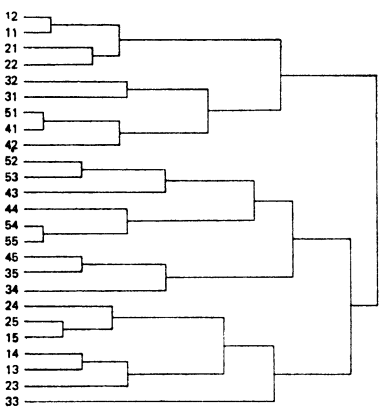
De carré il n'en est plus question, de structure non plus. Seules quelques remarques peuvent être faites qu'une quelconque matrice de similitude eût permis de faire à simple lecture. Essayons encore deux méthodes : les approximations supérieure et inférieure d'une ultramétrique [1], en prenant une ressemblance entre variables qui supporte l'inversion du codage, leur coefficient de corrélation.

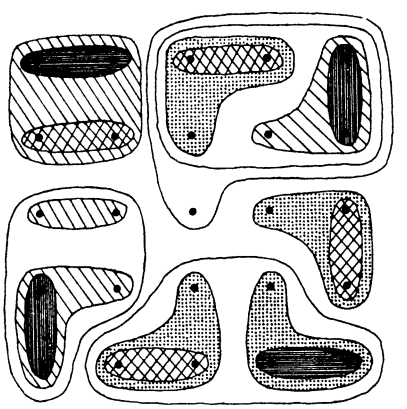
3. Approximation par une ultramétrique supérieure

L'arbre obtenu est porté sur la Figure 17 et, puisque l'ultramétrique supérieure procède par agglomération de « grumeaux » (groupes fortement liés), il était légitime de conserver la représentation par courbe de niveau qui confrontait bien la hiérarchie obtenue et la « réalité ». Le résultat est assez voisin de celui de l'analyse hiérarchique en codage inverse, il serait même plutôt pire, une sorte d'hésitation apparaissant entre le rôle des côtés du carré et celui des coins (principalement pour le bas du carré).

4. Approximation par une ultramétrique inférieure

L'arbre obtenu se trouve sur la Figure 18. Pour la confrontation avec la réalité, nous avons changé la représentation car l'ultramétrique inférieure coïncide avec l'arbre minimum[2]. On peut donc tracer cet





Figures 17 et 18. — *Ultramétrique supérieure : mêmes représentations que pour les 2 premières analyses hiérarchiques*

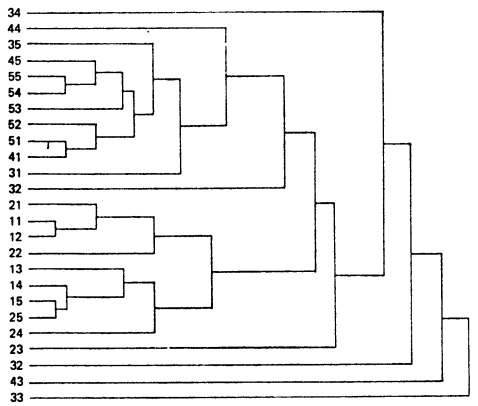
arbre sur le carré lui-même (les sommets sont les différents tests) et ceci a été fait Figure 19. Le résultat est bon. Si l'on se souvient de la représentation obtenue en composantes principales il s'explique aisément : les cases de l'intérieur étaient plus éloignées entre elles que celles de l'extérieur, ainsi le contour a pu être obtenu par cet arbre minimum.

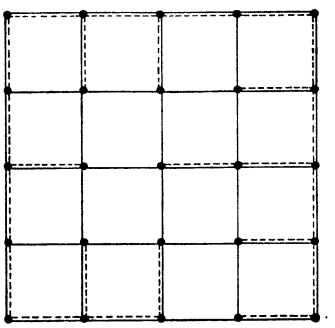
---

1. *Introduction à la classification automatique.* Lerman, Paris, Gauthiers-Villars — *Principles of numerical taxonomy,* Sokal et Sneth, Londres, Freeman, 1963.

2. Berge C., *Théorie des graphes et ses applications,* Paris, Dunod, 1966.

De nombreuses méthodes existent et nous n'allons pas les appliquer : les segmentations, plus frustes, les hiérarchies descendantes, et toutes sortes d'indices de similarité ou de dissemblance peuvent être choisis (surtout avec des variables binaires). Mais nous en savons assez et nous le savions déjà : on ne reconstituera jamais une structure euclidienne en utilisant une représentation en arbre. Tout au plus félicitons-





Figures 19 et 20. — *Ultramétrique inférieure : représentation de l'arbre et représentation de l'arbre minimum porté sur la structure réelle*

nous que les arbres n'aient pas procédé à des regroupements aberrants et qu'ils se soient bien comportés, étant donné leurs faibles moyens. Imaginons enfin que certaines variables manquent et que certaines qui restent aient leur codage inversé, il ne reste plus aucun élément de notre structure réelle.

Jusqu'ici nous étions encore dans un cas satisfaisant à plusieurs points de vue : nous avions 1 000 malades, et pour chacun exactement 15 cases atteintes par la lésion. Certes, il est fréquent de traiter d'un univers complet ou presque, mais il arrive que l'on ne possède qu'un échantillon de cet univers. Aux imperfections des analyses précédentes, s'ajoute alors un « brouillage » causé par le hasard. Nous allons terminer par une illustration de ce phénomène qui perturbe assez profondément les analyses qui avaient bien reconstitué la « réalité » et achève de détruire celles qui s'éloignaient de cette réalité. En outre, nous modifierons le modèle de croissance cellulaire de manière à obtenir des lésions de tailles diverses : nous arrêterons la croissance de la tumeur lorsque le nombre de cases qui ont été menacées au cours de la croissance aura atteint une valeur donnée.

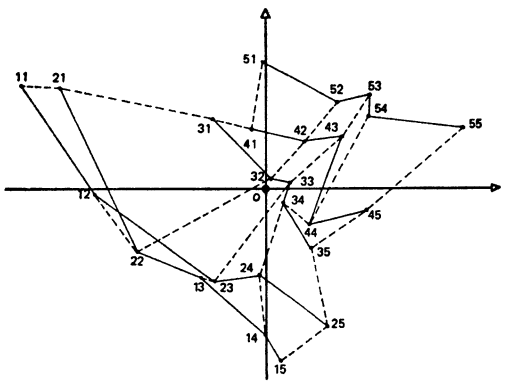


Figure 21 a. — *Premier plan de l'analyse des correspondances en codage direct sur 70 malades*

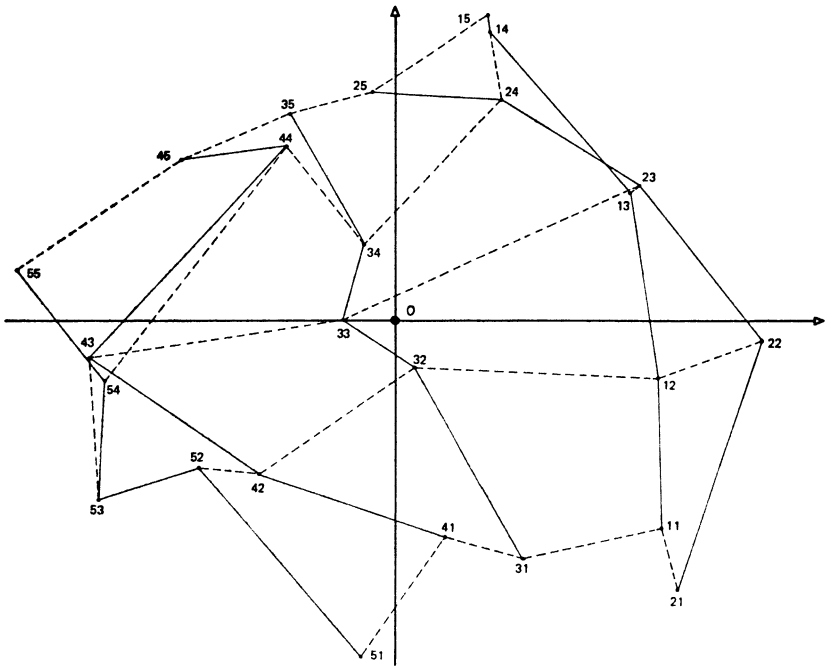


Figure 21 b. — *Premier plan de l'analyse d'interaction sur 70 malades*

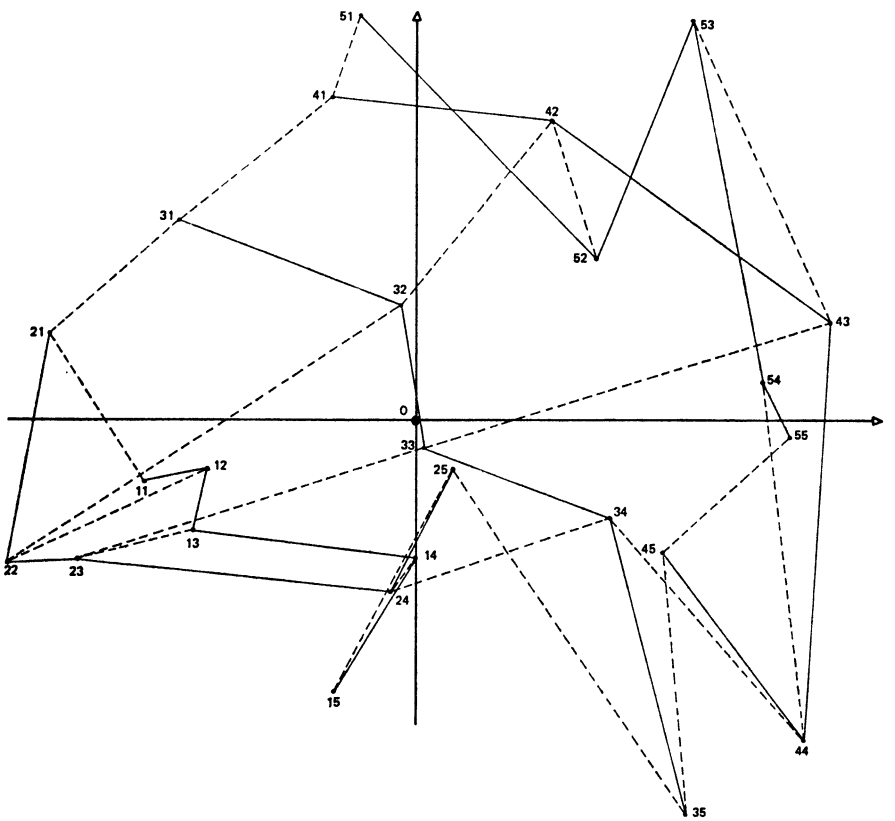


Figure 21 c. — *Premier plan de l'analyse des correspondances en codage inverse sur 70 malades*

Les Figures 21 a, b, c, montrent les résultats de l'analyse d'interaction et de l'analyse des correspondances en codage direct et inverse sur 70 malades avec un seuil d'arrêt tel que les lésions soient importantes. Les deux premières analyses sont fortement perturbées par les fluctuations du hasard, l'aspect géométrique disparaît mais l'aspect planaire reste assez fort. L'analyse des correspondances en codage inverse détruit maintenant toute la structure, on ne peut même plus recomposer mentalement les déformations ou pliages qu'aurait pu subir le carré initial. En fait, les deux dernières analyses sont intéressantes surtout pour l'inertie de leurs deux premiers axes : 43 % en codage direct et 26 % en codage inverse, transition très nette entre le second et le troisième axe en codage direct, pas de transition en codage inverse. Ce critère d'inertie est donc bien douteux, de nombreux auteurs le soulignent. Tout douteux qu'il soit, il est cependant au cœur de toutes les méthodes de représentation euclidienne puisque c'est grâce à lui que les axes successifs sont choisis. Il est donc difficile de le mettre en doute dans les résultats lorsqu'il a été la cheville ouvrière dans la méthode. Il est tout aussi difficile d'accepter une telle variation d'inertie pour une simple inversion de codage. (En composantes principales, l'inertie des axes n'est modifiée ni par inversion globale ni par inversion locale du codage.)

Terminons par un exemple sur 100 malades avec un seuil d'arrêt plus faible dans la constitution des lésions. Les Figures 22 a, b montrent le premier plan de l'analyse des correspondances pour le codage direct puis inverse. Nous sommes maintenant habitués au phénomène et nous pouvons cependant admirer la bonne reconstitution opérée par l'analyse en codage direct, malgré le faible nombre de données (100 lésions seulement pour des milliards de lésions possibles avec le modèle). Les Figures 23 a, b montrent le résultat de l'analyse d'interaction et celui d'une analyse des correspondances après inversion du codage de certaines variables (premiers plans). L'analyse d'interaction a été plus sensible au faible nombre de malades que l'analyse des correspondances. L'inversion de codage produit les mêmes effets néfastes que dans les exemples précédents.

A la lumière de cette petite expérience, nous pouvons argumenter ce que nous affirmions au début de ce travail :

Nous assignons deux buts aux analyses multivariées : décrire une structure parfaitement connue et préciser une structure pressentie. Nous avons peu parlé du premier but, mais il apparaissait en filigrane tout au long de l'article : nous connaissions presque parfaitement la structure puisque nous l'avions créée, alors chacune des analyses jetait un éclairage intéressant sur cette structure et permettait de l'illustrer dans sa généralité (analyses euclidiennes) ou dans ses finesses (analyses hiérarchiques). Bien d'autres méthodes

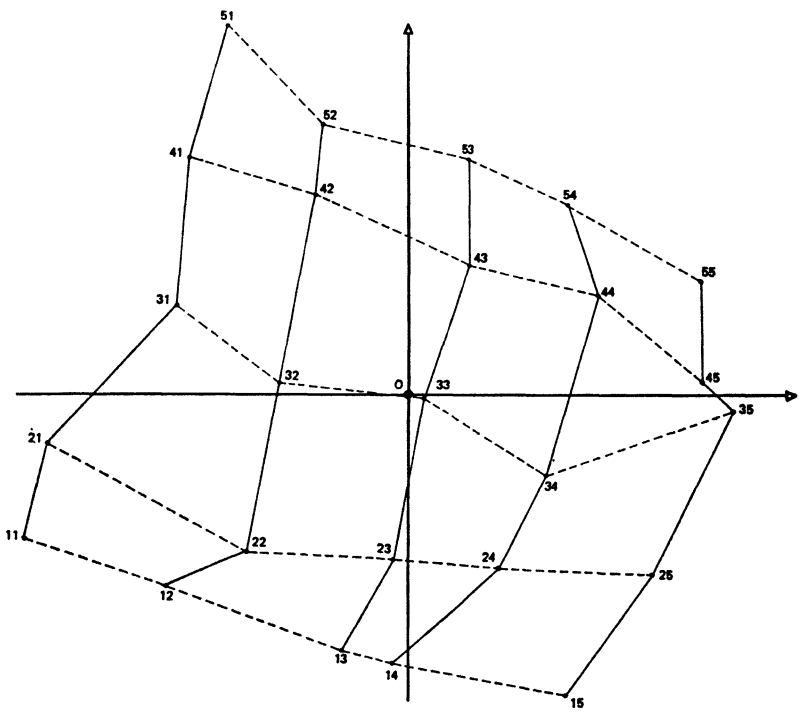


Figure 22 a. — *Premier plan de l'analyse des correspondances en codage direct sur* 100 *malades*

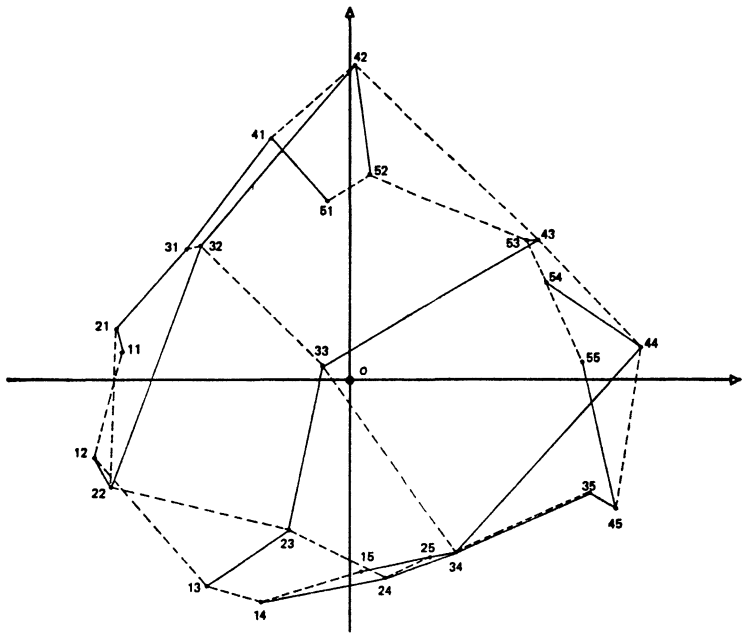


Figure 22 b. — *Premier plan de l'analyse des correspondances en codage inverse sur* 100 *malades*

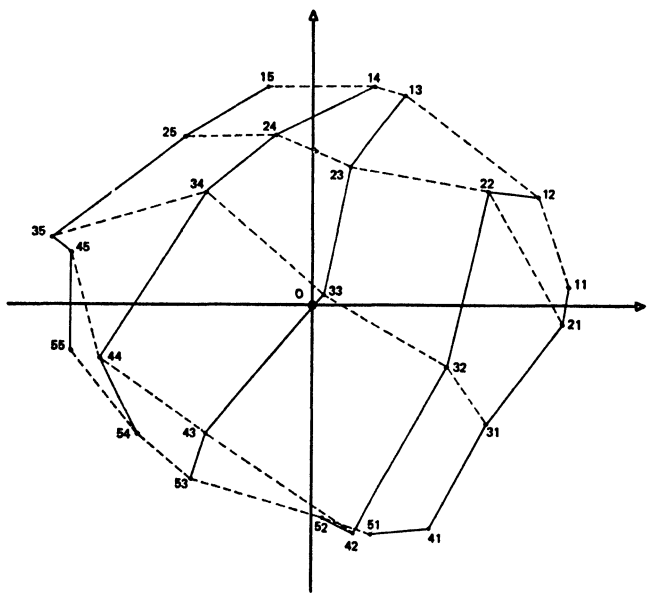


Figure 23 a. — *Premier plan principal de l'analyse d'interaction sur* 100 *malades*

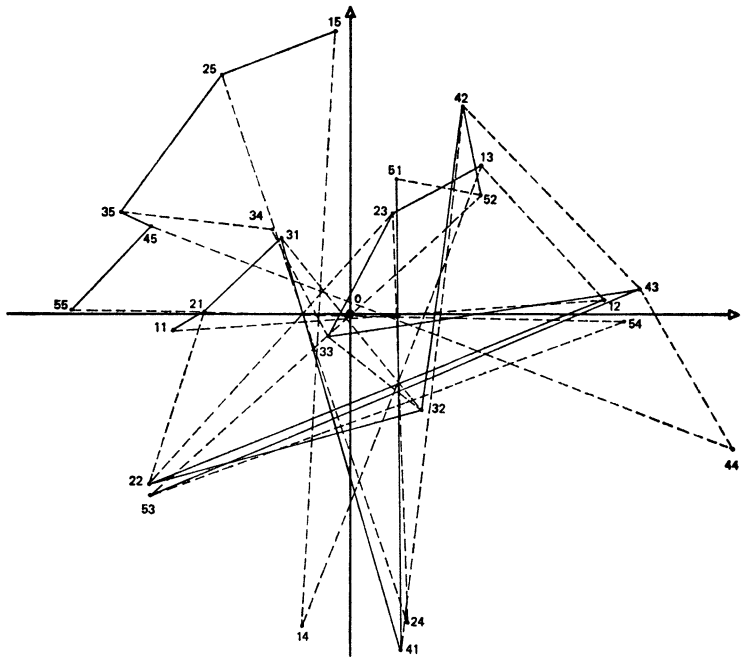


Figure 23 b. — *Premier plan de l'analyse des correspondances avec inversion locale de variables sur* 100 *malades*

auraient représenté d'autres aspects de notre structure, mais le résultat de celles que nous avons employées est déjà excellent. Par contre, pour découvrir la structure, nous avons rencontré d'énormes difficultés. Elle ne peut apparaître que si nous avons de très bonnes informations en dehors des données. Même si, ne connaissant rien, nous effectuions toutes les analyses possibles, nous ne pourrions décider laquelle est la bonne.

Cette information nécessaire sur la structure semble interdire son émergence à partir des seules données. Nous pouvons préciser la structure grâce à tout ce que nous connaissons du phénomène avant de recueillir les données. Ainsi, ces deux buts se ramènent à un seul : illustrer une structure. Dans le premier cas, tout est connu sur la structure, dans le second cas, il faut rapprocher deux sortes de connaissances : les données et la nature de la structure.

Nous retrouvons ainsi la démarche suivie depuis longtemps en analyse de variance : ce que l'on sait de la structure est traduit par la succession des tests et l'emboîtement des hypothèses, les données servant à vérifier ou infirmer la structure prévue. Et tout statisticien connaît bien avec quelle facilité on peut, sur des données quelconques, bâtir après coup une analyse de variance dont les résultats soient significatifs. Cette prudence nécessaire à l'analyse de variance apparaît souvent comme un carcan inutile, comme une bride à l'imagination. Le désir de garder une façade rationnelle pousse à utiliser les analyses multivariées.

Le mathématicien ou le statisticien qui agit ainsi travaille à sa propre élimination puisqu'il élude le problème de la formalisation et de la modélisation.