STATISTICAL ANALYSIS AS A TOOL TO MAKE PATTERNS EMERGE FROM DATA

Jean-Paul Benzecri

UNIVERSITY OF PARIS
PARIS, FRANCE

Forward

Before we say anything about recognition, it should be fair to speak of patterns. The most common idea we have of a pattern is that given by a word of some natural language. Although dictionaries have been available for centuries, none has succeeded to give a clearcut algorithm by which it is to be decided when and to which objects any word can be applied. Some logicians claim that semantical definitions do exist, yet unstated, which provide lists of features determining the class of things corresponding to each meaningful sound. However, it is wise to notice that such a simple noun as, e.g., a neighbor, can be understood only in a complex situation where one can ask such questions, as: to whom is he a neighbor?, or: is it enough to be at walking distance from each other to be called neighbors? After all your neighbor is just someone who seems presently closer to you than a number of others.

The same difficulty is met when we are to distinguish a letter d from a letter a in a manuscript: is a d which it makes sense to call a d, and besides is taller than a class of resembling letters which we have advantage to read as a. Indeed a very efficient reading machine, and possibly a not too complicated one, should be built on this program; segment what is handed to you into resembling units which you label from 1 to at most 100; and just let another program guess, possibly from frequency tables, possibly by use of a dictionary or even a grammar, that item No. 13 is capital D, and item No. 29 is small z.

This is enough I think to suggest that what has been contributed those last five years in France and mainly in my laboratory in the field of automatic classification and multivariate analysis is of interest to the science of shapes, and that the page you have just read is a promising introduction to my coming lecture!

1. Data Analysis Versus Models and Discriminant Analysis

When faced with intricated sets of data, human scientist or biologist calls for help from the statistician, the former generally has some hypothesis, or even an accepted theory, while the latter has none. Besides, only in the last decade has statistical science received from electronic devices due to help to cope with the huge mass of digits that has to be handled in order to submit the flesh of data to the soul of formulae. It is therefore quite natural that the opinions, which the sociologist or psychologist brought with him, were made the main ingredients of the mathematical answer he received. In some cases, common-sense-concepts were built into a model with just enough parameters in it to outnumber the data. In others, an a priori classification was recovered up to but not always much less than 50% error-rate, using an ad hoc discriminant function; or else which in a sense is the worst of all, an evidently true statement was rejected on the basis of a parametric test which involved many more hypotheses and by far less natural ones than did the statement itself. But since those good old days when it was thought that it was enough to have two to make out an average, the statistician has acquired certain rights not seldom stronger than those of evidence.

That such a practice takes us back to fictions which we thought had been ruled out of science by Newton's "Hypotheses non fingo" (which could be properly paraphrased: "I don't use hypotheses as an ingredient in their own proof") should be clear to any scientist. However, there is a growing evidence, too, that Baconian methodology does not suffice to make out a knowledge in a context of numerous fluctuating variables. If one is willing to study sociology by letting all parameters constant but one, it is to fear that nobody will have neither power nor time to produce and observe all of the various situations relevant to the field. The notion of a shape or a pattern should emerge from a sea of data, not through a priori nominalistic postulates or axioms or by unduly fragmentary measurement of isolated facts which in themselves are meaningless since they depend on the environment and quickly change their mode but through the simultaneous synthesis, in the ethymological sense of putting together a number of elementary facts which help us ascend a step in the hierarchy of causes. But multidimensional synthesis cannot be achieved inside a human brain is without many arbitrary choices which often make the result devoid of significance. Therefore, the help of a computer is needed to apply to the data previously collected a set of quasi-universal computations or rather transformations which give them such a shape that the man of the field may unarbitrarily read on the output what was undecipherable in the input.

This certainly raises a question: To what extent are we able to fulfill this magnificient program? If factor analyses strongly rely upon arbitrary rotations to extract anything significant, how can it be claimed that the statistician should do nothing more to the data than the multidimensional analogue (or even the non-euclidean generalization) of computing an average or adjusting a straight line? that determinate purely mathematical handling of data tables is enough to make emerge such significant results that the man of the field may read them and accept them with as little doubt as he welcomes a nose in the middle of a face?

Let us not boast of too great a success. Alternating the study of the numerical data and of the concepts of the field is sound and besides unavoidable. Already several times the results of our computational analysis did suggest extending the measurements or rebuilding the experimental framework. But too intricate a mixture between what we observe and measure and what we think to be the underlying structure of it is an obvious danger against which we claim our new statistical methods – a few years old as they are, they provide a significant weapon. And without indulging in more philosophizing, we will try to put before you a few mathematical principles illustrated by examples. Colleagues interested in more details, including listings of programs, are welcome to write to Paris Statistical Institute and will receive our articles and reports.

2. Automatic Classification and Similarity Indexes

Although we don't believe that a splendid future is promised to taxonomic hierarchies, it seems unavoidable to start our review with automatic classification techniques.

Usually a classification algorithm consists of two fairly independent parts: (1) the computation of a similarity index, (2) the formation of clusters. Let us study these two points successively and state recent French contributions to the art.

In their now classical treatise devoted to the "Principles of Numerical Taxonomy," Sokal and Sneath offer a great number of similarity indexes, most of which are computed from data of the following (or a resembling) format: an I x C table (I = {i}, set of items; C = {c}, set of characters) is filled with ones and zeros. There is a one at the intersection of column C with line i, if item C possesses character c, and there is a zero when the contrary happens to be true. Actually, set C may very well be a set of vegetal or animal species, and set I a set of patches of land with character c possessed by item i if animal c is found in land i . . . but these semantic peculiarities are not relevant to the mathematical treatment which is concerned only with format.

Given two items i and i', four numbers may be defined

$$Y\,Y(i,i');\ Y\,N(i,i');\ N\,Y(i,i');\ N\,N(i,i')$$

of which the first is the number of c's possessed jointly by i and i', etc. It is clear at first that Y Y and N N correspond to match while Y N and N Y number differs. It is not clear, however, whether a N O N O match is just as meaningful as a Yes Yes match. Indeed whales and slugs both lack feathers and shell which do not make them look much the same. While the fact of having non-dark hair being just equivalent to that of having light hair, it is a negation which amounts to the same as an assertion. This is only one among the problem involved in coding characters and computing similarities.

Inspired by a paper of R.N. Shepard, about which we will speak fully later on (Cf. No. 7), we thought that the notion of similarity-index (or pseudo-distance) should be replaced by that of "ordonnance" or set of inequalities between the distances and consider themselves irrelevant. This

implies, of course as will be seen immediately, that the clustering algorithms should be devised to rely only on the inequalities and *not* on the distances. But it has various advantages empirically tested by S. Regner and M. Roux:

(1) The host indexes enumerated by S. and S. is essentially reduced to two, which differ mainly (and do definitely differ!) according to the part played by N N matches.

(2) Cluster techniques become much less sensitive to arbitrary choices since a monotonic transformation on the index now becomes totally irrelevant.

Let us come now to clustering. Although ascending techniques, i.e., grouping elements into islets and these islets into islands or various clumpings have been suggested, we think that, as M.S. Watanabe made it already clear years ago, the partitioning of a set I into p distinct non-overlapping classes should, and to a certain extent can, rely on the following criterion: To each partition is associated a test quantity T and that partition is chosen which gives T the highest value. Unless we run into combinatorial infinity, not all partitions can be considered, but the following algorithm, pointed out to us by S. Regner (then at Maison des Sciences de l'homme, Paris), is convenient. Starting from one partition (coded as a function on set I with values ranging from 1 to p), we consider only those which can be obtained from it by changing the class of only one item i in I (that is, by changing only one value at a time of the function). The algorithm stops when no improvement of T can be obtained by such a restricted change in the partition. As for computing test quantity T, there is an ingenuous principle for which we are again indebted to S. Regner. Consider the data (a similarity matrix, or an ordonnance) as an element of a set S of mathematical structures to which the unknown (a partition) also belongs and then compute T as the inverse of a distance (of some kind or another) in set S. Indeed a partition may readily be interpreted as a (partial) ordonnance. It suffices to consider distances between elements inside one class as smaller than distances between elements belonging to different classes. There is still some choice left for the T-test distance used on set S of all (partial or total) ordonnances and only after some mathematical trials and practical experiments was de la Vega (again an investigator from M.S.H.) able to define a T better than the one we first suggested. This led to a program by M. Renaud which has been satisfactorily used to process archaelogical, psychiatrical, ecological data.

Whatever may have been this success, of which it is difficult to give here proper account in a few words without an explanation of the intricate data involved, we do think that the part played by automatic classification in data analysis should be modest. Let us state our opinion in few points.

(1) Some automatic ascending clustering seems necessary when too large sets of items are considered, in order to come down, let us say, from 5,000 elements to 500 islets.

(2) When processing a moderately large table, the very partitioning technique should use as an input not distances or ordonnance but the very table I x C itself. If both I and C are simultaneously partitioned, the former in a set Q of classes, the latter in a set K, there comes out a Q x K table:

cell (q,k) contains the total sum n_{qk} of the number of qualities in class k, each item i in class q possesses. For such a table a χ^2 may be computed:

$$\Sigma_{q,k}\ (n_{qk})^2/n_q\ n_k$$

The bigger this quantity, the more correlation there is between classes in I and C. It can therefore properly be chosen as T with this definite advantage that the data have been submitted to as little arbitrary treatment as possible. Since this χ^2 program has not been tested yet, we cannot however assert anything in its favor.

(3) After previous classifications have reduced the initial set from a few thousands to a few hundreds and then from a few hundreds to a dozen, it is probably a nuisance to partition further, since for the tables of which the least dimension is a few dozens, we have a very satisfactory and cheap factoring technique, about which we will speak in the next paragraph. Indeed, classification always involves much arbitrariness, much more than does a continuous geometrical representation of the data using, as coordinates in the map the factors extracted by some suitable analysis (or should we rather say, see above, synthesis). The only reason why classification has played such a prominent part in the history of science is that it is quite convenient for human memorizing and processing. But now we have machines, the storage as well as computation facilities of which allow us to give some rest to our brains!

3. The Analysis of Correspondences (Principles)

Let us start this paragraph with a typical example borrowed from linguistics. Let set I be the set of all characters appearing in a play and set J the set of the 40 (e.g.) words most frequently used in that play. We can fill an I x J table with integers by letting n(i,j) be the number of times character i has used word j in the whole course of the play. From such a table we wish to derive some compact information about both characters and words by which we mean the following. We think that words j and j′ are close to each other, irrespective of their absolute frequencies, if they are predominantly used by the same characters, that is, of columns j and j′ in our table are close to being proportional to each other. The same can be said of two characters, i and i′, although one may speak much more than the other. If both have the same sociocultural status or personal taste, they will prefer or avoid the same words. Therefore, lines i and i′ in our table will be similar in profile, whatever different their total weights may be. Again it is natural to say that word j and character i are close to each other, if the relative frequency with which i uses j is above the average. All these notions are precisely what we intend to extract from our table and to display in a compact diagram. We want both sets I and J to be inserted in the same map (be it linear, plane or spacial) so that all their mutual proximities be immediately visible. Before we go into more details, let us give the result of our analysis as applied to Racine's Phèdre (Figure 1). Here, the main contrast is between Phèdre, the passionate queen, and Théramène, a respectful servant: between characters who say tu and assez (enough!) and those who address the latter

by Vous, Madame and Seigneur.

How can such an analysis be achieved and in what does it differ at first sight from many variants of factor analysis you may have heard of? Let us indulge here in a few mathematical formulae:

$$n(i) = \Sigma_j\, n(i,j);\ n(j) = \Sigma_i\, n(i,j);\ n = \Sigma_{i,j}\, n(i,j);$$

$$f(i/j) = n(i,j)/n(j);\ f(j/i) = n(i,j)/n(i)$$

$$f(i) = n(i)/n;\ f(j) = n(j)/n;\ f(i,j) = n(i,j)/n$$

As is well known to all of you, the f(i) f(j) and f(i,j) are called frequencies and the f(i/j) f(j/i) conditional frequencies (of i when j and of j when i respectively). If we had the exact identity:

$$\forall\ i,j:\ \ f(i,j) = f(i) \times f(j)$$

that is, if we had statistical independence between i and j (all subjects use all words with the same conditional frequencies), nothing could be said, according to our viewpoint, about the data-table submitted to our analysis. In statistical terms, one can say that we have selected independence as a null hypothesis, H_0. But the use we make of a null hypothesis is quite different from what is common place in applied statistics. We are not interested in the question of accepting or rejecting a null hypothesis (indeed if H_0 has to be accepted, it means that our data are devoid of significance which we don't want! But that H_0 has to be rejected is only an indication that something may be done. It is not an interesting output in itself.) We want to see displayed as clearly as possible *how* our data differ from H_0, in what direction they deviate from independence. Again our null hypothesis differs from those commonly accepted in factor analysis where one would rather seek to what extent and along which directions do the lines and columns deviate from mere equality, while we are only interested in differences between *profiles*. By doing so we eliminate from the scope of our study the differences in weight (the n(i), n(j) or f(i), f(j)), which generally bring in an amount of noise of the same order of magnitude as the signal we are interested in and therefore prevent the factors extracted from being readily interpretable, even if the situation is obviously undimensional! Let us indicate here before we are misunderstood that we don't restrict our method to cases where f(i) is indeed just noise. It may very well happen that raw marginal frequencies are in themselves a useful information, which will be wise to compare with the factors; but even then it is a great advantage to compute the factors as independently as possible from the marginal frequencies.

And now what do we have to compute? How expensive is it to perform our analysis; what is the rationale for those computations? The computation cost first is as low as possible; it amounts to looking for the eigenvectors (as many eigenvectors as we extract factors) of some symmetrical square matrix, the dimension of which is equal to the number of elements in the smallest of sets I and J. Actually, if I is the smallest set, the matrix is the table of the s(i,i') given by the following formula:

$$s(ii') = \Sigma_j (n(i,j).n(i'j)/n(j).(n(i).n(i'))^{1/2}) \ .$$

As for the rational, the situation is more complicated; and it is quite fortunately so, since our formulae happen to be a crossroad to which one is led from several different approaches. Let us explain it as briefly as possible.

Our first principle was that of distributional equivalence. Actually we were mainly concerned with linguistic applications and included our statistical studies in a course on mathematical linguistics. The very word distribution was borrowed from Z.S. Harris, a famous linguist who in many a stimulating but indeed not universally approved paper, has been claiming that statistical analysis of the distributional facts (namely, environments, co-occurrences of words, etc.) could suffice to give all, or almost all of the structure of a language without any consideration of meaning being involved. An extreme corollary, as you may see, to the statement that to build science is to decode the universe . . . a bold project in pattern recognition. I am afraid it would not be easy to find a sponsor for! Anyway, even without a sponsor, a mathematician can define a formula for a so-called distributional distance, and so we did, letting the square of the distance between i and i' be:

$$d^2(i,i') = \Sigma_j (f(j/i) - f(j/i'))^2 / f(j)\ ;$$

a formula in which a statistician recognizes immediately the χ^2 distance between two laws of frequencies, and to which we were led not so much by analogy than by the requirements that the distance be quadratic and that melting into one two individuals with the same conditional frequencies (two elements i and i' distributionally equivalent or synonymous in the sense that

$$\forall\ j:\ f(j/i) = f(j/i')$$

would not change anything to the whole pattern. After a distance had been defined, it was only natural to do multidimensional scaling (that is briefly stated to make out an euclidean map, extracted axis by axis that would gradually adjust as closely as possible to the computed distances), and so we did. Besides, we also devised some conjectural principle in order to put together both sets I and J on the same map, and with no more theorems behind us, we started writing programs and testing them on data.

The situation gradually grew better; we first recognized that the program was efficient, and then the programmer (B. Cordier) noticed a number of peculiarities she could prove to be theorems and afterwards the mathematician was somewhat ashamed he did not notice at first sight.

Here are those theorems. If the various factors F_1, F_2, etc., relative to the eigenvalues λ_1^2, λ_2^2 (all of which between 0 and 1) are considered simultaneously as functions on both sets I and J, one has the following expansion formula:

$$f(i,j) = f(i)\ f(j)(1+\lambda_1\ F_1(i)\ F_1(j)+\lambda_2\ F_2(i)\ F_2(j) + \ldots \qquad (1)$$

and factors on sets I and J are related to each other by a center of mass principle:

$$F(i) = \lambda^{-1} \Sigma_j F(j)\ f(j/i). \qquad (2)$$

Formula (1) means that marginal frequencies f(i), f(j) plus factors $F_1, F_2 \ldots$ is indeed equivalent to the sum of the information contained in the data table. Besides from some statistical considerations (χ^2 test!), it may be roughly stated how many factors are needed to represent the significant part of this information. Formula (2) quite satisfactorily asserts that up to a λ factor an individual i will appear on the map (our output) with coordinate F(i) at the center of masses of the j, themselves given as masses the conditional frequencies f(j/i).

Besides, now this formula had been proven, it was possible to compare our method with Eckart and Young Factor analysis (this comparison was suggested to us in 1965 by Douglas Carrol, then at Bell Laboratories), and with Guttman views on principal components in scalogram analysis. The latter comparison was especially fruitful since it revealed that, starting from Formula (2) as a rationale, Guttman had first contemplated doing the same analysis we did (and of which scalogram analysis is a particular case) but never went to computations, probably simply because in 1941 there were no computers available! From contacts with scalogram analysis, we benefited the following:

(1) Our attention was focused on the fact that a factor may be an algebraic function of another: e.g., F_2 may be a parabolic function of F_1. This actually happened several times and once to an amazing degree of accuracy (see below).

In the work of Guttman, successive factors depending algebraically on each other were given names such as intensity closure; such rigid interpretations do not always agree with facts, but they may be helpful if carefully handled.

(2) When a parallelogrammatic situation underlies the data table (i.e., when through due permutation of the lines and columns, the proeminent coefficients in the table could be clustered roughly into parallelogrammatic neighborhood of the diagonal), it sometimes happens that unlike what could be predicted through the theory of perfect scalograms, the second factor F_2 is *not* a quadratic function of the first one. This second factor may be used to extract a perfectly parallelogrammatic submatrix and eventually gives a second interpretable dimension.

(3) When the data matrix is filled not with various integers but only with one's and zeroes (Yes and No; Cf. Paragraph 2 above), it is convenient to give two columns to each character (as Guttman did) – a yes column and a no column – so that the underlying factor can be recovered most clearly.

Before we illustrate what has been said by a set of examples, let us speak briefly of the relationship between correspondence analysis on one side and some models on the other.

Let us first suppose that there exists two partitions, $\{I_1, \ldots, I_q, \ldots, I_p\}$ and $\{J_1, \ldots, J_q, \ldots, J_p\}$ of sets I and J respectively into p disjoint sets such that n(i,j) is non zero only if i and j belong to classes numbered

by the same integer [i.e., if $n(i,j) \neq 0$ and $i \in I_q$, then $j \in J_q$]. This may be paraphrased by saying that the data matrix may be (through suitable permutations) rewritten as a sequence of rectangular blocks along the diagonal. Then these partitions are discovered using the first factors extracted. Classes $I_1, I_2, \ldots$ and $J_1, J_2, \ldots$ are concentrated into points on the output map, with classes I_q and J_q exactly at the same point. Such a situation, which was actually discovered from an example before it was proved, is especially important since when the hypothesis is only approximately fulfilled, the program is still able to recover an interesting structure.

Let us now suppose that there exists a normally distributed random vector with two components (x,y) (actually the situation readily generalizes to any number of components) underlying our data matrix in the following sense: to the i's (resp. j's) are associated successive intervals on the X (resp. Y) axis; and n(i,j) is proportional to the probability that the random vector is in the i x j rectangle. Then the first factor extracted F_1 will approximate x on set I ($F_1(i)$ will be the abscissa of a middle point in interval i) and y on set J. And the following factors (assuming that there is no noise, or rather no other interesting structure involved) approximate the successive Hermite polynomials.

4. Analysis of Correspondence (Examples)

In the preceding paragraph we took as a standard example of a correspondence the case where the n(i,j) were integers, representing roughly speaking, how many times individual i from set I had been found to match individual j from set J in same process, be it textual or psychological. However, we also suggested that our technique could also be used when the n(i,j) were but zeroes and one's corresponding to No and Yes answers. And actually we did consider successfully as correspondence matrices the tables of lines of which corresponded to individuals in some set I and the columns to continuous non-negative variables [n(i,j) being, e.g. the length of some part j of a plant i].

Our first three examples will be concerned with confusion matrices; in which case set I and J happens to coincide: n(i,j) being the number of times item i (stimulus) has been mistaken for item j (response). Example 2 is actually concerned with pair-wise comparison: how many times is i (first presented) judged similar to j (second). Then there will come two cases of logical (Yes–No) description matrices and finally continuous variables.

We can only mention here that in ecology our analysis of correspondence has proven to compete successfully with classification algorithms.

Example 1: A classical experiment in learning (Data by le Ny) – six sounds of increasing frequencies have to be associated with letters. On the output (a tridimensional map) each sound appears twice, once as a stimulus and once as a response. It should be noticed that the answers are systematically lower than the stimuli; besides, all points appear on a parabola; the second factor is a quadratic function of the first (Guttman, effect, see above). [See Figure 2]

Example 2: Comparing morse code signal pairs (Data by Rothkopf) – A 36 x 36 matrix is analyzed; signals as i and j almost perfectly coincide (because the comparison matrix is close to being symmetrical). The first factor correlates to .95 with the total duration of the signal (measured in Bodes); the second factor is a perfect parabolic function of the first (Guttman effect); the third factor is interpretable as the opposition of dots versus dashes.

The total duration and dot-dashes being relevant have already been found by Shepard (No. 7), but his technique gave a two dimensional chart without separating factors, and nothing like a .95 correlation with a physical quantity (the duration) *not* involved in the data (integers: how many times has i been judged equal to j) could have emerged. [See Figure 3]

Example 3: Inventing a code (experiment by J.F. Richard and J.P. Benzecri) – It is impossible to explain in a few words the experiment. Let us say that subjects were set by pairs and had to communicate to each other the names of European countries using only as signals a point marked on a 12 centimeters long straight-line; the psychologist said them if the communication had been successful. From this experiment we got:

(1) a confusion matrix–How many times had this country been answered for that;

(2) an average code–By which point on the line did the subjects code Spain, Italy, etc. (in most cases the subjects simply enumerated the countries from left to right, i.e., from Portugal to Russia; but very few errors affected Russia).

The output was brilliant. The first factor (labeled f_1 with index R for responses and D for questions) separated Russia from all other countries (since few confusions have arisen, this suggested proving the partition theorem as stated above) The second factor (f_2 on the diagram) had .94 correlation to the average code (labeled 1) (of all countries excluding Russia). Again a physically measurable quantity not included in the data matrix was recovered, and this times as a second factor (of course, with *no* rotation!). [See Figure 4,5,6]

Example 4: A descriptive matrix (Chinese pottery data by Eliseev)– This author suspected the existence of a parallelogrammatic structure; he applied with some success to Guttman's technique. Factor analysis revealed that the second factorwas not a quadratic function of the first: Characters (having an ear with flowers like that and so on) with non-zero second factor were excluded; a perfect scalogram appeared with no pottery rejected.

Example 5: Are you afraid of . . . (see below) – Thirty-four questions were asked to 150 children. By duplicating each question into two columns (Yes and No) a 68 x 150 matrix resulted. The first two factors extracted are readily interpretable. The psychologist, who thought only galvanic responses were meaningful and verbal (Yes–No) behaviour was just noise, was much astonished at what we recovered. The first factor accounts for the no-yes opposition: No, I don't fear any damage; Yes, I do fear. It is very remarkable that both yes and no answers were coded exactly the same. The analysis

could distinguish them through their mutual correlations. The second factor concerns the agent: a natural little known and actually not much damageable agent as a mouse, or even thunder versus family problems (Data from Zlotowicz). [See Figure 7]

QUESTIONS

1. Are you afraid your father will fall sick?	1. Est-ce que tu as peur que ton pere tombe malade?
2. Are you worried when your father is late to come back home?	2. Est-ce que tu es inquiet quand ton pere est en retard pour rentrer a la maison?
3. Are you worried when your mother is late to come back home?	3. Est-ce que tu es inquiet quand ta mere est en retard pour rentrer a la maison?
4. Are you afraid your mother will fall sick?	4. Est-ce que tu as peur que ta mere tombe malade?
5. Are you afraid of what father will tell you?	5. Est-ce que tu as peur de ce que ton pere te dira?
6. Are you afraid of what your mother will tell you?	6. Est-ce que tu as peur de ce que ta mere te dira?
7. Are you frightened of rats?	7. Est-ce que tu as peur des rats?
8. Are you afraid of spiders?	8. Est-ce que tu as peur des araignees?
9. Are you frightened of mice?	9. Est-ce que tu as peur des souris?
10. Are you afraid of snakes?	10. Est-ce que tu as peur des serpents?
11. Are you frightened to be bitten by a dog?	11. Est-ce que tu as peur d'etre mordu par un chien?
12. Are you afraid of the wolf?	12. Est-ce que tu as peur du loup?
13. Are you afraid of thieves?	13. Est-ce que tu as peur des voleurs?
14. Are you frightened when you have to come into a dark room?	14. Est-ce que tu es effraye quand tu dois entrer dans une piece ou il fait noir?
15. Are you worried when you have to go to sleep by night?	15. Est-ce que tu es inquiet quand tu dois aller te coucher le soir?
16. Are you worried when you are alone at home and somebody is knocking?	16. Quand tu es tout seul a la maison et qu'on frappe à la porte, est-ce que tu es inquiet?
17. Are you afraid when you are alone in a room and you are hearing some noise?	17. Quand tu es tout seul dans une piece et que tu entends du bruit est-ce que ca te fait peur?
18. Are you afraid when you are alone at home at night?	18. Est-ce que tu as peur quand tu es tout seul à la maison la nuit?
19. Are you afraid of coming home alone by night?	19. Est-ce que tu as peur de rentrer tout seul chez toi la nuit?
20. Are you afraid of what may happen?	20. Est-ce que tu as peur de ce qui va arriver?
21. Are you worried in situation which do not matter?	21. Est-ce que tu t'inquiètes pour des choses qui n'en valent pas la peine?
22. Are you worried about what people think of you?	22. Est-ce que tu es inquiet de ce que les gens pensent de toi?
23. Are you frightened of storm?	23. Est-ce que tu as peur de l'orage?
24. Are you afraid of falling sick?	24. Est-ce que tu as peur de tomber malade?

25. Are you frightened of seeing a dead animal?	25. Est-ce que ca t'effraye de voir un animal mort?
26. Do you feel bad when you touch a sharp edge?	26. Est-ce que ça te fait un drôle d'effet de toucher quelque chose de très coupant?
27. Are you worried about your work at school?	27. Est-ce que tu t'inquiètes pour ton travail a l'école?
28. Are you frightened when you are at a high place and you are looking down?	28. Est-ce que tu es effrayé quand tu es dans un endroit élevé et que tu regardes par terre?
29. If you have to climb a ladder, will you feel afraid of falling?	29. Si tu devais grimper a une échelle est-ce que tu aurais peur d'en tomber?
30. Don't you like to fight because you are afraid of being hurt?	30. Est-ce que tu n'aimes pas te battre parce que tu as peur d'être blesse?
31. Are you afraid of being hurt in an accident?	31. Est-ce que tu as peur d'être blessé dans un accident?
32. Do you have sweaty hands?	32. Est-ce que tes mains transpirent?
33. Have you often a lump in your throat?	33. Est-ce que tu as souvent la gorge serrée?
34. Are you often feeling your heart beating like mad?	34. Est-ce que tu sens souvent ton coeur battre très fort?

Example 6: Testing 4,186 children (finishing what roughly correspond in France to High School and averaging 16 years old). (Data from Benedetto and Reuchlin; analysis by Lacourly)

Let C be a set of 6 knowledges:

C_1 = grammatical analysis	C_4 = French literature
C_2 = Syntax I (phrases to be completed)	C_5 = Arithmetic and Algebra
C_3 = Syntax II (time concordances)	C_6 = Geometry

A 6 x 6 correspondence matrix k(C,C′) was built, the analysis of which provides the same factors as that of the full 6 x 4186 matrix, namely:

$$k(C,C') = \Sigma_e \, [C(e)\, C'(e) \,/\, \Sigma_{e''} \, C''(e)]$$

where C(e) is the note obtained by the pupil e ∈ E (E is a set of 4186 children) to the knowledge C. The analysis of this matrix showed that the two first factorial axes gave 82% of the total inertia; a quite agreeable and interpretable organization of the whole set of test was obtained. [See Figure 8]

Example 7: Factor analysis of medical data (Data from Hotel-Dieu Rennes; Prof. Bourel, Prof. Lenoir; analysis by M. Kerbaol) – The rates of six substances have been measured in the blood of 613 patients. In order to give to all substances approximately the same importance, each column was divided by its average and its deviation. The data matrix was then analyzed according to our standard program. Two significant factors were obtained

and on the map (Cf. Figure 7) hemochromatosis of first and second kind appeared to be well-grouped which encouraged us to resume an analysis with only 180 patients including all concerned with hepatic pathology (namely, ethilism, cirrhosis, hemochromatosis, 1° and 2°). Up to slight changes the same factors were extracted, a stability which increased the doctor's confidence in the method. The separation between various classes were satisfactory (Cf. Figure 8) and some anomalous positionings could be explained either by errors in punching data (!) or by the curation of the patient, who traveled across the diagram (isolated black squares on the left). An interesting feature is that when a new set of data comes in, it can be inserted readily in the map by computing its factors through linear combination of the various rates; in various terms, the doctors took advantage of this facility.

Data for new investigations are now being collected and coded at Hotel-Dieu de Rennes. [See Figures 9 and 10]

5. Factor Analysis of a Random Process

Classical statistics have been concerned with such questions: Is the process Markovian; and if not, to what depth should we go back to get all information relevant to the prediction of the future? Or, can the hypothesis be accepted that the process is Markovian and belongs to a certain parametric class; and if this is true, how to estimate properly the parameters, etc.?

Much has been done to define the spectrum of a fluctuating random variable, such as a pressure or a magnetic field. Our problem here is quite different: Given a finite a priori structureless set I and a long string of items i(1), i(2), . . . , i(n), all belonging to set I, define a structure on set I from the statistical analysis of string {i(n)}. To achieve this, we use, of course, the frequencies of the various pairs of elements in I; that is, we compute how many times element i has been found coming put after element i′; it is possible to use also the number of times element i is found coming after element i′ with just one element in the between. Very beautiful things can be dreamt of, I am sure, if one consents to use triplets; but since a set with a decent number of elements as 50 may generate an undecently memory consuming three dimensional table, it is wise to be careful about triplets – my fellow mathematicians asking for n-uplets anyway.

So we have one program (a variant of the correspondence analysis program) working – honestly working I mean, with no rule of thumb such as rotation whatsoever, and we had up to now the opportunity to have it run on three sets of data.

(a) A monkey on the typewriter: As a typewriter, we actually used the keyboard of an IBM 026, which was equipped with an aluminum screen which let visible only one line of ten characters. As a "monkey" we took one of our youngest programmers; we instructed him to type at random with only one finger. The result was a hundred of cards representing a sequence of about 10 thousand elements which we call digits. The sequence was analyzed by using the standard program; and the first factor extracted gave, up to two inversions, the correct order of the digits on the keyboard.

This can actually be explained by the fact that when a boy is instructed to type at random, he will alternately stay around the same spot on the keyboard or jump from one end to the other. In both cases, the probability of the pairs of digits following each other will depend continuously on the position of the keys. [See Figure 11]

(b) A phonemically coded Spanish text: Since the pronunciation of the Spanish language is rigorously determined from the spelling and the set of phonemes is among the shortest found among natural languages, it is possible to write a simple program to transform a Spanish-written text into a sequence of phonemes. Such a sequence was analyzed from a corpus of 4,200 letters using the standard program, and the first factor was found to separate quite clearly the vowels (one side of the null point) from the consonants (clustered on the other side, with the semi-vowel jod identified to i). With the second factor, it was possible to get a moderately distorted vowel-triangle; but this point deserves further study. It is presumable that much better results will be achieved, taking advantage of the already obtained vowel-consonant opposition. The main interest of such an elaborate study would be to get evidence from the frequential data of the evolution of a sound system and also to compare different languages. Analogous results have been obtained for Italian [See Figures 12 and 13].

(c) A phonemically coded Portuguese text: One may wonder why Portuguese was chosen as a second language to study, since its sound system is almost as intricate as that of French or English, but as we started from Spanish . . .

The separation between vowels and consonants was again successfully achieved (with a corpus of 8,000 letters). But nothing has emerged up to now about the vowel triangle which obviously requires more elaborate tools, especially in Portuguese where nasal vowels should have to be recognized.

Let us notice that since a vowel is more frequently followed by a consonant than by another vowel, sequential frequencies should suffice to distinguish between vowels and consonants; and an ad hoc classification program divised by SUKHOTIN (in *Problemy Strukturnoi Linguistiky,* Moskva 1962) could actually achieve this distinction. But here a standard program was used which gives besides a continuous coordinate.

6. Analysis of Preference Ranking

From psychological, sociological, and other investigations, there frequently emerge such data as the following: all subjects from a set S rank from first to last (or possibly they rank only a few first) the items of a set I. Such preferences rankings may be coded into points representing each subject choice as exemplified on Figure 6a in the two dimensional case. Now the set of subjects and also the items become a cloud of points in a euclidean space; and therefore, it is only natural to achieve a multidimensional scaling. What comes out of such a scaling is first a center of masses = the average preference ranking, which very often is not significantly different

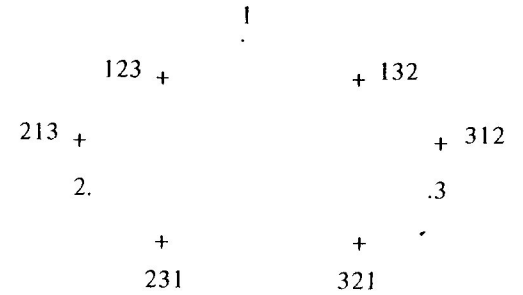


Figure 6a: Geometrical coding of preference rankings; dots represent items and plus signs represent possible rankings (and therefore subjects).

from indifference, or is absurd (whatever may be your feelings in favour of collegial decisions, from a mixture of divergently motivated choices quite absurd an average is liable to emerge); but sometimes it indicates the general acceptation or rejection of this or that item. And second, which in all cases we had the opportunity to study turned out to be genuinely significant, dispersion factors expressed under what respect the various choices were opposed to each other. Here are a few examples:

(a) Ranking six pictures: Two hundred and nine subjects were asked to rank six paintings reproduced as postcards and all concerned with flowers. The pictures were respectively:

(1) A common place "Chromo" picture printed in Italy;
(2) by a Japanese artist;
(3) Sun flowers by Van Gogh;
(4) by Dufy;
(5) by Douanier Rousseau (a famous "naive");
(6) by Paul Klee (a thoroughly abstract piece where it was impossible to diagnose flowers, except from the title).

The only information given by the center of masses was that almost all subjects rejected (6).

Four factors were extracted of which the first two accounted for more than 75% of the total inertia and could *without* rotation be readily interpreted as follows:

—The first factor accounts for the sensibility of subjects to the figurative qualities of the pictures.

–The second factor accounts for the opposition between Rousseau at one end, and Dufy, Van Gogh at the other (Klee is out of the question because almost everybody refuses it; and 1, 2 are not concerned with this factor). Actually although naive painting is considered by critics as distinct from figurative and close to abstract, it may be accepted by people naive enough themselves to welcome the naive style as a variant of the figurative one. As a matter of fact, Rousseau considered himself a figurative painter!

–Factor 3 is concerned with the splitting of Western from Eastern figurative art.

–Factor 4 is with the opposition between Dufy and Van Gogh. It is indeed remarkable that those two pairs (1,2) and (3,4) of pictures were most generally kept close by generating an axis, just as the latent conflict of the naive (5) and the abstract (3,4).

The correlation of factor one with the sociocultural background was as could be expected. The sociologist who had previously used with little success the methods available previously to ours (e.g., Coombs unfolding techniques) was very satisfied with our treatment. But she saw with disappointment that Rousseau so strongly interfered with the abstract–realistic opposition which she intended to study. [See Figure 14]

(b) Multiple-choice Tests: A few years ago, multiple choice test was introduced in France as a standard procedure to grade students in medical schools and Faculties of Medecine. Although most teachers were reluctant, the system was widely accepted to cope with the increasing number of students. As you may know, various types of questions may be used and the teachers were asked to state their preference order among eight possible types.

A – Compléments simples,
B – Compléments groupés,
C – Variation relative,
D – Association simple,
E – Association composee,
F – Relation de cause a effet,
G – Comparaison quantitative,
H – Schéma, planche,

I state those types in French, since the English precise equivalents, which are unknown to me, may certainly be guessed better from the original than from a poor translation!

As far as the average order is concerned, all that can be said is that A is universally accepted while G, H are universally rejected. The remainder is, statistically speaking, non-significant and happens to be poorly interpretable.

Factor one *(extracted without rotation)* accounts satisfactorily for the opposition between questions relying on memory, and question relying on judgment. Factor two (again *without* rotation) accounts for the opposition between simple and compound (or group).

Information brought by factor three and four was found to be non-negligible but it will not be explained here.

The analysis was also resumed, including only the teachers of morphological disciplines anatomy, etc., and those of physiology, chemistry and physics. It turned out that (.86 x second factor + .14 x first factor) .79 correlation with the partition function = 1 to morphologists (who use preferably type B), –1 to physicists, etc. (who use preferably type F). This discrimination or recognition was achieved by processing only the behavioral information (the rankings) and not the fields of the subjects (teachers), which were only introduced afterwards on the map generated by the first two factors (of which the second could almost have been used as an optimal discriminent function). [See Figures 15 and 16].

(c) Why do you practice sport? The subjects (about 1,500 sportsmen) had to state and rank the first six of a list of 24 motivations. The analysis, again completed using the standard program for the analysis of preferences, gave as first two factors (non-rotated) two dimensions considered classically as most relevant = . . . ; . . .

(1) Nature and beauty versus *Combativity*
(2) Friendship and play versus *Self-affirmation*

[See List of Motivations]

LIST OF MOTIVATIONS

1.	Attraction for a natural element	Attrait pour un élement de la nature
2.	Instinct of competition	Instinct combatif
3.	Self-affirmation	Affirmation de soi
4.	Cultivate one's cleverness	Cultiver son habileté
5.	Taste for beauty	Goût du beau
6.	Social accomplishment	Réussite sociale
7.	Attraction for danger	Attract du danger
8.	Fame	Briller–gloire
9.	Live for game	Amour du jeu
10.	Compensation	Compensation
11.	To build up one's character	Se forger le caractère
12.	To bustle about	Se remuer
13.	To fill up one's spare time	Remplir les loisirs
14.	Money	Gain lucratif
15.	To make friends	Avoir des camarades
16.	To spare no troubles	Se dépenser à fond
17.	To cultivate one's strength	Cultiver sa robustesse
18.	Evasion	Evasion
19.	Health	Hygiène–Santé
20.	Spirit of competition	Esprit de compétition
21.	Relax	Détente–Distraction
22.	Sense of corporeal dignity	Sentiment de dignité corporelle
23.	To be gifted	Possession de dons
24.	Adventure	Aventure

7. On the Factor Analysis of Proximities

R. N. Shepard (1,2) has published, in a few recent papers, an interesting discovery about the following geometrical problems.

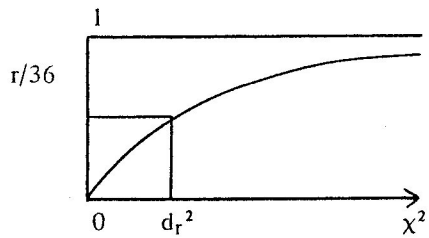
Given a set I of elements (of any nature) for which not distances but only inequalities between distances are known (i.e., for any four i_1, i_2, i_3, i_4 in I, we know whether i_1 is closer to i_2 than i_3 to i_4 , or the contrary). We construct a model J for I in an Euclidean space having as little dimensions as possible, and such that the given inequalities be realized (i.e., if i_1 is said to be closer to i_2 than i_3 to i_4, then $|j_1 j_2| < |j_3 j_4|$).

It turns out, Shepard noticed, that approximately the solution to this problem is unique; especially if one started from the inequalities between the interpoint-distances of a set I in the plane or the ordinary space, this set can be recovered (and first its dimension discovered) with good approximation. Shepard writes: " . . . an underlying Euclidean configuration . . . can be metrically recovered by an analysis based essentially upon the rank order . . . This conversion from a non-metric to a metric representation seems to be tantamount to getting something from nothing."

Shepard's construction involves rather complicated iterative computations since it requires, he says, a few minutes on a 7090 computer. Relying on a mathematical theory devised for the purpose (3), we are now able to go much faster and even by hand (in 15 minutes!) especially in the cases solved by Shepard in his paper.

It turns out that the squared interpoint-distances between the points of a finite subset taken at random in R^p are a fairly good sample of a p-degrees of freedom χ square distribution. Therefore, it is easy to recover approximate distances (up to a constant multiplier) by just using the proximity–inequalities and a χ-square table.

Given the number of points (e.g. 9), we have the number of interpoint distances (36). We assign to each a *precise* value d_r (taking into account only its rank r) by reading the squared distance on the χ^2–axis against the coordinate r/36 on the cumulative probabilities axis.



It is very easy, in case p = 2, to put the points in a diagram just using those approximate distance d_r and working from point to point with a compass. We solved this problem without having to resort to any iterative technique.

However, a program for multidimensional scaling can be used to extract a p-dimensional picture from the approximate distances computed according to the χ^2 rule. And this is actually the only possible solution when p is greater than (or even equal to) three.

Here is an example taken from Shepard. On a table there are found the original distances between 9 points in the plane. Then the same is multiplied by a constant so that they have an average of 100. Then the d_r is computed from χ^2 table and then the same is averaged to 100. At last, the distances on the diagram are reconstructed with a compass.

Actually, just try to do it yourself: put 9 points "at random" on the plane; call them 1 . . . 9; measure the interpoint distances 1–2, 1–3, . . . , 8–9; rank them first 3–6, second 7–9, third 8–2, etc.; take 3–6 = 146, 7–9 = 125, 8–2 = 115, etc. (middle column on the table); and then work a diagram out of those distances using a compass. There is a high probability that you nearly come back to the same shape you started from (But be aware of one thing–there is one chance out of two, just as if you had had the precise distances, that the reconstructed diagram is a mirror-image of the original one.).

However, although, from the viewpoint of the geometer, to recover distances from mere inequalities is an achievement, the factor analysis of proximities as an applied technique is not so useful as could have been expected. The reasons for this are the following:

(1) The statistician is not a man trapped by a Sphynx who asks him geometrical puzzles, but he is a man faced with natural phenomena whose structure he has to decode;

(2) From what is known of natural phenomena, inequalities between proximities cannot be straightforwardly computed. Various formulae are available, as stated above, may sometimes lead to opposite results. There is no sphynx on the way to provide him at least with the appropriate inequalities;

(3) In non-artificial cases where factor analysis of proximities has been successfully used (e.g., Rothkoff's morse-code confusion matrix, analyzed by R.N. Shepard; or a set of similarities between wave lengths given in Coombs' book and successfully analyzed by our all-hand technique), the data as seen above have a beautiful correspondence matrix structure which it is wiser to keep than to convert into proximities.

8. Conclusion

I understand I should now bring my report to an end. Certainly I have omitted many important mathematical arguments or positive examples that could have helped you to appreciate the real scope of our work. On the contrary, I have more than once indulged in mixing dreams with facts because

with the little time allowed it is so much easier to suggest a pleasant potentiality than to describe what precisely has been achieved. But we all suspect that if our congresses have been called symposia (after a Greek word which means a banquet) since the days of Platon, it is because scientists reporting on their own work or that of their teachers, fellows or pupils are indeed to be somewhat intoxicated, in order that they may speak more freely.

On this confession I will stop, since my conclusion could but repeat what my introduction said, and you certainly all still have in mind.

9. Acknowledgement

The work of our Laboratory at the Institut de Statistique de l'Universite de Paris has been made possible by support received from the French University and by contract grants D.R.M.E. No. 66–34–557 and D.G.R.S.T. No. 66 FR 001.

BIBLIOGRAPHY

Section 2

J.P. Benzecri–Problemes et methodes de la taxinomie (1965).

J.P. Benzecri–Sur les algorithmes de classification (1965) (Both papers included in "Lecons sur l'analyse factorielle et la reconnaissance des formes; 1965–66, I.S.U.P., Paris).

S. Regner, M. Renaud, W.F. de la Vega–technical reports some of which include programs issued by Laboratorie de Calcul de la Maison des Sciences de l'Homme. (M.S.H. Paris,.rue Puiseux); work now continued by I.C. Lerman.

M. Roux–Sur l'application de quelques methodes de classification en phytosociologie, in Revue de statistique appliquee, 1967.

Section 3

J.P. Benzecri–Cours de linguistique mathematique; 3eme et 4eme lecon; Rennes 1963–64. Mimeographed notes available from I.S.U.P. Paris.

B. Cordier–L'analyse des correspondances; These Rennes 1965. This work includes program-listings and the proofs of various theorems; it is available from I.S.U.P. with successive paper by the same author.

Ch. Rousse–Lagordaire & W. Moussatat–(unpublished) proof of the theorem stated above regarding the normal case.

Section 5

B. Cordier (Mme Escofier)–Sur l'analyse des correspondances, These, Faculte des Sciences de Rennes (1965).

Programs and analysis by B. Cordier and Ph. Marano.

Section 6

On the method:

(1) J.P. Benzecri–Analyse des preferences (I.S.U.P. 1965).

(2) J.P. Fenelon–id–in Sciences (Hermann edit.), Paris, 1967.

On the data:

(3) Y. Bernard–C.N.R.S. (1966), Esthetique experimentale.

(4) M. Bouet–Thesis to appear, On the sociology of sport.

(5) J.J. Guilbert & Ch. Rousse Lacordaire–Opinions des enseignants sur les examens en faculte de Medecine (avec un appendice de J.P. Fénelon). Paris–Ministère de l'Education Nationale et Imprimerie Nationale (1967).

Programs by J.P. Fénelon; Analysis by J.P. Fénelon, Ch. Rousse Lacordaire & M.O. Lebeaux.

Section 7

(1) Shepard, Roger N. – *The Analysis of Proximities,* Scaling with an Unknown Distance Function I, published in Psychometrika, June, 1962, Vol. 27, No. 2.
(2) Shepard, Roger N. – *The Analysis of Proximities,* Scaling with an Unknown Distance Function II, published in Psychometrika, September, 1962, Vol. 27, No. 3.
(3) Benzecri, J.P.–*Analyse Factorielle des Proximités,* Publication de l'Institut de Statistique de l'Universite de Paris, 1964–65.

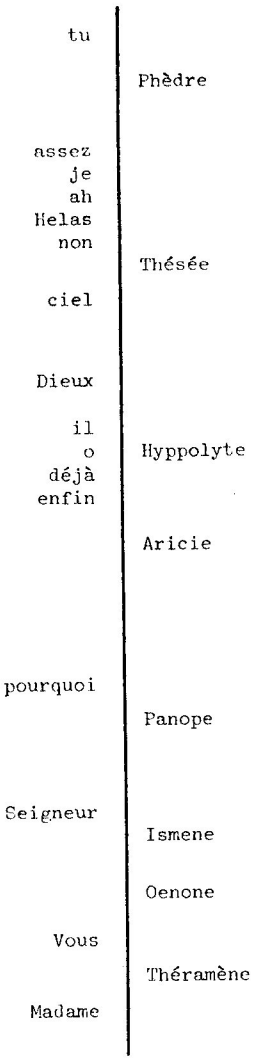


Figure 1

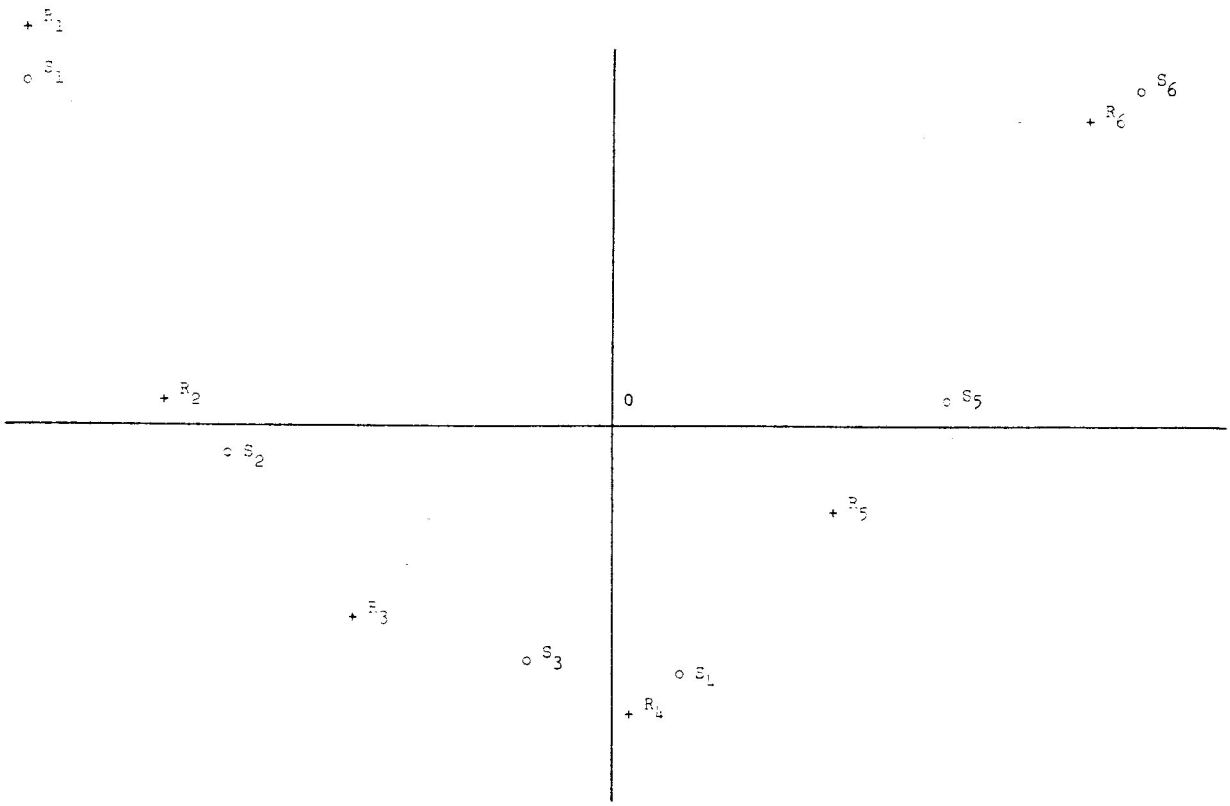


Figure 2: Experience d'apprentissage d'association de lettres de sons

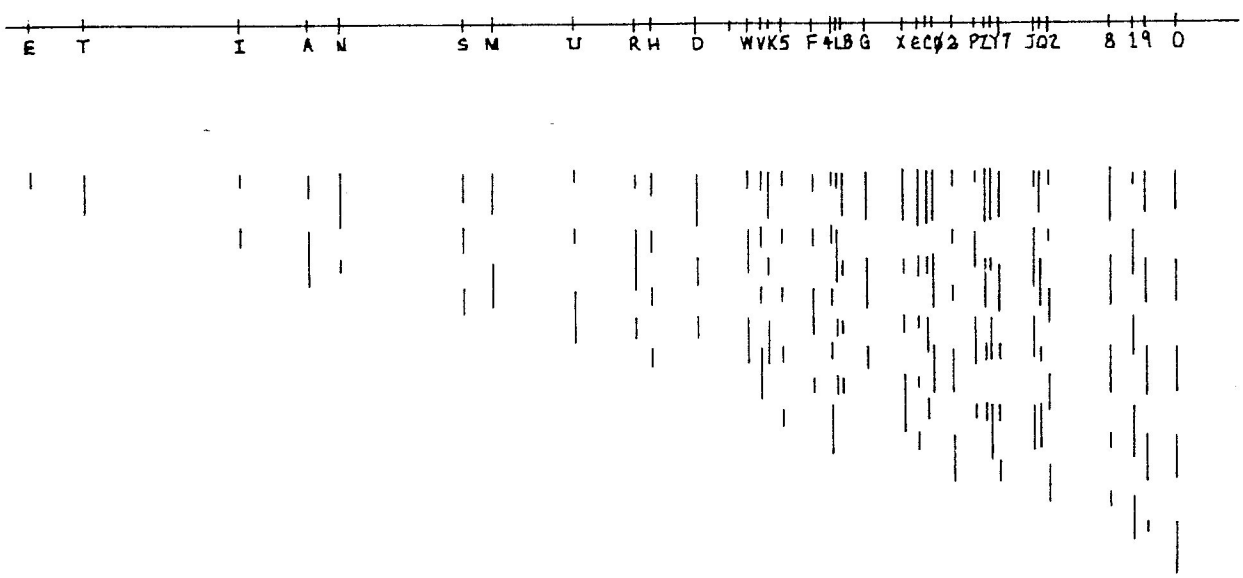


Figure 3

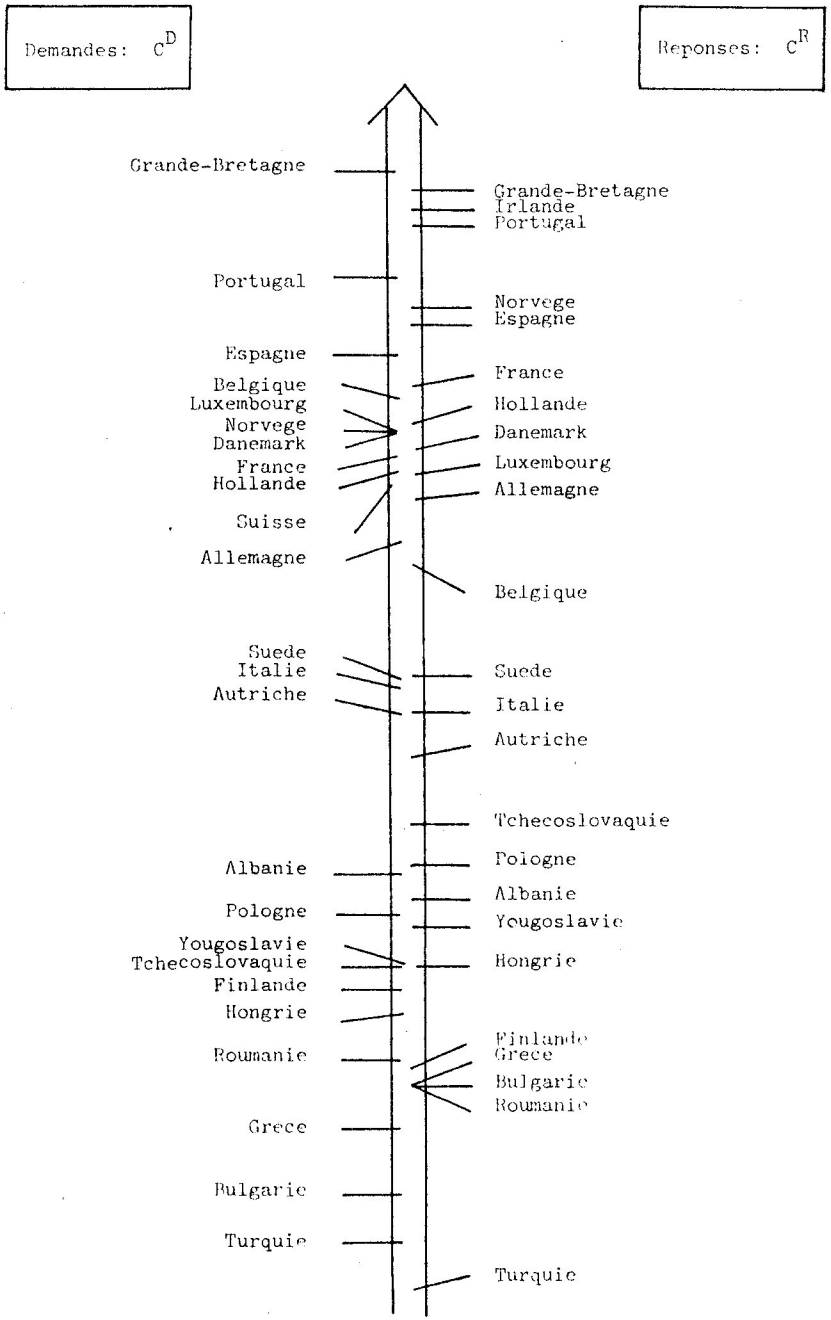


Figure 4: Le code moyen

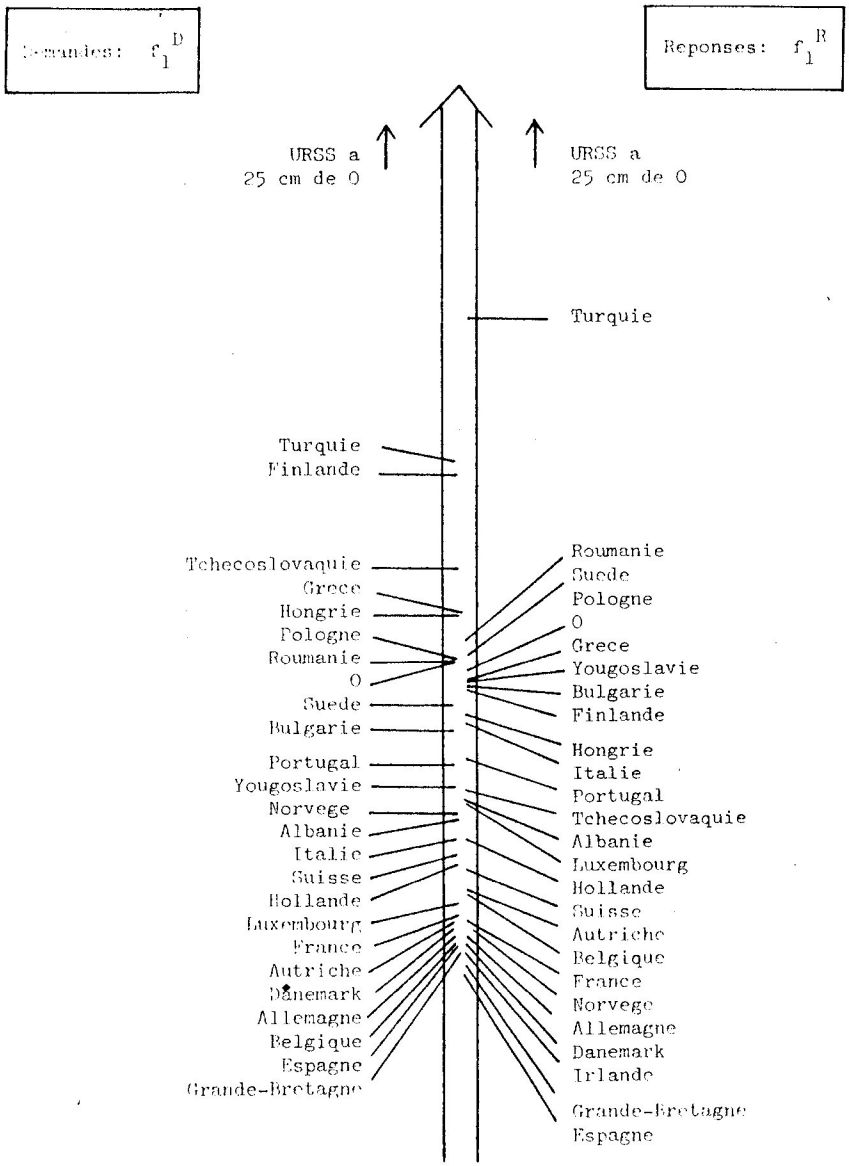


Figure 5

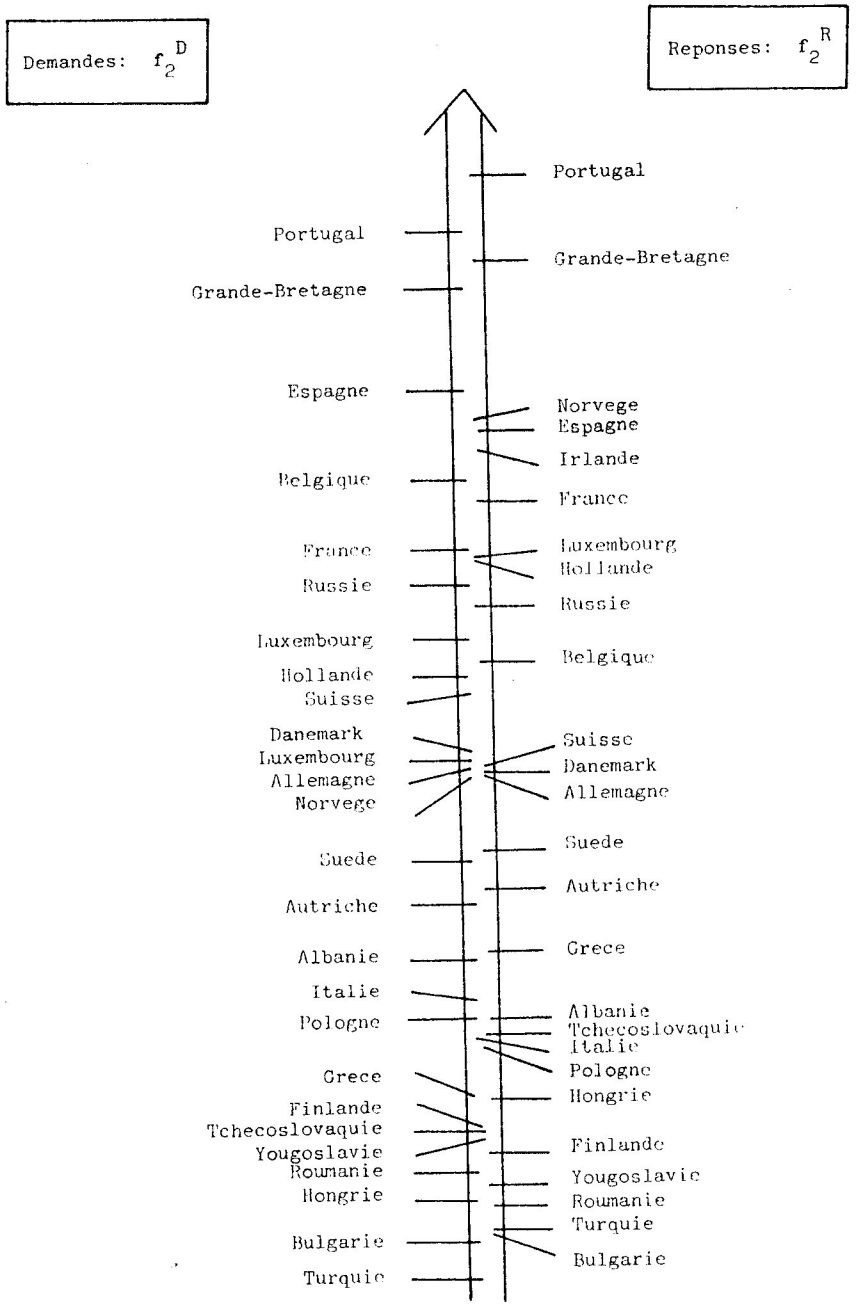


Figure 6

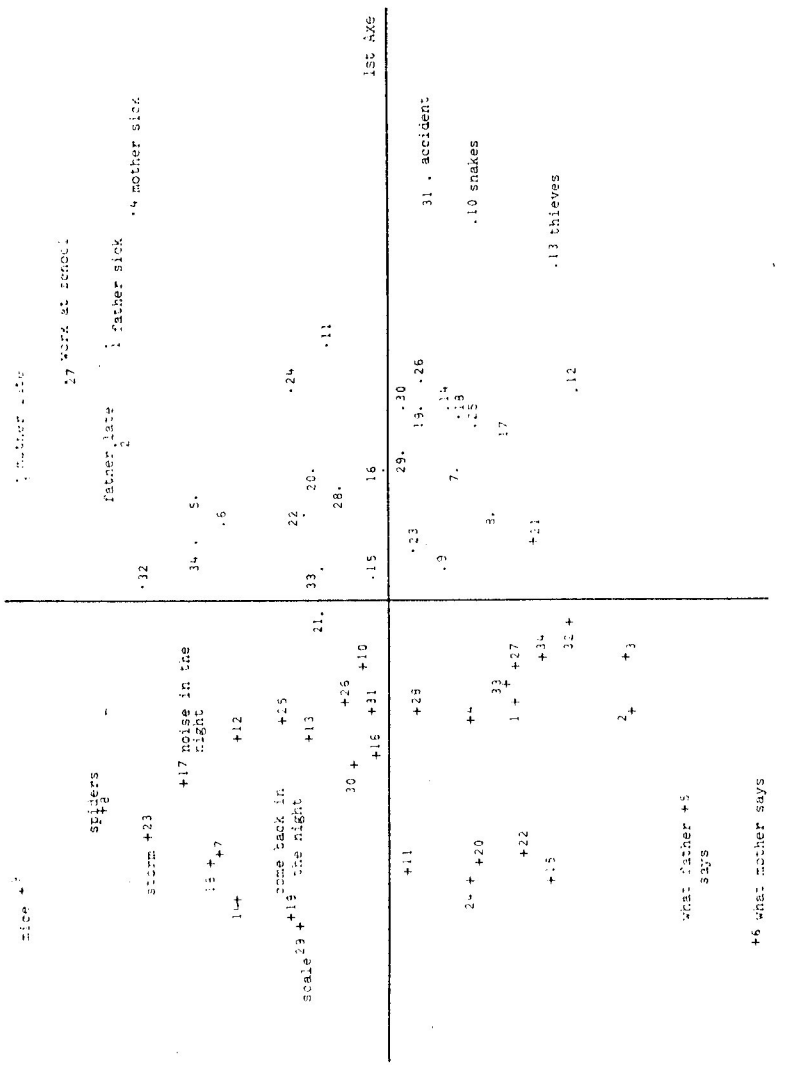


Figure 7

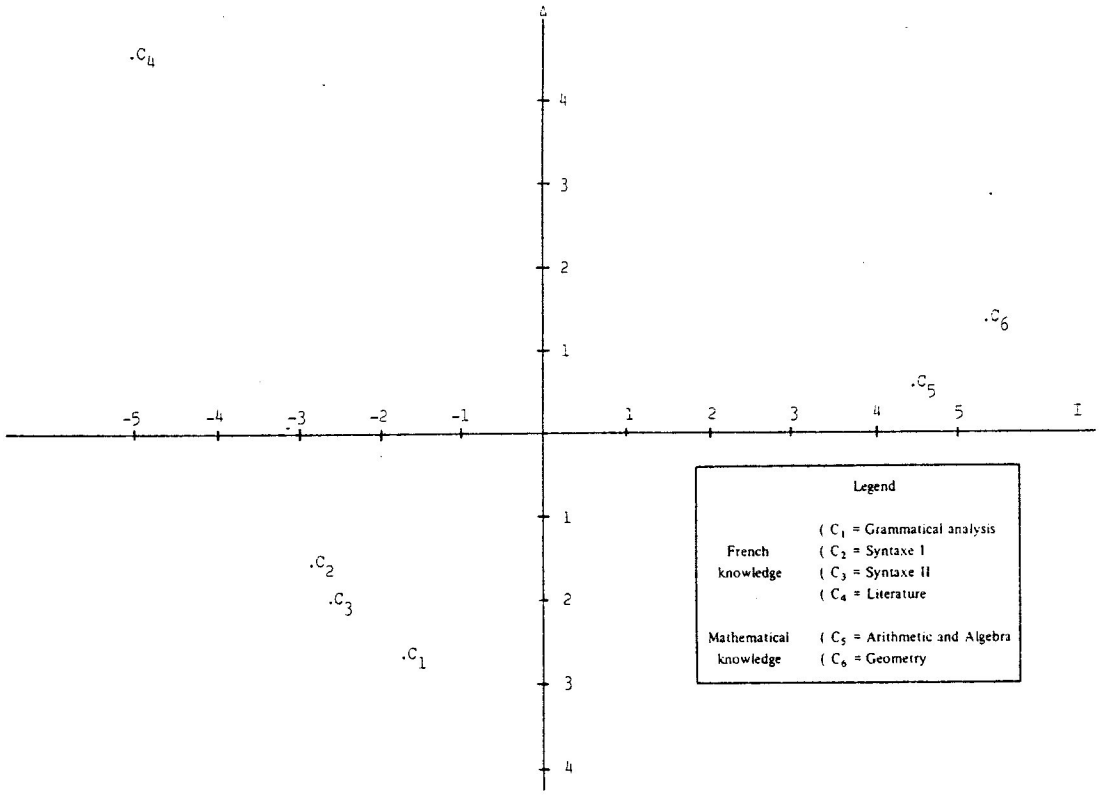


Figure 8

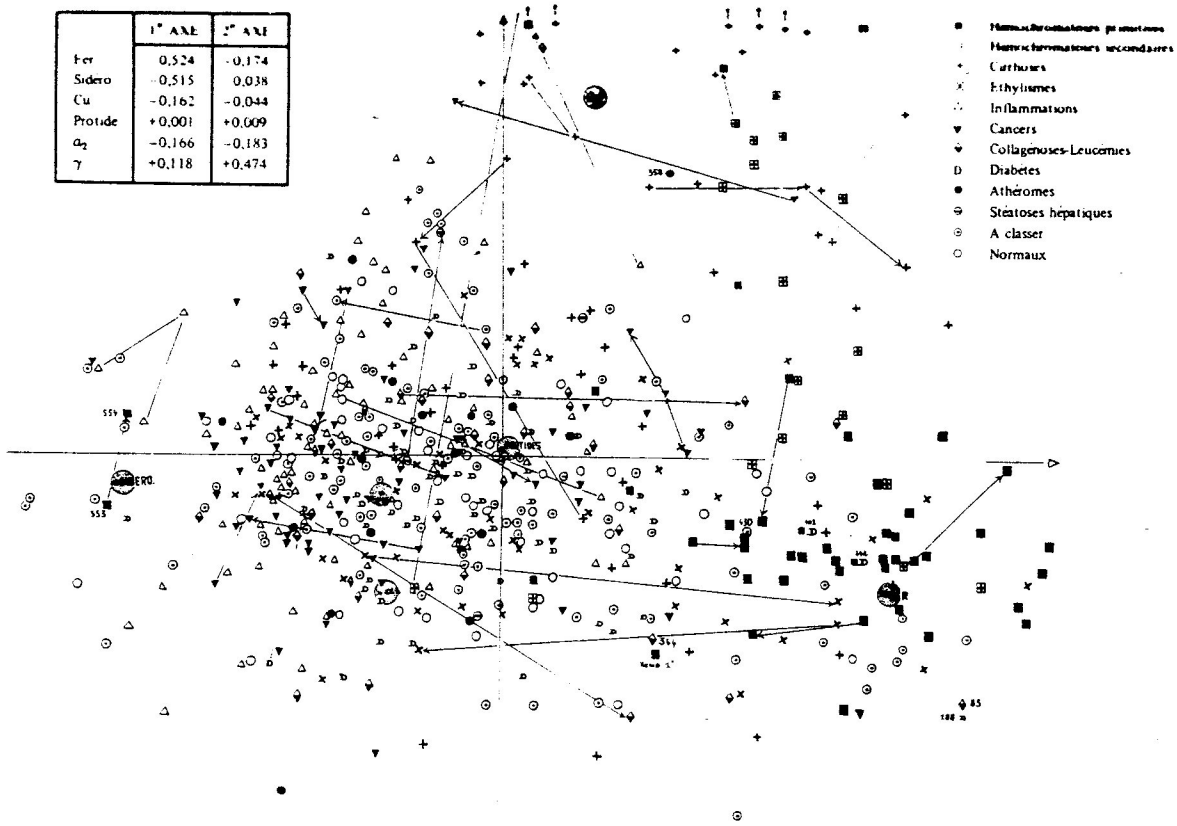
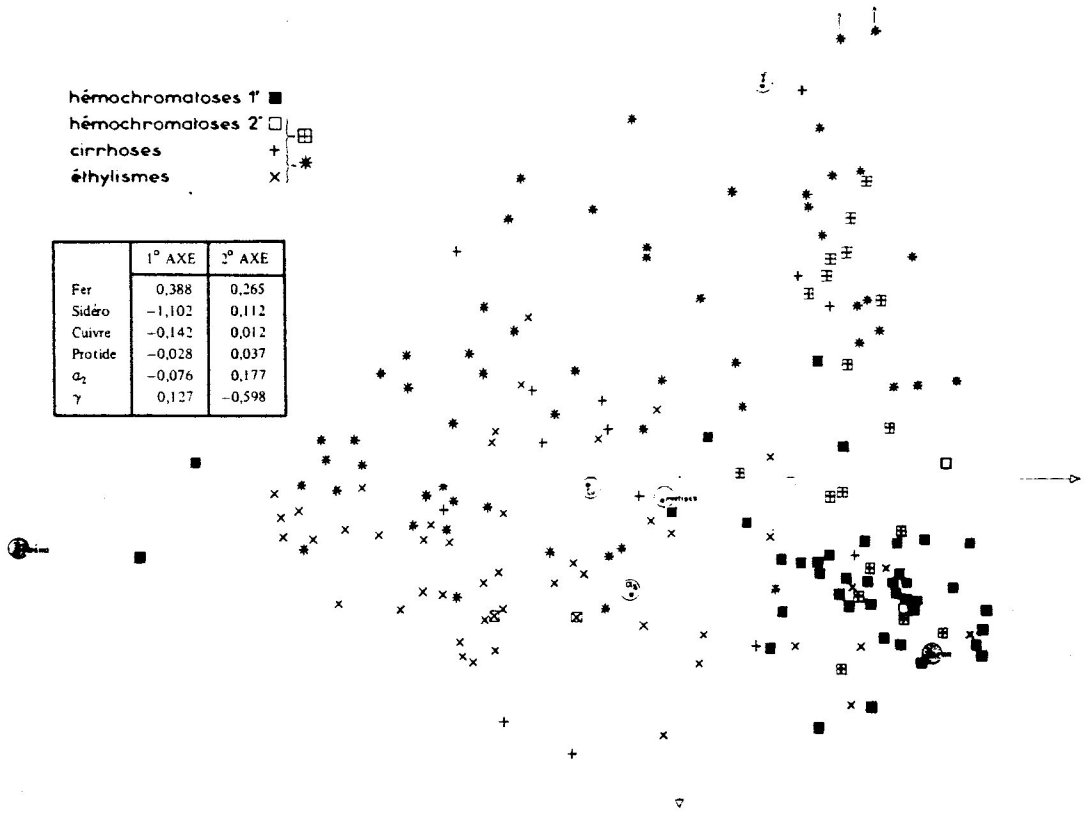


Figure 9: Analyse Factorielle Globale (613 cas)



	1° AXE	2° AXE
Fer	0,388	0,265
Sidéro	−1,102	0,112
Cuivre	−0,142	0,012
Protide	−0,028	0,037
α_2	−0,076	0,177
γ	0,127	−0,598

Figure 10: ANAEMO (180 cas)

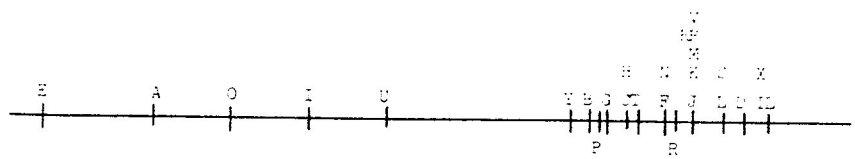


Figure 11

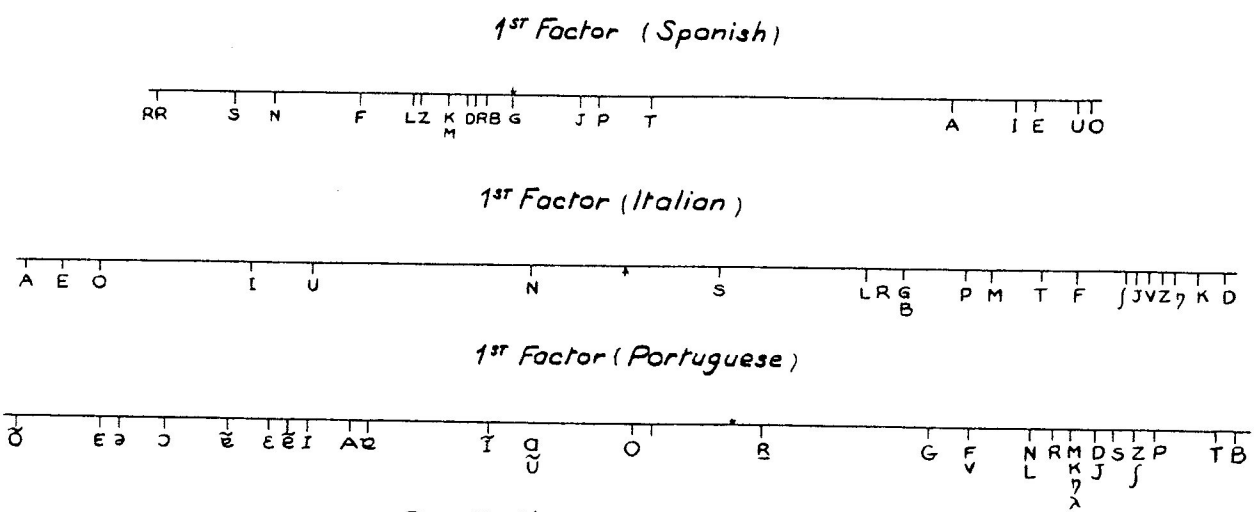


Figure 12: Séparating Vowels from Consonants

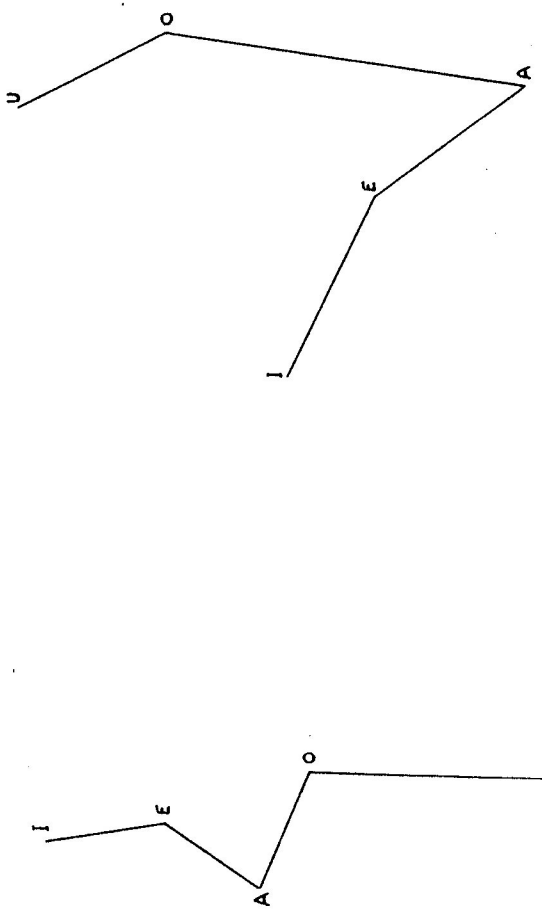


Figure 13: Recovering an Approximate Vowel Triangle

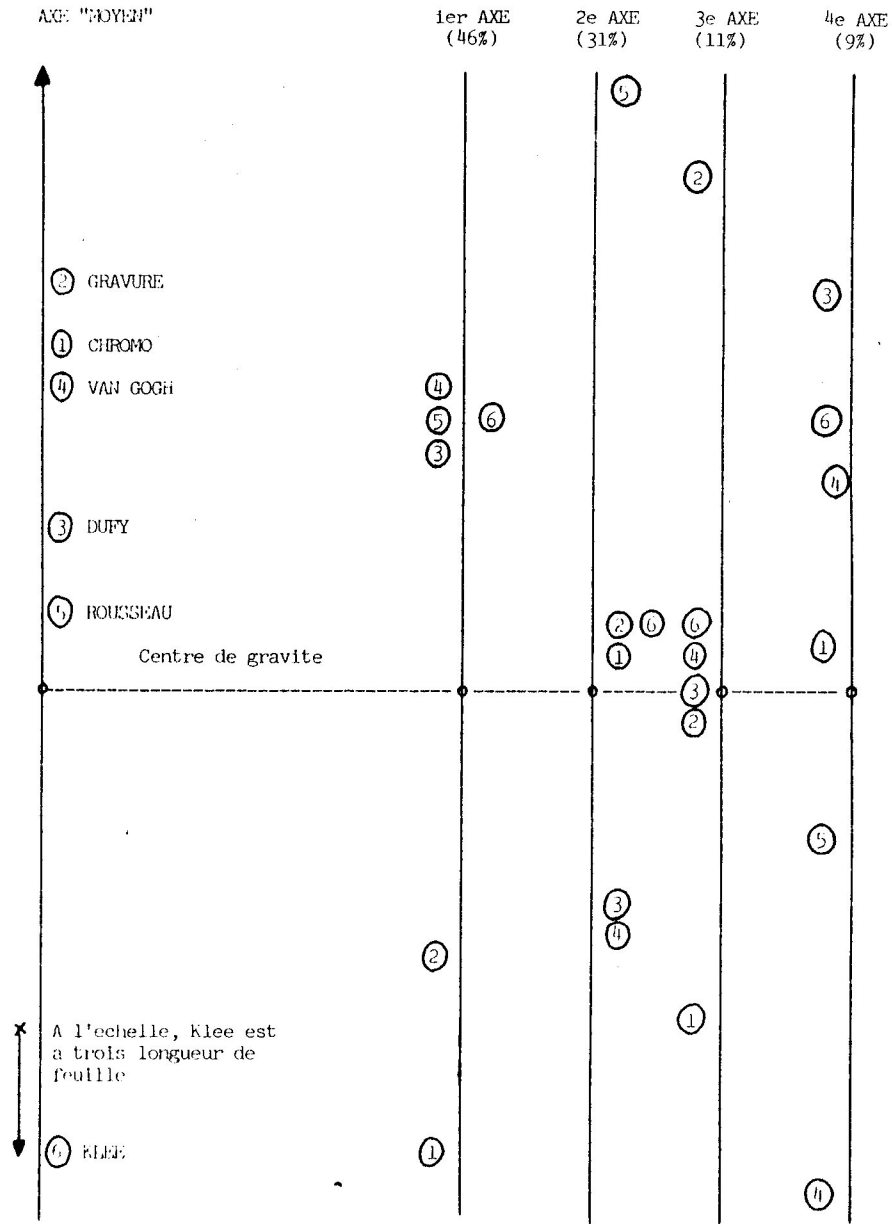


Figure 14: Les Axes

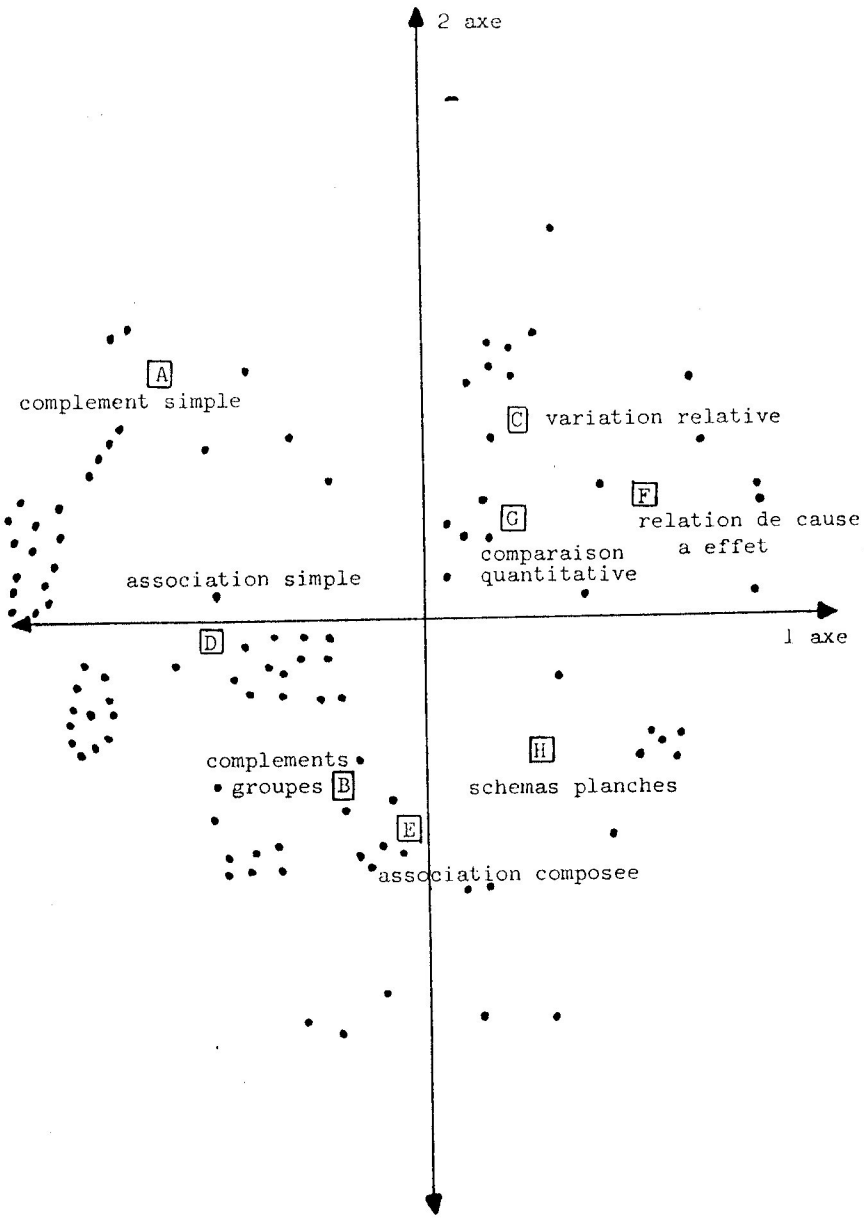


Figure 15

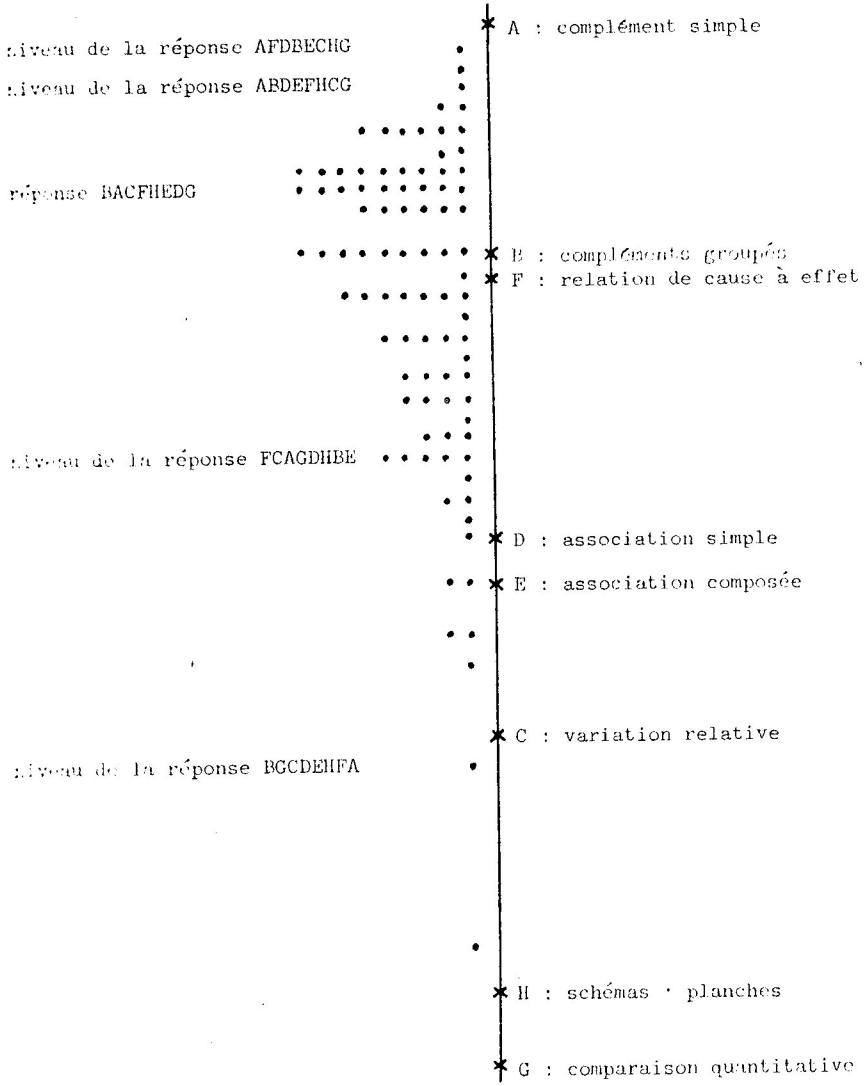


Figure 16

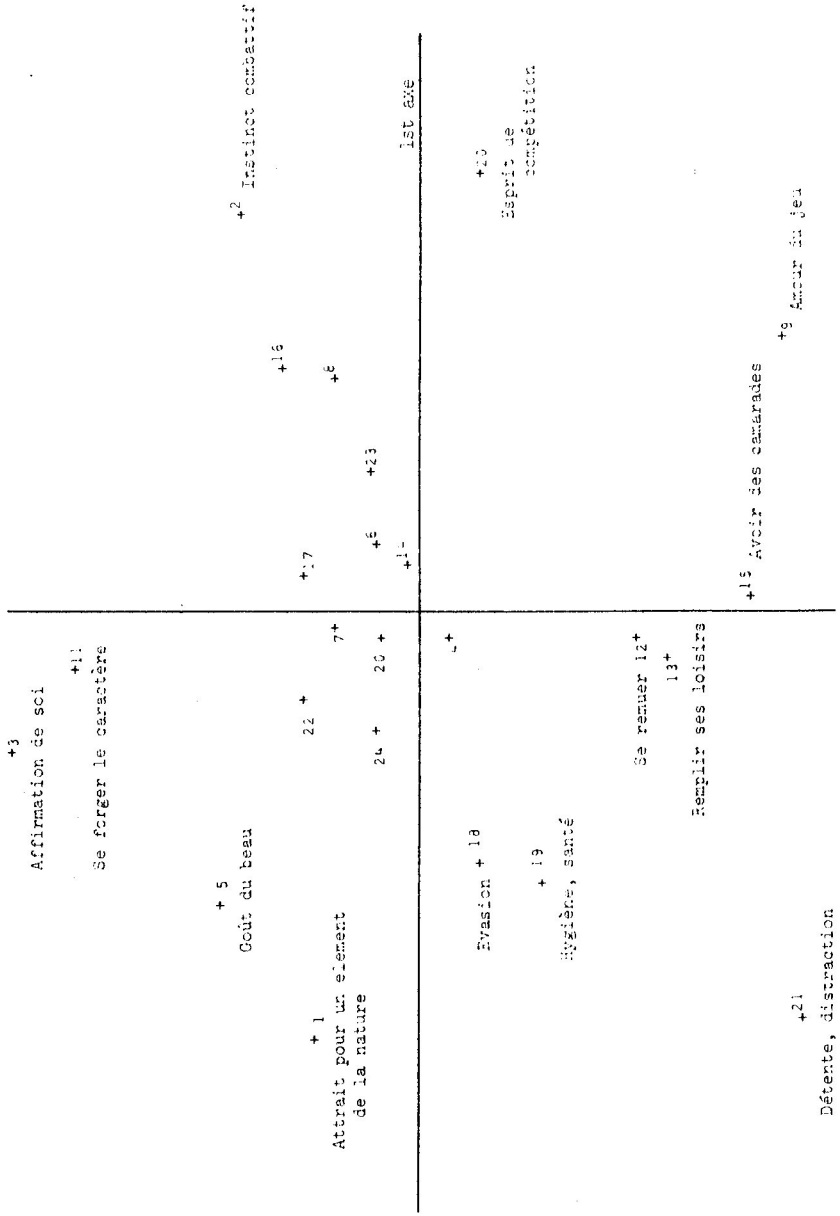


Figure 17

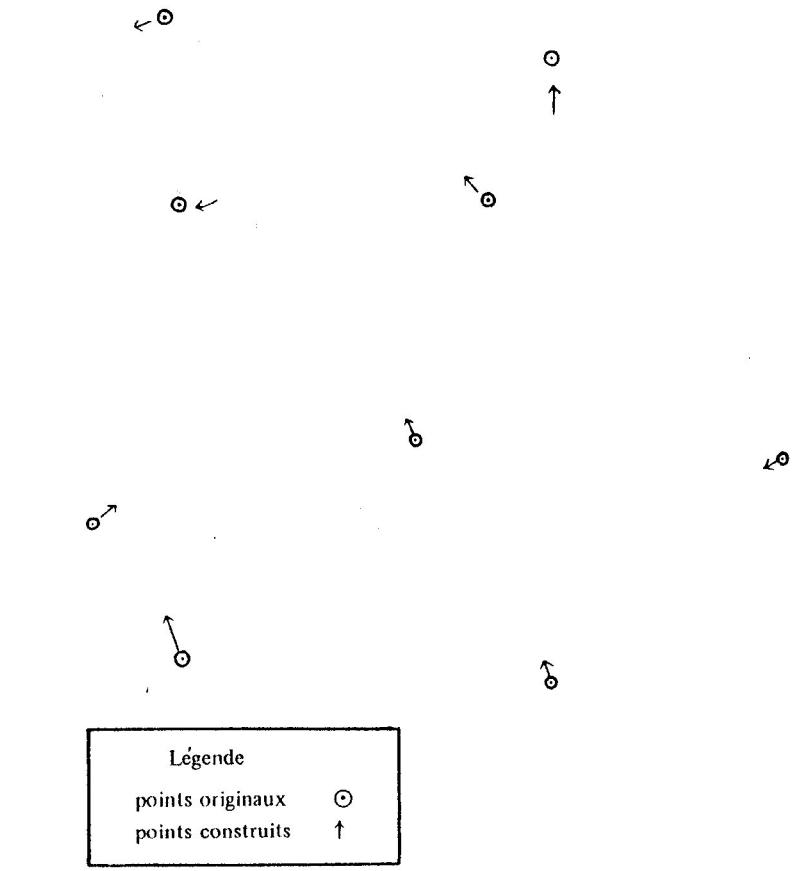


Figure 18

TABLE 1

An example from R.N. Shepard, solved by our method.

No	Nom	d - mesuree	d normalisee dx 100	x mesure	x normalise x=100	d experimentale
1	6-9	67	182	146	235	175
2	1-9	66	180	125	201	172
3	3-9	64	174	115	185	165
4	1-8	57	155	109	175	158
5	1-6	56	153	102	164	158
6	2-9	53,5	146	98	158	136
7	4-6	52	142	93	146	136
8	2-6	48,5	132	89	143	130
9	1-7	48	131	85	137	121
10	6-7	47,5	130	81	130	124
11	5-9	46	125	77	124	123
12	2-8	45	123	74	119	121
13	6-9	44	120	71	114	117
14	3-7	42,5	116	68	109	105
15	1-4	38,5	105	66	106	102
16	3-4	38,5	105	64	103	102
17	1-5	37	101	62	100	100
18	3-8	37	101	60	97	99
19	4-8	36	98	57	92	97
20	2-7	34	93	55	89	82
21	3-6	29,5	81	53	85	81
22	4-9	29	79	51	82	70
23	7-8	27,5	75	48	77	78
24	1-3	27	74	46	74	76
25	5-6	27	74	44	71	72
26	2-5	25	68	41	66	66
27	4-5	25,0	68	39	63	66
28	2-4	24,5	67	37	60	66
29	6-8	24,5	67	34	55	61
30	5-7	24	65	31	50	60
31	2-3	23	63	29	47	60
32	7-9	22,5	61	25	40	60
33	5-8	20,5	56	22	35	57
34	3-5	19	52	18	29	50
35	1-2	14	38	14	23	39
36	4-7	12	33	8	13	23