Quality and Quantity, 14 (1980) 431–469
Elsevier Scientific Publishing Company, Amsterdam – Printed in The Netherlands

COMBINATORIAL ANALYSIS IN THE STATISTICAL TREATMENT OF BEHAVIORAL DATA

I.C. LERMAN

Laboratoire de Statistique, I.R.I.S.A., Université de Rennes I, 35042 Rennes Cédex, France

I. Introduction

Most of the methods of Data Analysis in current use involve a geometrical representation. However, the descriptive variables, as they appear in the Humanities and Natural Sciences, are rarely numerical variables; it then becomes important that the method of synthesis take into account the natural representation of data. On the other hand, the most convenient structure of condensation for the problem submitted by the specialist has, more often than not, finite character: partition, sequence of partitions represented by a tree, order, . . . Therefore, the criterion of synthesis has to respect the mathematical nature of the searched structure; on the other hand, it must be clearly founded from a statistical point of view. Finally, the proposed algorithms must be justified by the criterion they have to optimize. These are the ideas that we have tried to develop in a systematic way in our work that is, on the whole, dominated by the classificatory approach in the synthesis of the information given by a table of data.

Our aim in this paper is to give a general description of the system of methods we have elaborated that permit a combinatorial and statistical strategy for the treatment of large tables of data.This strategy is based on the recognition and classification of finite structures on the set of objects defined by the population under study. This approach has enabled us in manifold real examples, from different disciplines of the Humanities, to clearly extract the principal tendencies of behaviour (of the studied population) and the different components of each one of these tendencies. The reason for this success is certainly the way of stating precisely the notion of proximity between structures of the same type α and α'. For this, we have used an old and well-established idea of statistical approach: the proximity between α and α' must only

0033-5177/80/0000–0000/$02.25

take into account what may be significant with respect to a hypothesis of non-link, having regard for cardinal characteristics of α and α'. This notion intervenes at each step in constructing the classification or hierarchical classification or, more generally, of an algorithm of synthesis of the information on a given structure.

Given a table of data which crosses a finite set V of descriptive variables and a finite set E of individuals or objects, two perspectives of synthesis by the hierarchical classification approach appear. The first one concerns E and the second V. For the first, we hope to obtain a family of groups and subgroups of similar objects such that the higher the level of the hierarchy where two given objects are joined, the higher the degree of resemblance between them. As regards the second, the construction of the hierarchy of classifications on the set V of descriptive variables based on the proximities between them leads to the discovery of the principal tendencies of behaviour of the studied population. This second approach is generally the most fruitful since it reduces and then simplifies our means of recognition by organizing the links between the numerous questions asked by the specialist to understand his population that constitutes the large set E. Though the method we shall describe is more concerned with clustering the variables, it permits us with equal flexibility to classify the set E of objects. Otherwise in the important case of a symmetric table of data where the representation of V with respect to E is of the same nature as that of E with respect to V (e.g. incidence or contingency tables), the crossing of a couple of trees of classifications of V and of E gives a very rich interpretation of the data. We suppose, even though we have to make some transformations, that V is composed of variables of the same nature; that is to say, determining a unique type of structure on E. The definition of measure (or coefficient or index) of proximity (also referred to as association or similarity or resemblance) of V or of E will closely depend on the type of this structure. Hence it is necessary to start by describing the different types of descriptive variables as they appear in the Humanities and Natural Sciences, according to the structure induced on E, and thus specifying the mathematical representation of each type of variable. The measure of proximity between two variables will thus be a measure of proximity between the two structures of the same type induced on E by the variables. The definition of a measure of similarity between two objects of E will also depend on their mathematical representation with respect to V.

II. Types of Descriptive Variables

We distinguish between two principal types of descriptive variables: one represented by a subset of the set E of objects or by a weighting on E; and the other whose representation is a subset of $E \times E$ or a weighting on $E \times E$.

To the first category we can assign:

1. The attribute of description (i.e. feature)

This variable is defined by a mapping a of the set E of objects into the set of the two codes $\{0, 1\}$ where to each $x \in E$ is associated $a(x) = 1$ (0) if the feature is present (absent) in the object x. The subset of E indicated by this variable is $E_a = a^{-1}(1)$, which is composed of all elements of E that posess the attribute.

2. The numerical variable

This variable is defined by a mapping a, assigning to each element x of E a real number $a(x)$ which is the measure of the variable on the object x. Mostly, the variable takes naturally positive values. We can distinguish the case where the variable takes the values in the set $\mathbb{N}$ of integers; but the statistical treatment of such a variable is generally the same as that of the real-valued variables. This variable is considered as defining a weighting on E.

To the second category we can assign:

1'. The "ranking" variable

This variable is generally associated with a scale of measure of aptitude which is discriminating enough to make the mapping injective that assigns to each subject his measure of ability. This variable, that we have for technical reasons distinguished from the following one, determines a total order o on E that we represent by its graph $R(o)$ in $E \times E$;

$$R(o) = \{(x, y) \in E \times E / x < y, \text{ for } o\}$$

Its cardinality is $n(n-1)/2$.

2'. The descriptive character with a completely ordered set of values

Let $C = \{c_1, c_2, \ldots, c_k\}$ be the set, totally ordered by preference relation, of the modalities of the character; we suppose that $c_1 < c_2 < \ldots < c_k$. The mapping of E onto C ($x \to a(x) = c_i$, if the object x has the ith modality of the character) defines the variable. This variable indicates a total preorder on E with k classes; the ith being determined by $E_i =$

$a^{-1}(c_i)$. This last structure is equivalent to that of a partition with completely ordered classes. This variable is distinguished from the preceding one by the fact that there exists at least one subscript i for which the class E_i contains more than one element; it is distinguished from the following one because the different classes E_i are in this case ranked. We represent a total preorder w on E by the subset $R(w)$ of $E \times E$ defined as follows:

$$R(w) = \{(x, y) \in E \times E / x < y, \text{ and not } y < x \text{ for } w\}$$

The set $R(w)$ may be written as

$$R(w) = \sum_{i<j} E_i \times E_j$$

where E_i is the ith class of the preorder.

It is natural to associate the sequence $(n_1, n_2, ..., n_k)$ of the cardinality of the classes ranked according to the preorder. This sequence is called "composition" of the preorder.

The cardinality of $R(w)$ is $\sum_{i<j} n_i n_j$.

3'. The character with a set of modalities without any structure

Let $C = \{c_1, c_2, ..., c_k\}$ be the set of values of the character. We do not assume here any structure on C. The variable is defined, as above, by a mapping of E onto C, where $x \to x(x) = c_i$, if the object x has the ith modality of the character. This variable induces a partition on E, $\pi = \{E_1, E_2, ..., E_k\}$ where $E_i = a^{-1}(c_i)$. We naturally associate with the partition the sequence $(n_1, n_2, ..., n_k)$ of the cardinalities of its different classes (n_i = card (E_i)). Representing this type of variable is equivalent to representing the partition π associated with it: considering the reflexive and symmetric character of the equivalence relation defined by the partition, the smallest set of representation is $F = P_2(E)$; where $P_2(E)$ is the set of unordered object pairs: $P_2(E) = \{\{x, y\}/x \in E, y \in E, x \neq y\}$. The partition π will be represented by a subset of F. In the same way as above (cf. *2'*), we will represent π by the following subset:

$$S(\pi) = \{\{x, y\} \in F / \exists\ i \neq j, x \in E_i \text{ and } y \in E_j\}$$

which is composed of the pairs of objects separated by the partition and the cardinality of which is exactly equal to that of $R(w)$, where w

is a total preorder having the classes E_i. But the closest representation of the graph of equivalence relation is given by the subset $R(\pi)$ of pairs of objects joined by the partition (i.e. belonging to the same classes of):

$$R(\pi) = \{\{x, y\} \in F / \exists\ j,\ x \in E_j \text{ and } y \in E_j\}$$

$R(\pi)$ is the complement of $S(\pi)$ in F, and we have:

$$\text{card}(R(\pi)) = \sum_{1 \leqslant i \leqslant k} n_i(n_i - 1)/2;\ \text{card}(S(\pi)) = \sum_{i<j} n_i n_j$$

and card $(F) = n(n-1)/2$, where $n = \text{card}\ (E)$.

In fact we need not worry about the choice of the representation for comparing two characters. By choosing either of the two representations for comparing two characters we shall obtain the same measure of proximity.

4'. The variable "weighting" on $E \times E$

This case is more rare than the preceding. The representation of this variable is a square matrix $\{\mu_{xy}/(x, y) \in E \times E\}$ where μ_{xy} is a weight attached to the couple (x, y).

Unlike many methods where the discrete scale of values of the variable is plunged more or less arbitrarily into a richer scale, generally numerical, we have attached a simple mathematical representation, *translatable* into natural language, which respects exactly the poverty of scale of the descriptive variables to be treated. This allows us to respect faithfully the nature of the data and thus to be closer to the meaning of a variable introduced by the expert.

However, the *same* variable can admit more than one faithful mathematical representation; thus an attribute of description (i.e. feature) a may also be considered as defining a partition of E in two classes $\{E_a, E_a^c\}$ where E_a^c is the complement in E of E_a that has the same sense as in (1) above. It results for a in the representation given in (3') of a descriptive character. It is equally possible to consider the attribute a as defining on E a total preorder with two classes E_a^c and E_a where $E_a^c < E_a$ for the quotient order. In this situation, it results for a in the same representation as that of a descriptive character with a completely ordered set of modalities (cf. 2' above). Thus the set A of descriptive attributes, *indexing* one side of the incidence table of data, can be considered as defining a sample either from the set of all subsets of E, or from the set of the partitions with two classes of E, or from the set of the total preorders with two classses on E. The constitution of a measure of proximity according to the principle stated in the introduction may depend closely on the mathematical representation retained for

the feature; there arises here the question of the influence of these different representations on the *form* of the types which will appear by clustering A. The most recent treatments show indisputably that the results of our methods are very stable when we replace the first representation defined above in *1* by those defined above in *2'*.

Otherwise, it is possible, for technical reasons, to limit deliberately *the descriptive richness* of the scale of a given variable. Thus, for example, to retain from a numerical variable a, the associated "ranking" variable defined as follows:

$$\forall (x, y) \in E \times E,\ x < y \Leftrightarrow a(x) < a(y)$$

The problem is to study the influence of the impoverishment of the descriptive structure on the organization of their links.

III. Measure of Proximity between Descriptive Variables

Here we explore more completely the relevance of the classification of the different types of descriptive variables into two categories: the first, where the variable can be represented by a subset of E or by a weighting on E, and the second, where it can be represented by a subset of $E \times E$ or by a weighting on $E \times E$.

1. Measure of proximity between attributes

If (E_a, E_b) is the pair of subsets of E, representing a given pair (a, b) of descriptive attributes, let us consider $s = \text{card}\ (E_a \cap E_b)$. It is obvious that the statistic s (number of objects which possess the two attributes a and b) must play an important part in the construction of the measure of proximity. As a matter of fact, the presence of two features in the same object can be significant of their association. s will play the role of the "rough" index of proximity; but the value of s alone is certainly a biased index of the resemblance between the two attributes a and b and to obtain a rather high (or low) value of s it is indeed sufficient to have two frequently-occurring (or rare) features, irrespective of the relative positions of E_a and E_b. Therefore we introduce the hypothesis N of non-link whose one specific form consists of fixing the subset E_a and associating to E_b a random element Y taken from the set of all subsets of E, having the same cardinality $n_b = \text{card}\ (E_b)$, this set having a uniform measure of probability. What is required is to evaluate s with respect to the distribution of the random variable $S_a = \text{card}\ (E_a \cap Y)$. We must here emphasize that the distribution of S_a is exactly *the same* as that of $S_b = \text{card}\ (X \cap E_b)$ where instead of fixing E_a and

associating to E_b a random subset Y, we fix E_b and associate to E_a a random subset X from the set of all equally probable subsets of E with the same cardinality n_a = card (E_a). The common distribution (of S_a and S_b) is hypergeometric, its mean and variance being respectively $\mu = n_a\, n_b/n$ and $\sigma^2 = n_a(n - n_a)n_b(n - n_b)/n^2(n - 1)$.

To be in accordance with the principle stated in the introduction and to take out of s only what can be "significant" with respect to the hypothesis N of non-link, we consider the measure of proximity between the descriptive attributes a and b, defined by:

$$P(a, b) = \mathrm{P}_{\mathrm{r}}^{N}\{S_a < s\} \tag{1}$$

in other words, the bigger the value of s compared to hypothesis N, the higher the degree of resemblance between a and b. The index $P(a, b)$, which takes values in the interval [0, 1], is here exactly the proportion of subsets Y, in the set of all subsets of E with the same cardinality n_b, for which the cardinality of the intersection with E_a is lower than s. There may exist a form of the hypothesis of non-link where the index has a more probabilistic nature. In the Appendix, we compare two classifications obtained according two close hypothesis of non-link N and N' where for N' the measure associated of proximity $P'(a, b)$ has probabilistic character. In any case, to calculate $P(a, b)$, we assume normal approximation (which is generally very good) of the distribution of S_a. According to this reference to the normal distribution, we can propose a coefficient which is obtained by standardizing s with respect to N; this index is then:

$$S(a, b) = (s - \mu)/\sigma$$

with

$$P(a, b) = \Phi(S(a, b) \tag{2}$$

where μ and σ^2 are the mean and variance of the normal distribution $\mathcal{N}(\mu, \sigma^2)$ which approximates the distribution of S_a, and where Φ is the distribution function of $\mathcal{N}(0, 1)$. The coefficient $S(a, b)$ can be expressed as follows:

$$S(a, b) = (st - uv)/\sqrt{(s + u)\,(s + v)\,(t + u)\,(t + v)} \tag{3}$$

where t = card $(E_a^{\mathrm{c}} \cap E_b^{\mathrm{c}})$, u = card $(E_a \cap E_b^{\mathrm{c}})$ and v = card $(E_a^{\mathrm{c}} \cap E_b)$; E_a^{c} and E_b^{c} being the complements in E of E_a and E_b respectively. Eqn. (3) is nothing other than the association coefficient of Pearson's K, of which the square is exactly the chi-square χ^2 attached to the contingency table that crosses the two partitions $\{E_a, E_a^{\mathrm{c}}\}$ and $\{E_b, E_b^{\mathrm{c}}\}$. The symmetry of the form of eqn. (3) shows that we shall be led to the

same index if instead of starting with the number s of "positive associations," we start with the number t of "negative associations" as the basis of the statistic of proximity.

2. Measure of proximity between numerical variables

In the comparison of a pair (a, b) of descriptive attributes, the basis s of the measure can be put in the form

$$s = \sum_{1 \leqslant i \leqslant n} \alpha_i \beta_i \text{ where } (\alpha_1, \ldots, \alpha_i, \ldots, \alpha_n) \text{ or } (\beta_1, \ldots, \beta_i, \ldots, \beta_n)$$

is the characteristic vector of the subset E_a (or E_b): $\alpha_i = 1$ (or $\beta_i = 1$) if the object coded i is in possession of the feature a (or b), and 0 if not. To s we have associated the two dual random variables S_a and S_b having the same distribution, that we can write here in the following form:

$$S_a = \sum_{1 \leqslant i \leqslant n} \alpha_i \beta_{\tau(i)}, S_b = \sum_{1 \leqslant i \leqslant n} \alpha_{\tau(i)} \beta_i \qquad (4)$$

where $(\tau(1), \ldots, \tau(i), \ldots, \tau(n))$ is a random permutation of $(1, 2, \ldots, i, \ldots, n)$. In other words, τ is a random element in the set $\mathcal{C}_n$, provided with a uniform measure of probability, of all permutations on $(1, 2, \ldots, i, \ldots, n)$. In the comparison of a pair (a, b) of descriptive numerical variables, the basis of the measure will also be $\sum_{1 \leqslant i \leqslant n} \alpha_i \beta_i$ where now α_i (β_i) is of the value of the variable a (b) on the ith object. In an analogous way, to calculate an index such as given by eqn. (2), we use the normal approximation of the common distribution of S_a and S_b given by eqn. (4). The asymptotic normality of this distribution is the subject of the celebrated theorem of Wald and Wolfowitz (1944) in non-parametric statistics. The decisive statement was given by Noether (1949) and more recently Hàjek (1961) has contributed to the theorem in a very interesting way. The corresponding standardized statistic $((s - \mu)/\sigma)$ which generalizes K. Pearson's coefficient is nothing other than $\sqrt{(n-1)}\,\rho_{ab}$ where ρ_{ab} is the correlation coefficient between the two variables a and b.

These remarks justify the use of a same measure of proximity on the set V of descriptive variables if the latter consists of numerical variables or attributes. Moreover, the case of the corresponding table of data is specially common.

1'. Measure of proximity between "ranking" variables

Let (o_a, o_b) be the pair of total orders on the set E of the individuals or objects, defined by a pair of "ranking" variables. The representation of a total order o being a subset $R(o)$ of $E \times E$ (cf. Section II.1'), and

considering the comparison between descriptive attributes, the basis of the measure to start with is naturally

$$s = \text{card}\,(R(o_2) \cap R(o_b)) = \sum_{1 \leqslant i \leqslant n-1} g(i)$$

where i indicates the object of which the rank is i for o_a and where $g(i)$ is the number of objects j on the right of i according to o_b such that $j > i$.

The hypothesis N of non-link associates to one of the two total orders o_a (o_b) a random element o in the set O, provided with a uniform measure of probability, of all total orders on E; we have card $(O) = n!$. The mean and variance of the common distribution of card $(R(o_a) \cap R(o))$ and card $(R(o) \cap R(o_b))$, are $\mu = n(n-1)/4$ and $\sigma^2 = n(n-1)(2n+5)/72$, respectively.

The convergence of this distribution to the normal one has been established by M.G. Kendall (1970) whose coefficient τ_b is obtained by centring the basis index s given by eqn. (5) and reducing it with the maximum of the absolute value of the numerator. The reference to the normal distribution permits one to calculate the measure of association as the following probability:

$$P(a, b) = \phi\left(\frac{s-\mu}{\sigma}\right) \tag{6}$$

where ϕ is the distribution function of $\mathcal{N}(0, 1)$.

2'. Measure of proximity between descriptive characters with totally ordered sets of values

If (w_a, w_b) is the pair of total preorders on E determined by the pair of variables of the type considered, then the basis of the measure can be put in the following form:

$$s = \text{card}(R(w_a) \cap R(w_b)) = \sum_{\{i<p, j<q\}} n_{ij} n_{pq} \tag{7}$$

where $\{n_{ij}/(i, j) \in I \times J\}$ is the contingency table of crossing the two total preorders; I (J) indicates the set of codes of modalities of the first (second) character.

The hypothesis N of non-link fixes one of the two preorders (w_a or w_b) and associates to the other a random preorder in the set – provided with a uniform probability measure – of all total preorders having the same composition. A theorem of duality permits us to establish that the distribution of card $(R(w_a) \cap R(w'_b)$ is the same as that of card $(R(w'_a)$ $\cap R(w_b))$; that is to say, it does not depend on the fixed total preorder

w_a or w_b. The calculation of the moments of this distribution (Lerman, 1973a; 1976) permit one to standardize the above-mentioned index s and to refer to a probability scale defined from the normal distribution to obtain an index as eqn. (6) above. It is shown in this analysis that the statistic proposed by Kendall in this case is biased in the sense that the expected value of the associated r.v. under the non-link hypothesis N is other than zero. In fact, Kendall has defined his index by simply assigning, in the algorithm of computation of τ_b the value of an ordinal function to different objects belonging to the same class of the pre-order (treatment of ties).

3'. Measure of proximity between descriptive characters of which the respective sets of modalities are without any structure

The path leading to this new measure is entirely analogous to the previous one. The material support of the information is still the contingency table which crosses the pair (π_a, π_b) of partitions defined by the pair of variables. Taking into account the representation retained (cf. Section II.3'), the basis of the measure can be put as follows:

$$s = \text{card}(R(\pi_a) \cap R(\pi_b)) = \sum_{(i,j) \in I \times J} n_{ij}(n_{ij} - 1)/2 \tag{8}$$

The hypothesis N fixes one of the two partitions (π_a or π_b) and associates to the other a random partition in the set, provided with a uniform measure of probability, of all partitions having the same type; the type of a partition being defined as the decreasing sequence of the cardinalities of its classes. The same above-mentioned theorem of duality permits one to establish the corresponding result; that is to say, the distribution of card $(R(\pi_a) \cap R(\pi'_b))$ is exactly the same as that of card $(R(\pi'_a) \cap R(\pi_b))$ where π'_a (π'_b) is the random element associated to π_a (π_b). The calculation of the moments of this distribution permits one to centre and to reduce the above-mentioned index s and to refer to a probability scale defined from the normal distribution and hence, to obtain an index such as that defined by eqn. (6) above.

This coefficient is essentially different from the χ^2 associated to the contingency table where the two partitions π_a and π_b are crossed. The experimental work of comparison of the two statistics have already been undertaken.

4'. Measure of proximity between variables defining weightings on $E \times E$

We have shown above (cf. 2) how the measure of proximity between numerical variables generalizes K, Pearson's coefficient of association

between two descriptive attributes. Henceforth we can try to extend the different previous indices (*1'*, *2'* and *3'*) of proximity between two discrete variables of the same type, represented by subsets of $E \times E$, to the comparison of two weightings on $E \times E$ of the form:

$$\{\mu_{ij}/(i,j) \in I^{[2]}\} \text{ and } \{\nu_{ij}/(i,j) \in I^{[2]}\}$$

with $I^{[2]} = (I \times I - \Delta)$, Δ being the diagonal of $I \times I$; and where I is the set $\{1, 2, ..., i, ..., n\}$ of subscripts coding E.

It is natural to consider the rough index:

$$s = \sum_{(i,j) \in I^{[2]}} \mu_{ij}\nu_{ij} \tag{9}$$

and to associate the dual random variables:

$$s = \sum_{(i,j) \in I^{[2]}} \mu_{ij}\nu_{\tau(i)\tau(j)} \text{ and } T = \sum_{(i,j) \in I^{[2]}} \mu_{\tau(i)\tau(j)}\nu_{ij} \tag{10}$$

shown to have the same distribution, where τ is a random permutation defined as before (cf. *2* above).

We have in fact proved that the different cases of comparison of discrete variables considered above (*1'*, *2'* and *3'*) appear formally as particular cases of the present situation.

Inspired by a previous paper of Daniels (1944), Lecalvé (1976) had the idea of this extension which led to a new measure. Nevertheless, we have resumed this approach in a more precise way, especially in the calculation of the moments of S (T) to better justify the normal approximation of S (T) (Lerman, 1976).

5'. Conclusion

Although the random element considered is a permutation, the presentation of M.G. Kendall's τ_b generally done in non-parametric statistics does not seem to be well integrated with the permutation statistics. We are content with making sure that the statistic corresponding to τ_b is not linear with respect to the one of the form $\Sigma_i \xi_i \eta_{\tau(i)}$, subject to a very minute analysis (Wald and Wolfowitz, 1944; Noether, 1949; Hàjek, 1961).

Between S considered by us and the latter is situated the statistic studied by Motoo (1957) which can be written as:

$$V = \sum_{1 \leqslant i \leqslant n} \gamma_{i\tau(i)} \tag{11}$$

such a statistic makes it possible to compare two weightings on $E \times E$ with respect to the random permutation of rows or (exclusively)

columns of the square table representing the weighting. In fact, in this situation we are led to the statistic:

$$W = \sum_{1 \leqslant j \leqslant n} \sum_{1 \leqslant i \leqslant n} \xi_{ij} \eta_{i\tau(j)} \tag{12}$$

where we have put:

$$\gamma_{j\tau(j)} = \sum_{1 \leqslant i \leqslant n} \xi_{ij} \eta_{i\tau(j)} \tag{13}$$

On the other hand, the study of the statistics attached to the contingency tables (which occur when we have to cross two partitions or total preorders) appear in the literature of non-parametric statistics as quite separate from the permutational statistics. Now we have seen how our measure combines the two aspects in a very natural way.

We must insist that relative to such tables our point of view remains essentially different from that developed in a series of articles by Goodman and Kruskal (1954; 1963): let I (J) be the set indexing the rows (columns) of the contingency table, and let:

$$\{\pi_{ij}/(i,j) \in I \times J\} \tag{14}$$

be the table of proportions or theoretical probabilities for the entire or hypothetical population.

The above-mentioned authors, after some intuitive and somewhat arbitrary considerations propose different measures which are functions of the numbers π_{ij}; each of them is supposed to measure a certain type of relationship. In the centre of these works is the asymptotic behaviour of the sampling distributions or estimators of each of those measures.

The distribution of the permutation statistics is studied in favour of the development of the non-parametric theory of the statistical tests where the hypothesis of independence or non-link is tested. But in our approach to the data in Humanities, there is always a link, however subtle it may be, which can be "measured" from what is observed. However, the hypothesis of non-link plays a crucial role, because it enables one to establish the scale of reference for the evaluation of the link. These steps have led us to a uniform approach at the different stages of organization of the links between different variables in the form of a tree; the most significant nodes of which can be recognized from similar considerations.

IV. Measure of Proximity between Classes of Variables and Associated Tree of Classifications

Whatever the type of the pair (a, b) of descriptive variables, the index $P(a, b)$ that we retain takes its values in the interval $[0,1]$. It arises as a proportion or probability and it introduces the notion of "likelihood" within the notion of "resemblance". To be able to establish a classification tree on the set V of descriptive variables, it is necessary to extend the notion of proximity between two variables to that between two classes C and D of variables. The basis of the index will be here:

$$p(C, D) = \max \{P(c, d)/(c, d) \in C \times D\} \tag{15}$$

The index of proximity that we will retain is defined with respect to the hypothesis N of non-link which takes into account characteristics of cardinalities. It is the probability with respect to N, that the largest value of the proximity $p(C', D')$, where C' (D') is randomly associated to C (D), be lower than the observed value $p(C, D)$. The coefficient of association to which we are led takes the following form:

$$P(C, D) = p(C, D)^{l \times m} \tag{16}$$

where l = card (C) and m = card (D).

The definition of a measure of proximity between disjoint subsets of V makes possible the first stage of condensation in the form of detailed hierarchy of classifications (fig. 1) which is obtained step-by-step by successive *unions* of classes, where at each step we join together the nearest two classes. The algorithm will start from the discrete partition where each class contains exactly one element, to end with the rough partition, where all the variables are joined together in the same class.

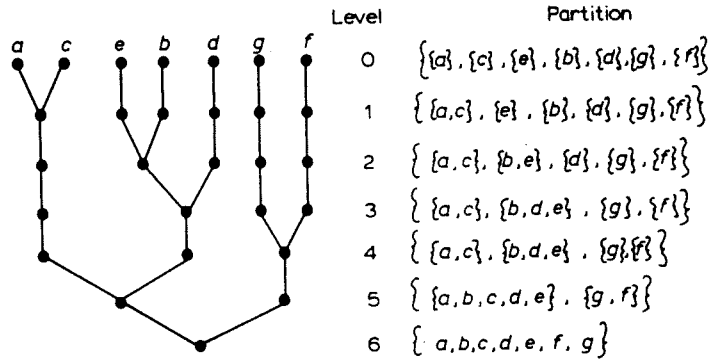


Fig. 1.

We called the algorithm corresponding to the proximity measure (16) the "Likelihood of the Link Algorithm". It has been programmed and studied in a third-cycle thesis by Nicolaü (1972). In a very general way the program provides the possibility of proceeding from a measure of proximity between disjoint subsets of V, to the polished representation of the detailed tree of the classifications.

What is new with respect to other clustering methods of hierarchical classification is that here the distribution of $P(C', D')$ is apprehended under the hypothesis N of non-link; and that permits the reference to a clearly expressed scale to evaluate the measure of proximity between two classes and to compare without bias the values of the proximities attached to two pairs of classes.

A hierarchy of classifications (i.e. ordered chain of partitions) such as those on the right-hand side of Fig. 1, is represented by a tree such as that in the figure which represents a binary tree since the partition of a given level is deduced from the preceding one by joining together two classes. It is clear that the algoritm gives the possibility of joining together more than one pair of classes at a given level.

V. Condensation of the Tree to its Significant Nodes

A decisive stage of the method consists of condensing the classification tree to the levels where a "significant" node appears, and this is done by means of a proximity coefficient between the association corresponding to the node and an adequate structure retained from the resemblances between the elements of the set D to be classified. The principle of contructing this index is exactly the same as that leading to the definition of a measure of proximity between descriptive variables. For comparing two structures of a unique type, we are led to retain from the structure of the information concerning the resemblances only the related total preorder on the set F of unordered object pairs from D (i.e. on the set of all subsets with two elements of D), called *preordonnance*, associated to the similarity index S defined on D, in the following way:

$$(p, q) \in F \times F; p < q \Leftrightarrow S(p) < S(q) \tag{17}$$

Mostly the preordonnance associated with our proximity measure is almost an *ordonnance*: total order on the set F. The preordonnance $w(D)$ will be represented by the subset $gr(w)$ of $F \times F$ defined as follows:

$$gr(w) = \{(p, q) \in F \times F / p < q, \text{ and not } q < p \text{ for } w\} \tag{18}$$

To evaluate the fit of a partition on D according to the proximities between its elements, we consider the partition as defining a preorder of F with two classes S and R where S is the set of pairs of objects separated by the partition (i.e. a pair $\{x, y\}$ belonging to S has its components x and y distant, according to the partition) and where R is the set of pairs of objects joined together by the partition (i.e. a pair $\{x, y\}$ belonging to R has its components x and y close, according to the partition). S precedes R for the quotient order; and the partition will be represented by the Cartesian product $S \times R \subset F \times F$.

The basis of the measure of proximity will then be:

$$\text{card}(gr(w) \cap (S \times R)) \tag{19}$$

Benzecri (1974) has introduced this cardinality but in the following metrical form: "number of inequalities between the distances specified by the partition and compatible with the ordonnance w," w being a total order on F. We prove elsewhere (Lerman, 1973) that the distribution of card $(gr(w) \cap (S(\pi) \times R(\pi)))$ is asymptotically normal, where π is a random element in the set, provided with a uniform measure of probability, of all partitions having a given type. The mean and variance of this distribution are respectively $rs/2$ and $rs(f + 1)/12$, where $r = \text{card}$ (R), $s = \text{card}$ (S) and $f = r + s$. The statistic:

$$\Sigma = \left\{\text{card}(gr(w) \cap (R \times S)) - \frac{rs}{2}\right\} / (rs(f + 1)/12)^{1/2} \tag{20}$$

will measure the cohesion of the classes of a given partition of D that can for example be produced at a given level of the classification tree. Σ will be called the "total" statistic.

The behaviour of the distribution of Σ, on the increasing sequence of the levels of the classification tree, observed through manifold real examples, shows a general tendancy to go on increasing until a total maximum is slowly reached, followed by an abrupt decrease (Fig. 2).

Over the beginning sequence of levels, where the general tendancy of Σ is to grow, it is possible to observe fluctuations that indicate local maxima and minima as the graph suggests. This graph shows the behaviour of the distribution of Σ over the increasing sequence of the levels of a classification tree established according to the proximities.

A given level i, where Σ shows a diminution, indicates that the corresponding partition π_i fits the ordonnance w, less well than the preceeding one $\pi_{(i-1)}$.

Experience shows that in this case it is difficult to interpret the link between the two classes of $\pi_{(i-1)}$, joined together to form π_i, at this degree of synthesis given by the classification π_i. Therefore, the special-

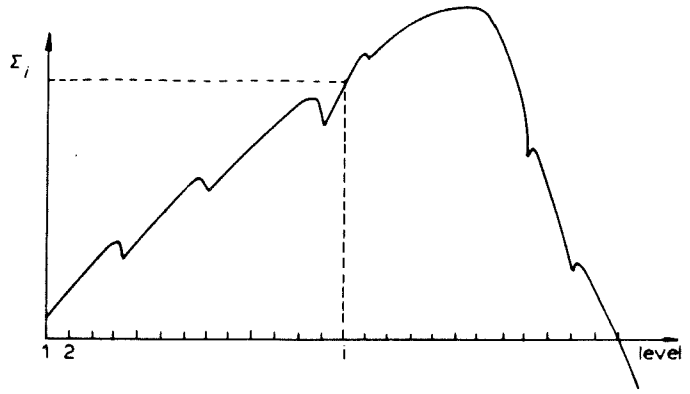


Fig. 2.

ist who wants a classification with a number of classes of around k will retain the level of tree associated to a local maximum of Σ and corresponding to about k classes.

However it is not the behaviour of the "total" statistic Σ, but that of a "local" statistic which allows the reduction of the tree to its most significant levels. This "local" statistic θ can be directly deduced from the total one by defining it as the rate of variation of Σ; hence, the value of θ at the ith level is:

$$\theta_i = \sum_{i} - \sum_{(i-1)} \tag{21}$$

Another way of establishing the "local" statistic attached to the association of two classes is to start with the rough index card $\{gr(w) \cap (R'(\pi_i) \times S(\pi_i))\}$ where $R'(\pi_i)$ is the set of object pairs joined together for the first time at the level i and where $S(\pi_i)$ is the set of object pairs still separated by the partition π_i. The final local statistic τ_i is obtained by centring and reducing the preceding rough index with respect to a relevant hypothesis of non-link (Lerman, 1973a). In fact, the behaviour of τ is nearly the same as that of θ. We retain the most significant nodes as these corresponding to the local maxima of the distribution of the local statistic (θ or τ) on the increasing sequence of the tree's levels. The experiment has in fact shown that the value of this statistic increases when a class, which is being formed, grows; and decreases perceptibly when such a class, having some consistency, drops off in favour of a rising embryo of another class. Therefore the significant nodes indicate completion stages of the different classes appearing in the tree.

We have in Fig. 3 undertaken the graph of the distribution of the local statistic τ on the increasing sequence of the tree's levels obtained

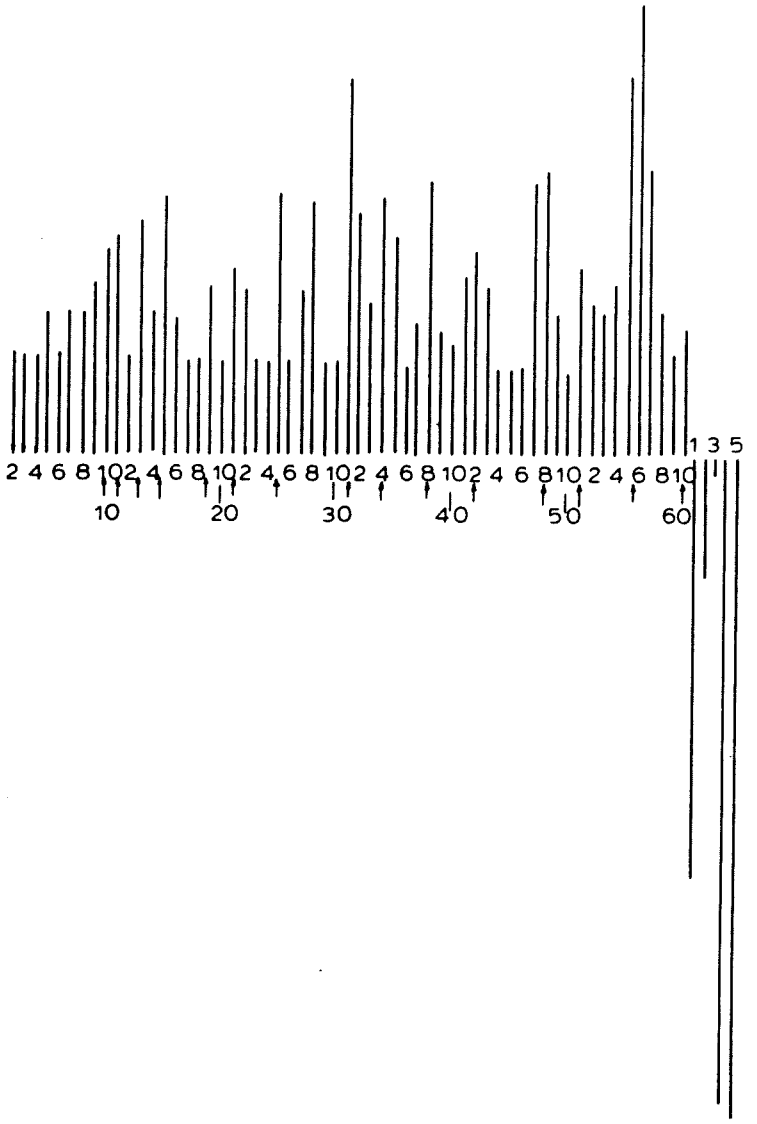


Fig. 3. Distribution of τ_i in a real case.

in a real case, to which we associate simulations in the corresponding random case defined by the hypothesis N of non-link. What was clearly observed, unlike in the simulated cases, is the noteworthy persistence of the rather important values of τ up to a high level of the tree. From this

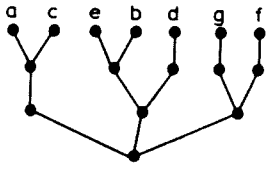


Fig. 4. The reduction of the tree of Fig. 1 to the levels 2 and 4.

the abrupt falls in τ correspond to associations of classes essentially distinct from the point of view of their interpretations.

VI. General Classifiability and Neutral Character of a Given Element

We have in the Introduction emphasized the property that our own classification is able to extract the principal tendencies of behaviour of the studied population. But this decomposition by tendencies can be more or less marked. To solve this problem, we have characterized and measured the ability of a set D to be organized according to a hierarchical classification which respects in a satisfactory way the system of the inequalities between the resemblances of object pairs of D, defined by the preordonnance (section V above and Lerman, 1970) show easily in the latter that the structure of an increasing chain of partitions on D is equivalent of an "ultrametric" preordonnance on D; that is to say, a total preorder on the set F of element pairs of D, which is characterized by the following property: whatever the subset $\{a, b, c\}$ with three elements of D, such that $\{a, b\} \leqslant \{b, c\} \leqslant \{a, c\}$ according to the preorder, the median pair $\{b, c\}$ and the upper one $\{a, c\}$ are necessarily in the same class of the preorder. The degree of classifiability is defined by a distribution which characterizes the distortion of the structure of the preordonnance on D associated with the index of similarity, with respect to its ultrametric structure.

With each pair p of F, we associate the ratio $\varphi(p)$ of triplets of elements of D, such that the median pair and the upper one of each of these triplets are strictly separated by p; the mentioned distribution is then defined by the decreasing sequence of the weighted values of $\varphi(p)$; a given value α of $\varphi(p)$ is weighted by the proportion in F of elements p for which $\varphi(p) = \alpha$.

Example: Let $D = \{a, b, c, d, e\}$ and let w be the following preordonnance on D:

$\{a, d\} \sim \{a, c\} < \{a, e\} < \{c, e\} < \{b, d\} \sim \{c, d\} < \{b, c\} < \{d, e\} < \{a, b\} < \{b, e\}$

The classifiability can be represented by a graph (Fig. 5) which consists only of horizontal lines, one of which represents the subset of pairs which intervene to strictly separate the median pair from the upper one of *the same number of triplets*; the length of a given line represents the proportion of these pairs, and its elevation the proportion of triplets of which the median and upper pair are strictly separated by any one of these pairs.

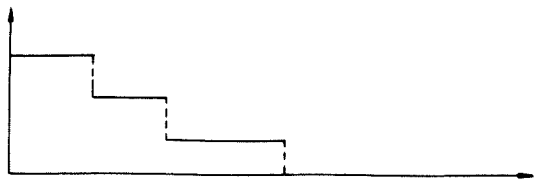


Fig. 5.

The classifiability of D is closely linked to the relative removal of each of its elements with respect to the sequence of the others, established by increasing dissimilarity. The data analysis prior to a classification leads us to measure the neutral character of a given element a of D, with respect to a classificatory goal, by the smallness of the observed variance of the proximities to a; that is to say, the variance of the distribution:

$$\{S(a, x)/x \in D - \{a\}\} \tag{22}$$

which can be written:

$$\mathcal{V}(a) = \frac{1}{(m-1)} \sum_{x \in D - \{a\}} (S(a, x) - S(a))^2 \tag{23}$$

where m = card (D) and where $S(a)$ is the mean of the distribution (22).

If, in a table of data, there are some elements corresponding to rows or (exclusively) columns such that their degree of neutrality is rather high to disturb the classifiability nature of the whole set corresponding to the rows (columns) we are able to detect these elements by means of the above statistic (23) and to extract them before any classification. There is generally no reason for this treatment if the set D to be classified is the set V of descriptive variables and if the set E of objects has, as is mostly the case, very important cardinality with respect to the cardinality of V. As a matter of fact, in this situation, the dispersion of the representation of V with respect to E is generally too large, (cf. section II). However, the bigger the value of $\mathcal{V}(a)$, the stronger the intervention of a in the definition of the class where it appears. Therefore, the difficulty of interpreting in the context of one class the presence of a given variable a can be understood by the weakness of the value of $\mathcal{V}(a)$.

We have previously (Lerman, 1972) studied a geometrical representation of V or rather, of a subset of V corresponding to a homogeneous class, by means of simultaneous analysis of the mean and variance of proximities to each element in the class. This representation gives the geometrical organization – for the metric attached to the considered proximity measure S – of the different elements of that class from its

extremal points, the least linked, that we call "attraction poles". If W denotes the class to be figured, the first attraction pole p_1 is determined by maximizing the variance:

$$\mathcal{V}_w(c) = \frac{1}{(l-1)} \sum_{y \in W - \{c\}} (S(c, y) - S(c))^2 \tag{24}$$

where $c \in W$ and where l = card (W).

The second pole p_2 which, simultaneously, must have important discriminant value and must be weakly linked to p_1, is determined by the rule consisting of maximizing the criterion:

$$\{\mathcal{V}_w(c)/S(c, p_1)\}^2 \tag{25}$$

and so on; the $(k+1)$th pole is determined from the first k, by the rule:

$$\max\left\{\min\left[\left(\frac{\mathcal{V}(c)}{S(c, p_1)}\right)^2, \left(\frac{\mathcal{V}(c)}{S(c, p_2)}\right)^2, \ldots, \left(\frac{\mathcal{V}(c)}{S(c, p_k)}\right)^2\right]\right\} \tag{26}$$

If unidimentional scale is underlying the different items forming the class W it is possible by this technique (in fact by the calculation of the first two poles) to restitute the ordination of W (Guttmann's old problem). In the same way, if we apply on analogous technique to a class G of objects it is possible to detect "seriation" (archeological problem). More generally, the prior determination of a sequence of attraction poles in the whole set D to be classified leads to a new and rich family of clustering algorithms which is obtained by:

(i) varying the criterion, as (26), by working for example with the distances instead of the proximities and with the 2nd absolute moments instead of the variances. The expression of the criterion depends on the nature of the set to be classified;

(ii) varying the criterion of assignation of a given element to one of the classes being constituted around the attraction poles. This criterion can be defined from the distance of the given incomplete classes with respect to the element to be assigned.

(iii) attaching to the classification with k classes ($k = 1, 2, \ldots$) obtained at the k^{th} stage of the algorithm, the value of the significance statistic such as the one used in section V or as the explained variance.

The analysis of this family of algorithms has just been completed by Leredde (1979).

VII. Classifiability and Cohesion of the Classes

One may point we wished to explore was the relation between the degree of classifiability and the cohesion of classes obtained by a good

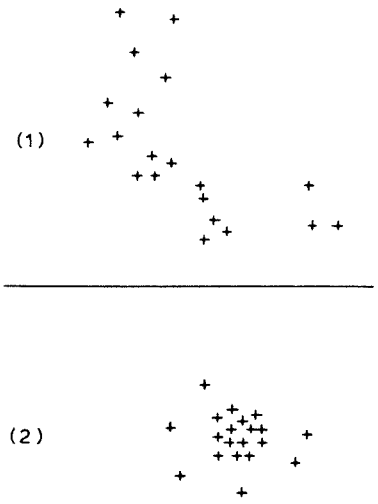


Fig. 6.

algorithm. The better the classifiability the more flattened the graph consisting of horizontal lines, on the horizontal axis; thus, it can be measured by the smallness of the area between the two perpendicular axes. The cohesion of classes determined by the partition will be measured by the statistic Σ (cf. eqn. (20)).

Therefore we are interested in the distribution of card $\{gr(w) \cap (S(\pi) \times R(\pi))\}$, where π is a random partition of a fixed type, for a given value of the classifiability measure of the preordonnance w. If we consider the biggest value obtained by this distribution we notice that it clearly decreases when w becomes weakly classifiable. On the other hand, the slope of the graph of the distribution function becomes steeper when we proceed from a good to a bad classifiability of w. The variance of this distribution remains almost constant if w is well classifiable and becomes very unstable if not: it is rather small if the different classes have the same cardinality, and rather large if the dispersion of classes' cardinality is strong. These results which had been obtained by a medium-size experimental analysis completely support the theoretical study (Lerman, 1973a). However apart from very extreme cases, we must not believe that a relative amount of classifiability increase will necessarily imply stronger cohesion for the formed classes. We realized this fact from random simulations of an incidence table of data, associated to a real case, under the hypothesis N of non-link. To illustrate this geometrically, let us consider the two sets of points (1) and (2) in Fig. 6. The set (1) of points seems to be more suitable than the set (2) for organizing in classes; nevertheless, the classification of (2) in to seven classes, of which the first groups all the central points and where

each of the six others contains exactly one peripheral point, will have a stronger cohesion than any classification of the set (1).

VIII. Analysis of a Class of Ordinal Scales

In the case where the set V of descriptive variables is composed of characters having totally ordered sets of values (cf. section II.2'), the research of unique ordinal scale, taking into account the different variables belonging to a given class $C \subset V$, can be justified if C has a strong cohesion. The specialists in multidimensional scaling put the problem of recognizing unidimensionalscaling behind a group C of variables determined after some intuitive and somewhat arbitrary considerations. We propose here to fix C as corresponding to a class which appears from a suitable classification algorithm on V. The class C can be represented by a product of chains, each of which is associated to the totally ordered set of values of a given variable of C. A unidimensional scale can be represented by a maximal chain which determines a total order on the whole set of modalities of the different items, except those which occupy the initial position of the variables constituing C. This problem has been previously analysed (Lerman, 1967) where we developed an algorithm of combinatorial nature which permits us to establish, in a quick and optimal way, the unidimensional scale which fits best the set of weighted vertices of the representation's lattice. A given vertex corresponds to a possible pattern and the weight attached corresponds to its observed frequency. The fit is done according to a statistically relevant distance.

On the other hand, we propose in this work an approach which permits us to link in a very natural way the unidimensional scaling of Guttman (1950) to the latent structure analysis of Lazarsfeld (1950).

IX. Return to the Individuals

We have pointed out that our principal method of classification is rather oriented towards the organization of the relations between descriptive variables in a system of classes and subclasses. Nevertheless, it can be adjusted to the classification of the set E of individuals or objects, whatever the common scale of the different variables constituting V. This adjustment presents no problems in the particular but very important case of the incidence table of data crossing A and E, where the representation of the set A of descriptive attributes with respect to E

(sample in the set $\mathcal{P}(E)$ of subsets of E) is analogous to the representation of E with respect to A (sample in the set $\mathcal{P}(A)$ of subsets of A). In the case of such tables of data, if we shade the elements of the table containing 1 and corresponding to the pairs $(a, x) \in A \times E$ for which $a(x) = 1$, and if we permute rows and columns according to the crossing of two reduced classification trees on A and E respectively, we obtain a very rich interpretation of the data. This technique had been successfully used in an experiment in genetic psychology by Pierraut-le-Bonniec and Van Meter (1976) where the principal genetic stages of cognitive development have clearly appeared.

Let $\{A_i/1 \leqslant i \leqslant k\}$ and $\{E_j/1 \leqslant j \leqslant h\}$ be a pair of significant classifications, determined on A and E respectively. The previous technique that we have applied in different treatments shows that two partitions are dual in the following sense:

(i) each class A_i refers to a subset of classes $\{E_j/j \in J(i)\}$, where $J(i)$ is the subset of subscripts associated to i such that each feature of A_i is possessed by a "large" proportion of objects of $U\{E_j/j \in J(i)\}$;

(ii) each class E_j refers to a subset of classes $\{A_i/i \in I(j)\}$, where $I(j)$ is the subset of subscripts associated to j such that each object possesses a "large" proportion of attributes of $U\{A_i/i \in J(i)\}$.

Relative to the significant classification $\{A_i/1 \leqslant i \leqslant k\}$ which is generally indicated at one of the most significant levels of the classification tree, Nicolaü (1972) studied the following geometrical representation of the set E of individuals or objects. We begin by attaching to each element x of E the vector of proportions $(f_1^x, f_2^x, \ldots, f_i^x, \ldots, f_k^x)$, where f_i^x is the proportion of features of A_i possessed by x. If f_l^x, f_m^x and f_u^x are the three biggest components of the latter, we reduce the preceding vector to the three components:

$$p_l^x = f_l^x/(f_l^x + f_m^x + f_u^x)$$

$$p_m^x = f_m^x/(f_l^x + f_m^x + f_u^x)$$

$$p_u^x = f_u^x/(f_l^x + f_m^x + f_u^x) \tag{27}$$

distinct from zero. It is then possible to adopt a barycentric representation of E with respect to three given classes A_r, A_s and A_t, represented by the three vertices R, S, T of an equilateral triangle; thus the individual x will be represented by the point $P(x)$ defined as follows:

$$P(x) = p_r^x R + p_s^x S + p_t^x T \tag{28}$$

where p_r^x, p_s^x and p_t^x define, respectively, the degrees of definition of x, by the classes of attributes A_r, A_s and A_t. This technique has been used successfully in some interesting examples.

It is possible to associate to each type, defined by a class A_i of attributes, the individual who fits "the best" to the type. This one is to be found from those who are the most "responsible" of the classification $\{A_i/1 \leqslant i \leqslant k\}$. The degree of responsability of x in the forming of types can be measured by the following ratio of two variances:

$$\frac{1}{k} \sum_{1 \leqslant i \leqslant k} \alpha_i (f_i^x - f^x)^2 / [f^x(1 - f^x)/n] \tag{29}$$

where α_i denotes the proportion in A of attributes belonging to A_i (α_i = card (A_i)/card (A)) and where

$$f^x = \sum_{1 \leqslant i \leqslant k} \alpha_i f_i^x \tag{30}$$

is the proportion of atrributes possessed by x.

The individual who will be considered representative of the class A_i, must have the largest value of (29) caused by a high value of f_i^x and by a low value of f_l^x, for $l \neq i$.

The discovery of the most "typical" individuals can be of great importance for the psycho-sociologist since it gives him the possibility of furthering the research by studying more accurately each subject determined in the above way.

We can attempt to generalize the preceding considerations whatever the common type of descriptive variables. On the other hand we can study the case where the respective roles of V and E are inverted; that is to say, where we find V from a classification on E.

X. Conclusion

We now give an account of our collaboration in the use of our methods with various specialists. These methods had been applied several times and we contacted several scientists involved with different branches of the humanities. Their very rich data enabled us to test different aspects of our methods. Generally, the collaboration necessitates an intermediate whose experience in computer programming and statistics is of importance in treating the data in depth. Cohen (1977) enabled us to contribute to a very interesting data analysis with Berge and Denjean (1974), based on a large psychopedagogical investigation where we attempt to establish a relation between development of infancy, feeding behaviour and intellectual performance.

It is obvious that the main goal of the classification method is to replace the multiple and uncertain hypothesis about the associations

between the descriptive variables by the observation of general behaviors which confirm some and attenuate others of the basic hypotheses that have led to the inquiry. In any case, the treatment gives unexpected synthetic vision of the data; we brought out this fact several times by asking the specialist to predict the classes that would appear. We were surprised that the specialist often forgets many of his hypotheses established before the treatment when he examines the results. Nevertheless, the method cannot be usefully recommended to the researcher in humanities who is "without prior hypotheses"; the hypotheses arise directly from the knowledge (even partial) of the object studied and intervene crucially in the definition of descriptive variables. The method cannot also be usefully recommended to the specialist who has a rigid hypothesis because he is only able to find at the end of the treatment what he was willing to put in: his hypothesis. Otherwise, the method can be used with great interest and profit by the specialist who examines the object of his study from multiple points of view, according to some fundamental hypothesis.

Our contacts with the specialist were generally successful and mostly led the latter to an important publication. The tendency of the specialists is to look for the things that would confirm his own hypothesis, through the results. He is however very much influenced by the indications not referring to his assumptions. From a formal point of view, the relative ease of our contacts is due to the finite character of the representation of the data and their synthesis. In fact, the entire work is based on the conviction that if we adopt a rather poor structure of condensation we can go deep into the interpretation by respecting the finite structure of the data, provided that we define the proximities in a "relevant" statistical way.

The acquired confidence in the method is firstly due to the fact that the results support, at least in part, the fundamental hypothesis put forward in the elaboration of the inquiry survey. These results can point towards the specialist's wildest hypothesis, but can also bring to light important and unsuspected behaviors which enables the specialist to enlarge his prior hypothesis. But we cannot hide the fact that he is still distrustful if the technical aspects of the algorithms used are not very clear to him. Thus, he may be at a loss when faced with two good data analyses showing two slightly different aspects which can, for example, proceed from two classification trees obtained by changing the probabilistic nature of the non-link hypothesis, with respect to which the proximities are established.

These remarks relate to a general impression about the specialist's attitude. This opinion, which has a relative value, must be confirmed by

considering other observations of the collaboration between the statistician and the scientist of a given discipline in the humanities. On the other hand, there are probably different types of behavior of the specialist, depending on his training and experience. We do not dare to suggest here a psycho-sociological survey where we would show by classification theses types of behavior; it would bring us back to the beginning.

References

Benzecri, J.P. (1974). *L'Analyse des Données*. Paris: Dunod.

Berge, A. and Denjean, G. (1974). "Comportement digestif et fonctionnement intellectuel," *Revue de Neuropsychiatrie Infantile* 22(6): 355–370.

Chombard de Lauwe, M.J. and Bellan, C. (1979). *Enfants de l'image. Enfant personnage des medias, enfant réel.* Paris: Payot,

Cohen, I. (1977). Classification d'une famille d'echelles au moyen d'un nouveau indice. Comparaison avec le traitement par l'analyse des correspondances. Application à des données en psycho-pédagogie et sociologie rurale. Thèse de 3ème cycle, Université de Paris VI, (I.S.U.P.).

Daniels, H.E. (1944). "The relation between measures of correlation in the universe of sample permutations," *Biometrika* 33: 129–135.

Goodman, L.A. and Kruskal, W.H. (1954). Measures of association for cross classifications," *J.A.S.A.* 49 (December): 732–764.

Goodman, L.A. and Kruskal, W.H. (1963). Measures of association for cross classifications, approximate sampling theory," *J.A.S.A.* 58 (June): 310–364.

Hájek, J. (1961). "Some extensions of the Wald–Wolfowitz–Noether theorem," *Ann. Math. Stat.* 32: 506–523.

Hartigan, J.A. (1975). *Clustering Algorithms*. New York: John Wiley.

Jardine, N. and Sibson, R. (1971). *Mathematical Taxonomy*. New York: John Wiley.

Kendall, M.G. (1970). *Rank Correlation Methods*. London: Charles Griffin, fourth edition.

Lecalvé, G. (1976). Problèmes d'analyse des données, 2nd part of thesis, Université de Rennes I.

Leredde, H. (1979). La méthode des "pôles d'attraction"; une nouvelle famille d'algorithmes de classification, thèse de 3ème cycle, Université de Paris VI, Institut de Programmation.

Lerman, I.C. (1967). "Analyse hiérarchique," *Revue Mathématique et Sciences Humaines* 17: 37–46. Also chpt. 8 of (Lerman, 1974/75).

Lerman, I.C. (1970). *Les bases de la classification automatique*. Paris: Gauthier-Villars, collection Programmation.

Lerman, I.C. (1972). "Analyse du phénomène de la "sériation", *Revue Mathématiques et Sciences Humaines* 38: 39–57. Also chpt. 7 of (Lerman, 1974/75).

Lerman, I.C. (1973a). "Etude distributionnelle de statistiques de proximité entre structures finies de même type; Application à la classification automatique," *Cahiers du B.U.R.O.* 19: 1–52.

Lerman, I.C. (1973b). "Introduction à une méthode de classification automatique, illustrée par la recherche d'une typologie des personnages enfants à travers la littérature enfantine," *Revue de Statistique Appliquée*, XXI(3): 23–49.

Lerman, I.C. (1974/75). Cours sur la reconnaissance et classification des structures finies en analyse des données, Université de Rennes I.

Lerman, I.C. (1976). "Formal analysis of a general notion of proximity between variables," in J.R. Barra and al. (eds.) (1977), *Recent Developments in Statistics*. Amsterdam: North Holland.

Motoo, M. (1957). "On the Hoeffding's combinatorial central limit theorem," *Ann. Inst. Stat. Math.* 8: 145–154.

Nicolaü, F. and Nicolaü, M.H. (1972). Analyse d'un algorithme de classification " and "Contributions au traitement automatique de données," thèses de 3ème cycle, Université Paris VI, I.S.U.P.

Noether, G.E. (1949). "On a theorem by Wald and Wolfowitz," *Ann. Math. Stat.* 20: 455–458.

Pieraut-le-Bonniec, G. and Van Meter, K. (1976). "Etude génétique de la construction d'une propriété relationnelle: la relation de passage," *Monographies Françaises de Psychologie*, No. 35. Paris: C.N.R.S.

Wald, A. and Wolfowitz, J. (1944). "Statistical tests based on permutations of the observations," *Ann. Math. Stat.* 15: 358–372.

Appendix – Types of Child Characters through Children's Literature

The real example that we deal with here only illustrates the hierarchical classification method. Indeed it is not a trivial example; but certainly, it is one of the easiest examples that we had to treat. The problem arises in the psycho-sociological field and is situated in the framework of research led by Chombard de Lauwe (1979). The purpose is to determine the models created by the adults and proposed to the child in the literature concerning the child. The data consists of 1500 child personages provided by a sample of books published on this literature; each of them being determined by criterion put forth by the psychosociologist. These publications are edited between 1880 and 1960; as a matter of fact, the author of the research wanted to recover three consecutive periods; before the first war, between the two wars and after the second war. Each subject is described by a set A of attributes (cf. section II.1) which establish the following:

(a) The portrait

Aptitudes (knowledge, memory, particular gifts . . .); the fields where the aptitudes are demonstrated (art, sport, command, adventure, first aid, practical jokes, way of life, scholastic results, . . .); religion (Chris-

tian, believes in God, not specified); temperament (patience, anger, coquetry, impulsiveness, frankness, gaiety, . . .); relations with others (submission, equality or command with respect to another child, equality or command with respect to an adult, . . .); the role of the subject with respect to action (active, passive); great topics where he evolves (childhood drama, moral teaching, vocation, adventure, . . .); general appreciation of the subject (positive, firstly negative but becoming positive, essentially negative).

(b) The material environment

Town, castle, vehicle habitation (e.g. house on wheels, caravan), vehicle action (e.g. car, aeroplane), wild nature, road track, living abroad . . .

(c) His close surroundings

Father, mother, substitute father, substitute mother, presence of a rival child, fried of the same or opposite sex, animal, presence of a policeman (representative of "law"), presence of a traitor or bandit (representative of "bad"), . . .

(d) His family atmosphere

Not described, absence of family, incomplete or replacing family, marked by a type of adult surrounding (e.g. army), community of children.

(e) His social environment

Labourer, aristocracy, personage impossible to locate in the described social context, . . .

Thus, 110 attributes were retained for the processing that followed. The basic information is contained in an incidence table T with 1500 rows (set E of child personages) and 110 columns (set A of attributes). We consider the data of this table to be that of the family or the sample of the subsets $\{E_a/a \in A\}$ of the set $P(E)$ of all subsets of E, where it is clear that E_a is the subset of the subjects having the attribute a.

As mentioned in section III.1 we put in two forms N_1 and N_2 the non-link hypothesis with respect to which the proximity index is calculated in the form of a probability. For N_1, we associate with the subsets E_a and E_b of objects possessing attributes a and b respectively two independent random subsets X' and Y' defined in such a way that each individual has a probability n_a/n (n_b/n) of belonging to X' (Y') independently of others, where n_a = card (E_a) and n_b = card (E_b).

For N_2, we have already developed (cf. section III.1) one of the subsets E_a or E_b – suppose, for example, E_a – which is fixed and with E_b is associated a random subset Y belonging to a set of all equally probable subsets of E having the same cardinality n_b. Let us recall that the distribution of the r.v. $S_a = \text{card}\ (E_a \cap Y)$ is the same as that of $S_b = \text{card}\ (X \cap E_b)$ obtained by fixing E_b and associating with E_a the random subset X. Denoting by S_2 this common variable and by S_1 the r.v. card $(X' \cap Y')$, S_1 and S_2 have the same mean $\mu = n_a n_b/n$.

The variance of S_1 is approximately equal to its mean, and that of S_2 can be put in the form $\sigma^2 = n_a(n - n_a)\ n_b(n - n_b)/n^2(n - 1)$. The corresponding measures obtained by standardizing the rough index $s = \text{card}\ (E_a \cap E_b)$ are written as:

$$S_1(a, b) = (s - \mu)/\sqrt{\mu} \text{ and } S_2(a, b) = (s - \mu)/\sigma$$

It can be verified that the first measure gives more weight to the positive associations than to the negative ones whereas the second one is symetric (i.e. positive and negative associations have the same importance. (cf. eqn. (3)).

Referring to a probability scale, we may take the following measures:

$$P_1(a, b) = P_r^{N_1}\{S_1 < s\} \text{ or } P_2(a, b) = P_r^{N_2}\{S_2 < s\}$$

It is easy to calculate $P_1(a, b)$ or $P_2(a, b)$, if we notice that $(S_1 - \mu)/\sqrt{\mu}$ and $(S_2 - \mu)/\sigma$ are approximately distributed as standard normal variates. However the observed variance may be so high as to necessitate the replacement of the hypotheses N_1 and N_2 by N_1' and N_2', respectively, where $(S_1 - \mu)/\sqrt{\mu}$ and $(S_2 - \mu)/\sigma$ are distributed normally with mean 0 and with variances λ_1, and λ_2 respectively, such that the highest observed value of $|S_i(a, b)/\lambda_i|$; $(i = 1, 2)$ be less than or equal to 2.5 (Nicolaü, 1972).

This allows a sufficient degree of discrimination with the scale defined by normal distribution, the measures of proximity being calculated by the following formulae:

$$P_1'(a, b) = \int_{-\infty}^{(s-\mu)/\sqrt{\mu\lambda_1}} \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-t^2}{2}\right) dt,$$

$$P_2'(a, b) = \int_{-\infty}^{(s-\mu)/\sigma\sqrt{\lambda_2}} \tfrac{1}{2} \exp\left(\frac{-t^2}{2}\right) dt$$

In connection with the data presented above we consider the reduction of two "trees" obtained by the Likelihood Link Algorithm (L.L.A.) and associated, respecively, with the hypotheses N_1' and N_2'.

These "trees" were produced and interpreted by Mrs. M.H. Nicolaü with the assistance of Cl. Bellan.

(a) THE TREE ASSOCIATED WITH L.L.A. ESTABLISHED WITH RESPECT TO N_1'

This is a binary tree, having 109 levels (there are 110 elements). The levels retained for condensing the tree and which correspond to the local maxima of the increment rate θ of the global statistic Σ are numbers 8, 12, 41, 55, 71, 86, 90, 93, 99, 102, 104, 106, 107 (giving the absolute maximum of Σ) and number 108.

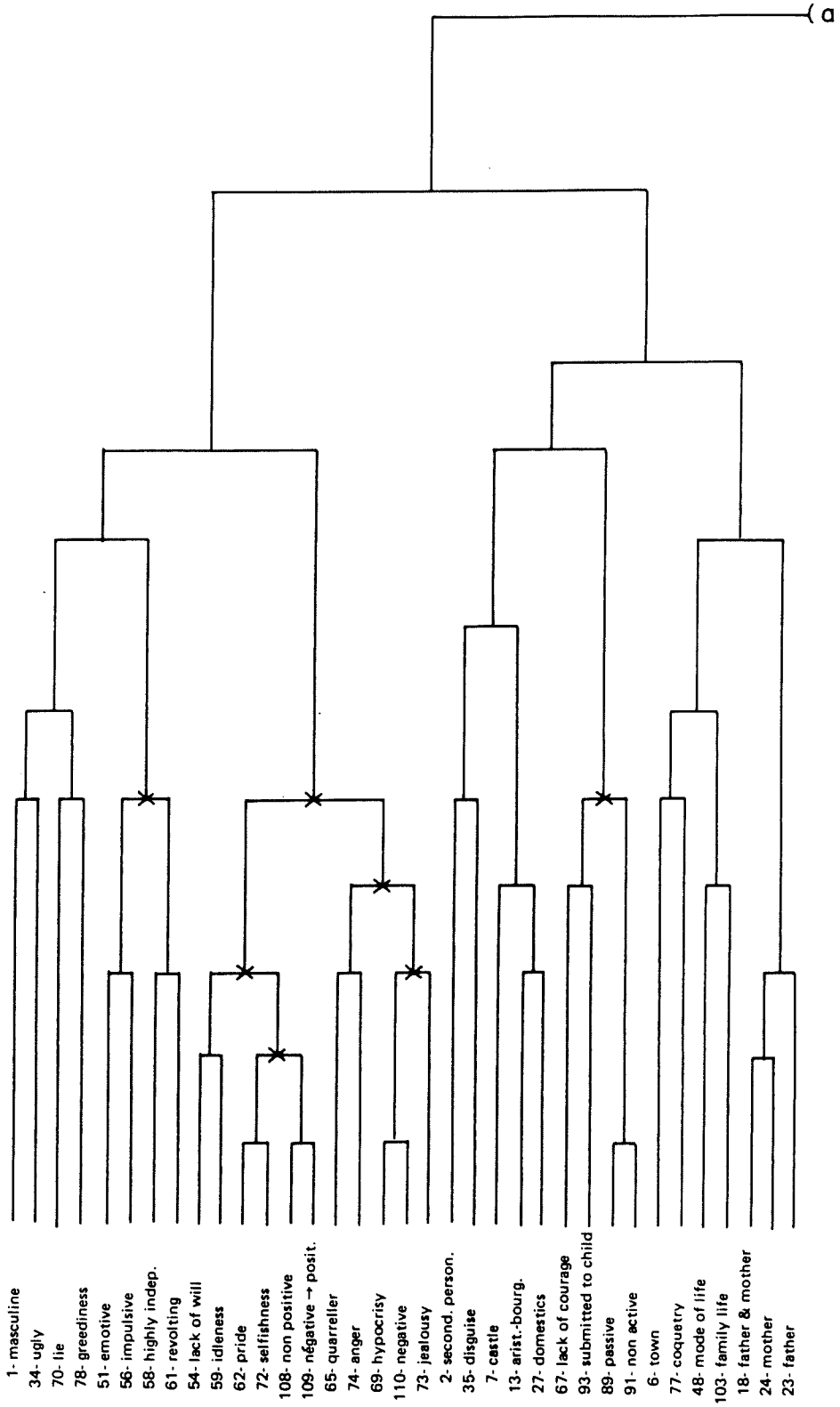
From the 41st level onwards some interesting types start emerging; there appears a group consisting of the attributes such as pride and selfishness, non-positive and negative character evolving towards the positive one, idleness and lack of will defining a type of personage whose character is "full of defaults and weak," but not fundamentally bad because he evolves towards the good. At the 55th level another form of negative character becomes clear, made up in the beginning by the kernel of negative, hypocrisy, and jealousy, to which are joined the disagreeable external manifestations of anger and quarelling; it concerns the "most abject aspect of negativeness". These two classes are merged at the 71th level giving rise to a more general negative type. Meanwhile many kernels are formed. At the same level (71) "agressive" appears with the rebellious and highly independent associated with impulsive and emotional. At the 86th level one perceives clearly a group of "minor defaults": greediness and tendency to lie being peculiarly associated with an ugly boy. The "aristocrate bourgeois" composed of the features: aristocracy, great "bourgeoisie", servant, castle; the "droll" passive in the action, lack of courage, obeying a child; the "comic": comic personage, caricatural, set in time, joke, funny; the "authoritative": authority, commandment, commanding a child, equal to the adult, commanding an adult; the "adventurer": adventure, adventure-mission, traitor-bandit, police investigation; the "celebrated": distant past, vocation, art, future perspectives, aspiration, plans his future, very gifted; the "brilliant": knowledge, memory, temperate, scolarly success, child community; the "sensitive": sensitive to the environment, likes nature, presence of an animal, animal care; the "good Christian": Christian, faith, believes, modesty, patience, perfect manners, obeying an adult; the "unhappy": sadness, lonely, fragile, tries to gain affection, moral advice, childhood drama, family absence (difficulties), incomplete or substitute family. Besides, there appeared many less marked aggregates (clusters) consisting mostly of the rela-

tively neutral attributes (cf. section 3); moreover it should be noted that some of the unions that gave rise to the above-mentioned classes correspond to the levels at which the global statistic Σ shows a decrease, which means that considered on the whole the partition of this level is less in accordance with the initial "ordonnance" than that of the preceding level. In this case we prefer not to insist upon the significance of such groupings, an example of this being the association of greediness with an ugly boy. Between the levels 86 and 99 the small clusters will join together and give birth to some more general types, the meaning of which will become clearer.

The 99th level is the most important next to the most significant (107th) level; because on the one hand, the increment θ of the statistic Σ at this level is high compared to the set of values of θ at all levels and because the unions at the following levels (100 and 101) lead to clear fall of this statistic, on the other. We shall therefore take as the most relevant classifications those given by the 107th and the 99th levels. To the latter is first assigned a large class resulting from the aggregation of the "agressive" type and the "minor defaults" with the "negative" to form a quite general "negative" type; the "droll', "comic" and "adventurer" types are enriched with some attributes; the "sportive" and the "authoritative" now appear together; the command seems to be present in the sport, or more generally in the action. "Celebrated" and "brilliant" join together to give rise to "model-child" type; elsewhere the "good Christian" is connected with the "sensitive", the "unhappy" with the "fragile". The last two types are formed: the one that will be called the "scout hero" defined by: first-aid practice, rescue, justice maker, courage, controlled, initiative, volunteer, stubborn; the other one representing the "loyal friend, good comrade" composed of candour, honesty, generosity, positive, gaiety, friend of opposite sex, substitute father, substitute mother. These two classes, in fact, have a very weak internal cohesion, the attributes of which they are constituted being sufficiently neutral.

Finally at the 107th level a classification with three wide classes appears. The first one regroups the "negative" and the "droll" and thus represents a very general "negative" type. The second one embodies the "comic", the "sportive-authoritative", the "adventurer" and the "scout hero" defining the usual character of escapist literature. The third class gathers together the "model child", the "good Christian", the "unhappy" and the "loyal friend" defining a general "normative" type. The last levels where these classes are joined by two's, lead to the abrupt fall in the global statistic.

(continued on p. 465)


a
1- masculine
34- ugly
70- lie
78- greediness
51- emotive
56- impulsive
58- highly indep.
61- revolting
54- lack of will
59- idleness
62- pride
72- selfishness
108- non positive
109- négative → posit.
65- quarreller
74- anger
69- hypocrisy
110- negative
73- jealousy
2- second. person.
35- disguise
7- castle
13- arist.-bourg.
27- domestics
67- lack of courage
93- submitted to child
89- passive
91- non active
6- town
77- coquetry
48- mode of life
103- family life
18- father & mother
24- mother
23- father

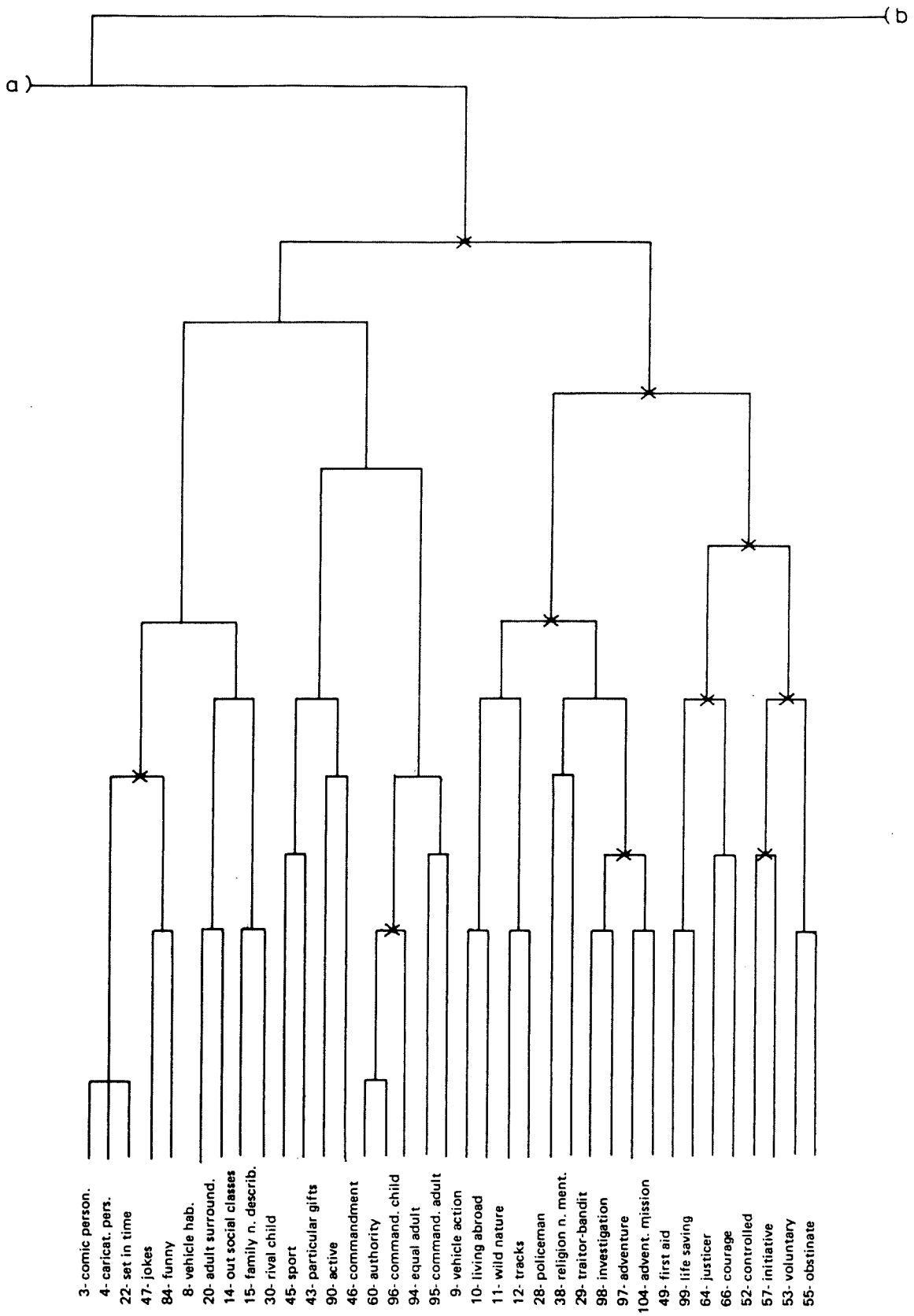
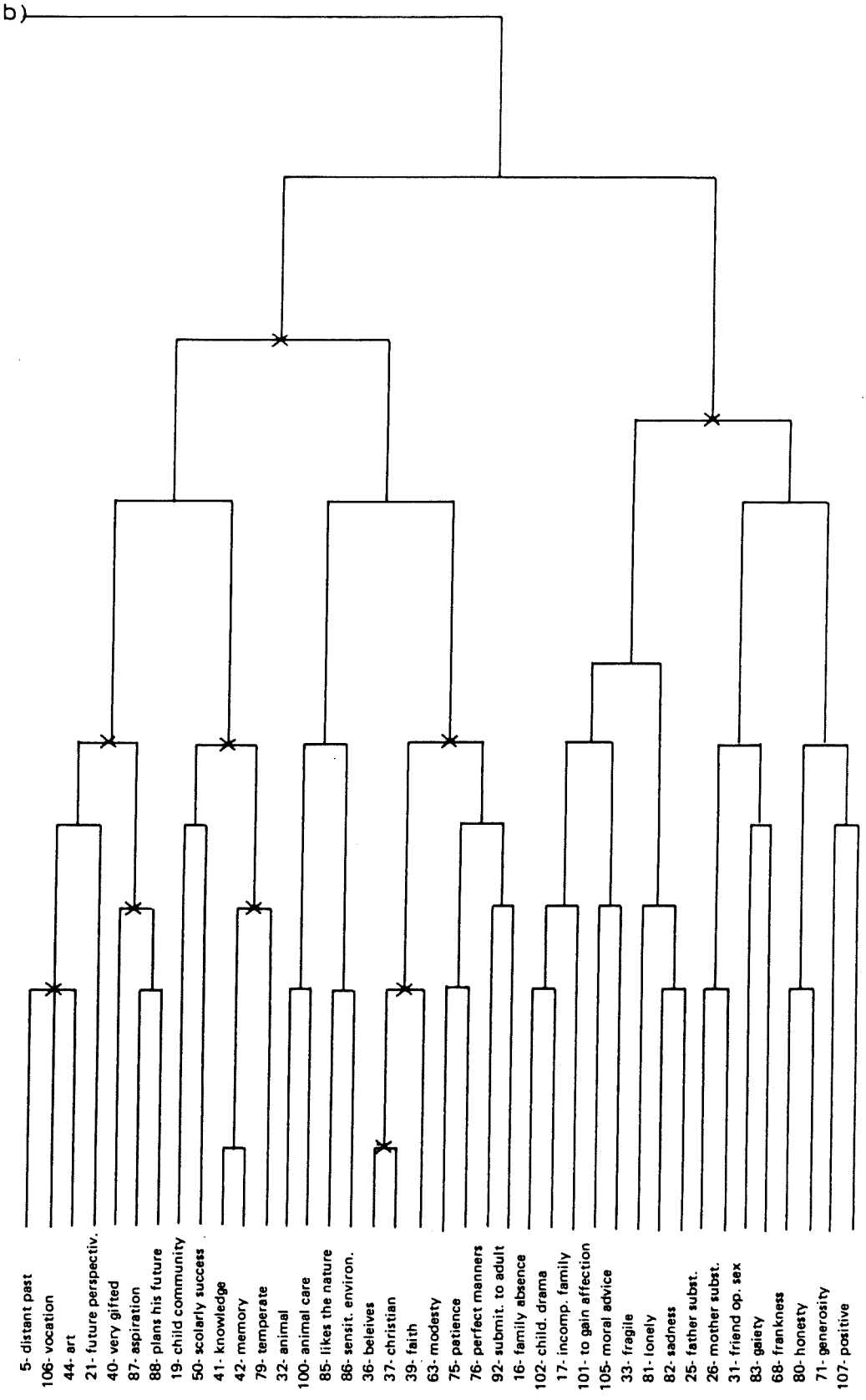


Fig. 7. Reduced tree associated with L.L.A. established with respect to N_1'. The sign X indicates "significant node".

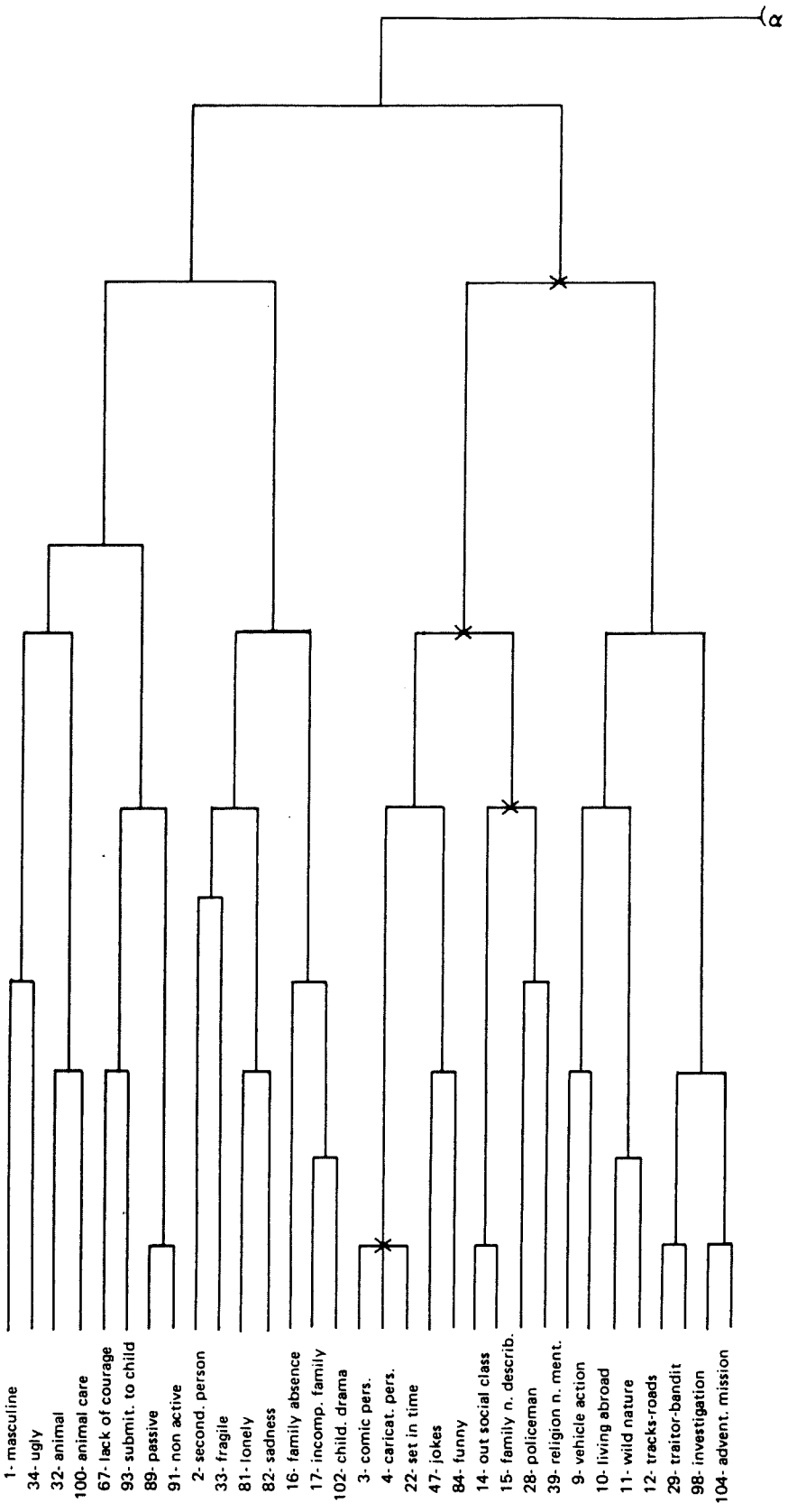
b)


5- distant past
106- vocation
44- art
21- future perspectiv.
40- very gifted
87- aspiration
88- plans his future
19- child community
50- scolarly success
41- knowledge
42- memory
79- temperate
32- animal
100- animal care
85- likes the nature
86- sensit. environ.
36- beleives
37- christian
39- faith
63- modesty
75- patience
76- perfect manners
92- submit. to adult
16- family absence
102- child. drama
17- incomp. family
101- to gain affection
105- moral advice
33- fragile
81- lonely
82- sadness
25- father subst.
26- mother subst.
31- friend op. sex
83- gaiety
68- frankness
80- honesty
71- generosity
107- positive

(b) TREE ASSOCIATED WITH THE L.L.A. WITH RESPECT TO N_2'

We examine the reduction of this tree at the significant levels (viz. 23, 32, 49, 60, 78, 82, 91, 96, 98, 100, 103, and 107). The most significant level is the 105th, the 103rd gives the last pertinent local maximum of the global statistic Σ before its absolute maximum. At the 103rd level emerges the "negative" type composed of the following subtypes: "character full of defaults and feeble" found in the preceding tree; the "bad"; the "agressive" which seems more successfully to complete itself with the attribute anger that was associated with the "bad" in the preceding tree; and finally a subtype difficult to define that seems to represent the child's setting in the family and town marked by unconcern and lightness (family life, practical life, town, disguise, coquetry, greediness . . .). At this 103rd level the "comic" and the "adventurer" are already joined. The "comic" has given the vehicle abode and adult surrounding to the "sportive-authoritative" and has taken the policeman and religion not mentioned from the "adventurer". It concerns the attributes which, according to the psychologists, appear in the literature associated with any one of these three types. The "adventurer" is composed of two groups as before; the first characterises the atmosphere in which he lives; adventure, adventure-mission, traitor-bandit, research and the second one his physical environment: action-vehicle, lives in a foreign country, wild nature, road-tracks. The "unhappy" is here a more global type that contains the "droll" and a group with low cohesion: masculine, ugly, animal, animal care. Association of the "droll" with the "unhappy" might be explained by obedience in misfortune. The whole represents the unfortunate character, submissive and ugly who seeks compensation by taking care of an animal. Masculine and ugly that belonged to the "negative" in the previous tree have joined the "unhappy" and that corresponds to the fact that in child literature the ugliness of a little boy is generally linked with bad character or bad fortune, sometimes also with the "comic" (the "clown"; result of a tree not mentioned here). Similarly the attribute of the secondary character (which in the first tree were joined to the "droll", and now to the "fragile") represents according to the psychologist a character established to bring out a weak main personage. The "good Christian example" is split into two parts: the set of qualities modesty, patience, sensitive to the environment is aggregated to the "brilliant" and the "gifted" to form the type of "the model child" and the embryo "believes", "Christian", "faith" unites with the "celebrated" and the "aristocratic bourgeois" giving rise to a type that seems to represents the personage of the "celebrated

(continued on p. 469)


α
1- masculine
34- ugly
32- animal
100- animal care
67- lack of courage
93- submit. to child
89- passive
91- non active
2- second. person
33- fragile
81- lonely
82- sadness
16- family absence
17- incomp. family
102- child. drama
3- comic pers.
4- caricat. pers.
22- set in time
47- jokes
84- funny
14- out social class
15- family n. describ.
28- policeman
39- religion n. ment.
9- vehicle action
10- living abroad
11- wild nature
12- tracks-roads
29- traitor-bandit
98- investigation
104- advent. mission

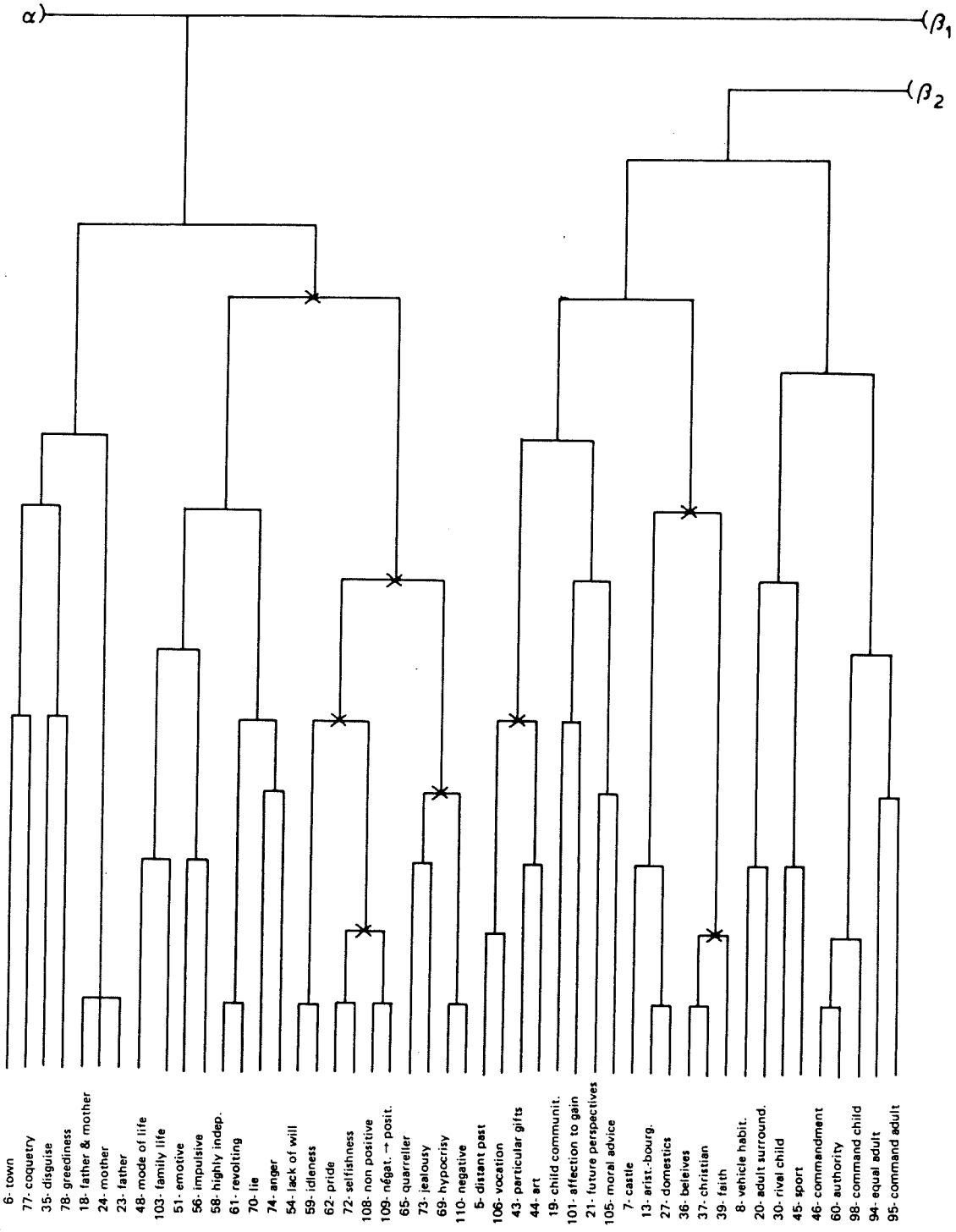
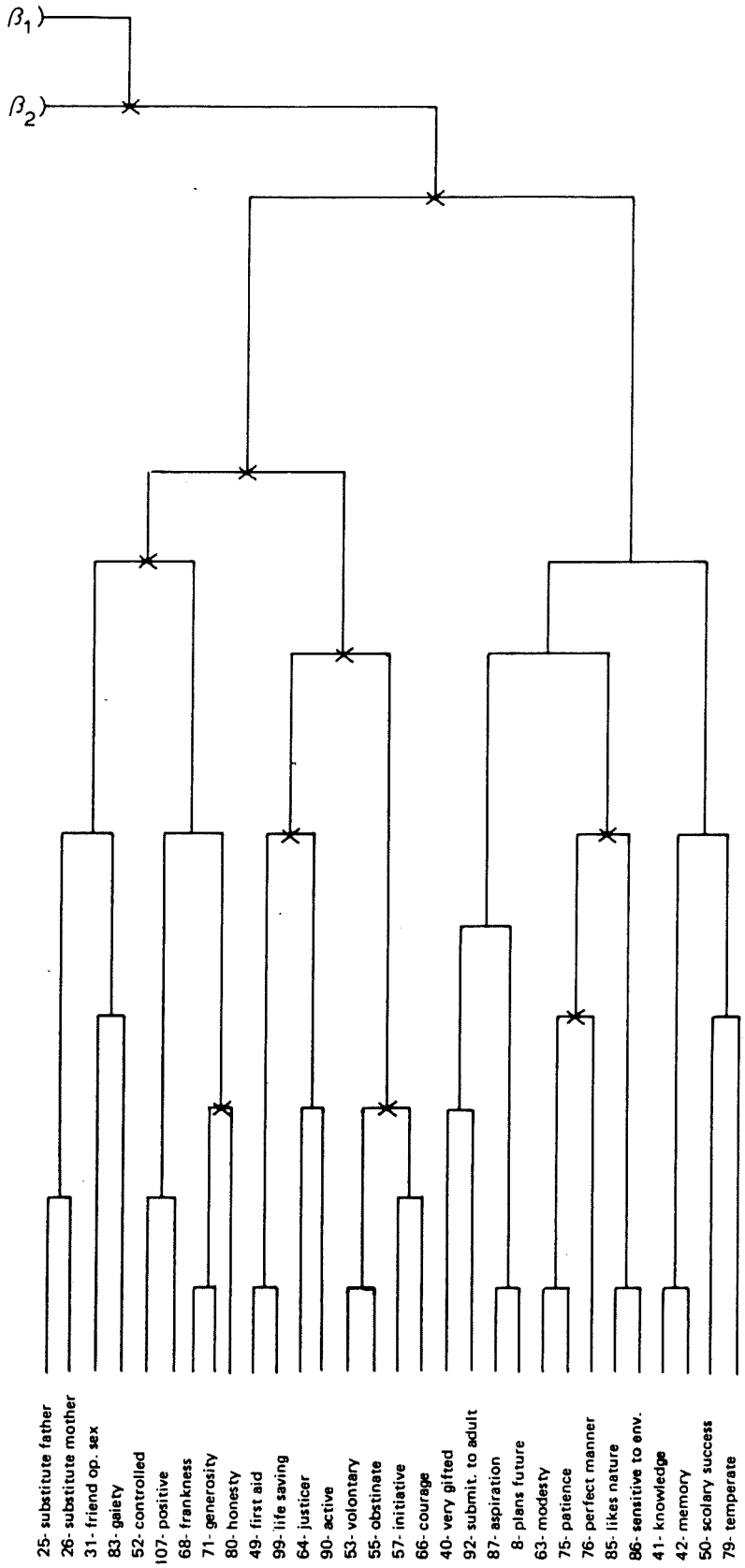


Fig. 8. Reduced tree associated with L.L.A. established with respect to N_2'. The sign X indicates "significant node".


β_1)
β_2)
25- substitute father
26- substitute mother
31- friend op. sex
83- gaiety
52- controlled
107- positive
68- frankness
71- generosity
80- honesty
49- first aid
99- life saving
64- justicer
90- active
53- volontary
55- obstinate
57- initiative
66- courage
40- very gifted
92- submit. to adult
87- aspiration
8- plans future
63- modesty
75- patience
76- perfect manner
85- likes nature
86- sensitive to env.
41- knowledge
42- memory
50- scolary success
79- temperate

children's biographies". Finally the "scout hero" and the "loyal friend, good comrade" join giving an account of another kind of model child; if the first form is centred on the spiritual qualities, the second one is centred rather on the sentiments. At the most significant level (105th), we have a partition in five classes: the "misfortunate", the "comic-adventurer", the "negative", the "model" and a group resulting from the union of the "celebrated"-"bourgeois"-"Christian" with the "sportive-authoritative", the previous constitution of which was accompanied by a local minimum of the global statistic.

If we are now interested in the sequence of values of Σ associated with each of the trees, it is verified that the groupings responsible for the most visible dissimilarities between the two trees correspond, in general, to the local minima of Σ that is the case, for example when:

(i) {Masculine, ugly} joins in the first tree with {lie, greediness} and in the second with {animal, animal care}.

(ii) In the first tree "droll" is associated with "aristocratic bourgeois" and when the whole is joined to town, coquetry, practical life . . . , in the second tree when "droll" is associated with "the ugly little boy . . ." and then with the "fragile".

In the first tree the union of "loyal friend-good comrade" with the "misfortunate" is also accompanied by a minimum of Σ, as well as that of the "unhappy" with the "comic adventurer" in the second tree.

We are here concerned with the associations which are not in perfect accordance with the information on the resemblances defined by the preordonnance and where the neutral attributes play an important part. Other groupings where Σ does not have the local minima, but that vary from one tree to the other, are due to the presence of the "neutral" attributes (cf. section VI); the classes then give rise to interpretations equally justified but faintly visible. One realizes that the identification of the neutral attributes in the set A is a precious aid for the interpretation of the results.