STICHTING
MATHEMATISCH CENTRUM
2e BOERHAAVESTRAAT 49
AMSTERDAM
REKENAFDELING

MR 108

On the hierarchical decomposition of complexity

by

M.H. van Emden

July 1969

Printed at the Mathematical Centre, 49, 2e Boerhaavestraat, Amsterdam, The Netherlands.

The Mathematical Centre, founded the 11-th of February 1946, is a non-profit institution aiming at the promotion of pure mathematics and its applications; it is sponsored by the Netherlands Government through the Netherlands Organization for the Advancement of Pure Research (Z.W.O) and the Central Organization for Applied Scientific Research in the Netherlands (T.N.O), by the Municipality of Amsterdam, by the University of Amsterdam, by the Free University at Amsterdam, and by industries.

ACKNOWLEDGEMENT

The reserach reported here was supported by the Hugo de Vries Laboratory for Systematic Botany in the University of Amsterdam as part of its program to develop mathematical methods for analysis of plant-ecological data. The author wishes to thank professor Dr.A.D.J. Meeuse, the Director of the Laboratory and Dr.S. Segal for the freedom and encouragement they have given him during the course of the work.

ON THE HIERARCHICAL DECOMPOSITION OF COMPLEXITY

0. SUMMARY

Classification methods for quantitative data have received more attention than those for qualitative data. Excess-entropy, which may be interpreted as a measure of complexity, enables us to formulate existing methods for normally distributed data in such a way as to be applicable also to qualitative data.

After an introductory section 1, section 2 defines excess-entropy and provides some information-theoretical background. It then treats the qualitative case by methods analogous to principal components and clustering respectively. The first of these is much like the existing technique of Association Analysis.

Section 3 is concerned with the multivariable normal distribution. The well-known method of principal components is given a simple interpretation in terms of entropy. Furthermore, excess-entropy is shown to be identical with the log likelihood ratio statistic applicable when testing for dependence between sets of random variables.

In section 4 it is shown that in Markov chains excess-entropy provides a measure of clustering. In order to be able to do so, an operation on a Markov chain has to be defined: that of "fusing" two states.

1. THE ANALYSIS OF COMPLEX SYSTEMS.

We may think of a "system" as a set of variables influencing each other. Complexity may arise in two ways: the presence of a large number of variables and the fact that most of these influence many others.

There exist situations where the simultaneous treatment of all variables presents a computational problem that is too large by any standard. Yet in such a situation it is sometimes possible to decompose the whole system into a few subsystems with relatively weak interactions between them. At this level we have a system of manageable complexity where the subsystems are treated as "black boxes".
In their turn, each of these subsystems may be subjected to the same treatment, and so on. This process is just a particular case of the well-known principle: "divide and rule"; or, as we shall encounter it as a recurrent theme: "hierarchical decomposition of complexity".

In this study we want to see what can be done by viewing the interaction between subsystems as "information transfer". The decomposition of complexity then corresponds to the decomposition of the total amount of information transfer.

1.1. EXAMPLE: A SET OF LINEAR ALGEBRAIC EQUATIONS.

Let

(1)... $Ax = b$

represent a set of n linear algebraic equations in n unknowns. A is an $n \times n$ - matrix and $x^T = (\xi_1, \ldots, \xi_n)$, $b^T = (\beta_1, \ldots, \beta_n)$ are n-component vectors.

Consider a partition $\begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}$ of A, where A_{11} is a $k \times k$ - matrix.

With (x_1^T, x_2^T), (b_1^T, b_2^T) as the corresponding partitions in x^T and b^T, we can write (1) as:

(2)... $$\begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}.$$

Suppose that variables $\xi_1, \ldots, \xi_k$ are but "weakly" represented in the last n - k equations, i.e. the elements of A_{21} are small compared to those of A_{11} or A_{22}. In such a situation it may be advantageous to use the following iterative scheme for solving (1):

(3)... Start with: $x_2^0 := b_2/A_{22}$ (In this notation we denote the solution vector of: $A_{22}\, x_2^0 = b_2$);

For k = 0, 1, 2, ... do:

(4)... $x_1^{k+1} := (b_1 - A_{12}\, x_2^k)/A_{11}$;

(5)... $x_2^{k+1} := (b_2 - A_{21}\, x_1^{k+1})/A_{22}$;

The sequence of : $\begin{pmatrix} x_1^1 \\ x_2^1 \end{pmatrix}$, $\begin{pmatrix} x_1^2 \\ x_2^2 \end{pmatrix}$, ... is regarded as a sequence of approximations to the solution of (1).

If A_{21} consists of zeroes only, the solution is obtained after (3) and a single execution of (4). It seems reasonable to suppose that this scheme converges faster when the elements of A_{21} are smaller. In general, the solution is not obtained after a finite number of steps because a change in x_2 is transmitted to x_1 via A_{12} in step (4), and then the change in x_1 is transmitted to x_2 via A_{21} in (5), and so on. Viewed in this way, A_{12} and A_{21} represent the interactions between the subsystems A_{11} and A_{22}.

It is desirable to find a quantitative description of this interaction. In a similar situation (see next example) an "information transfer" may be defined between subsystems.

In section 3.4 and 4.2. we exhibit special systems of linear equations where we can express the interaction between subsystems as a quantity of information in the usual interpretation of this concept (see 2.2).

1.2 EXAMPLE: A MODEL OF THE DESIGN PROCESS.

Let us consider the following abstraction of a complicated design problem: a designer has to construct a "form" which has to satisfy a large number of conditions.

For example, we might think of the design of a human settlement where the number of conditions may run in the hundreds, many of which are conflicting. Here again complexity may arise in two ways: the number of variables is large and there occur many interactions between them. In general there may not exist a form which satisfies all conditions to the required extent, so the designer should aim at maximizing goodness-of-fit with respect to all conditions simultaneously.
The designer cannot keep in mind all of the conditions at once; suppose he finds an iterative design process by first concentrating on some subset A_{11} of the conditions, finding a provisional form that maximizes goodness-of-fit locally and then proceeding with another subset A_{22}. Interaction between condition i and condition j arises in the following way: In adapting the form to condition i, it may be modified in such a way that goodness-of-fit with respect to condition j decreases.
We see that the iterative design process sketched above is analogous to the iterative method of solving a system of equations. Suppose that conditions are partitioned into subsets A_{11} and A_{22}, then the designer first ignores A_{22} and then A_{11}. When there is no interaction between the two, he is done. In general he finds on returning to A_{11}, that, while concentrating on A_{22}, he has undone some of the good properties his provisional form had with respect to A_{11} and he will start a following cycle of the iterative process. Even when there is some interaction between A_{11} and A_{22}, this process may succeed in yielding a satisfactory form after an acceptable number of cycles.
Alexander [1] has studied the problem of finding subsets in the set of all conditions in such a way that the amount of interaction between is small compared to interaction within subsets. He quantified "interaction" by regarding it as "information transfer". To this end he constructed a model consisting of a set of random variables corresponding to conditions. He was then able to define "information transfer" as a difference between entropies. He reports the existence of computer programs for the hierarchical decomposition of the set of conditions, where their interactions are specified pair by pair.

2. ENTROPY AND OBJECT-PREDICATE TABLES.

2.1 THE OBJECT-PREDICATE TABLE.

Suppose we have a certain set of "objects" and each of these may be described by stating whether it does or does not have any of a fixed (same for all objects) set of "predicates". In this way each object is identified with a certain subset of the predicates; when two objects have an identical subset there is, in this context, no way to tell them apart.

This situation may be represented by an "object-predicate table": a rectangular array of noughts and crosses. The j-th cell of the i-th row of this array shows whether the i-th object does (when it contains a cross) or does not (when it contains a nought) possess the j-th predicate.

predicates →

objects ↓

i \ j	1	2	3	4	5	6
1	0	0	x	0	0	0
2	0	x	0	0	x	x
3	0	x	x	x	x	0
4	x	0	0	0	x	0

AN OBJECT-PREDICATE TABLE

The object-predicate table is a rather general scheme for exhibiting relations between objects, either via (common) predicates or, directly, by identifying the i-th predicate with the i-th object. Nought or cross then indicates whether the one object is dependent on the other. An example of a "system" would then be a set of objects related as specified by their object-predicate table.

2.2 THE ENTROPY FUNCTIONAL

In order to provide a conceptual framework and a terminology for what follows, we will first review some important properties of the entropy functional H (Khinchin [3]).

Suppose there are two sets of descriptions of events

$A = \{a_1, \ldots, a_m\}$ and $B = \{b_1, \ldots, b_n\}$.

A probability is assigned to each description:

$P_r\{a_k\} = p_k \geq 0 \ , \sum_k p_k = 1, \qquad k = 1, \ldots, m$ and

$P_r\{b_l\} = q_l \geq 0, \ \sum_l q_l = 1, \qquad l = 1, \ldots, n.$

Thus a_k and b_l are sets of events having an identical description, i.e. in this context events belonging to the same set cannot be distinguished. In the sequel we will therefore denote such a set of events as "event". The entropy functional H associated with $\{p_1, \ldots, p_m\}$ is defined as:

(1) ... $H(A) = - \sum_k p_k \log p_k$.

H may be interpreted as a measure of uncertainty with respect to the outcome of an experiment A, which is the event a_k with probability p_k. The following two properties of H justify such an interpretation:
H is never negative and vanishes if $p[k] = 1$ for some $k = 1, \ldots, m$. In this case a_k is certain to occur; the uncertainty vanishes. The other property is that H attains its maximum when all events are equally probable, which corresponds to the situation of maximum uncertainty. This property follows from the well-known inequality:

(2) ... $f(\sum_k \lambda_k x_k) \leq \sum \lambda_k f(x_k)$

where $\lambda_k \geq 0, \ \sum_k \lambda_k = 1$ and f is a continuous and convex function of x. We now choose $\lambda_k = {}^1/m$, $f(x) = x \log x$ and $x = p$.

$$f(\sum_k \lambda_k x_k) = \frac{1}{m} \log \frac{1}{m} \leq \sum_k \frac{1}{m} p_k \log p_k \implies$$

$$H(A) = - \sum_k p_k \log p_k \leq \log m.$$

Let us also consider the Cartesian product set A x B of the two sets. On A x B a two-dimensional array of probabilities is defined as: $P_r\{a_k \text{ and } b_l\} = r_{kl}$. The associated conditional probabilities are: $P_r\{b_l \mid a_k\} = q_{kl} = r_{kl}/p_k$.
(the probability that b_l will occur under condition that a_k has occurred).

The two sets are said to be independent when $r_{kl} = p_k\, q_l$. In that case $q_{kl} = q_l$ which means that the probability of the occurrence of b_l is independent of which a_k, k = 1, ..., m, has occurred. For the entropy of the product scheme we have:

$$H(A \times B) = -\sum_{kl} r_{kl} \log r_{kl}.$$

In the case of independence this reduces to:

$$H(A \times B) = -\sum_{kl} p_k\, q_l (\log p_k + \log q_l).$$

$$= -\sum_{l} q_l \sum_{k} p_k \log p_k - \sum_{k} p_k \sum_{l} q_l \log q_l$$

$$(3) \ldots \quad H(A \times B) = H(A) + H(B).$$

In case A and B are dependent, this relation generalises to:

$$H(A \times B) = -\sum_{kl} r_{kl} \log r_{kl} = -\sum_{kl} p_k\, q_{kl} \log p_k\, q_{kl}$$

$$= -\sum_{kl} p_k\, q_{kl} \log p_k - \sum_{kl} p_k\, q_{kl} \log q_{kl}$$

$$= H(A) + \sum_{k} p_k\, H_k(B)$$

$H_k(B)$ is regarded as the outcome of a random variable : The entropy of the conditional scheme $\{q_{k1}, \ldots, q_{kn}\}$ under condition that a_k has occurred. The second term is then the mathematical expectation of H(B) in the scheme A, which we shall designate by $H_A(B)$:

$$(4) \ldots \quad H(A \times B) = H(A) + H_A(B) \text{ and similarly}$$

$$H(A \times B) = H_B(A) + H(B)\ .$$

$H_A(B)$ never exceeds $H(B)$. This is a consequence of the inequality (2) where this time we take $\lambda = p$ and $f(x) = x \log x$.

$$- \sum_k p_k q_{kl} \log q_{kl} \leq -(\sum_k p_k q_{kl}) \log(\sum_k p_k q_{kl}).$$

Summing both sides over l gives:

(5) ... $H_A(B) \leq H(B)$.

From (3) and (4) we find that equality is attained in the case of independence.

If we view the entropy functional as a measure of uncertainty this may be interpreted as the fact that prior knowledge of the outcome of A never increases the uncertainty in the outcome of B.

The inequality (5) is an important one: In a study on the interactions of nucleons, S. Watanabe introduced in 1939 a measure of dependence between random variables based on a difference between entropies. In a later paper [6] this idea is elaborated.

From (4) we find that:

(6) ... $$\begin{aligned} H(A) + H(B) - H(A \times B) &= H(A) - H_B(A) \\ &= H(B) - H_A(B) \\ &= C(A,B) \text{ bij definition.} \end{aligned}$$

The quantity C defined in this way is never negative according to (5) and it vanishes only when A and B are independent. Watanabe [6] proposed to use C as a measure of dependence between A and B. In this report the quantity C will be called the "excess-entropy".

2.3 ENTROPY IN OBJECT-PREDICATE TABLES

2.3.1. Entropy and excess-entropy in partitions.

A k-partition of a set S is a set of k mutually disjoint subsets (called "cells") whose union is S. Suppose the i-th cell has n_i elements and $\sum_i n_i = n$. We may associate with a k-partition the set $\{^{n1}/n, \ldots, ^{nk}/n\}$ of non-negative numbers whose sum is 1.

This analogy to the discrete probability scheme caused Rescigno and Maccacaro [5] to define the entropy of a partition as:

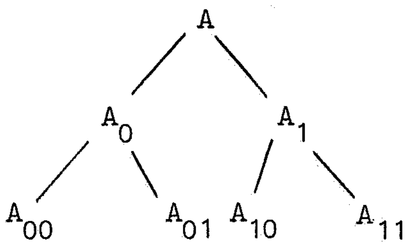
(1) ... $$H = - \sum_i \frac{n_i}{n} \log \frac{n_i}{n} = \log n - \frac{1}{n} \sum_i n_i \log n_i$$

To every pair of partitions there corresponds a product partition (which is again a partition): if a partition is defined on S, so also it is on every subset of S and therefore also on each of the cells of the other partition. Accordingly, there corresponds an excess-entropy to every pair of partitions A and B:

$$C(A,B) = H(A) + H(B) - H(A \times B) .$$

Let us consider partitions of the set A generated by subjecting every cell to a 2-partition. One of the subcells is denoted by putting a 0, the other by putting a 1 behind the name of the cell. Starting from the trivial partition {A} of A we get successively:

(2) ...



Now let the elements of A be partitions. We are going to study the entropy-relations between the product partitions of the partitions of a subset of A: H and C will denote entropy and excess-entropy again, with indices to indicate to which subset of A they apply. We found for the excess-entropy between the two product partitions ΠA_0 and ΠA_1 :

$$C(0,1) = H_0 + H_1 - H.$$

It will be found useful to extend this definition to apply to more than 2 partitions, for instance the 4-partition of the lowest level of (2):

$$C(00,01,10,11) \overset{\text{def}}{=} H_{00} + H_{01} + H_{10} + H_{11} - H$$

This 4-way excess-entropy can be expressed in 2-ways entropies as follows:

$$C(00,01,10,11) = H_{00} + H_{01} - H_0 + H_{10} + H_{11} - H_1 + H_0 + H_1 - H$$
$$= C(00,01) + C(10,11) + C(0,1)$$

This can be represented in a hierarchical diagram:

$$H$$

$= H_0 \quad + \quad H_1 \quad - C(0,1)$ First level

$= H_{00}+H_{01}-C(00,01) \quad = H_{10}+H_{11}-C(10,11)$ Second level

Thus the multi-way excess-entropy of a certain level may be hierarchically decomposed into two-way excess-entropies of all levels not below it. We regard this as a method in compliance with the principle of dealing with systems by hierarchical decomposition of complexity. The analogous procedure for a set of random variables has been described by Watanabe [5].

2.3.2 Data compression in an object-predicate table

Let us now study the object-predicate table as directly as possible from the point of view of the information provided by the predicates about the objects. This may be illustrated by a guessing game: One person takes an object in mind and has to answer yes or no to another person's questions about it in the form: Does it have predicate p_i? The answers to questions concerning a subset of the predicates define a partition in the set of objects. Following the classical definition of Shannon's , a suitable definition for the information provided by a set of predicates is the entropy of their product partition as defined in the previous section. The set of n predicates defines a partition of 2^n cells and the maximum entropy of such a partition is n bits. When the actual entropy is less than this, we say there is "redundancy" in the set of predicates.

When we realise that there exists an object-predicate table with n predicates and 2^n objects where every cell of the partition contains exactly one object and which therefore does not contain any redundancy, it is apparent that in tables with moderately large numbers (between, say, 10 and 1000) of objects and predicates, enormous amounts of redundancy are usual.

Thus we are led to the problems of "data compression" (see the articles by Tou and Heydorn, Watanabe and others in [7]):

1. For given $k < n$ find a subset $\{p_{i_1}, \ldots, p_{i_k}\}$ of the predicates such that $H(p_1, \ldots, p_n) - H(p_{i_1}, \ldots, p_{i_k})$ is a minimum.
2. For $k = 1,2, \ldots, n$ find the k such that the data compression achieved in 1. is, in some respect, optimum.

It will be interesting to encounter, in a later section, an analogous problem for an n-dimensional normal probability distribution.

2.3.3 Hierarchical decomposition of excess-entropy: Association Analysis

In plant ecological studies data may be obtained in the following way. In the geographical area to be treated, a number of plots, called "quadrats", are staked off and of each of these it is noted which species of plants are present. Williams and Lambert [9] (the quotations are from this paper) have proposed "Association Analysis" as a method for sorting quadrats into groups.

Data of this origin may be presented as an object-predicate table where it is immaterial whether species (quadrats) are identified with the objects (predicates)."The basic problem is to subdivide the population so that all associations disapear ..." . Here "association" is to be used in its "statistical sense". It seems desirable to give a more precise interpretation of "association".

In 2.3.1. we have defined the excess-entropy of a set of partitions. A predicate effects a 2-partition in the set of objects (the objects that do and those that do not have the predicate); a set of predicates therefore corresponds to a set of partitions in the objects.

Likewise, an object effects a 2-partition in the set of predicates (those it does and those it does not have) and, by the previous sentence, this object corresponds to two sets of partitions in the set of objects. To these two sets of partitions there corresponds an excess-entropy and this we may call the "entropy loading" of that object.
Now the set of all predicates together define a product partition in the set of objects and this has a "collective" entropy. Every predicate on its own defines a 2-partition and the "individual" entropy of this partition. The difference between the sum of individual entropies and their collective entropy is the (multi-way) excess-entropy defined in 2.3.1. Its hierarchical decomposition may be used to analyse the structure of the interrelations existing in the set.
Let us identify objects as species and predicates as quadrats. The purpose of the rest of this section is to show that the excess-entropy of a set of predicates has the properties that Williams and Lambert [9] expect the undefined concept of association to have.

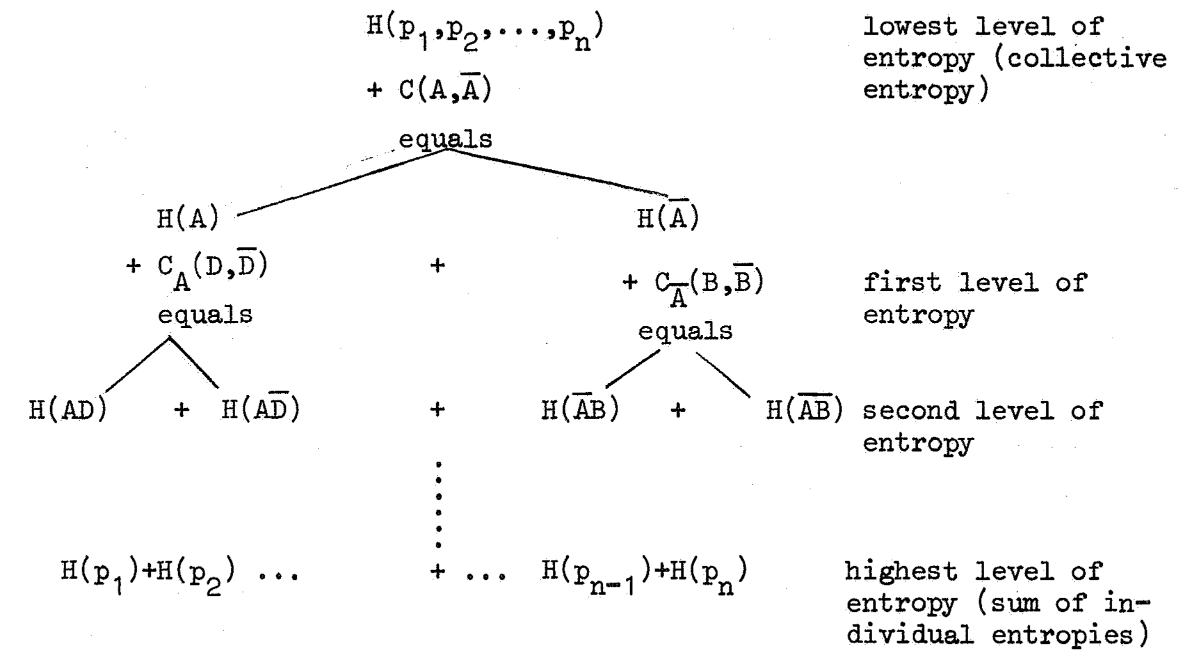
a) Williams and Lambert [9] argue that "positive" as well as "negative" associations are to be taken into account. From this we may infer that, roughly speaking, if two species are positively (negatively) associated, then the presence of the one makes occurrence of the other more (less) likely. Therefore association without sign is something that is expected to give a positive contribution in both cases. This is just what excess-entropy does: it is never negative and, independently of the sign of the interaction, indicates whether species influence each other more or less strongly.

b) Apparently, association should not only be defined between a pair of species, but somehow all associations present in a set of species should be pooled. This is the reason why we have used the extension from a pair to an arbitrarily large set of partitions.

c) The objective of association analysis is to subdivide the set of quadrats so that all associations disappear. This is done by taking a particular species and partitioning the set of quadrats into those in which it did and those in which it did not occur.

The species is chosen so that the pooled association in the subsets is as small as possible. Stating it in terms of excess-entropy in the object predicate table, we can say that we must find the object with the highest entropy loading and repeat the process on each of the cells of predicates.

If we therefore interpret association as excess-entropy, we find that association analysis is the hierarchical decomposition of the total excess-entropy in such a way that the largest component of excess-entropy is produced first.



<u>Diagram showing hierarchical subdivision of excess-entropy.</u>

$\{p_1,p_2,\dots,p_n\}$: set of all predicates.
A is the set of predicates possessed by object "a".

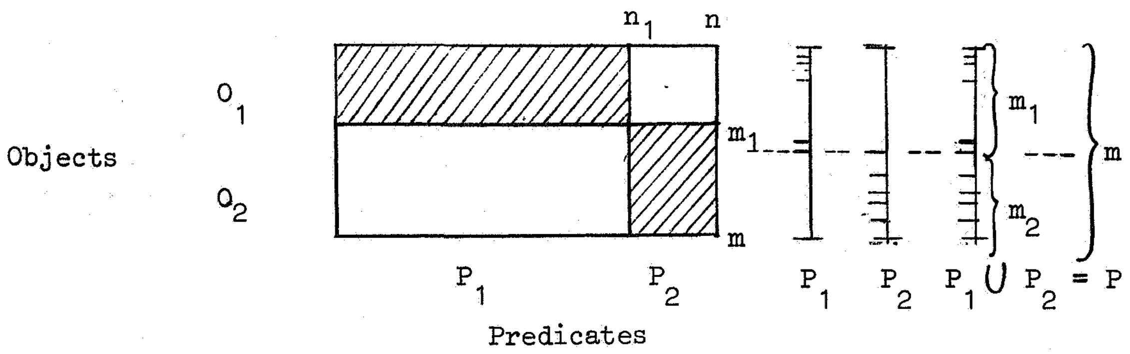
$\bar{A}(\bar{B},\bar{D})$ is the complement of $A(B,D)$

2.3.4. <u>Clustering</u> .

In the first section we mentioned the possibility of a system of many variables consisting of a few subsets of variables with interactions between subsets weak relative to those within subsets.

In such a case it is possible to study one aspect of the whole system by regarding a simple system consisting of these subsets as "black boxes".

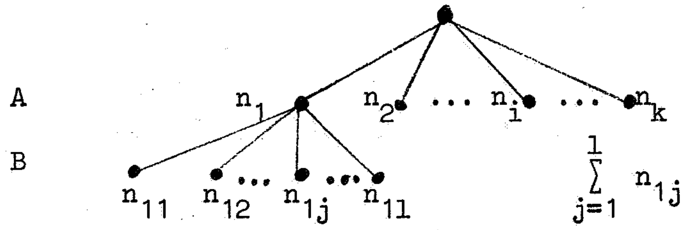
It is of course necessary to describe in a more specific fashion the interactions between variables. To a certain extent this is possible in the object-predicate table. We will define the situation in which the table is considered to be "completely decomposed" into two subsets of objects and predicates and we will show that in that case the excess-entropy between the subsets is minimal. This gives a more flexible way of describing a table, which is of practical importance because a table "almost" decomposed is more likely to occur in practice than one completely decomposed.



Object-predicate table (without actual entries) with examples (at right) of partitions in the set of objects induced by subsets P_1 and P_2 of the set P of all predicates.

If there are no crosses outside the shaded area, that is, when none of the predicates of P_1 is possessed by any of the objects in O_2 and vice versa, we say that the table is completely decomposed. The partition P is the product partition of P_1 and P_2. This product is of a peculiar kind: P_1 subdivides only one cell of P_2 and vice versa.

Let us consider a related special form of product partition: that of hierarchical subdivision. Suppose that we have a partition A and that partition B acts only on one cell of A; without loss of generality we may suppose this one to be the first.



$$\sum_{i=1}^{k} n_i = n \qquad \sum_{j=1}^{l} n_{1j} = n_1$$

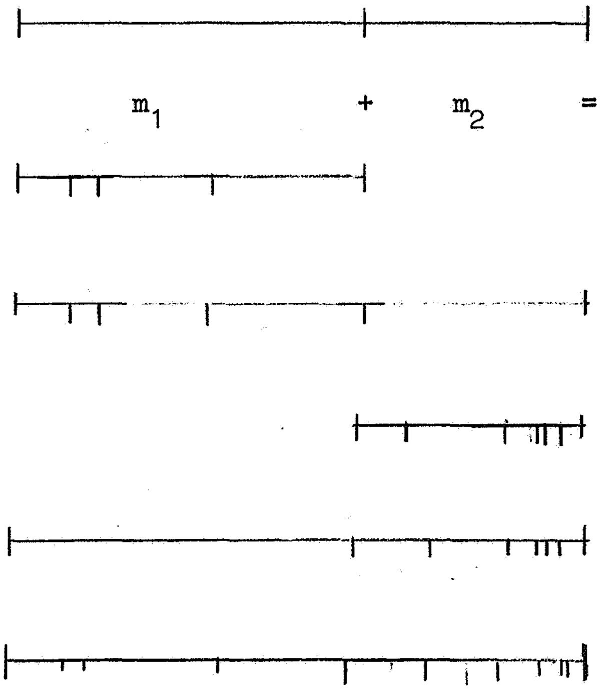
$$H_A = \log n - \frac{1}{n}\sum_{i=1}^{k} n_i \log n_i \;;\; H_B = \log n_1 - \frac{1}{n_1}\sum_{j=1}^{l} n_{1j} \log n_{1j} \;;$$

$$H_{AxB} = \log n - \frac{1}{n}\{\sum_{i=2}^{k} n_i \log n_i + \sum_{j=1}^{l} n_{1j} \log n_{1j}\}$$

$$= \log n - \frac{1}{n}\{\sum_{i=1}^{k} n_i \log n_i - (n_1 \log n_1 - \sum_{j=1}^{l} n_{1j} \log n_{1j})\}$$

$$H_{AxB} = H_A + \frac{n_1}{n} H_B$$

We use this formula to find the excess-entropy that exists between sets of partitions P_1 and P_2 completely decomposing the table. It seems convenient to derive the entropies of the partitions P_1 and P_2 by hierarchical subdivision of the partition $\{m_1, m_2\}$ that they have in common.



$$H = \log m - \frac{1}{m}(m_1 \log m_1 + m_2 \log m_2)$$

$$m_1 + m_2 = m$$

$$H_1'$$

$$H_1 = H + \frac{m_1}{m} H_1'$$

$$H_2'$$

$$H_2 = H + \frac{m_2}{m} H_2'$$

$$H_{1x2} = H + \frac{m_1}{m} H_1' + \frac{m_2}{m} H_2'$$

$$H_1 + H_2 - H_{1x2} = C(P_1, P_2) = H.$$

Any additional subdivision in the left half of H_1 or in the right half of H_2 leaves the excess-entropy $C(P_1,P_2)$ unchanged at H.
Any additional subdivision in the right half of H_1 or in the left half of H_2 makes the excess-entropy $C(P_1,P_2)$ greater than H.
Therefore the excess-entropy of an object-predicate table completely decomposed with respect to two mutually disjoint subsets of predicates P_1 and P_2 of m_1 and m_2 elements respectively is minimal and equal to $H = \log m - \frac{1}{m}(m_1 \log m_1 + m_2 \log m_2)$.

3. ENTROPY AND THE NORMAL PROBABILITY DISTRIBUTION.

3.1 Variance and entropy.

Let A be a positive definite symmetric matrix with propervalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$ and corresponding propervectors $x_1, x_2, \ldots, x_n$.

Theorem (Bellman, [2], p. 117):

If A is positive definite,

$$(1) \ldots \quad \frac{(2\pi)^{k/2}}{\sqrt{|A|_k}} = \max_{R_n} \int_{R_k} e^{-\frac{1}{2}(z,Az)} \, d\,V_k \quad \text{, where}$$

$|A|_k = \prod_{i=n-k+1}^{n} \lambda_i$, the product of the k smallest proper values and dV_k is the k-dimensional element of volume in R_k.

Taking k = n and noting that $|A|_n = |A|$, the determinant of A, we obtain the well-known equality:

$$\frac{(2\pi)^{\frac{1}{2}n}}{\sqrt{|A|}} = \int_{R_n} e^{-\frac{1}{2}(z,Az)} \, dV_n,$$ which allows us to define as the n-dimensional normal probability density:

$$(3) \ldots \quad f(x) = \frac{|A|^{\frac{1}{2}}}{(2\pi)^{\frac{1}{2}n}} \, e^{-\frac{1}{2}(x,Ax)}$$ with x some n-dimensional vector and A a positive definite symmetric matrix.

$V = A^{-1}$ is the covariance matrix of the distribution.

The determinant of V is called the generalized"variance";hereafter we will refer to it as the "variance". The motivation of this treatment of the normal distribution is the fact that here, too, we may define the entropy functional. The practical use of the normal distribution is limited by the fact that in many situations the assumption that the data arise from a normal distribution is difficult to justify. The object-predicate table is of wider applicability. In both cases the entropy functional may be defined and this allows us to formulate analogous problems .

For the entropy of the normal distribution we find:

$$H = -\int_{R_n} f(x)\ \ln f(x)\ d\ V_n$$

$$H = -\int_{R_n} \frac{|A|^{\frac{1}{2}}}{(2\pi)^{n/2}}\ e^{-\frac{1}{2}(x,Ax)}\ \{-\ \frac{n}{2}\ \ln(2\pi) + \frac{1}{2}\ \ln|A| - \frac{1}{2}(x,Ax)\}d\ V_n$$

$$= -\ \frac{1}{2}\ \ln|A| + \frac{n}{2}\ \ln(2\pi) + \frac{|A|^{\frac{1}{2}}}{(2\pi)^{\frac{1}{2}n}} \int_{R_n} \frac{1}{2}(x,Ax)\ e^{-\frac{1}{2}(x,Ax)} d\ V_n$$

$= \frac{n}{2}\ \ln(2\pi e) + \frac{1}{2}\ \ln|V|$. If we express entropy in bits, we get the usual formula:

$H = \frac{n}{2}\ \log(2\pi e) + \frac{1}{2}\ \log|V|$ where the logarithms are to the base 2.

Thus we see that there is a relationship between variance and entropy. Suppose now that R_k is spanned by $x_n, \ldots, x_{n-k+1}$. Any vector x in R_n may be decomposed into an $x_1 \in R_k$ and an $x_2 \perp R_k$ such that $x = x_1 + x_2$. This implies that $Ax = Ax_1 + Ax_2$.

Because R_k and its orthogonal complement $R_k^{\perp}$ are spanned by proper vectors (these are orthogonal because A is symmetric)

$Ax_1 \in R_k$ and $Ax_2 \perp R_k$ for all x, that is, R_k reduces A into a matrix A_1 of order k and a matrix A_2 of order n - k. A_1 acts only within R_k, A_2 only within $R_k^{\perp}$

A consequence of this decomposition of A by R_k is the decomposition of the n-dimensional distribution f(see 3.1.1.3) into a k-dimensional distribution

$$f_1(x_1) = \frac{|A_1|^{\frac{1}{2}}}{(2\pi)^{k/2}}\ e^{-\frac{1}{2}(x_1,A_1x_1)}$$

and a(n-k)-dimensional distribution

$$f_2(x_2) = \frac{|A_2|^{\frac{1}{2}}}{(2\pi)^{(n-k)/2}}\ e^{-\frac{1}{2}(x_2,A_2x_2)}$$

The decompostion of the density function f:

$f(x) = f_1(x_1).\ f_2(x_2)$ results in similar decompositions for variance and entropy:

$|V| = |V_1|.|V_2|$ and $H = H_1 + H_2$.

3.2 Data compression.

In section 2.3.2 we discussed the possibility of a small subset of predicates saying almost as much as the whole set. In such a case we spoke of "data compression". An analogous problem may be posed for the normal distribution:
Suppose we have a projection of x on an arbitrary k-dimensional subspace, is it possible to choose this subspace so that the variance of this projection is almost as much as that of x? Or, equivalently, that its entropy is almost as much as that of x? In that case we have a redundancy of dimensions and we achieve data compression by substituting the projection for x itself.
Bellman's result (3.1.1) now becomes useful: it states that of all k-dimensional subspaces the one containing the largest part of total entropy is the one spanned by the proper vectors belonging to the k largest proper values of V. Whether the largest part is actually large, depends on the distribution of the proper values. The more nearly they are equal, the less data compression is possible. x_1, the projection of x on R_k, is a linear combination of projections on the proper vectors $x_n, \ldots, x_{n-k+1}$. These were called by Hotelling the "principal components": They decompose R_n in such a way that in the corresponding factorization of $|V|$ one factor is the largest and therefore the other the smallest.
The fact, that the k-dimensional subspace containing the maximum part of the total entropy is the subspace spanned by the proper vectors belonging to the k largest proper values of V, was derived in a paper by J. Tou and R. Heydorn in [8]. They did not seem to be aware that their problem and solution are only a restatement of Hotelling's well-known result on principal components.
A more general result has been obtained by Watanabe [7] by showing that the Karhunen-Loève expansion has a similar entropy-extremizing property. The greater generality lies in the fact that this expansion may be used for samples as well as for distributions.

3.3 Excess-entropy and likelihood ratio.

The "likelihood-ratio", an often used test statistic, may be interpreted as an excess-entropy. Kullback ([4], pp. 4-5) gave an information-theoretical interpretation of the likelihood-ratio. In this section we will show that Kullback's I(0:1) is the same as an excess-entropy as we have introduced it before.

Let $H_0(H_1)$ be the hypothesis that the random variable X is from the population with probability density function $f_0(f_1)$. From the definition of conditional probability:

$$\Pr\{H_i|x\} = \frac{\Pr\{H_i \wedge x\}}{\Pr\{x\}} = \frac{\Pr\{x|H_i\}.\Pr\{H_i\}}{\Pr\{x \wedge H_0\} + \Pr\{x \wedge H_1\}} \Rightarrow$$

$$\Pr\{H_i|x\} = \frac{f_i(x).\ \Pr\{H_i\}}{f_0(x).\Pr\{H_0\} + f_1(x)\ \Pr\{H_1\}} \qquad \text{for } i = 0,\ 1.$$

$$\frac{\Pr\{H_0|x\}}{\Pr\{H_1|x\}} = \frac{f_0(x).\Pr\{H_0\}}{f_1(x).\Pr\{H_1\}} \Longrightarrow$$

$$\log \frac{f_0(x)}{f_1(x)} = \log \frac{\Pr\{H_0|x\}}{\Pr\{H_1|x\}} - \log \frac{\Pr\{H_0\}}{\Pr\{H_1\}}$$

The last formula says that the log likelihood-ratio is the difference of log-ratios of "a posteriori" and "a priori" probabilities. This is interpreted as the information present in the observation x in favour of the null hypothesis H_0. When this quantity is averaged over the distribution with density f_0 we have

$$I(0:1) = \int_x f_0(x).\log \frac{f_0(x)}{f_1(x)}.\ dx\ .$$

Let us now consider the case where we have a set x of random variables partitioned as $x = \{x_0, x_1\}$. f is supposed to be the probability density function of x; g and h are the marginal probability density functions of x_0 and x_1 respectively.

Consider the null hypothesis H_0 that x_0 and x_1 are dependent and the alternative hypothesis

$H_1 : f(x) = g(x_0).h(x_1)$ for all x. In this case:

$$I(0:1) = \int_x f(x).\ \log \frac{f(x)}{g(x_0)h(x_1)}\ dx$$

$$= \int_x f(x) \log f(x)dx - \int_x f(x) \log g(x_0)dx - \int_x f(x) \log h(x_1)dx$$

$$= - H(f) - \int_{x_0} \left[\int_{x_1} f(x)dx_1\right] \log g(x_0)dx_0$$

$$- \int_{x_1} \left[\int_{x_0} f(x)dx_0\right] \log h(x_1)dx_1$$

$$= H(g) + H(h) - H(f)$$

$I(0:1) = C(x_0,\ x_1)$.

We find that the excess-entropy is equal to the average (under the null hypothesis of dependence) information present in the observations in favour of the null hypothesis. We may regard this as a measure of dependence. This measure is used as a test statistic for the log likelihood ratio test for independence.

3. Clustering.

Let f(X) be the normal density function of an n-dimensional random vector $X = (x_1, \ldots, x_n)$. Let X be partitioned as (X_1,X_2) with $X_1 = (x_1, \ldots, x_k)$ and $X_2 = (x_{k+1}, \ldots, x_n)$ and let the corresponding partition of V be:

$$V = \begin{pmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{pmatrix} .$$

For the excess-entropy between X_1 and X_2 we find

$$C(X_1,X_2) = H(X_1) + H(X_2) - H(X).$$

$$= \tfrac{1}{2} \log \frac{|V_{11}| \cdot |V_{22}|}{|V|}$$

This quantity is defined for any nonsingular matrix V and any partition in it. When V is reduced into V_{11} and V_{22}(when there are only zero elements in V_{21}), it is zero. It may therefore be used to indicate to what extent V is almost reduced as is in fact done when using the likelihood ratio test for independence between X_1 and X_2. In 1.1 we saw that the fact that the blocks A_{11} and A_{22} do not reduce the matrix manifests itself as the transfer of the error in one partial approximation to the other and vice versa. Especially in view of 1.2, we tentatively described this phenomenon as "information transfer". In the special case where the matrix is symmetric and positive definite we have shown that this description is compatible with the mathematical definition of information.

4. <u>ENTROPY IN MARKOV CHAINS</u>.

Let us consider a Markov chain M with a finite number n of states s and a discrete time parameter t; $s_t = j$ means that M is in state j at time t. For every value t of the time parameter there is a probability distribution over the states:

$\Pr\{s_t = j\} = a_j^t$.

M must be in some state: $\sum_{j=1}^{n} a_j^t = 1$ for all t.

We will also use the matrix P of transition probabilities whose elements are:

$p_{ji} = \Pr\{s_t = j \mid s_{t-1} = i\}$. These we will suppose to be independent of time. P connects successive distribution vectors $A_t^T = (a_1^t, \ldots, a_n^t)$ in the following way:

(1) ... $a_j^t = \sum_{i=1}^{n} p_{ji}\, a_i^{t-1}$ for $j = 1, \ldots, n$ or $A_t = PA_{t-1}$.

Columns of P add up to unity (if M is in any state i at time t-1, it is certain to be in some state at time t), so we may define the <u>conditional entropy</u>

$H_i = -\sum_{j=1}^{n} p_{ji} \log p_{ji}$ under condition that M is in state i.

If we have any probability distribution $A^T = (a_1, \ldots, a_n)$ over the states of M we may consider its elements as weights to produce a weighted average of the conditional entropies H_i:

$$(2) \ldots H = \sum_{i=1}^{n} a_i H_i \ .$$

Many interesting Markov chains have the property that $\lim_{t\to\infty} A_t$ exists for every probability distribution and is independent of it. The entropy (2) obtained by taking $A = \lim_{t\to\infty} A_t$ is, in information theory, defined to be <u>the</u> entropy of the Markov chain (see, for instance, Khinchin[3]).

4.1 FUSING TWO STATES.

Suppose it can no longer be decided whether $s_t = j$ or whether $s_t = k$, but only whether $s_t = j$ or $s_t = k$. Then we say that states j and k are <u>fused</u>, say into j'. We see at once that:

$$(3) \ldots a_{j'} = a_j + a_k \quad \text{and}$$

$$(4) \ldots p_{j'i} = p_{ji} + p_{ki}$$

$p_{ij'}$ is obtained from (1) as follows:

$$a_i = \sum_{m\neq j,k} p_{im} a_m + p_{ij} a_j + p_{ik} a_k \qquad \text{for } i \neq j,\ i \neq k.$$

If we put

$$(5) \ldots p_{ij'} = \frac{p_{ij} a_j + p_{ik} a_k}{a_j + a_k} \quad \text{, we get}$$

$a_i = \sum_{m\neq j'} p_{im} a_m + p_{ij'} a_{j'}$, as it should be.

Similarly for the case i = j or i = k:

$$a_j = \sum_{m\neq j,k} p_{jm} a_m + p_{jj} a_j + p_{jk} a_k$$

$$a_k = \sum_{m\neq j,k} p_{km} a_m + p_{kj} a_j + p_{kk} a_k \quad +$$

$$a_{j'} = \sum_{m\neq j'} p_{j'm} a_m + p_{j'j} a_j + p_{j'k} a_k$$

If we put

$p_{j'j'} = \dfrac{p_{j'j}a_j + p_{j'k}a_j}{a_j + a_k}$, the last two terms may be replaced

by $p_{j'j'}a_{j'}$, as they should be.

To summarize, the effect of fusing two states, j and k, is to replace a_j by $a_j + a_k$, the j-th row of P by the sum of the j-th and k-th rows and the j-th column of P by the weighted sum of the j-th and k-th columns with weights $a_j/(a_j+a_k)$ and $a_k/(a_j+a_k)$ respectively. Finally, a_k ,the k-th row of P and the k-th column of P are deleted.

4.2 EXCESS-ENTROPY AS A MEASURE OF CLUSTERING.

Fusion of two states may occur in any chain having not less than that number of states. The result is a chain again, where two states, if available,may be fused again. In short, as many states as are present may be fused.

Consider the sets of states

$X = \{1,2, \ldots, j\}$ and $Y = \{j+1, \ldots, n\}$. Besides the original chain M with states $\{X \cup Y\}$ we also consider the chain M_x with states $\{X,y\}$ and the chains M_y with states $\{x,Y\}$ where y(x) is the state resulting from fusion of all states of Y(X) .

$$
\begin{array}{c|cccc|cccc}
 & \multicolumn{4}{c|}{X} & \multicolumn{4}{c}{Y} \\
\hline
 & p_{11} & \cdots & p_{1j} & & p_{1,j+1} & \cdots & p_{1n} & \\
X & \vdots & & \vdots & & \vdots & & \vdots & \\
 & p_{j1} & & p_{jj} & & p_{j,j+1} & & p_{jn} & \\
\hline
 & p_{j+1,1} & \cdots & p_{j+1,j} & & p_{j+1,j+1} & \cdots & p_{j+1,n} & \\
Y & \vdots & & \vdots & & \vdots & & \vdots & \\
 & p_{n1} & \cdots & p_{nj} & & p_{n,j+1} & \cdots & p_{nn} &
\end{array}
$$

M's matrix of transition probabilities

$$\begin{matrix} p_{11} & \cdots & p_{1j} & p_{1y} \\ \vdots & & \vdots & \vdots \\ p_{j1} & \cdots & p_{jj} & p_{jy} \\ p_{y1} & \cdots & p_{yj} & p_{yy} \end{matrix}$$

M_x's matrix of transition probabilities (states of Y fused into y)

M_y's matrix of transition probabilities (states of X fused into x).

$$\begin{matrix} p_{xx} & p_{x,j+1} & \cdots & p_{xn} \\ p_{j+1,x} & p_{j+1,j+1} & \cdots & p_{j+1,n} \\ \vdots & \vdots & & \vdots \\ p_{nx} & p_{n,j+1} & & p_{nn} \end{matrix}$$

Applying formulae (3)-(6) for fusing states of X and also for states of Y we find:

$$a_x = \sum_{i \in X} a_i \quad ; \quad a_y = \sum_{i \in y} a_i$$

$$p_{iy} = \sum_{j \in y} \frac{a_j}{ay} p_{ij} \ , \ i \in X \text{ and } p_{ix} = \sum_{j \in X} \frac{a_j}{ax} p_{ij} \ , \ i \in Y.$$

Let us now introduce the quantities:

$$T_{xx} = - \sum_{i \in X} a_i \sum_{k \in X} p_{ki} \log p_{ki}, \quad T_{xy} = - \sum_{i \in X} a_i \sum_{k \in Y} p_{ki} \log p_{ki},$$

$$T_{yx} = - \sum_{i \in Y} a_i \sum_{k \in X} p_{ki} \log p_{ki}, \quad T_{yy} = - \sum_{i \in Y} a_i \sum_{k \in Y} p_{ki} \log p_{ki}.$$

According to (2) we then have for the entropy of M:

$$H = T_{xx} + T_{xy} + T_{yx} + T_{yy}.$$

We will now obtain inequalities for the "cross terms" T_{xy} and T_{yx}.

$$T_{xy} = - \sum_{k \in Y} a_x \sum_{i \in X} \frac{a_i}{a_x} p_{ki} \log p_{ki}$$

Application of 2.2.2 to the inner sum, where this time $\lambda_i = \frac{a_i}{a_x}$ and $f(x) = x \log x$, yields:

(7) ... $$T_{xy} \leq - \sum_{k \in Y} a_x p_{xx} \log p_{xx} \text{ and similarly}$$

$$T_{yx} \leq - \sum_{k \in X} a_y p_{ky} \log p_{ky}$$

"H(X)" and "H(Y)" will be used to denote the entropies M_x and M_y respectively. According to (2) we have:

$$H(X) = T_{xx} - \sum_{k \in X} a_y \, p_{ky} \log p_{ky} - \sum_{i \in X} a_i \, p_{yi} \log p_{yi} - a_y \, p_{yy} \log p_{yy}$$

Because of (7) it follows that $H(X) \geq T_{xx} + T_{yx}$

Similarly we find that $H(Y) \geq T_{yy} + T_{xy}$, whence the main result:

(8) ... $H \leq H(X) + H(Y)$.

Again, as in the case of probability schemes and object-predicate tables, we may regard the concomitant excess-entropy:

(9) ... $C(X,Y) = H(X) + H(Y) - H \geq 0$

as a measure of dependence, this time between states of X and states of Y.

A Markov chain may have a "clustering" structure in the sense that if it is in a state of X(Y) at time t, it has a very small probability of being in Y(X) at time t + 1. It is clear that this is the more so as p_{xx} and p_{yy} are closer to 1. The excess-entropy defined in (9) is one possible measure of such clustering. When we consider all Markov chains with a partition {X,Y} of the set of n states, where $p_{xx} < 1$ and $p_{yy} < 1$, then the equality (9) is sharp, that is, 0 is the greatest lower bound of C(X,Y). Thus we see that C(X,Y) may be used as a measure of clustering, smaller values corresponding to stronger clustering.

If we are given the probability matrix P of some Markov chain, the stationary probability distribution A may be obtained by solving the system of linear equations

$(P - I)A = 0.$

Suppose that we have a clustering structure in the above sense, namely that p_{xx} and p_{yy} are near unity. Then the system may profitably be solved by the iterative method mentioned in 1.1. Again, as in the case of a positive definite matrix, we see that the cause of continuation of iteration, which we tentatively called "information transfer", may be explained in terms of entropy which is fundamental to the mathematical definition of information.

5. LITERATURE REFERENCES

[1] C. Alexander, Notes on the Synthesis of Form, Harvard U.P., 1967.

[2] R. Bellman, Introduction to Matrix Analysis, McGraw-Hill, 1960.

[3] A.I. Khinchin, Mathematical Foundations of Information Theory, Dover, 1957.

[4] S. Kullback, Information Theory and Statistics, Dover, 1968.

[5] A. Rescigno and G.A. Maccacaro, Information Content of Biological Classifications, in: C. Cherry (ed.), Information Theory: Fourth London Symposium, Butterworths, 1961.

[6] S. Watanabe, Information Theoretical Analysis of Multivariate Correlation, IBM Journal of Res. and Dev., 1960, pp. 66-82.

[7] S. Watanabe, Karhunen-Loève Expansion and Factor Analysis, Proc. 4th Conf. on Information Theory, Prague, 1965, pp. 635-659.

[8] J. Tou (ed.), Computer and Information Sciences-II, Academic Press, 1967.

[9] W. Williams and J.M. Lambert, Multivariate Methods in Plant Ecology I, The Journal of Ecology, 47(1959), pp. 83-101.