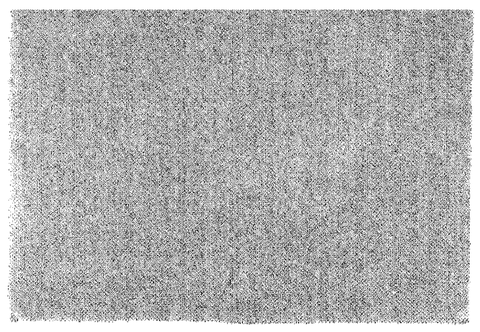
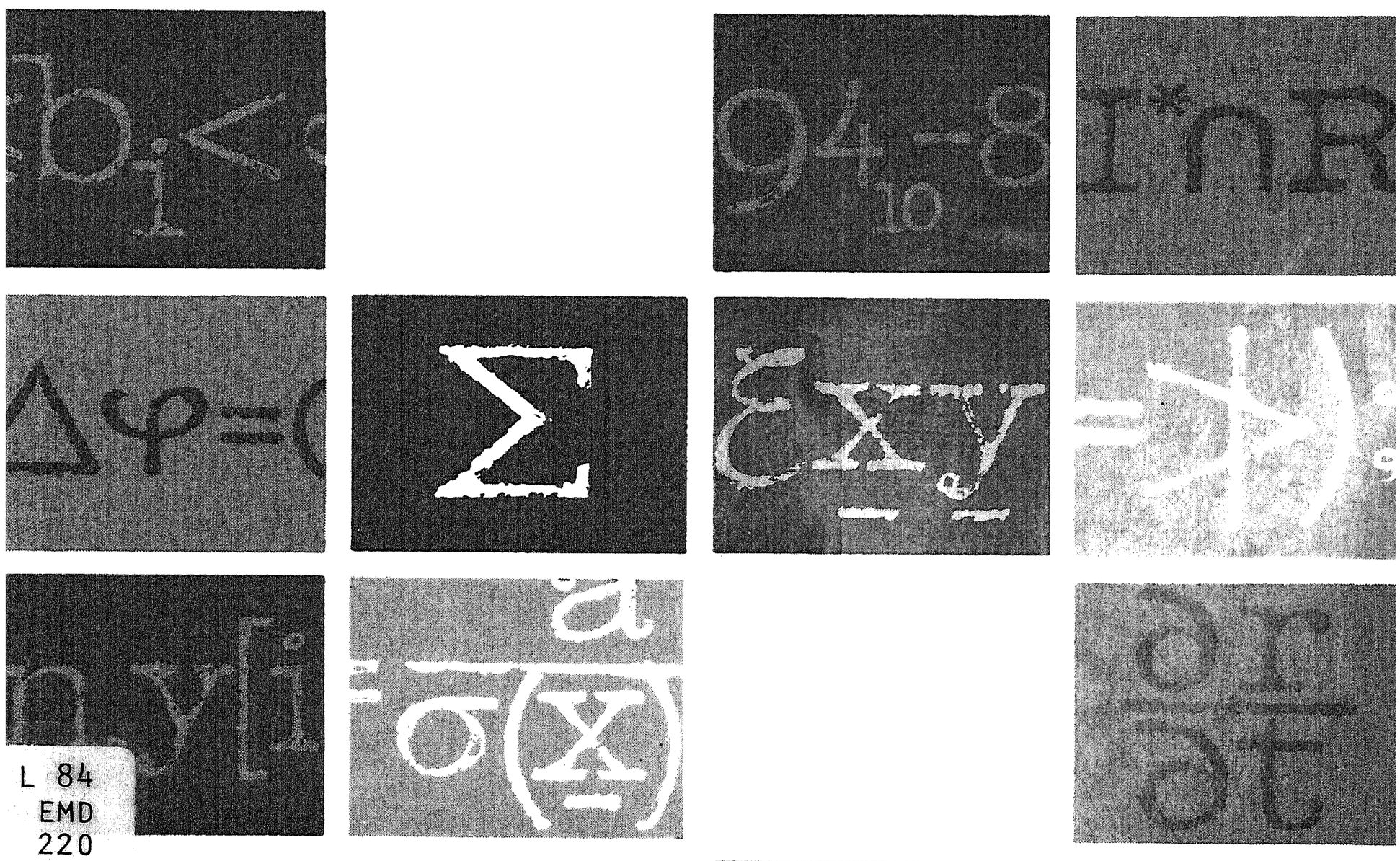


AN ANALYSIS OF COMPLEXITY

M.H. VAN EMDEN



MATHEMATICAL CENTRE TRACTS 35

MATHEMATICAL CENTRE TRACTS

35

AN ANALYSIS OF COMPLEXITY

BY

M.H. VAN EMDEN

MATHEMATISCH CENTRUM AMSTERDAM

1971

ACKNOWLEDGEMENTS

To professor A. van Wijngaarden I owe a debt of gratitude for the near-optimal combination of freedom and encouragement with which he saw me through the research of which this tract is an outcome. To professor J.Th. Runnenburg I am greatly indebted for the expertness and energy expended by him on an early draft of the tract. His help has ranged from constructive criticism on its organization and its conceptual background to detailed assistance resulting in simplified proofs and an improved English text. I was much helped by J.D. Alanen and T.J. Dekker who critically read chapter 3.

The research for this tract was carried out at the Mathematical Centre. The Hugo de Vries Laboratory for Systematic Botany in the University of Amsterdam has provided substantial support as a part of its program to develop numerical methods for classification. I am especially grateful for the support and encouragement of S. Segal which made this possible.

The Edinburgh University Press kindly gave permission for including in section 3.2 a part of my paper "Optimal Data Compression" which they published early in 1971 in "Machine Intelligence 6" (edited by B. Meltzer and D. Michie).

Finally, many thanks to miss Astrid Fasen and messrs. D. Zwarst and J. Suiker who did an expert and speedy job in the typing and printing, respectively, of this tract.

SUMMARY

Recently, numerous methods have been proposed for using a computer for the classification of objects. If all methods give substantially the same classification, there is the theoretical problem of explaining this. If they give different classifications, there is the practical problem of deciding which of these, if any, gives a "good" classification. If a "good" classification means that it should be "meaningful" or that it should "explain as much as possible", the problem is caused by the difficulty of bridgeing with mathematical reasoning the gap between criteria of this form and an algorithm suitable for execution by computer.

When attempting to solve a large system of equations, the problem of classification arises in such a way that the criterion for a good classification can be formulated precisely. It is also naturally expressed in terms of the "complexity" of a system if this is interpreted to be the totality of interactions within it. This suggests that the phenomenon of complexity is worthy of being studied in its own right and that it provides a conceptual foundation for classification.

In chapter 1 we propose a mathematical definition of complexity based on a definition of interaction in terms of the theory of information. In chapter 2 we discuss the analysis of qualitative data. Pairwise interactions between entities to be classified may be used to define a distance function without, however, supposing that the qualitative data themselves constitute a metric space. This allows a model of classification to be formulated in terms of information and to discuss its relation to clustering.

In chapter 3 data are discussed that describe objects that can be represented by points in n-dimensional inner-product space, and the covariance matrix of the set of points is studied. The several criteria, according to which the principal components approximation of multivariate statistics is optimal, are related to data compression. In connection with this a maximum-entropy characterization of the multivariate normal distribution is given. With the aid of this characterization, we propose a measure of the complexity of a covariance matrix, and we study how particular coordinate systems give special representations of complexity. The condition number of the covariance matrix, a quantity which is important in numerical computa-

tion, is related to its complexity. Finally, an iterative method for solving a system of linear equations, of which the matrix of coefficients is the covariance matrix, is treated. It is shown that, if the variables have a strong clustering in the sense of information theory, the solution by means of the iterative method is expedited if the variables are classified according to this clustering.

CONTENTS

1. A QUANTITATIVE ANALYSIS OF COMPLEXITY

1.1. INTRODUCTION

Early computer applications have been concerned mainly with theoretically well-understood subjects like numerical analysis, administrative data processing, or linear programming. More recently, computer programs have been constructed for a great variety of problems in which the theoretical basis is less firm or even non-existent. In the following paragraphs some of these problems will be described very briefly to give the reader, who is not acquainted with them, an idea of the background of this tract.

A bottleneck in the practical use of a computer is the preparation of input data in a form readable by machine. The laborious conversion of hand or typewritten material to punched cards or magnetic tape would not be necessary if there were a machine for optically reading conventionally written characters. One of the possible designs for such a machine provides for the formation of a suitably enlarged image of a character on a two-dimensional array of devices sensitive to light, each of which generates a voltage corresponding to a shade of grey. The problem is that different occurrences of the same character generate, in general, different sets of voltages and, yet, must be recognized by the machine as being sufficiently similar without, however, confusing sets of voltages arising from different characters.

Written characters are but one example of a *pattern*; the more general problem of *pattern recognition* is to find basic techniques that may be applied to the recognition of such diverse patterns as electric signals from a microphone exposed to human speech, microscopic images of chromosomes, photographs of events in a bubble chamber, and so on. Most workers in these areas are trained in physics or in electrical engineering. A survey covering much work in this field is found in [39] and in [59]. A unified presentation of the store of ideas relevant in this, and also in a wider, context may be found in [63].

Quite a different "culture" is numerical taxonomy, so called after the most influential publication in this field (Sokal and Sneath [57]). The impression of a complete lack of communication between numerical taxonomists

and the technologists mentioned above is strengthened by the fact that none of the papers in a recent symposium on numerical taxonomy [15] contains a reference to, for instance, any of the articles by S. Watanabe that appeared from 1960 onwards containing much material relevant to and, apparently, unknown in numerical taxonomy. Most of this material may now be found in [63].

Sokal and Sneath were concerned with classification in a biological context; they understood classification to mean "the ordering of organisms into groups (or sets) on the basis of their relationships, ..." and taxonomy to mean the theoretical study of classification. In numerical taxonomy the relationship considered is that of similarity and its distinguishing method is the *numerical evaluation* of similarity. The outstanding aims of numerical taxonomy are *repeatability* and *objectivity* of the resulting classification; the lack of these they consider the most important failure of the "natural system".

Older than, but closely related to, numerical taxonomy is the use of systematic methods by plant ecologists to characterize different vegetation units. The so-called Franco-Swiss school following J. Braun-Blanquet recognizes vegetation units on the basis of their floristic composition. According to the typical method employed by this school, the basic data are collected in the following way. Throughout a fairly large geographical region small representative sampling areas are selected. For each of these areas each occurrence of a species belonging to a certain category (often, that of the vascular plants) is noted. It is then required to recognize certain sets of sampling areas as belonging to a particular vegetation unit. The original form of the method stressed the characterization of vegetation units according to the occurrence of "faithful" species. Although this method has received much adverse criticism, we expect that its spirit can be preserved in a more acceptable formulation by means of such information-theoretic concepts as are discussed in the present tract. For a recent survey of methods in plant ecology, together with a particular application, the reader is referred to Segal [52].

We shall proceed on the assumption that it is fruitful to search for principles that are equally relevant to endeavours such as pattern recognition, numerical taxonomy, and plant ecology. In each of these the problem

is that of classification, which attempts to group a set of objects into different classes such that objects of the same class are, in general, similar to each other and those in different classes are not.

There is an important difference between the problem of automatic classification as encountered in character recognition on the one hand and as encountered in biology on the other hand. In the first case, the criterion, according to which a method is to be judged, has an obvious property: it is to be a decreasing function of the cost of a machine that implements the method and of the average incidence of its misclassifications. Only the precise specification of this function presents a problem. In biology the requirement seems to be that the resulting classification be as "meaningful" as possible, or should "explain" as much as possible. It is much more difficult to compare any two out of the numerous methods proposed with respect to such a criterion. The difference may be summarized by saying that in character recognition one's aim is to save money and in biology one's aim should be to advance science.

It is a very unsatisfactory state of affairs in automatic classification in biology that there are numerous methods that demand consideration (for a survey of some methods studied by plant ecologists, see [33] and [34]) and that, in a particular situation, there is little on which to base a choice. It seems that the absence of a clearly definable criterion for a successful classification, which is clearly related to its purpose, is at the root of the difficulty.

The use of a numerical criterion does not by itself, as has been suggested, represent an advance over traditional methods if this criterion is not related to the *purpose* of the classification. The situation in numerical taxonomy may be illustrated by the following analogy. Imagine a situation where the different technologies would have evolved with a relative speed much different from the one actually observed, such that there would have been a computer technology as we at present have, but that engineering mechanics would still be at the medieval level. Then bridges would still be built, but on an appropriately smaller scale, and with little understanding of the mechanical principles involved. In such a situation, it may well be imagined that computer programs would be used to try different ways of putting stones or wooden beams together to form a bridge (just as numer-

ical taxonomists now use computer programs to construct classifications) and that a numerical criterion would be used to compare different designs.

Although this is reminiscent of the present situation in the automatic construction of classifications, it seems hard to push the analogy so far as to imagine that the numerical criterion used in bridge-design would have no clear connection with the purpose of the bridge; it is rather obvious that various designs would be compared by means of a numerical criterion which is a function of the weight the bridge can carry. This is obvious only insofar as the purpose of the bridge is obvious and is used as the guiding principle in its design.

Perhaps, something may be learned by studying some other situations where the criterion for the success of the classification is clear. One such situation is where a set of numerical variables have to satisfy a number of conditions, each of which is expressed as an equation in which one or more variables occur. In other words, it is required to solve a set of simultaneous equations. The fact that two variables occur in the same equation means that they interact: if one is changed the equation will, in general, no longer be satisfied unless the other undergoes a compensating change. In this context classification means that strongly interacting variables should be in the same class and variables in different classes should interact at most weakly. The success of a classification is unambiguously defined if the equations may be solved in such a way that the classification is taken advantage of.

Consider, as an example of such a method, a simple classification with only two classes V_1 and V_2 of variables. Suppose that the equations are also partitioned into disjoint classes E_1 and E_2. Solve the set E_1 for variables in V_1 assuming those in V_2 to be fixed at the value previously obtained (or, if no such values are available, at an arbitrary value). Then, keeping the variables V_1 fixed at the values just found, solve E_2 for variables V_2. Repeat this iteration until successive partial solutions have not changed. In this situation a good classification would be one where V_1, V_2, E_1, and E_2 are chosen in such a way that the variables of V_1 occur at most weakly in the equations of E_2, and vice versa. In this situation the better of two classifications would be the one requiring fewer iterations.

As a last example where a classification problem arises, consider the situation where a designer has to construct a *form* which is determined by a large number of variables, which have to satisfy a large number of conditions. The form might be the lay-out of a human settlement (Alexander [1]) determined by about a hundred variables which have to satisfy a number of conditions of the same order of magnitude. In general, there may not exist a form which satisfies all conditions to the required extent, and the designer aims at maximizing a goodness-of-fit criterion with respect to all conditions simultaneously. The designer cannot pay attention to all variables at once; suppose he finds an iterative design process by first concentrating on some subset V_1 of the variables and a suitable subset E_1 of the conditions, finding a provisional form that maximizes goodness-of-fit locally, and then proceeding with other subsets V_2 and E_2. Interaction between two variables occurs when both are involved in the same condition. If, initially, the condition is satisfied and one of the variables is changed, the condition is, in general, no longer satisfied unless the other variable is subjected to a compensating change.

Thus, if there is interaction between V_1 and V_2, the designer may have to start anew, because, when concentrating on V_2, he has made changes that necessitate compensating changes in V_1, and vice versa. If there is not too much interaction between V_1 and V_2, the successive approximations become more satisfactory. Thus we see that here, too, success of the method depends on good classification. Alexander [1] was concerned with architectural and industrial design; Brams [11,12] recognized that Alexander's method may be used for classification in another context.

In the last two examples the criterion of successful classification is quite clear: it is the number of iterations required to attain a satisfactory solution (or form). As we said, it seems to us that the use of mathematical methods cannot be attempted if the criterion is of the form "most meaningful" or "explaining as much as possible". It does not seem to be possible to bridge with mathematical reasoning the gap between criteria of this form and the operation of an algorithm suitable for execution by computer. Therefore, we propose to study in its most general aspect the iterative solution of systems of equations. What makes the system difficult to solve, or as it may also be said, what makes the system *complex*, is inter-

action of variables. In the above examples, if the two sets of variables do not interact, one iteration suffices. Now, what can one do if a problem is too complex? One can try to apply the strategy of "divide, and rule"; in this situation it means to find out whether the complex system is, perhaps, composed in a *simple* way of *complex* subsystems. Each of these subsystems presents a smaller problem, and it may be attacked in the same way. Such a decomposition of a complex system into a simple system of complex subsystems seems to us a classification of which the success may be tested in an unambiguous way.

By studying the phenomenon of complexity in its own right, we shed light on the problem of classification in this context. Because complexity also occurs in a wider context, we expect that a better understanding of complexity will contribute to a conceptual foundation of classification in biology. To do this, criteria of the form: "most meaningful" or "explaining as much as possible" would have to be expressed in terms of complexity and operations, mathematically defined in terms of complexity, can then be expressed unambiguously in terms of an algorithm that can be executed by computer. In this tract, only a small part of this program is carried out: a mathematical definition of complexity is attempted; consequently, it is shown that the classification of qualitative data and of quantitative data may be expressed in terms of complexity; in the case of quantitative data, the corresponding concept of complexity in a covariance matrix turns out to be of mathematical interest; finally, the solution of a certain type of system of equations is treated along the lines sketched in this introduction.

1.2. INTERACTIONS AS ADDITIVE CONTRIBUTIONS TO COMPLEXITY

We intend to make precise the concept of *complexity* and the investigations to be described in the sequel are motivated by the desire to see whether *entropy* is as useful for the measurement of complexity as it has proved to be (in the mathematical theory of communication) for the measurement of the possible information content of a signal. A definition of complexity which is simple and not yet precise enough to motivate a mathematically defined measure, is:

Complexity is the way in which a whole is different from the composition of its parts.

Let us supplement this by postulating that the complexity of the whole is more than the sum of the complexities of the parts if it is different from their composition and, if not different, equal to it. In the first case, we say that there is interaction between the parts and that the amount of interaction equals the difference between the complexity of the whole and the sum of the complexities of the parts. Therefore, if we can find a suitable measure for these interactions, we shall regard them as additive contributions to complexity; that is, for every partition of the whole, its complexity should equal the sum of the interactions between the parts plus the sum of the complexities of the parts considered by themselves.

Let us call the whole a *system* and its parts *subsystems*. Suppose, for instance, that a system S has been partitioned into subsystems S_1 and S_2, which are sub-partitioned into S_{11}, S_{12} and S_{21}, S_{22}, S_{23}, respectively, and so on. We shall restrict ourselves to those domains of investigation where there is some stage where the partitioning can be carried no further; the corresponding subsystems are called the *components* $V_1,\ldots,V_n$, which are considered to be the elements of the set S.

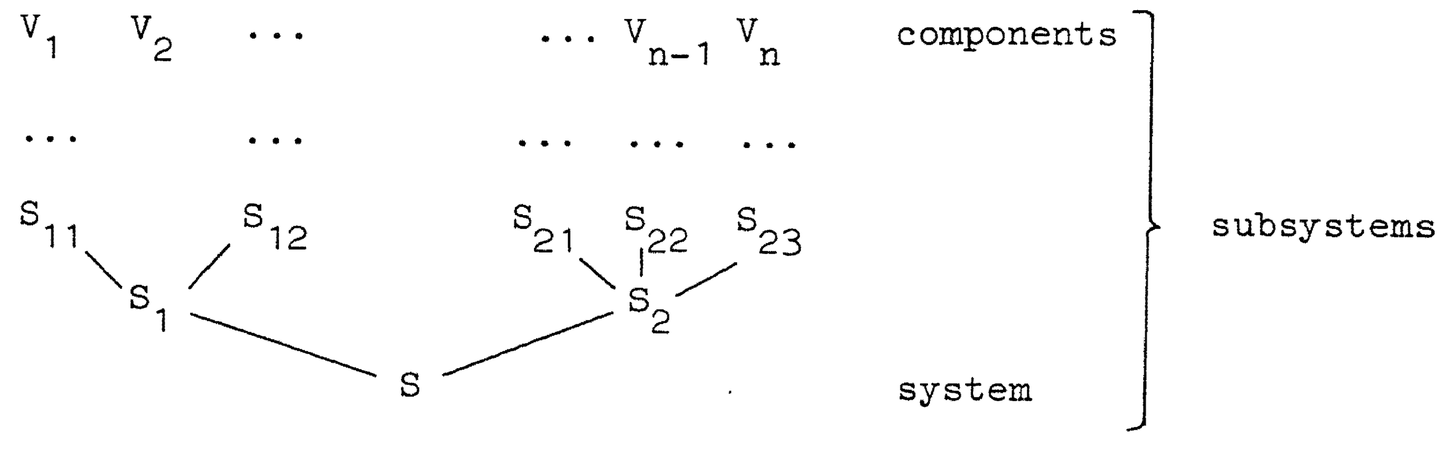


figure 1.1

A decomposition of the system S

This may be pictured as in figure 1.1, where the number of partitions between V_i and S may depend on i. Let us write C for the complexity of a subsystem and R for the interaction between subsystems. We can express

the above in a formula as follows:

$$C(S) = R(S_1,S_2) + C(S_1) + C(S_2) =$$
$$= R(S_1,S_2) + R(S_{11},S_{12}) + R(S_{21},S_{22},S_{23}) +$$
$$+ C(S_{11}) + C(S_{12}) + C(S_{21}) + C(S_{22}) + C(S_{23})$$
$$\ldots$$
$$C(S) - C(V_1) - \ldots - C(V_n) =$$
$$= R(S_1,S_2) + R(S_{11},S_{12}) + R(S_{21},S_{22},S_{23}) + \ldots$$

This last expression gives the difference between the complexity at the level of S and the sum of the complexities at the level of the V's in terms of the interactions at the partitions needed to obtain the decomposition of S into $V_1,\ldots,V_n$. Within a certain domain of investigation it may well be the case that the decomposition of subsystems can only be done a finite number of times. When a certain system is studied, this number depends, in general, on the means with which decomposition is carried out. For instance, in the study of matter, the level at which subsystems appear as atomic components depends on the maximum energy of the disturbances taken into account. We shall take for granted that, in our case, $V_1,\ldots,V_n$ cannot be decomposed any further. Accordingly, we shall take $C(V_1) + \ldots + C(V_n)$ as the zero level of complexity and this yields an expression for the complexity of S which is a sum of interactions only.

To define, as we have done, complexity as the sum of the complexities of the parts plus the interaction between them is like defining an onion as a smaller onion with a skin around it. After taking away the skin, it turns out that the smaller onion also has a skin around it. When we continue to separate onion from skin, we end up with all skin and no onion. In the same way, when we try to separate complexity from interaction, we keep on finding interactions and the complexity itself is elusive. Unlike the example of the onion, the decomposition process may not have an end. Atoms of matter turned out to be composed of elementary particles, and these, in turn, proved to be composite.

Now that complexity has been interpreted as a sum of interactions, we

only have to express interaction in a mathematical definition, which we shall find in information theory. Inevitably, perhaps, the result is much more restricted in applicability than the intuitively understood concept of complexity. Such is also the case with, for instance, "force". When we say "By the sheer force of his personality, ..." something at once richer and more vague is denoted than the product of mass and acceleration, which is the meaning of "force" in physics. "Complexity" will have undergone an equally great and, we hope, an equally useful change by the end of this chapter.

Of course, the considerations in this section are only of interest if it is possible to give a mathematical definition of a system to which the above description is applicable. Such a definition is given in 1.4, where it is also shown that it allows an amount of interaction to be defined which has the properties discussed above. In chapter 2 such a system is used as a mathematical model for qualitative data.

1.3. ELEMENTS OF THE QUANTITATIVE STUDY OF INFORMATION

1.3.1. INFORMATION AND ENTROPY

In the previous section we argued that complexity can be reduced to a sum of interactions. We should, therefore, find a measure of interaction applicable to subsystems. Information theory provides such a measure which is applicable between subsystems of a very general nature. It derives from information theory (Shannon [53]), where entropy was introduced as a measure of uncertainty. In this section we present the elements of information theory in such a way that the relation between information and uncertainty is emphasized.

Let there be random variables x and y with outcomes $x_1,\ldots,x_m$ and $y_1,\ldots,y_n$, respectively, and with the joint probability distribution $\Pr(x=x_i \text{ and } y=y_j) = r_{ij} > 0$. Let the marginal distributions be

$$s_i = r_{i1} + \ldots + r_{in} \qquad \text{for } i = 1,\ldots,m \text{ and}$$

$$t_j = r_{1j} + \ldots + r_{mj} \qquad \text{for } j = 1,\ldots,n.$$

We shall first consider the case where n = 2 and we shall study the situation where only the outcome of x can be observed. In such a situation one may be interested in the information contained in an outcome of x about the corresponding unknown outcome of y. Let the conditional probability $Pr(x=x_i|y=y_1) = p_i$ and $Pr(x=x_i|y=y_2) = q_i$, i = 1,...,m. Suppose that the outcome of x is known to be x_i and that p_i is much greater that q_i. Then one would consider the outcome y_1 of y more likely than without this knowledge: one can say that the outcome of x contains information about the corresponding outcome of y. This can be made more precise with the aid of the formula for conditional probability:

$$\begin{aligned} &Pr(x=x_i|y=y_j) = \\ &= Pr(x=x_i \text{ and } y=y_j)/Pr(y=y_j) \\ &= Pr(y=y_j|x=x_i)Pr(x=x_i)/Pr(y=y_j) \end{aligned}$$

for j = 1,2, where $Pr(x=x_i) = s_i$ and $Pr(y=y_j) = t_j$ are the marginal probabilities. Hence,

$$\frac{Pr(y=y_1|x=x_i)}{Pr(y=y_2|x=x_i)} = \frac{Pr(y=y_1)\ Pr(x=x_i|y=y_1)}{Pr(y=y_2)\ Pr(x=x_i|y=y_2)}\ .$$

Good [20] introduced the quantities:

$$\begin{aligned} &O(y=y_1|x=x_i) = \\ &= Pr(y=y_1|x=x_i)/Pr(y=y_2|x=x_i) \quad \text{and} \\ &O(y=y_1) = Pr(y=y_1)/Pr(y=y_2) \end{aligned}$$

to represent, respectively, the *odds* of $y = y_1$ given $x = x_i$ and the initial odds of $y = y_1$. Their quotient he called the factor in favour of $y = y_1$ in virtue of the observation $x = x_i$. This gives rise to the discrimination information (Kullback [31]) contained in the observation $x = x_i$ for the discrimination between $y = y_1$ and $y = y_2$:

$$\ln \frac{p_i}{q_i} = \ln \frac{Pr(x=x_i|y=y_1)}{Pr(x=x_i|y=y_2)} = \ln(O(y=y_1|x=x_i)) - \ln(O(y=y_1)).$$

The mean of this expression under condition $y = y_1$ is called the discrimination information of the distribution $p = (p_1,\ldots,p_m)$ against the distribution $q = (q_1,\ldots,q_m)$:

$$(1.1) \qquad I(p;q) = \sum_{i=1}^{m} p_i \ln(p_i/q_i).$$

The less the distributions differ, the less information an outcome of x contains about an outcome of y. In fact, the discrimination information satisfies Gibbs' inequality:

$$(1.2) \qquad I(p;q) = -\sum_{i=1}^{m} p_i \ln(q_i/p_i) \geq -\sum_{i=1}^{m} p_i(q_i/p_i-1) = 0,$$

which holds in virtue of the fact that, for $a > 0$, $\ln(a) \leq a-1$ with equality if and only if $a = 1$. This shows that Gibbs' inequality is an equality if and only if the distributions are the same: $p_i = q_i$ for $i = 1,\ldots,m$. Thus, the discrimination information may be interpreted as a measure of the difference between the two distributions.

In his mathematical theory of communication, Shannon [53] introduced the notion of uncertainty in the outcome of a discrete random variable. He sought to express it as a function H of the probability distribution $(p_1,\ldots,p_m)$ of the random variable, which he required to have the following properties:

(1.3) H is continuous in the p_i.

(1.4) If $p_i = 1/m$ for $i = 1,\ldots,m$, then H should be a monotonic increasing function of m.

(1.5) $H(p_1,\ldots,p_m) = H(p_1,\ldots,p_{m-2},a) + aH(p_{m-1}/a,p_m/a)$, for all permutations of the p_i's and where $a = p_{m-1} + p_m$.

Shannon showed that such a function H must have the form $H = -K\sum_{i=1}^{m} p_i \ln(p_i)$, where K is a positive constant. We shall suppose the units to be chosen such that $K = 1$.

Now, we can express the information contained in an outcome of x about the corresponding outcome of y as the difference between $H(t_1,t_2)$, the prior uncertainty of y, and the average posterior uncertainty. Under the condition that $x = x_i$, the posterior uncertainty is

$$H(Pr(y=y_1|x=x_i),Pr(y=y_2|x=x_i)) = H(r_{i1}/s_i,r_{i2}/s_i).$$

Because $Pr(x=x_i) = s_i$, we have for the average posterior uncertainty:

$$\begin{aligned}
&\sum_{i=1}^{m} s_i\, H(r_{i1}/s_i,r_{i2}/s_i) = \\
&= -\sum_{i=1}^{m} (r_{i1}\ln(r_{i1}/s_i) + r_{i2}\ln(r_{i2}/s_i)) = \\
&= \sum_{i=1}^{m} (s_i\ln(s_i) - p_i\,t_1\ln(p_i t_1) - q_i\,t_2\ln(q_i t_2)) = \\
&= H(t_1,t_2) - (H(t_1p+t_2q) - t_1\,H(p) - t_2\,H(q)).
\end{aligned} \tag{1.6}$$

We use Jensen's inequality to show that the second term cannot be negative, which implies that the average posterior uncertainty is not greater than the prior uncertainty. Let f be a concave function of one real argument. Let $a_1,\ldots,a_n$ be non-negative and let $w_1,\ldots,w_n$ also be non-negative and have unit sum. Then Jensen's inequality states that (see, for instance, Hardy, Littlewood, and Pólya [24], theorem 86):

$$f(w_1a_1+\ldots+w_na_n) \geq w_1f(a_1) + \ldots + w_nf(a_n).$$

If we put $f(x) = -x\ln(x)$ for $x > 0$ and $f(0) = 0$, we have

$$\begin{aligned}
&H(t_1p+t_2q) - t_1H(p) - t_2H(q) = \\
&= \sum_{i=1}^{m} (f(t_1p_i+t_2q_i) - t_1f(p_i) - t_2f(q_i)) \geq 0.
\end{aligned}$$

We have $0 < t_1 < 1$ because all r_{ij} were assumed to be positive. Then we can have equality only if $p_i = q_i$ for $i = 1,\ldots,n$. In that case, the average posterior uncertainty equals the prior uncertainty and the discrimination information of the p_i against the q_i vanishes. Therefore, we say that there is no information contained in an outcome of x about the

corresponding outcome of y. Also, if $p_i = q_i$ for $i = 1,\ldots,n$ then $r_{ij} = s_i t_j$, which means that x and y are statistically independent.

Suppose now that p and q are not the same distribution. For which prior probabilities t_1 and $t_2 = 1-t_1$ is the difference (in terms of information) between p and q greatest? That is, what can one say about the maximum of $H(t_1p+t_2q) - t_1H(p) - t_2H(q)$ if t_1 is allowed to vary between 0 and 1? To see what happens, let us consider the first and second derivatives.

$$(d/dt_1)(H(t_1p+t_2q) - t_1H(p) - t_2H(q)) =$$

$$= -\sum_{i=1}^{m} (1 + \ln(t_1p_i+t_2q_i))(p_i-q_i) - H(p) + H(q) =$$

$$= \sum_{i=1}^{m} (p_i\ln(p_i) - p_i\ln(t_1p_i+t_2q_i) - q_i\ln(q_i) + q_i\ln(t_1p_i+t_2q_i)) =$$

$$= I(p;t_1p+t_2q) - I(q;t_1p+t_2q).$$

$$(d^2/dt_1^2)(H(t_1p+t_2q) - t_1H(p) - t_2H(q)) =$$

$$= (d/dt_1) \sum_{i=1}^{m} - (p_i-q_i) \ln(t_1p_i+t_2q_i) =$$

$$= \sum_{i=1}^{m} - (p_i-q_i)^2/(t_1p_i+t_2q_i).$$

For $0 < t_1 < 1$ the second derivative is negative, which implies that a maximum of $H(t_1p+t_2q) - t_1H(p) - t_2H(q)$ is unique and it occurs for that value of t_1 for which (1.7) vanishes. It is the value for which $t_1p + t_2q$ is as much different (in terms of discrimination information) from p as it is from q.

The relation between discrimination information and uncertainty is illustrated in the following situation. Let $(x_1,\ldots,x_m)$ be the outcomes of a random variable x and $Pr(x=x_j) = p_j$, $j = 1,\ldots,m$ with $p_1 + \ldots + p_m = 1$. Suppose that the outcome of x cannot be observed with certainty; that is, if the outcome is x_i, the probability is r_{ij} that x_j is observed, $r_{1j} + \ldots + r_{mj} = 1$ for $j = 1,\ldots,m$. In the special case where there is no uncertainty in the observation, we have $r_{ii} = 1$, and therefore also

$$r_{1j}p_1 + \ldots + r_{mj}p_m = p_j, \qquad \text{for } j = 1,\ldots,m. \tag{1.8}$$

We shall consider a more general situation, where we do not necessarily have $r_{ii} = 1$, but where condition (1.8) still holds. If the outcome is x_i, the posterior uncertainty is $-\sum_{j=1}^{m} r_{ij} \ln(r_{ij})$. This occurs with probability p_i; hence, the average posterior uncertainty is

$$\sum_{i=1}^{m} p_i \sum_{j=1}^{m} - r_{ij} \ln(r_{ij}).$$

If the outcome is x_i, then we are interested in the discrimination information between the distributions $(r_{i1},\ldots,r_{im})$ and $(p_1,\ldots,p_m)$ which is $\sum_{j=1}^{m} r_{ij} \ln(r_{ij}/p_j)$; hence, the average discrimination information is

$$\sum_{i=1}^{m} p_i \sum_{j=1}^{m} r_{ij} \ln(r_{ij}/p_j) =$$

$$= - \sum_{j=1}^{m} \ln(p_j) \sum_{i=1}^{m} r_{ij}\, p_i + \sum_{i=1}^{m} p_i \sum_{j=1}^{m} r_{ij} \ln(r_{ij}) =$$

$$= \sum_{j=1}^{m} - p_j \ln(p_j) - \sum_{i=1}^{m} p_i \sum_{j=1}^{m} - r_{ij} \ln(r_{ij}). \text{ (by condition (8))}$$

Thus, we have shown that the average discrimination information in the experiment is the prior uncertainty minus the average posterior uncertainty.

1.3.2. INTERACTION AS A MEASURE OF DEPENDENCE

Suppose that a random variable x has outcomes $(x_1,\ldots,x_m)$ and that a random variable y has outcomes $(y_1,\ldots,y_n)$ and that x and y have joint probability distribution

$$r_{ij} = \Pr(x=x_i \text{ and } y=y_j), \quad i = 1,\ldots,m, \quad j = 1,\ldots,n.$$

Let the marginal distributions be $p_i = \sum_{j=1}^{n} r_{ij}$ for $i = 1,\ldots,m$ and $q_j = \sum_{i=1}^{m} r_{ij}$ for $j = 1,\ldots,n$.

Consider the discrimination information between the jointly distributed variables, denoted by x,y and the variables distributed according to the product of the marginal distributions, denoted by $x \times y$:

$$(1.9)\qquad I(x,y;x\times y) = \sum_{i=1}^{m}\sum_{j=1}^{n} r_{ij}\,\ln(r_{ij}/(p_i q_j)).$$

According to the previously given interpretation of discrimination information, this is the mean information in favour of the joint distribution against the hypothesis of independence. According to Gibbs' inequality (1.2), I is zero in case $r_{ij} = p_i q_j$ for all i and j, which implies statistical independence, and is positive otherwise. This quantity may therefore be called the *informational measure of dependence* between x and y as implied by their joint distribution. It will often be called the *interaction* R in the joint distribution, or between the random variables distributed according to it:

$$R(x,y) = I(x,y;x\times y) \geq 0.$$

We may express this in entropies as follows:

$$(1.10)\qquad R(x,y) = H(x) + H(y) - H(x,y) \geq 0.$$

If R attains its minimum, then H(x,y) = H(x) + H(y); R can therefore also be regarded as the amount by which the information in x,y is short of its maximum, hence, we shall also call R the *redundancy* in x,y because it is the information contained in x and y separately that is redundant. Shannon [53] calls R(x,y)/H(x,y) redundancy; we shall refer to this quotient as the *relative* redundancy.

If we have k random variables $x_1,\ldots,x_k$ we have analogously to (1.10):

$$(1.11)\qquad \begin{aligned} R(x_1,\ldots,x_k) &= I(x_1,\ldots,x_k;x_1\times\ldots\times x_k) \\ &= H(x_1) + \ldots + H(x_k) - H(x_1,\ldots,x_k). \end{aligned}$$

For instance, we find for the interaction between x_1 and the joint distribution $x_2,\ldots,x_k$:

$$\begin{aligned} R(x_1,(x_2,\ldots,x_k)) &= I(x_1,x_2,\ldots,x_k;x_1\times(x_2,\ldots,x_k) \\ &= H(x_1) + H(x_2,\ldots,x_k) - H(x_1,x_2,\ldots,x_k)\ . \end{aligned}$$

To obtain more insight into the interaction between random variables, we shall study not only the entropy of a marginal distribution, like $H(x)$ in (1.10), but also of a conditional distribution. For the entropy of x under condition that $y = y_j$ we have:

$$H(x|y=y_j) = -\sum_{i=1}^{m} (r_{ij}/q_j) \ln(r_{ij}/q_j)$$

$$= \ln(q_j) - (1/q_j) \sum_{i=1}^{m} r_{ij} \ln(r_{ij}).$$

If we average this expression over the outcomes y_j of y, we obtain $H(x|y)$, which is called the conditional entropy of x given y:

$$H(x|y) = \sum_{j=1}^{n} q_j \ln(q_j) - \sum_{j=1}^{n} \sum_{i=1}^{m} r_{ij} \ln(r_{ij}) \tag{1.12}$$

$$= H(x,y) - H(y).$$

Thus we find

$$H(x,y) = H(y) + H(x|y) \quad \text{and, similarly}$$

$$= H(x) + H(y|x); \text{ hence,}$$

$$R(x,y) = H(x) - H(x|y) = H(y) - H(y|x). \tag{1.13}$$

These results were obtained by Shannon [53], who used them in his model for the transmission of information through a noisy channel. In this model, the input is represented as a random variable y and the output as a random variable x. The mean information transmitted between a pair of outcomes of y and x is then $R(x,y) = H(y) - H(y|x)$ which is the information contained in y (the input) minus the uncertainty in y given x. This last quantity, the equivocation, is zero if x and y are identically distributed, which means no noise in the channel, and positive otherwise. $R(x,y)$ can be interpreted as the information about y contained in x. Indeed, in our introductory example, we found that (see (1.6)) $H(y) = H(t_1,t_2)$ and $H(y|x) = H(t_1,t_2) - H(t_1p+t_2q) + t_1H(p) + t_2H(q)$ and, hence, $R(x,y) = H(t_1p+t_2q) - t_1H(p) - t_2H(q)$.

We saw that $H(x,y) - H(y)$ is the entropy of the conditional distribution averaged over the outcomes of y. This is the reason for writing $H(x|y)$ for $H(x,y) - H(y)$. Intuitively, it is apparent that $H(x|y) \geq H(x|y,z)$; the intuition being that the uncertainty of x cannot be increased by the knowledge of another random variable z, however irrelevant it may be. Let $p(x,y,z)$ denote the joint probability of x, y, and z; $p(x,y)$ the marginal distribution $\sum_z p(x,y,z)$; $p(z|x,y)$ the conditional distribution $p(x,y,z)/p(x,y)$; and so on for the other variables.

THEOREM 1.2

(Khinchin [29]; the shorter proof given here is similar to Gallager [17]).
$H(x|y) - H(x|y,z)$ vanishes if $p(x|y) = p(x|y,z)$ for all values of (x,y,z) such that $p(y,z) > 0$, and is positive otherwise.

PROOF

$$
\begin{aligned}
& H(x|y) - H(x|y,z) = \\
& = H(x) - H(x|y,z) - (H(x) - H(x|y)) = \\
& = H(x) + H(y,z) - H(x,y,z) - (H(x) + H(y) - H(x,y)) = \\
& = \sum_{x,y,z} p(x,y,z) \ln(p(x,y,z)/(p(x)p(y,z))) + \\
& \quad - \sum_{x,y} p(x,y) \ln(p(x,y)/(p(x)p(y))) = \\
& = \sum_{x,y,z} p(x,y,z) \ln(p(x|y,z)/p(x|y)) = \\
& = - \sum_{y,z} p(y,z) \sum_{x} p(x|y,z) \ln(p(x|y)/p(x|y,z)),
\end{aligned}
$$

where summation is understood to involve only those (x,y,z) for which $p(y,z) > 0$. According to Gibbs' inequality (1.2), the inner sum vanishes only if $p(x|y) = p(x|y,z)$ for all values of x, and is negative otherwise. This concludes the proof.

1.3.3. DISTANCE IN TERMS OF INTERACTION

In the previous section we saw that the interaction R between the discrete random variables x and y vanishes if they are statistically independent. It will be useful to have a function of two random variables that vanishes, for instance, when these random variables are the same and that does not, in general, vanish when they are statistically independent. We shall see that there is a function that has these properties.

Consider a function f that is defined for pairs of arbitrary entities X, Y, and Z that has the following properties:

(1.14) $f(X,Y) \geq 0$ with equality if $X = Y$,

(1.15) $f(X,Y) = f(Y,X)$,

(1.16) $f(X,Y) + f(Y,Z) \geq f(X,Z)$.

Such a function has the most important properties that a distance function should have; more precisely, such a function f is known as a *pseudometric*. It would be called a metric if it would also have the property that $f(X,Y) = 0$ implies $X = Y$.

LEMMA 1.1

If x and y are random variables as introduced in the beginning of 1.3.2, then

(1.17) $D(x,y) = H(x,y) - R(x,y)$

is a pseudometric.

PROOF. By (1.4), (1.10), and (1.12) we have

$$D(x,y) = 2H(x,y) - H(x) - H(y),$$

(1.18)
$$= H(y|x) + H(x|y), \text{ and}$$
$$= \sum_{i=1}^{m} p_i \, H(y|x=x_i) + \sum_{j=1}^{n} q_j \, H(x|y=y_j).$$

The entropy of a discrete probability distribution is non-negative and vanishes only if one of the probabilities is 1. Therefore, D vanishes only if, for each outcome x_i of x with positive probability, there is one outcome of y that has positive probability, and vice versa. In such a case x and y are said to be functionally dependent and we can conclude that D(x,y) vanishes if and only if x and y are functionally dependent discrete random variables. Any discrete random variable is functionally dependent upon itself and, therefore D satisfies (1.14). It is trivial to verify that D also satisfies (1.15).

To show that D also satisfies the triangle inequality (1.16), suppose that x, y, and z are discrete random variables for which a simultaneous distribution function exists.

$$D(x,y) + D(y,z) - D(x,z) =$$

$$= 2H(x,y) - H(x) - H(y) + 2H(y,z) - H(y) +$$

$$- H(z) - 2H(x,z) + H(x) + H(z) =$$

$$= 2(H(x,y) + H(y,z) - H(x,z) - H(y)) \geq$$

$$\geq 2(H(x,y) + H(y,z) - H(x,y,z) - H(y)) = \qquad \text{by (1.5)}$$

$$= 2H(x|y) - 2H(x|y,z) \geq 0 \qquad \text{by theorem 1.2.}$$

This completes the proof of Lemma 1.1.

If we replace the entropies in (1.18) by entropies conditional on a third variable, say z, we obtain an expression that may be called the conditional distance D(x,y|z). This may be shown not to exceed the corresponding unconditional distance:

$$D(x,y) - D(x,y|z) =$$

$$= 2H(x,y) - H(x) - H(y) - 2H(x,y|z) + H(x|z) + H(y|z) =$$

$$= 2R((x,y),z) - R(x,z) - R(y,z).$$

$$R((x,y),z) - R(x,z) =$$

$$= H(x,y) + H(z) - H(x,y,z) - H(x) - H(z) + H(x,z) =$$

$$= H(z|x) - H(z|x,y) \geq 0, \qquad \text{by theorem 1.2.}$$

In a similar fashion we may verify that

$$R((x,y),z) - R(y,z) \geq 0, \quad \text{whence our result}$$

$$D(x,y) \geq D(x,y|z).$$

This completes the proof of Lemma 1.1.

Note that $D(x,y) \leq H(x,y)$; this suggests a normed distance d to be defined as

$$d(x,y) = D(x,y)/H(x,y) \quad \text{if } H(x,y) > 0$$
$$= 0 \quad \text{if } H(x,y) = 0. \tag{1.19}$$

THEOREM 1.3

d is a pseudometric.

PROOF. It is readily verified from the definition that d is non-negative and that it is symmetrical in its arguments. We also have $d(x,x) = 0$ if $H(x) = 0$, by the definition, if $H(x) > 0$ because $D(x,x) = 0$.

We shall now show that d satisfies the triangle inequality (1.16) for any simultaneously distributed discrete random variables x, y, and z. If, for any two pairs from (x,y,z), the joint entropy vanishes, x, y, and z are pairwise functionally dependent and the triangle inequality is trivially satisfied. In the case where only $H(x,z) = 0$, (1.16) is also trivially satisfied. Suppose that the joint entropy vanishes for another pair, say (x,y). In that case we have $H(x) = H(y) = 0$, $H(x,z) = H(y,z) = H(z)$, and $d(x,y) = 0$, $d(y,z) = 1$, $d(x,z) = 1$.

Therefore, we only have to consider the case where $H(x,y) > 0$, $H(y,z) > 0$, and $H(x,z) > 0$ are simultaneously satisfied. We shall distinguish the following possibilities:

A: $H(x,z)$ is a greatest among $H(x,y)$, $H(y,z)$, $H(x,z)$.

B: $H(x,z)$ is neither the greatest nor the smallest among $H(x,y)$, $H(y,z)$, $H(x,z)$.

C: $H(x,z)$ is a smallest among $H(x,y)$, $H(y,z)$, $H(x,z)$ and $H(y) \leq H(x,z)$.

D: $H(y) > H(x,z)$.

A:
$$d(x,y) + d(y,z) - d(x,z) \geq$$
$$\geq (2H(x,y) - H(x) - H(y) + 2H(y,z) +$$
$$- H(y) - H(z) - 2H(x,z) + H(x) + H(z))/H(x,z) =$$
$$= (D(x,y) + D(y,z) - D(x,z))/H(x,z) \geq 0, \qquad \text{by Lemma 1.1.}$$

B: Suppose that, in addition to the condition already mentioned, $H(x,y) \geq H(x,z) \geq H(y,z)$.

$$d(x,y) + d(y,z) - d(x,z) \geq$$
$$\geq (2H(x,y) - H(x) - H(y) + 2H(y,z) +$$
$$- H(y) - H(z))/H(x,y) - 2 + (H(x) + H(z))/H(x,y) =$$
$$= 2(H(x,y) + H(y,z) - H(y))/H(x,y) - 2 =$$
$$= 2(H(y,z) - H(y))/H(x,y) \geq 0.$$

The assumption that $H(y,z) \geq H(x,z) \geq H(x,y)$ gives a completely analogous derivation.

C:
$$d(x,y) + d(y,z) - d(x,z) \geq$$
$$\geq 2 - (H(x) + H(y))/H(x,z) +$$
$$+ 2 - (H(y) + H(z))/H(x,z) +$$
$$- 2 + (H(x) + H(z))/H(x,z) =$$
$$= 2(H(x,z) - H(y))/H(x,z) \geq 0.$$

D: Suppose that, in addition to the condition already mentioned, $H(x,y) \geq H(y,z)$.

$$d(x,y) + d(y,z) - d(x,z) \geq$$

$$\geq 2 - (H(x) + H(y))/H(y,z) +$$

$$+ 2 - (H(y) + H(z))/H(y,z) +$$

$$- 2 + (H(x) + H(z))/H(y) =$$

$$= (2H(y,z) - H(x) - H(z))/H(y,z) +$$

$$- (2H(y) - H(x) - H(z))/H(y) +$$

$$+ 2(H(y,z) - H(y))/H(y,z) =$$

$$= (H(x) + H(z))/H(y) +$$

$$- (H(x) + H(z))/H(y,z) +$$

$$+ 2(H(y,z) - H(y))/H(y,z) \geq 0.$$

The assumption that $H(x,y) \leq H(y,z)$ gives a completely analogous derivation.

This completes the proof of theorem 1.3.

Jardine and Sibson [28] used the fact that d is a distance function. For a proof they referred to Rajski [45], but this is a mistake: Rajski, although, as far as we know, the first to state the fact, apparently thought a proof too tedious to write down.

1.4. COMPLEXITY IN TERMS OF ENTROPY

We shall discuss partitions in a finite set T so far as to be able to show that everything derived for random variables with a finite set of outcomes can also be interpreted in terms of partitions. Suppose an equivalence relation is defined among the elements of T. It is well-known (see, for instance, [36], p. 21) that such a relation corresponds to a partition

in T, which is a set (we shall call it X) of mutually disjoint subsets of T (we shall call them $X_1,\ldots,X_m$, the classes of the partition) whose union is T. The correspondence is that the equivalence relation between two elements of T holds if and only if they are in the same class of the partition. Let X_i contain n_i elements and let $n = n_1 + \ldots + n_m$, which is, then, the number of elements in T. We shall call $p(X_i) = n_i/n$ the relative frequency of X_i in T.

Suppose we have two partitions in T, say $X = (X_1,\ldots,X_m)$ and $Y = (Y_1,\ldots,Y_k)$. We shall define another partition in T which is called their joint partition (X,Y): Two elements t_i and t_j of T are in the same class of the joint partition if and only if they are in the same class of X and also in the same class of Y. In this way, each pair (X_i,Y_j) of a class from X and a class from Y defines a class in (X,Y), which we denote by $(X,Y)_{ij}$.

We shall also use the product partition X × Y, which is a partition in the set T × T, the set of all ordered pairs of elements of T. Two elements (t_i,t_j) and (t_i',t_j') of T × T are, by definition, in the same class of X × Y if and only if t_i and t_i' are in the same class of X and t_j and t_j' are in the same class of Y. In this way, each pair (X_i,Y_j) of a class of X and a class of Y defines a class of the partition X × Y, which we denote $(X\times Y)_{ij}$. This class has n_in_j elements and the total number of elements of T × T is n^2; therefore, we find for the relative frequency of $(X\times Y)_{ij}$ in T × T: $p((X\times Y)_{ij}) = n_in_j/n^2 = p(X_i)p(Y_j)$.

It is often possible to say that each element from a set is of one of a certain number m of different *kinds*. If being the same kind is an equivalence relation, this defines a partition in the set; each class of the partition corresponding to a kind. Then we can say that a certain amount of *variety* exists in the set. Suppose an amount H of variety were to have the following three properties:

a) H is a continuous function of the relative frequencies of the classes, and of these only.
b) In case each of the relative frequencies equals 1/m, H sould be a monotonic increasing function of m.

c) $H(p_1,\ldots,p_m) = H(p_1,\ldots,p_{m-2},a) + aH(p_{m-1}/a,p_m/a)$, where $a = p_{m-1} + p_m$ and where $p_1,\ldots,p_m$ are the relative frequencies of the classes. This should hold for all permutations of the p's.

These are the same properties that Shannon required the uncertainty of a random variable to have. We think that these are also properties that one would reasonably expect an amount of variety to have. The meaning of property c) in terms of variety is the following. Suppose that we have a partition X' in T such that $X'_1 = X_1,\ldots,X'_{m-2} = X_{m-2},X'_{m-1} = X_{m-1} \cup X_m$.

We could say that in X' the kinds indexed by m-1 and m have become indistinguishable, and we would expect the amount of variety to have decreased. According to property c) this is indeed the case; in effect, it says that the difference is the amount of variety in $X_{m-1} \cup X_m$ under the partition (X_{m-1},X_m) multiplied by the relative frequency of $X_{m-1} \cup X_m$ in T.

It seems reasonable to require an amount of variety to have these three properties, and we shall do so in the sequel. Then, as Shannon showed, the amount H of variety must have the form $-K \sum_{i=1}^{m} p_i \ln(p_i)$, the entropy of the set of relative frequencies. Again, we shall take the unit of entropy such that K = 1. Ashby ([5], Chapter 7) considers an amount of variety equal to ln(m), which corresponds to our definition in the case $n_1/n = \ldots = n_m/n = 1/m$.

To every partition $X = (X_1,\ldots,X_m)$ of a set T there corresponds a random variable x with outcomes $x_1,\ldots,x_m$ and with probabilities $Pr(x_i) = n_i/n$, the relative frequency of X_i in T. If we have another random variable y corresponding in the same way to a partition $Y = (Y_1,\ldots,Y_k)$, the joint partition (X,Y) corresponds to a joint distribution of x and y. In the same way, the product partition corresponds to a random variable having as outcomes pairs (x_i,y_j), i = 1,...,m, j = 1,...,k such that $Pr(x_i,y_j) = Pr(x_i)Pr(y_j)$. In this way, all results of information theory described in section 1.3 may be interpreted in terms of partitions of a finite set.

We may interpret the results as applying to certain *weight distributions*; the weights may be interpreted as either relative frequencies of the classes of a partition or as probabilities of a random variable. In particular, the entropy function is defined on a set of weights. If the weights are interpreted as probabilities, entropy is an amount of uncertainty. If

the weights are interpreted as the relative frequencies of a partition, entropy is an amount of variety.

Let us now take up the definition of a system where we left it at the end of 1.2. There we argued that the total amount of complexity equals the sum of the interactions corresponding to decompositions necessary to decompose the system into its set of components. We now specify the nature of the components $V_1,\ldots,V_n$ of the kind of system that we shall consider to be some partitions $X_1,\ldots,X_n$ of a set T of arbitrary objects. The system S is then defined to be the joint partition of $X_1,\ldots,X_n$ and the composition of the components to be their product partition. We assume that the interactions between the parts of the system are a function only of the relative frequencies of the partitions involved. Then, with respect to interactions and complexity, the system may be regarded as a set $S = (V_1,\ldots,V_n)$ of weight distributions, which are the marginal distributions of a given joint distribution. We are free to associate the relative frequencies of a partition with this joint distribution, or a set random variables.

We defined complexity to be "the way in which a whole is different from the composition of its parts". Henceforth, we shall consider the system to be different from the composition of its components if the joint distribution is not the same as the product of the marginal distributions. Apart from the *way in which*, we are also interested in the *amount by which* the system is different from the composition of its components, and this would be the amount C(S) of complexity in the system S; this amount we define to be the discrimination information of the joint distribution against the product of marginal distributions. According to (1.2), this quantity can only vanish if the distributions are identical and is positive otherwise. According to (1.10) we have

$$\begin{aligned} C(S) &= H(V_1) + \ldots + H(V_n) - H(S) = \\ &= \sum_i H(V_i) - H(S_1) + \sum_j H(V_j) - H(S_2) + \\ &\quad + H(S_1) + H(S_2) - H(S) = \\ &= C(S_1) + C(S_2) + R(S_1,S_2), \end{aligned}$$

where $\sum_i$ means summation over all $V_i \epsilon S_1$ and $\sum_j$ means summation over all $V_j \epsilon S_2$.

The set $(X_1,\ldots,X_n)$ of partitions of a finite set T corresponds to what is known in statistics as an n-dimensional contingency table. Each class of a partition X_i corresponds to a category in the contingency table. In chapter 2 we are especially interested in a situation (the "object-predicate table") where n is large (say, 100) and where, for each of these partitions, the number of classes is small (typically, two). In the resulting 2^{100} contingency table almost all cells are empty because the number of elements in T is in the same order of magnitude as n.

C(S) is, apart from a constant factor, the log likelihood-ratio applicable when testing the hypothesis of independence between all coordinates in the contingency table. The study of contingency tables from the point of view of information theory originated with McGill and Garner [35,18] and was continued by Kullback [31]. Although there are difficulties involved in testing of hypotheses in a table as described above (for difficulties arising in a two-dimensional table, see [37]), we think that such analyses of C(S) as discussed in 2.2.1 and in 2.2.2 are relevant to descriptive statistics.

2. ANALYSIS OF QUALITATIVE DATA

2.1. A "STRUCTURE" AS DEFINED IN QUALITATIVE DATA

In the literature of numerical taxonomy a distinction is usually made between qualitative and quantitative data. The latter may be regarded as values of continuous variables. We shall assume quantitative data to be real numbers with such an interpretation that the usual operations of addition and multiplication make sense. An object described by a set of such quantities may then be regarded as a point in linear vector space if it is assumed, in addition, that the postulates for such a space make sense. For instance, if objects x and y are represented by points in linear vector space, any linear combination of these points must represent a possible object.

These assumptions are rather restrictive and are often not justified for the sort of objects that biologists, sociologists, or planners are interested in. The criticism regarding the use of mathematical methods in these fields is sometimes based on the implicit assumption that objects are represented by sets of quantities as described above. In planning, the criticism takes the form that "values" are of overriding importance in human affairs and that mathematical methods necessarily "quantify" these, which is inadmissable. This criticism is partially answered by Negroponte who described the situation as follows: "The handling of qualitative information is too often presumed unsuitable for the constitution of machines. Or it is granted feasibility only through abortive techniques of quantification." [40, p. 62].

In the literature on numerical taxonomy an exact description of "qualitative data" is lacking, although it is generally agreed that they do not satisfy the above criteria for quantitative data; in particular, they are assumed to be values of a *discrete* variable. But it is often not stated whether these values are supposed to be ordered and, if so, what are the algebraic properties of the ordering. For instance, in the context of plant ecology, "qualitative" data indicate presences or absences of species of plants. These may be regarded as rounded-off quantities, but this cannot be said of the qualitative data considered in sociology or planning.

Before methods for the treatment of qualitative data are considered, we shall assume that they are values of a variable having only a finite set of possible values; we shall not assume any ordering between these values and we shall not assume that any operations are defined between values, irrespective of whether they are from the same variable. This means that all that can be said of a variable taking an object as argument is that it corresponds to a partition in the set of objects. This is the reason for our considering the components of the system (see 1.4) to be partitions in set T of arbitrary objects. Of this method Ashby [6] writes: "As its concepts are initially quite free from any implication of either continuity, or of order, or of metric, or of linearity (though in no way excluding them) the method can be applied to the facts of biology without the facts having to be distorted for merely mathematical reasons." In the literature two interpretations of qualitative data (which are different from the one described above), namely as truth values or as rounded-off quantities, are encountered. We shall discuss them briefly.

Qualitative data are often represented in the form of a rectangular array of zeroes and ones. One interpretation is the following. Each row of the array corresponds to an *object* and each column to a *predicate*. The j-th entry of the i-th row of the array shows whether the i-th object O_i does (when it is one), or does not (when it is zero), possess the j-th predicate P_j; that is, the entry is the truth value of the proposition "$P_j(O_i)$".

In [62] an array of zeroes and ones denoting truth values is introduced under the name "object-predicate" table. In order to obtain a set of "most significant" predicates the following procedure is carried out. Initially, the entries are identified with the real numbers denoted by the same symbols "zero" or "one". Subsequently, the columns are imbedded in a vector space over the real numbers and the matrix of inner products between pairs of columns is formed. Those predicates whose representative points have smallest distance to a subspace spanned by k first eigenvectors of the matrix are considered to be most significant. Considering that truth values are the original meaning of the entries of the object-predicate table, the validity of this procedure is at least not obvious; for some applications a justification may well exist, but, then, it should be given explicitly.

In some applications where object-predicate tables arise such a justification is possible because the entries may not only be interpreted as truth values, but also as real numbers rounded off in the extreme. For instance, in plant ecology an object may be a sampling area and a predicate then corresponds to the presence of a particular species of plant in that area. Often, if less than a certain percentage of the area is covered by plants of this species, its presence is considered negligible and a zero is entered. In such a case the entries zero or one may be regarded as rounded-off real numbers. But then we are dealing with quantitative data where the method of principal components is applicable (see 3.2). In such a case, rounding off is a (possibly) necessary evil and it should only be done where computational advantage outweighs loss of information. It is not to be expected that rounding off to two values always turns out to be optimal.

Moreover, truth values do not seem to be particularly suited for the representation of qualitative data. Suppose objects are animals and there are predicates like "smooth", "hairy", "prickly", and "meat-eating". Of each of these one can separately determine the truth value, but then the special relationship between the first three is lost. It seems better to consider a variable "texture of skin" that can take as values "smooth", "hairy", or "prickly". We shall not use "predicate" in the sense of the predicate calculus of logic, but we shall use it to denote a variable taking an object as argument and having a finite number of values. We do not assume that an ordering, partial or complete, exists among these values; also, we do not assume any operation to be defined on them. Therefore, we can only say that each predicate corresponds to a partition in the, supposedly finite, set of objects. For convenience, we shall exhibit object-predicate tables where the predicates assume only two values. In order to emphasize that the entries are arbitrary marks, we write them as nought ("0") or cross ("X").

In 1.4 we defined the components $V_1,\ldots,V_n$ of a system as partitions $X_1,\ldots,X_n$ in a finite set T. This system is an object-predicate table if the predicates are the partitions $X_1,\ldots,X_n$ and T the set of objects. Two objects are in the same class of the partition X_i, $i = 1,\ldots,n$, if the i-th predicate assumes the same value for them. Thus, an object-predicate table defines a set of partitions in the objects and this we define to be a

structure in the set of objects. In chapter 3 we shall consider the case where objects may be represented as points in an n-dimensional vector space with an inner product and we shall see how a set of objects induces a structure in this space in an analogous fashion.

2.2. DECOMPOSITIONS OF COMPLEXITY

2.2.1. HIERARCHICAL DECOMPOSITION OF COMPLEXITY IN AN OBJECT-PREDICATE TABLE

We believe that the identification of an object-predicate table with a system is of practical importance. When viewed as a system of interacting components it is possible to express the purpose of a method of analysis in terms of complexity. When viewed as an array of marks identifying the values of the predicates it is easy to compute the joint partition in the set of objects of any set of predicates and then to compute the interaction between sets of predicates, which is a contribution to complexity.

It is important that the computation of interaction be not restricted to pairs of predicates, because otherwise complicated patterns of interaction, that are not restricted to pair-wise effects, would be beyond analysis. Of course, the computational effort required is greatly reduced when most of the interaction is accounted for by pair-wise effects and it is important to be able to detect this.

To a diagram as in figure 1.1 there corresponds a hierarchical decomposition of the total amount C(S) of complexity. Hierarchical, because each proper subset of components has only one other subset (or the entire set) as its immediate predecessor; a decomposition, because at each split a part of the remaining amount of complexity is converted into an interaction. The effect of the complete decomposition is that C(S) is found to be equal to a sum of interactions.

Such a decomposition scheme may be constructed in many different ways. Which of these is to be preferred depends on the purpose of the analysis. One purpose could be to find a "natural" subdivision into subsystems, that is, a division such that between subsystems there is little interaction compared to the amounts of complexity within. This means that as much in-

teraction as possible should correspond to splits high in the decomposition scheme. We hope that a system, whose complexity is beyond us, is composed in a *simple* way of *complex* subsystems. Applying such a decomposition corresponds to the well-known tactic: divide, and rule. Such a system may also be called "near-decomposable" or it may be said to have a "clustering" structure. To find such a structure, if present, or else to show that none exists, entails considerable computational difficulties.

The quest for near-decomposable structure seems relevant to the purpose of "general systems theory" [9]. This theory abstracts from properties peculiar to physical, biological, or social systems in order to find properties applicable to all of them. However, little attention is paid by von Bertalanffy [9] to the significance of near-decomposability. On the other hand, Simon [55] studies hierarchic structure in a variety of systems. He argues that the very mechanism of evolution of complex systems, whether natural or artificial, makes for a near-decomposable structure (the parable of Tempus and Hora).

Another purpose may be to give as succinct as possible a summary of interactions present in the system. A decomposition scheme useful to this purpose would have strong interactions associated with splits low in the scheme; the summary is obtained by disregarding all interactions above a certain level. Williams and Lambert [68] introduced "association analysis", which was intended for use with quantitative data rounded off to two values: "presence" and "absence".

Their work is remarkable because the method is applicable to qualitative data. They used "association" instead of interaction, which is, perhaps, unfortunate, because most people think of it as something like positive correlation. They included the positive as well as the negative, in short, what we call interaction. They did not give a numerical definition of interaction, but used instead a numerical criterion for deciding which subdivision to effect. This criterion involves computing tail probabilities for testing independence in large numbers of 2×2 contingency tables. Unfortunately, these are about the only contingency tables (see, for instance, [37], p. 317) where the asymptotically-approximating chi-squared distribution gives poor results; so either prohibitively laborious exact calculations are called for, or else corrections must be applied. Without any

corrections, a mathematical analysis of the properties of this criterion seems rather formidable; with them, it seems hopeless.

2.2.2. DECOMPOSITION OF COMPLEXITY ACCORDING TO ORDER OF INTERACTION

In section 2.2.1 we described a way of writing complexity as a sum of terms, each of which is the interaction between two sets of components. Here, we shall first define the amount d_k of interaction of order k in the system, which involves all sets of components of size k. The complexity turns out to be the sum of such amounts of interaction where k runs through $2,\ldots,n$. We shall prove that the d_k are monotone non-decreasing with increasing k.

Let the components of the system be a set $(x_1,\ldots,x_n)$ of jointly distributed random variables. We define the average entropy of order k as:

$$\bar{H}_k = \binom{n}{k}^{-1} \sum H(y_1,\ldots,y_k),$$

where the summation is over all subsets $(y_1,\ldots,y_k)$ of $(x_1,\ldots,x_n)$. The average interaction in subsets of size k is defined as:

$$\bar{R}_k = \binom{n}{k}^{-1} \sum (H(y_1) + \ldots + H(y_k) - H(y_1,\ldots,y_k)) =$$

$$= -\bar{H}_k + \binom{n}{k}^{-1} \sum (H(y_1) + \ldots + H(y_k)),$$

where summation is over all subsets $(y_1,\ldots,y_k)$ of $(x_1,\ldots,x_n)$. Each element, say y_i, $i = 1,\ldots,n$, occurs at most once in a subset. There are $\binom{n-1}{k-1}$ subsets in which it occurs; therefore

$$\bar{R}_k = -\bar{H}_k + \binom{n}{k}^{-1} \binom{n-1}{k-1} (H(x_1) + \ldots + H(x_n)) =$$

$$= -\bar{H}_k + (k/n)\,(H(x_1) + \ldots + H(x_n)) =$$

$$= k\bar{H}_1 - \bar{H}_k.$$

We define

$$d_k = \bar{R}_k - \bar{R}_{k-1}$$

and find

$$d_k = k\bar{H}_1 - \bar{H}_k - (k-1)\bar{H}_1 + \bar{H}_{k-1} = \bar{H}_1 + \bar{H}_{k-1} - \bar{H}_k;$$

we find for the complexity in $(x_1,\ldots,x_n)$

$$R(x_1,\ldots,x_n) = H(x_1) + \ldots + H(x_n) - H(x_1,\ldots,x_n) =$$

$$= n\bar{H}_1 - \bar{H}_n =$$

$$= \sum_{k=2}^{n} (\bar{H}_1 + \bar{H}_{k-1} - \bar{H}_k) =$$

$$= d_2 + d_3 + \ldots + d_n.$$

THEOREM 2.1.

$$d_{k+1} - d_k \geq 0 \quad \text{for } k = 2,\ldots,n.$$

PROOF. We shall first show that d_k equals the average of

$$H(y_1) + H(y_2,\ldots,y_k) - H(y_1,\ldots,y_k) =$$

(2.1)

$$= H(y_1) - H(y_1|y_2,\ldots,y_k)$$

over all subsets $(y_1,\ldots,y_k)$ of $(x_1,\ldots,x_n)$. If, instead of averaging over subsets, we average over ordered k-tuples $y_1,\ldots,y_k$, we obtain the same result because each subset is repeated an equal number of times. We shall write $n^{!i}$ for $n(n-1)\ldots(n-i+1)$ and we shall use

$$(1/n^{!i}) \sum_{y_1,\ldots,y_i} H(y_1,\ldots,y_k) =$$

(2.2)

$$= (1/n^{!i})\,(n-k)\ldots(n-i+1) \sum_{y_1,\ldots,y_k} H(y_1,\ldots,y_k) =$$

$$= (1/n^{!k}) \sum_{y_1,\ldots,y_k} H(y_1,\ldots,y_k) = \bar{H}_k,$$

which holds for $n \geq i \geq k \geq 1$. For the average of (2.1) we find

$$(1/n^{!k}) \sum_{y_1,\ldots,y_k} (H(y_1) + H(y_2,\ldots,y_k) - H(y_1,\ldots,y_k)) =$$

$$= \bar{H}_1 + \bar{H}_{k-1} - \bar{H}_k = d_k.$$

Because of (2.1) and (2.2) we also have

$$d_k = (1/n^{!(k+1)}) \sum_{y_1,\ldots,y_{k+1}} (H(y_1) - H(y_1|y_2,\ldots,y_k)),$$

which allows us to write

$$d_{k+1} - d_k =$$

$$= (1/n^{!(k+1)}) \sum_{y_1,\ldots,y_{k+1}} (H(y_1|y_2,\ldots,y_k) - H(y_1|y_2,\ldots,y_{k+1})).$$

According to theorem 1.2 every term in the sum in non-negative, which concludes the proof.

The quantities d_k were introduced by Watanabe. He showed [60] that $d_n = 0$ implies that $x_1,\ldots,x_n$ are statistically independent. Ashby [7] considers differences of $\bar{R}_k$ of any order, but does not derive inequalities for them.

From the definition of $\bar{R}_k$ we may conclude that the faster the sequence $\bar{R}_2,\ldots,\bar{R}_n$ increases, the greater is that part of the total amount of complexity that cannot be accounted for by interactions in small subsets. Although we have no absolute criterion that says when to consider the increase fast, theorem 2.1 implies that this increase must be at least linear. The computation of the sequence $\bar{R}_2,\ldots,\bar{R}_n$ allows us to compare, at least, two systems in this respect. This requires the computation of $\bar{H}_1,\ldots,\bar{H}_n$. The first and the last of these are easy to obtain; the calculation of $\bar{H}_2$ and $\bar{H}_{n-1}$, $\bar{H}_3$ and $\bar{H}_{n-2}$,... require rapidly increasing numbers of subsets. We should therefore revert to estimates from random samples of subsets, which need not be small if a high-speed computer is used.

2.3. CLASSIFICATION AND CLUSTERING

2.3.1. REMARKS ABOUT A MEASURE OF CLUSTERING

In order to be able to find a clustering structure one must, in the first place, be able to detect it; that is, there will have to be agreement as to which of two alternative decomposition schemes shows the more marked clustering structure. In that case, the problem of finding the optimum decomposition with respect to clustering may be formulated mathematically. The merit of this is but slight if, as in the present case, the solution of the problem presents great computational difficulties because the number of possible decomposition schemes increases so fast with the number of components. Yet, such a formulation seems to be not altogether superfluous. Several publications [10, 13, 33, 34, 57] have stressed the need for a mathematical approach to the problem of clustering and classification and have provided algorithms suitable for execution by computer. However, no mention is made of a criterion according to which the decomposition obtained can be compared to other decompositions. Thus, although elaborate computations are made, the problem is not stated for which the outcome is intended to be a solution; neither is it stated for which problem the outcome is meant to be an approximate solution.

We shall give some considerations relevant to a measure of clustering which allows different decomposition schemes to be compared with respect to degree of clustering. Such a scheme, partially shown in figure 1.1, may be regarded as the mathematically defined object called a *tree*. A tree consists of a set of *nodes* and a binary relation, called *successor*, among them. Each node, except one, the *root*, is the successor of exactly one node. The number of successors of a node is called the *degree* d of that node. With each node there is associated a real number q called the *flow* of that node. The flow of the root is 1; the successors of a node with flow q and degree d have flow q/d. With each node there is associated an integer r called the *rank*. The rank of the root is zero; the successors of a node have a rank which is greater by 1. Let T be the set of nodes that have no successor (the *terminal* nodes). Let I be the set of the remaining nodes (the *internal* nodes).

In the decomposition scheme each node corresponds to a subsystem. The successors of a node are disjoint sets of components and their union is this node. To each node i that has at least two successors there corresponds a real number R_i equal to the interaction between its successors. The decomposition scheme showing a high degree of clustering has little interaction between subsystems compared to amounts of complexity within. Eventually, all complexity is converted to interaction and in the ideal decomposition scheme, therefore, the stronger interactions should be associated with the higher ranks. This suggests a weighted average of the interactions as a measure of clustering. Suppose f is some increasing function of the non-negative integers, then the decomposition scheme for which

$$(2.3) \qquad m = \sum_{i \in I} f(r_i)\, R_i$$

is greater is considered to show stronger clustering, subject to the constraint discussed below.

We cannot maximize m over all decomposition schemes without constraint: m may also be increased by a peculiar structure of the tree, which we shall call "lopsidedness".

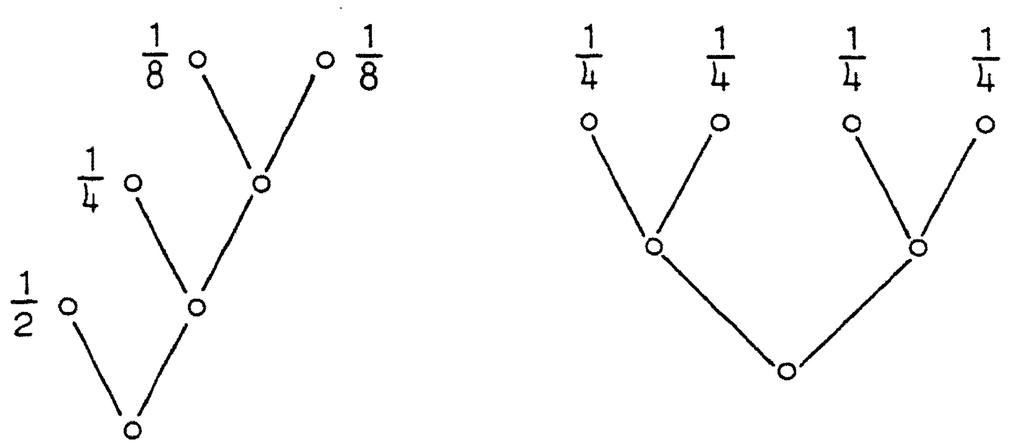


figure 2.1. A lopsided tree and a balanced tree

In figure 2.1 two trees are shown. A node is represented by a small circle. A successor of a node is drawn above it and connected to it with a line. The tree on the left we consider more "lopsided" than the one of the right, which is more "balanced". In a lopsided tree, we can have the situation that m is large, not because of a marked clustering, but because of nodes of a rank higher than any in a more balanced tree. A function that

indicates the degree of clustering would have to be not only an increasing function of m, but also a decreasing function of the lopsidedness of the tree.

As it happens, there is a real-valued function defined on trees that plays an important rôle in the theory of information and that may well be interpreted as a measure of balance. We take as criterion for a balanced tree the equality of the terminal flows; an obvious way to characterize it is to observe that the entropy of the set of terminal flows is maximal. Let us therefore define

$$h = -\sum_{i \in T} q_i \ln q_i. \tag{2.4}$$

Let us briefly indicate the rôle of this quantity in the theory of communication, which is the *capacity of a discrete noiseless channel*. To simplify the explanation, suppose that the tree is such that all non-terminal nodes have the same degree d. Then $q_i = d^{-r_i}$ and $h = \ln(d) \sum_{i \in T} q_i r_i$, a multiple of the average rank of the terminal nodes. The following brief remark should make it plausible that this is also the capacity of a discrete noiseless channel.

Consider an information source which emits symbols. Each symbol is encoded into a sequence of code symbols of which there are d, each of the same length. An encoding is represented as a terminal node in the tree. Optimum transmission of information is obtained if the probabilities of the source equal the flows of the corresponding terminal nodes. Then, the information contained in a unit length of encoded message is on the average, apart from a constant factor, equal to h. This is discussed as "coding for the discrete noiseless channel" in textbooks on information theory, such as [4]. For a monograph devoted to the informational study of trees, see [44].

We conclude that a measure of clustering should be an increasing function both of m in (2.3) and of h in (2.4). To specify the function further one would have to take into account what purpose the result is to serve; in the absence of such considerations the measure of clustering should be chosen such that the amount of computation needed to find the optimum decomposition scheme is minimum.

2.3.2. CLUSTERING AND THE EXTRACTION OF RELEVANT PREDICATES

The definition of a relevant subset of the set of predicates in an object-predicate table may be illustrated by a guessing game: one person takes an object in mind and has to answer another person who tries to identify the object with as few as possible questions of the form: "Which value does predicate p_i have for this object?". The answers to questions concerning a subset of the predicates define a partition in the set of objects. The information contained in answers to these questions is the entropy of the corresponding partition.

For instance, if all predicates have two values, the set of n predicates defines a partition of 2^n cells and the maximum entropy of such a partition is $\ln(2^n)$. When the actual entropy is less, there is redundancy in the set of predicates. When we realize 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 (between, say, 10 and 1000) and roughly equal numbers of objects and predicates enormous amounts of redundancy are usual.

Lance and Williams [32] have used information-theoretic considerations
n classification. They used the "information statistic" $I = H(p_1)+\ldots+H(p_n)$
o express the amount of information contained in the set of predicates.
ctually, this amount equals the entropy $H(p_1,\ldots,p_n)$ of the joint partition which is equal to I only if there is no interaction whatever between predicates. In this case, all interactions vanish and there is no clustering at all. Apparently, they were not concerned with classification in the sense described above. Moreover, it is difficult to understand in what sense their "classification" is to be interpreted. A hint is given in a later paper [33]:

"The agglomerative strategies ... can themselves be subdivided ...: by clustering strategies we imply those that optimize some property of a group of elements; by hierarchical strategies those that optimize the route by which groups are obtained".

The term "agglomerative" refers to an algorithm that generates a diagram as in figure 1.1 by starting with individual components and succes-

sively merging subsets. "Optimizing the route" seems to imply that with each route a certain number is associated, but no such number is defined and, as we saw in the previous section, it may not be easy to find an obviously satisfactory number. Also, in deciding which two subsets to merge, a measure of similarity between them is taken into account. Four of these are described (one is based on the information statistic mentioned above) and again, in the absence of a precisely defined optimal classification, there is little on which to base the choice of measure. Because Lance and Williams consider most of the combinations between one of the five strategies and between one of the four measures of similarity admissable, and because they can give only hints about which to use in a particular situation, the resulting classification has little to justify it, apart from the possible fact that the user likes it.

In our introduction 1.1 we argued that the problem of classification is unnecessarily complicated if, as is usually the case, the purpose of the classification cannot be expressed in terms of some simple criterion. It is preferable to study classification in a situation where such a simple criterion can be found and we gave two examples where useful classification is defined in terms of interaction between mutually disjoint sets of entities.

When we have qualitative data in the form of an object-predicate table, the interacting entities are the predicates and the classification problem is the same as that of finding "near-decomposable" or clustering structure. As explained in 2.2.1, this means that a decomposition scheme is wanted where as much as possible of the total amount of interaction is associated with internal nodes of high rank in the corresponding tree. However, it is out of the question to try all possible trees to find the best classification.

Watanabe [63, p.427] has proposed an economical method to find a clustering structure that proceeds in two steps. In the first step, a small subset of the predicates is found that gives almost as much information as the set of all predicates. Watanabe does this by means of the covariance matrix of the object-predicate table. As we explained in 2.1, this is not valid in the general case of qualitative data, nor is it, probably, meant to be. We shall define such a subset in terms of optimal data compression

and we shall discuss the accompanying computational difficulties. When such a subset is found, each of its elements is regarded as the representative of a different class. Each of the remaining elements is assigned to that class of which the representative has greatest interaction with it. We shall see that, if a good classification is present, this method does not always find it.

Considerations about the information content of a set of predicates suggest the problem: For a given $k \leq n$ (and $k \geq 1$), find a subset $(q_1,\ldots,q_k)$ of the predicates $(p_1,\ldots,p_n)$ such that the entropy H_k of their joint partition is close to H_n. Such a subset we shall call a *relevant subset*; a set of k predicates such that no other set of the same size has a larger joint entropy we shall call a *maximal* subset. We are interested in finding such a subset with k small compared to n; in that case we can neglect the predicates not in the relevant subset and yet incur only a small information loss (equal to $H_n - H_k$). This operation may be called *data compression*; the corresponding operation for quantitative data is discussed in 3.2. Consider the following inequalities:

$$\begin{aligned} H(p_1) + \ldots + H(p_n) &\geq H(p_1,\ldots,p_n) \geq \\ &\geq H(q_1,\ldots,q_k) \qquad \text{(by (1.10) and (1.5))} \end{aligned} \tag{2.5}$$

$$\begin{aligned} H(p_1) + \ldots + H(p_n) &\geq H(q_1) + \ldots + H(q_k) \geq \\ &\geq H(q_1,\ldots,q_k) \qquad \text{(by (1.10))} \end{aligned} \tag{2.6}$$

For data compression to be interesting, k must be small. If we take k as small as possible, namely $k = 1$, we rarely get a sufficiently informative subset. So we look for some constraint on k that prevents it from becoming too small. The inequalities (2.5) and (2.6) imply no ordering between the expressions in their middles. A natural way of preventing k from becoming too small seems to be to require that

$$H(q_1) + \ldots + H(q_k) \geq H(p_1,\ldots,p_n). \tag{2.7}$$

The maximal subset is *optimal* if it satisfies (2.7) and if it is the smallest that does so. Optimal data compression for qualitative data is the

replacement of a set of all predicates by an optimal subset.

In our problem Watanabe's method for the extraction of relevant predicates is not applicable. One must look for something else; an obvious try seems to be the following:

Suppose again that $(p_1,\ldots,p_n)$ is the set of predicates. Let S_j, $j=0,\ldots,k$ be a sequence of successive approximations to a maximal subset of size k. Take for S_0 the empty set and let S_{j-1} be a subset of S_j, which is formed by adding to $S_{j-1} = (q_1,\ldots,q_{j-1})$ the q_j such that $H(q_1,\ldots,q_j)$ + $- H(q_1,\ldots,q_{j-1})$ is maximum.

The reason for considering this procedure is that, by adding at each step the predicate that gives the greatest increase in entropy, one might end up with a maximal subset. That this is not necessarily the case is shown by the object-predicate table in table 2.1.

	objects →																	
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
predicates ↓ 1	x	x	x	x	x	x	x	x	x	o	o	o	o	o	o	o	o	o
2	x	x	x	o	o	o	o	o	o	x	x	x	x	x	o	o	o	o
3	x	x	x	x	o	o	o	o	o	o	o	o	x	x	x	x	x	x

table 2.1: An awkward object-predicate table

	o	x
p_1	9	9
p_2	10	8
p_3	8	10

	oo	ox	xo	xx
p_1,p_2	4	5	6	3
p_2,p_3	5	5	3	5
p_3,p_1	3	5	6	4

table 2.2: Showing the number of times a given configuration occurs in table 2.1

Suppose we require a maximal subset of 2 predicates. Table 2.2 shows that $H(p_1) > H(p_2) = H(p_3)$ and that $H(p_1,p_2) = H(p_1,p_3) < H(p_2,p_3)$. This implies that, if we follow the above procedure, the subsets (p_1) and (p_1,p_2) (or (p_1,p_3)) would have been obtained. Yet (p_2,p_3) is the maximal subset of size 2. Although it may well be possible to construct examples

in which the procedure gives the desired result, it is, apparently, not always the case.

Finding a maximal subset of predicates may be compared with the problem of the "travelling salesman" in which there is given a set of n cities, for each pair of which the mutual distance is given. The problem is to construct an itinerary that includes all cities and that has a minimum total length. One of the first things that a student of this problem discovers is that the required itinerary is not necessarily obtained by travelling at each stage to the closest city not already visited, which would be analogous to the above procedure. Much attention has been paid to this problem. The methods proposed are either not optimal or else require an amount of computing time that increases so fast with n that they are impracticable even on fast computers for n in the order of, say, a few hundred.

The second step in Watanabe's method seems to be based on the premise that, for any p, there must be some *single* q_j from among a relevant subset $(q_1,\ldots,q_k)$ such that $R(p,q_j)$ is large compared to others. It is true that, if $H_n - H_k$ is small compared to H_n, $R(p,(q_1,\ldots,q_k))$ must be large, but it is not necessarily the case that this is due to a single $R(p,q_j)$, $j = 1,\ldots,k$, being large.

$$R(p,(q_1,\ldots,q_k)) = H(p) + H_k - H(p,q_1,\ldots,q_k) \geq$$
$$\geq H(p) + H_k - H_n. \qquad (2.8)$$

If $H_n - H_k$ is small, then the inequality implies that any predicate p cannot have an interaction with $(q_1,\ldots,q_k)$ much less than its own entropy.

		objects →							
		1	2	3	4	5	6	7	8
predicates ↓	1	o	o	o	o	x	x	x	x
	2	o	o	x	x	o	o	x	x
	3	o	x	o	x	o	x	o	x
	4	o	x	x	o	o	x	x	o

table 2.3: An awkward object-predicate table

There is already a simple example where the assumption, on which the second step is based, is not justified. The object-predicate table in table 2.3 shows that, although the interaction between p_4 and a maximal subset (p_1,p_2,p_3) is large, this is not due to any *pairwise* interaction between p_4 and p_j, $j = 1,2,3$. Let us try to find a clustering among the predicates $p_1,\ldots,p_4$, that is, it is required to partition them into subsets C_1 and C_2, such that $R(C_1,C_2)$ is as small as possible. The total number of predicates is so small here that we shall allow subsets of size 1 and 3 as well as 2 and 2. According to Watanabe's method, one first finds a relevant subset of predicates and, subsequently, assigns any remaining predicate to the relevant predicate with which it has greatest interaction.

	p_1,p_2	p_1,p_3	p_2,p_3
oo	2	2	2
ox	2	2	2
xo	2	2	2
xx	2	2	2

	p_1,p_4	p_2,p_4	p_3,p_4
oo	2	2	2
ox	2	2	2
xo	2	2	2
xx	2	2	2

	p_1,p_2,p_3	p_1,p_2,p_4	p_1,p_3,p_4	p_2,p_3,p_4
ooo	1	1	1	2
oox	1	1	1	o
oxo	1	1	1	o
oxx	1	1	1	2
xoo	1	1	1	o
xox	1	1	1	2
xxo	1	1	1	2
xxx	1	1	1	o

table 2.4: Showing the number of times a given configuration occurs in table 2.3.

For (p_1,p_2,p_3) to be a relevant subset $H(p_1,p_2,p_3,p_4) - H(p_1,p_2,p_3)$ has to be small. With tables 2.3 and 2.4 one may verify that it even vanishes, so (p_1,p_2,p_3) certainly is a relevant subset. An application of

Watanabe's method would imply 3 clusters containing, respectively, p_1, p_2, and p_3. Which of $R(p_i,p_4)$, i = 1,2,3, is largest would decide to which of these classes p_4 belongs. However, each of these vanishes, which makes the clusterings $((p_1,p_4),p_2,p_3)$, $(p_1,(p_2,p_4),p_3)$, and $(p_1,p_2,(p_3,p_4))$ look equally good.

When we consider such *pairs* as can be selected from (p_1,p_2,p_3), we find that (see table 2.4) $R((p_1,p_2),p_4) = 0$, $R((p_1,p_3),p_4) = 0$, and $R((p_2,p_3),p_4) = H(p_4)$, the maximum amount. Thus we see that, although there is no *single* one relevant predicate that interacts strongly with p_4, there is a *pair*, namely (p_2,p_3). This suggests that (C_1,C_2), with $C_1 = (p_1)$, $C_2 = (p_2,p_3,p_4)$ is a clustering. Indeed, for these choices of C_1 and C_2 we have $H(C_1) + H(C_2) - H(C_1,C_2) = 0$ and equal to ln(2) for all other choices of C_1 and C_2.

To recapitulate, (2.8) implies that, if $(q_1,\ldots,q_k)$ is a relevant subset, then, for any p, $R((q_1,\ldots,q_k),p)$ is close to H(p). Whether such is also the case when we replace $(q_1,\ldots,q_k)$ by a proper subset C, or even a C consisting of just one element, such that R(C,p) is large, depends on the data and such a supposition cannot be relied on in a generally applicable clustering method. Therefore, if we were to use a relevant subset C_k = $= (q_1,\ldots,q_k)$ as a starting point for clustering, the work remaining after finding it is still enormous: to find a good home for some remaining predicate, not only single elements of C_k would have to be considered, but also pairs, triples, etc. The conclusion is that, if a feasible method can be found for finding relevant subsets (and the example of table 2.1 shows that this is at least a non-trivial problem), it is not necessarily a good starting point for finding a clustering.

2.3.3. CLASSIFICATION AND CLUSTERING IN METRIC SPACE

A classification (not necessarily "good" or meaningful) means a partition of a set of entities into mutually disjoint classes whose union is this set. A subset of the entities, one from each class, is called a set of *paradigms*. Suppose each of the remaining entities is assigned to the same class as the paradigm most similar to it. Then a classification, in general different from the initial one, is obtained. If we get the same, the set of

paradigms is said to be *representative* for the classification, and if every set of paradigms is representative, the classification is said to be *perfect.*

Notice that the above concepts are based entirely on similarities between *pairs* of entities. If interaction is interpreted as similarity, good classification means strong clustering. However, interaction is not only defined between pairs, but also between sets of arbitrary size. To assume that the total amount of interaction in a set is mainly due to pairwise interactions is a simplification that is justifiable only in special cases. With few exceptions, work in automatic classification has taken for granted the validity of this assumption and many methods use as their basic material a matrix of dissimilarities between the entities to be classified. Even then, such a fundamental distinction, as introduced above, is not made. In this subsection, we shall give some of its properties in a suitable model.

We think that a suitable model is obtained by regarding the entities to be classified as a finite set of points in a metric space. This model is more general than the one of a linear vector space with inner product which is used for most pattern-recognition research [39, 51, 59]. The vector-space model is appropriate for perception-like data, that is, data which are, even if they are outputs of threshold devices, basically quantities, although rounded-off in the extreme.

It seems to us that a measure of dissimilarity should have the following properties. It should be symmetrical in its arguments. Furthermore, a measure of dissimilarity of an entity with respect to itself should be zero and this should be less than with respect to any other entity. These two properties correspond to (1.15) and (1.14) in the definition of a distance function. Finally, for any triple of entities x, y, and z, an essential property of dissimilarity seems to be that, if x is rather similar to y and y is rather similar to z, then x and z cannot be very dissimilar. A simple way to ensure that a measure of dissimilarity has this property is to demand that it also satisfies the triangle inequality (1.16) and then it would be a distance function. If a measure of dissimilarity fails to satisfy the triangle inequality, then it must be shown in some other way that it has this property.

In the object-predicate table the informational distance introduced in (1.19) makes the set of predicates a metric space (if predicates effecting the same partition are considered identical). Rogers and Tanimoto [49] use the truth-functional interpretation of an object-predicate table in which there are only two different sorts of marks; s_{ij} is the number of objects that have predicate p_i and predicate p_j divided by the number of objects that have predicate p_i or predicate p_j. They use $-\log_2(s_{ij})$ as coefficient of dissimilarity between p_i and p_j. They are aware that the triangle inequality may fail and they make a virtue out of necessity by stating that a coefficient of dissimilarity should *not* have the intuitively formulated property that leads us to accept the triangle inequality (1.16). Many other coefficients are used [34, 57] that are not metric. In the truth-functional interpretation of the object-predicate table, the metric of Restle [48] is applicable.

Let $S_1,\ldots,S_k$ be the classes of a partition in a given subset S of a metric space where the distance function is called d. A set elements $(s_1,\ldots,s_k)$ is a *skeleton* of the partition if $s_i \epsilon S_i$ for $i = 1,\ldots,k$. Given such a skeleton, we define another partition of k classes by assigning each s_i of the skeleton to a different class and by assigning an arbitrary s, not already assigned, to a class that contains an s_i closest to it. In case there is more than one such s_i, decide by choosing, say, the s with smallest i. Thus, each skeleton of a partition defines a function mapping this partition on a, generally different, partition. If this partition is not different, we say that the skeleton is a *clustering skeleton.*

One can say that the existence of a clustering skeleton is the analogon in metric space of the condition of linear separability in inner-product space. Suppose that S is a subset of inner-product space, then S_i and S_j are linearly separable if there exists a linear form L(x) such that $L(x) \leq 0$ for $x \epsilon S_i$ and $L(x) \geq 0$ for $x \epsilon S_j$. The function $d(x,y) = (x-y,x-y)^{\frac{1}{2}}$ is a metric in inner-product space. $L(x) = d^2(x,s_i) - d^2(x,s_j)$ is a linear form in x and is non-negative for $x \epsilon S_j$ and non-positive for $x \epsilon S_i$. Therefore, S_i and S_j are linearly separable if they have a clustering skeleton. On the other hand, if S_i and S_j are linearly separable, they need not have a clustering skeleton.

A partition is said to be a *clustering* if it has at least one clustering skeleton. Therefore, if $(S_1,\ldots,S_k)$ is a clustering and if $(s_1,\ldots,s_k)$ is a clustering skeleton

$$d(s_j,z) = \min_{i=1,\ldots,k} d(s_i,z) \text{ if } z\epsilon S_j \text{ and}$$

$$z\epsilon S_j \text{ if } d(s_j,z) = \min_{i=1,\ldots,k} d(s_i,z)$$

and if j is the smallest for which this holds. Let the *diameter* D of a finite set be defined as the greatest distance between two of its elements:

$$D(S_j) = \max_{x\epsilon S_j} \max_{y\epsilon S_j} d(x,y), \quad j = 1,\ldots,k.$$

Let the *radius* R of a finite set of points be defined as:

$$R(S_j) = \min_{x\epsilon S_j} \max_{y\epsilon S_j} d(x,y), \quad j = 1,\ldots,k.$$

An x for which the minimum occurs is called a *centre* of the set of points S_j. The terms "diameter", "radius", and "centre" are justified by the following theorem.

THEOREM 2.2

For any finite set of points in metric space we have

$$D \leq 2R.$$

PROOF. Let u and v be points such that D = d(u,v). Let x be a centre of the set, then

$$R \geq \max(d(x,u),d(x,v)) \geq \tfrac{1}{2}d(x,u) + \tfrac{1}{2}d(x,v) \geq \tfrac{1}{2}d(u,v) = \tfrac{1}{2}D.$$

The inequality is sharp in the sense that there is a metric space and a set in it for which equality holds. For instance, take as metric space the real numbers with the distance function d(x,y) = |x-y|. Take as set of points (0,1,2).

A partition is said to be a *stable clustering* if every skeleton is a clustering skeleton. Examples of stable clusterings are the partition consisting of only one non-empty class and the partition where no class has more than one element. The *separation* T between two sets is defined as:

$$T(S_i,S_j) = \min_{x\epsilon S_i} \min_{y\epsilon S_j} d(x,y).$$

The concepts introduced up till now serve to characterize stable clustering.

THEOREM 2.3

A sufficient condition for a clustering to be stable is

$$T(S_i,S_j) > \max(D(S_i),D(S_j)), \quad \text{for all } i \text{ and } j. \tag{2.9}$$

A necessary condition for a clustering to be stable is

$$T(S_i,S_j) \geq \max(R(S_i),R(S_j)), \quad \text{for all } i \text{ and } j. \tag{2.10}$$

PROOF. To establish the sufficiency of (2.9), suppose that it holds. For any $y\epsilon S_j$, let $x\epsilon S_i$ be such that it minimizes $d(x,y)$. Then $d(s_i,y) \geq d(x,y) \geq T(S_i,S_j) > D(S_j) \geq d(y,s_j)$. This means that y is correctly classified whatever s_i and s_j we choose. The partition $(S_1,\ldots,S_k)$ must therefore be a stable clustering.

To establish the necessity of (2.10), suppose that it does not hold. Then for some i and j and for some $x\epsilon S_i$ and for some $y\epsilon S_j$ we have: $d(x,y) < R(S_j)$. Let $z\epsilon S_j$ maximize $d(y,z)$; then $d(y,z) \geq R(S_j) > d(x,y)$. If we choose $x = s_i$ and $z = s_j$, y is misclassified. There exists at least one skeleton which is not a clustering skeleton: the partition $(S_1,\ldots,S_k)$ is not stable.

3. ANALYSIS OF QUANTITATIVE DATA

3.1. A "STRUCTURE" IN INNER-PRODUCT SPACE

In this chapter we shall study the case where each object may be represented by a point in an n-dimensional linear vector space I with an inner product ("inner-product space"). We first have to propose how a "structure" in such a space is to be defined, and, preferably, such a definition should be closely analogous to the one for a structure in a set of objects as defined by an object-predicate table (section 2.1).

We have associated a predicate with a test to be performed on an object; the case where an object is represented by a point in vector space is reminiscent of the theory of observations in quantum physics. The following outline of it is due to Weyl [65]. By a vector s in I, quantum physics represents the *wave state* of the physical system under investigation. We suppose this vector to be normalized such that it has unit length. A *grating* $G = (I_1,\ldots,I_r)$ is a splitting of the total vector space into mutually orthogonal subspaces $I_1,\ldots,I_r$. The index j is called the *character* of I_j. If the system is in the wave state s, then its probability of having character j equals

$$w_j = ||I_j s||^2,$$

where $I_j s$ is the orthogonal projection of s on I_j and $||I_j s||$ its Euclidean norm. Pythagoras' theorem, and the fact that $I = I_1 + \ldots + I_r$, ensure that these probabilities add up to 1.

We can now make an obvious translation of the quantum-physical situation into ours. The state of the physical system corresponds to our object, a grating to a partition, a character to the value of a predicate, and the set of probabilities $(w_1,\ldots,w_r)$ to the set of weights of the partition.

However, we are not concerned with a single object s, but with a set of objects, or the set of vectors representing them. It will be easiest to interpret this set as a special case of a random vector x. The connection will be explained below, after we have first introduced some notation for random vectors.

Let the mean of a random vector x exist and let it be denoted by E(x) and suppose the origin is chosen such that E(x) = 0, the null vector. The covariance matrix of x is defined to be E(xx') = V(x) (which we suppose to exist and to be non-singular; the prime ' is used to denote transposition for vectors as well as for matrices; a vector without a prime we suppose to be a column vector). The eigenvalues of V (which are real and positive) are denoted by $\lambda_1(V) \geq \lambda_2(V) \geq \ldots \geq \lambda_n(V)$. A choice of corresponding normalized eigenvectors (which are orthogonal in the case of distinct eigenvalues and which are so chosen for a multiple eigenvalue) is denoted by $p_1(V), p_2(V), \ldots, p_n(V)$. Note that eigenvectors and eigenvalues are regarded as functions of the corresponding matrix. Sometimes, the argument will be omitted; in that case it is V. We suppose x has been multiplied by the scalar such that tr(V), the trace of V (which is defined as the sum of the eigenvalues), equals 1.

We may think of x as being associated with an n-dimensional probability distribution function, and in particular (following Okamoto [41]) with a set of $N \geq n$ vectors $s_1, \ldots, s_N$, each of which has a positive weight f_i, $f_1 + \ldots + f_N = 1$. Such a weight may be taken to be proportional to the number of times the corresponding object has been observed. If we define a random vector x by $Pr(x=s_i) = f_i$, $i = 1, \ldots, N$, then we have

$$V(x) = E(xx') = f_1 s_1 s_1' + \ldots + f_N s_N s_N' = SFS',$$

where s_i is a column of the n×N matrix S and f_i a diagonal element of the N×N diagonal matrix F.

We shall show how to assign a set of weights to every grating in I if a covariance matrix V of a random vector x is given. Let us choose a set of orthonormal coordinate vectors $(e_1, \ldots, e_n)$ in such a way that each I_j is spanned by a sequence of successive coordinate vectors $e_{j_1}, \ldots, e_{j_2}$.

Let us now suppose that V is the representation with respect to this coordinate system. The diagonal element v_{ii} of V is the variance of the orthogonal projection of x on e_i, $E(||I_j x||^2) = v_{j_1 j_1} + \ldots + v_{j_2 j_2} = w_j$, and, because $tr(V) = v_{11} + \ldots + v_{nn} = 1$, $w_1 + \ldots + w_r = 1$. In this way, a set of objects defines a set of weights for every grating $(I_1, \ldots, I_r)$ in I.

When considering qualitative data, we had a set of objects and, for
very predicate, a partition in it. A partition in T, which is a set of
utually disjoint subsets whose union is T, is analogous to a grating in I,
hich is a set of mutually orthogonal subspaces whose linear sum is I. For
ny two partitions, the joint partition (see 1.4) is defined and it is a
artition again. The analogous definition of a joint grating of two grat-
ngs $(I_1,\ldots,I_r)$ and $(J_1,\ldots,J_s)$ would be a set consisting of the operators
$_kJ_m$, k = 1,...,r and m = 1,...,s. However, only in a special case these
perators are orthogonal projections, even though I_k and J_m are. Only then
he joint grating is defined and we have $I_kJ_m = J_mI_k$ (see, for instance,
23]). In quantum physics a measurement on a system in wave state s is re-
resented by an orthogonal projection of the vector corresponding to s on
he subspace corresponding to the measurement. The special case where two
easurements are said to be compatible corresponds to cummutativity of the
orresponding projection operators. In that case we may define the joint
rating of any number of compatible observations, which we regard as a
tructure in inner-product space; this is analogous to our notion of a
tructure in a set of objects.

.2. OPTIMAL DATA COMPRESSION.

.2.1. DATA COMPRESSION AND PATTERN CLASSIFICATION

One of the possible approaches to pattern classification proceeds in
ıree principal steps. Let a "retina" denote an array of n sensitive ele-
ents. The retina is exposed to a pattern and the resulting (real-valued)
easurements constitute a point in n-dimensional vector space. Subsequent-
r, the "sensory cortex" transforms this into a point in k-dimensional vec-
or space (k < n) in such a way that enough information relevant to the
ext step is retained. Data compression is regarded as the activity of the
ensory cortex. Finally, in the "motor cortex" a decision mechanism assigns
ıe k-dimensional vector to one of the classes. This set-up is reminiscent
Rosenblatt's [50] "three-layer, series-coupled perceptron".

3.2.2. OPTIMAL APPROXIMATION TO A RANDOM VECTOR

From now on we need to discuss only a random vector x, which is regarded as the output of the retina. The sensory cortex transforms it to a k-dimensional random vector in such a way that information relevant to classification is preserved as much as possible. Two restrictions are imposed: the transformation is to be a perpendicular projection on a k-dimensional subspace and information relevant to classification is to be extracted only from the covariance matrix V(x). We interpret the problem of optimal data compression as the problem of optimal approximation to a random vector by one of given, lower dimension. We suppose x to be centered and normed such that $E(x) = 0$ and $tr(V(x)) = 1$.

In statistics an equivalent problem has been studied by Pearson [43]. Since Hotelling's work [25] on it, the method of approximating a random vector by its perpendicular projection onto a subspace spanned by a set of k first eigenvectors of the covariance matrix has become widely known as the "method of principal components". The optimality criteria used by Pearson and Hotelling are different, and Rao [46] has introduced yet another one; all three lead to the same approximation. Okamoto and Kanazawa [41,42] investigated the relation between these criteria. In the latter paper a theorem is presented that indicates a whole class of criteria that lead to the same approximation and of which the earlier are special cases.

In pattern recognition, the same problem of approximation to a random vector has been encountered, but different names were used: "feature selection" or "data compression". Possibly as a result of this, the problem was solved anew (Watanabe [61,64], Tou and Heydorn [58]). One of the results of Tou and Heydorn is a direct consequence of the properties of the principal components approximation. Watanabe uses a criterion that leads to the same approximation but is more powerful in the sense that it simultaneously characterizes all solutions for $k = 1,\ldots,n$.

3.2.3. WATANABE'S CRITERION

Let U be a square matrix whose columns are an orthonormal set $u_1,\ldots,u_n$, that is, $U'U = I$, the identity matrix. Then we have:

$$x = Ix = U'Ux = UU'x = u_1u_1'x + \ldots + u_nu_n'x.$$

Here, the scalar random variables $u_1'x,\ldots,u_n'x$ are the *components* of x with respect to the basis $u_1,\ldots,u_n$. Because of the invariance of the trace under a similarity transformation, we have:

$$tr(V(U'x)) = tr(U'V(x)U) = tr(V(x)) = 1.$$

This implies that, whatever orthonormal base we choose, the variances of the components add up to one.

Watanabe [61] chose as approximation to x its perpendicular projection onto a subspace spanned by k vectors of the basis. He selected the most "significant" k vectors, where significance of a basis vector was interpreted to be the variance of the corresponding component (a component with small variance gives little information about the difference between various occurrences of x, which is what we are interested in). Therefore, the subspace to be chosen is spanned by the basis vectors corresponding to the components that have the largest variances. The total amount of variance collected in this way is greater the more unequal the sum of variances is partitioned over the variances.

We must define precisely the conditions under which a set of nonnegative numbers $\rho_1,\ldots,\rho_n$ (which constitute the vector ρ) subdivides its sum more equally than does a set of nonnegative numbers $\lambda_1,\ldots,\lambda_n$ (which constitute the vector λ). The condition is

$$\rho_1 + \ldots + \rho_j \leq \lambda_1 + \ldots + \lambda_j \qquad \text{for } j = 1,\ldots,n-1$$

$$\rho_1 + \ldots + \rho_n = \lambda_1 + \ldots + \lambda_n,$$

where the indices are such that $\rho_1 \geq \rho_2 \geq \ldots \geq \rho_n$ and $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$. Hardy, Littlewood, and Pólya [24] said that under this condition λ *majorizes* ρ. They proved (as theorem 108) the following theorem.

THEOREM 3.1

Each of the conditions

1) λ majorizes ρ

2) there is an n×n matrix R with nonnegative elements whose row and column sums equal 1 such that $\rho = R\lambda$

is necessary and sufficient for

$$(3.1) \qquad \phi(\rho_1) + \ldots + \phi(\rho_n) \leq \phi(\lambda_1) + \ldots + \phi(\lambda_n)$$

to hold for all real convex functions ϕ. For the purposes of this tract it is adequate to define a function convex in a certain range if its second derivative exists and is positive. If this range contains the interval $[\lambda_n,\lambda_1]$, equality in (3.1) for some ϕ implies $\rho_i = \lambda_i$ for $i = 1,\ldots,n$.

Watanabe's result may be summarized as follows. A random vector is approximated by its perpendicular projection onto the subspace spanned by those basis vectors $u_1,\ldots,u_k$ for which the corresponding components have largest variances $\rho_1,\ldots,\rho_k$. The approximation is considered optimal for $k = 1,\ldots,n$ if the basis is chosen such that the entropy $H(\rho) = -\rho_1 \ln(\rho_1) - \ldots - \rho_n \ln(\rho_n)$ is minimal.

THEOREM 3.2 (Watanabe [61])

The minimum is attained if and only if $u_1 = p_1,\ldots,u_n = p_n$, where $p_1,\ldots,p_n$ are orthonormal eigenvectors of V(x).

PROOF. Let P be an orthonormal matrix of which the columns $p_1,\ldots,p_n$ are eigenvectors of V(x). Then V(x) = PDP', where D is the diagonal matrix with elements $\lambda_1,\ldots,\lambda_n$. Let U be an arbitrary orthonormal n×n matrix.

$$V(U'x) = E(U'xx'U) = U'V(x)U =$$

$$= U'PDP'U = QDQ',$$

where Q = U'P. If $\rho_1,\ldots,\rho_n$ are the diagonal elements of V(U'x) (and, hence, the variances of the components of U'x; without loss of generality we may suppose them to be ordered such that $\rho_1 \geq \rho_2 \geq \ldots \geq \rho_n$), then

$$\rho_i = \lambda_1 q_{i1}^2 + \ldots + \lambda_n q_{in}^2 \quad \text{for } i = 1,\ldots,n.$$

Let R be the n×n matrix whose (i,j)-th element equals q_{ij}^2. Then the row sums and the column sums of R equal 1. Hence, by Theorem 3.1, λ major-

izes ρ, and $H(\rho)$ is minimum if and (because $(d^2/dx^2)\ x\ \ln(x) = 1/x$ for $x > 0$) only if $\rho_1 = \lambda_1,\ldots,\rho_n = \lambda_n$. This implies that $u_1 = p_1,\ldots,u_n = p_n$, which concludes the proof.

Watanabe's criterion simultaneously characterizes the solution of the approximation problem for $k = 1,\ldots,n$, but it is rather an indirect criterion. A more direct derivation is obtained as follows. A random vector x is approximated by its perpendicular projection on a subspace of dimension k. Its difference with the approximation is the error vector, which is its perpendicular projection on the orthogonal complement, which is of dimension n-k. If the error vector is minimal, in a suitable sense, the approximation is optimal, in the corresponding sense. We shall compare different error vectors by means of a real-valued function of their covariance matrices. In order that these be independent of the coordinate system, such functions may only depend on the eigenvalues of the covariance matrix. The following theorem explains why different functions of the covariance matrix of the error vector result in the same optimal approximation: it snows that all eigenvalues of the error covariance matrix are minimized for a certain choice of subspace. This means that all functions of the eigenvalues that are monotone in each argument are minimized for this choice. Okamoto and Kanazawa [41,42] have used this method to show optimality for a larger class of approximations: initially, they do not suppose the projection to be perpendicular; the optimum approximation turns out to be a perpendicular projection. Although the result embodied in the following theorem is less general than theirs, we think it worthwhile to give a proof which shows it to be a simple consequence of the well-known Courant-Fischer max-min theorem.

THEOREM 3.3

Let U be an n×k matrix whose columns are orthonormal. In order to maximize each of the eigenvalues of U'VU, U must be chosen such that its columns are a basis of the subspace spanned by a set of k first eigenvectors of V. The maximum values are $\lambda_1(V),\ldots,\lambda_k(V)$.

In order to minimize each of the eigenvalues of U'VU, U must be chosen such that its columns are a basis of the subspace spanned by a set

of k last eigenvectors of V. The minimum values are $\lambda_{n-k+1}(V),\ldots,\lambda_n(V)$.

PROOF. We shall only prove the first part because the proof of the second part is completely analogous. Let R be a k×k orthonormal matrix such that R'U'VUR is diagonal with diagonal elements in non-increasing order of magnitude. Then $\lambda_j(U'VU)$ is the j-th diagonal element (j = 1,...,k). Let W be the k×j matrix consisting of the first j columns of UR. Then $\lambda_j(U'VU)$ is the smallest eigenvalue of W'VW, which is the minimum value of x'W'VWx = = (Wx)'V(Wx), where x varies over the j-dimensional vectors of length 1. According to the "max-min principle" [23], $\lambda_j(U'VU) \leq \lambda_j(V)$. Therefore, each of the eigenvalues is maximized by choosing U such that its columns are a set of k first eigenvectors of V. This completes the proof.

3.2.4. SOME CRITERIA SATISFIED BY THE PRINCIPAL COMPONENTS APPROXIMATION

Pearson [43] considered a set of $N \geq n$ points in n-space and sought a k-dimensional subspace that gives closest fit to this set, that is, a k-dimensional subspace such that the sum of squares of each of the perpendicularly-projecting lines from each of the points onto this subspace is a minimum. He concluded that the subspace sought is the one spanned by a set of first k eigenvectors of the covariance matrix of the set of points. Again, Tou and Heydorn [58] derived this result in the one of their approaches to feature selection that they called "estimation optimality".

This result is a consequence of theorem 3.3 if V is taken to be the ovariance matrix of the set of points. Then the sum of the squares of the erpendicularly-projecting lines from each of the points onto a k-dimensional subspace is tr(U'VU), where the columns of U span the orthogonal complement of this subspace. Each of the eigenvalues is minimal if the columns U are a basis for the subspace spanned by a set of n-k last eigenvalues and, therefore, also their sum.

Hotelling [25] considered the problem of approximating a random vector with covariance matrix V by its perpendicular projection onto the subspace spanned by an orthonormal set $u_1,\ldots,u_k$, which are the columns of an n×k matrix U. Then the perpendicular projection is U'x and the problem was to choose U in such a way that the determinant of its covariance matrix U'VU

is maximal. If each of the eigenvalues of U'VU is maximal, so also is its determinant, which is their product. Therefore, the solution is given by theorem 3.3.

This is closely related to a result about the entropy of a normal distribution derived by Tou and Heydorn [58]. Let y be a normally-distributed k-dimensional random vector. Its density is given by:

$$g(y) = |V(y)|^{-\frac{1}{2}} (2\pi)^{-\frac{1}{2}k} \exp(-\tfrac{1}{2}\mathrm{tr}((V(y))^{-1}yy')).$$

It may be verified that the entropy

$$H(y) = \int_{y_1=-\infty}^{\infty} \cdots \int_{y_k=-\infty}^{\infty} -g(y_1,\ldots,y_k) \ln(g(y_1,\ldots,y_k))dy_1 \cdots dy_k \tag{3.2}$$

$$= \tfrac{1}{2}k \ln(2\pi) + \tfrac{1}{2} \ln|V(y)| + \tfrac{1}{2}k.$$

If x is normally distributed, any perpendicular projection y = U'x is also normally distributed. The problem of choosing U such that H(U'x) is maximal reduces to maximizing $|V(U'x)| = |U'V(x)U|$ and, hence, to maximizing each of the eigenvalues of U'V(x)U.

Suppose that $U_{n-k} = (u_1,\ldots,u_{n-k})$ and we shall consider the perpendicular projection of x onto the subspace spanned by the columns of U_{n-k} as the error vector. Its covariance matrix is $V(U'_{n-k}x) = U'_{n-k} V(x) U_{n-k}$. The criterion which we consider now is the entropy H of the error vector, where $H(f) = \int -f(z) \ln(f(z)) dz$, where f is the probability density of $U'_{n-k}x$ and integration is over the subspace spanned by the columns of U_{n-k}. The problem is to choose U_{n-k} such that the entropy of the error vector $U'_{n-k}x$ is minimal. However, the entropy depends on the probability density function f. Good [21,22] argued that in many situations it makes sense to estimate probabilities in such a way that entropy is maximized under known constraints. He advocated the principle of "minimaxing entropy": maximize entropy to find a probability distribution and, when planning an experiment, which is analogous to our choice of U_{n-k}, minimize the maximum entropy. The minimax characterization of principal components to be given below is reminiscent of this. In our case, the constraint is that the distribution must have the same covariance matrix as the given error vector.

The maximization problem for the entropy was solved by Shannon, who stated [53] the following result.

THEOREM 3.4

Of all density functions having a given covariance matrix, the normal density with that covariance matrix has maximal entropy.

Using (3.2), we arrive at the following characterization of the principal components solution:

$$\min \max H(f(U'_{n-k}x)) =$$

$$= \tfrac{1}{2}(n-k)\ln(2\pi) + \tfrac{1}{2}\ln(\lambda_{k+1}\cdots\lambda_n) + \tfrac{1}{2}(n-k),$$

where maximization is over all distributions having the given covariance matrix and minimization is over $n\times(n-k)$ matrices U_{n-k}. The maximum occurs for the normal distribution and the minimum occurs for $U_{n-k} = (p_{k+1},\ldots,p_n)$, a set of n-k last eigenvectors of V.

3.2.5. A MAXIMUM-ENTROPY CHARACTERIZATION OF THE NORMAL DISTRIBUTION

Theorem 3.4 is not quite satisfactory because the entropy of an n-dimensional normal distribution with covariance matrix V turns out to be:

$$H = \tfrac{1}{2}n\ln(2\pi) + \tfrac{1}{2}\ln(|V|) + \tfrac{1}{2}n.$$

Apparently, not all covariances v_{ij} are necessary to specify the entropy, because this is already done by $|V|$ and Shannon's condition can be relaxed to stating this determinant. But then there is no unique distribution for which the maximum of entropy is achieved. In this section we are concerned with a less stringent constraint that leads to a uniquely determined maximizing distribution.

THEOREM 3.5

Let W be a positive definite real symmetric matrix of order n. Of all density functions f with zero average, of which the covariance matrix V

satisfies

$$\text{(3.3)} \qquad \operatorname{tr}(VW) \leq n,$$

$$\text{(3.4)} \qquad f(x) = |W|^{\frac{1}{2}} (2\pi)^{-\frac{1}{2}n} \exp(-\tfrac{1}{2}x'Wx)$$

has maximal entropy, which is

$$\text{(3.5)} \qquad H(f) = \tfrac{1}{2}n \ln(2\pi) - \tfrac{1}{2} \ln(|W|) + \tfrac{1}{2}n.$$

Furthermore, any distribution satisfying (3.3) and not identical to (3.4) has an entropy less than (3.5).

PROOF. Besides the *entropy* $H(f) = \int -f(x) \ln(f(x))\, dx$ of the density function f, we shall also consider its *energy* $U(f) = \int \tfrac{1}{2}f(x)\, x'Wx\, dx$. We first determine the density f that maximizes H under the constraints $\int f(x)dx = 1$ and $U(f) = \tfrac{1}{2}n$. This is equivalent to the maximization without constraints of:

$$H + \lambda(\tfrac{1}{2}n - U) + \mu\left(\int f(x)\, dx - 1\right),$$

where λ and μ are Lagrange multipliers.

$$H + \lambda(\tfrac{1}{2}n - U) + \mu\left(\int f(x)\, dx - 1\right) =$$

$$= \int f(x) \ln(1/f(x))\, dx - \int f(x)\, \tfrac{1}{2}\lambda\, x'Wx\, dx +$$

$$+ \int \mu f(x)\, dx + \tfrac{1}{2}n\lambda - \mu =$$

$$= \int f(x) \ln((\exp(\mu) \exp(-\tfrac{1}{2}\lambda\, x'Wx))/f(x))\, dx + \tfrac{1}{2}\lambda n - \mu \leq$$

$$\leq \int f(x)\, ((\exp(\mu) \exp(-\tfrac{1}{2}\lambda\, x'Wx))/f(x) - 1)\, dx + \tfrac{1}{2}\lambda n - \mu =$$

$$= \exp(\mu) \int \exp(-\tfrac{1}{2}\lambda\, x'Wx)\, dx - 1 + \tfrac{1}{2}\lambda n - \mu.$$

The maximum occurs if and only if in each point x we have

$$f(x) = \exp(\mu) \exp(-\tfrac{1}{2}\lambda\, x'Wx).$$

The choice $\lambda = 1$, $\mu = \tfrac{1}{2} \ln(|W|) - \tfrac{1}{2}n \ln(2\pi)$ satisfies the constraints. Then

we have

$$f(x) = |W|^{\frac{1}{2}} (2\pi)^{-\frac{1}{2}n} \exp(-\tfrac{1}{2}\, x'Wx) \quad \text{and} \tag{3.6}$$

$$H(f) = \tfrac{1}{2}\lambda n - \mu = \tfrac{1}{2}n + \tfrac{1}{2}n \ln(2\pi) - \tfrac{1}{2} \ln(|W|). \tag{3.7}$$

This derivation can be applied directly to the distribution of the velocity components of a molecule of an ideal gas to yield Maxwell's distribution. In that case W would be the identity matrix, but the full generality of W may well be useful to find the distribution in cases where the quadratic form for the energy is more complicated.

Note that $U(f) = \int f(x)\, \tfrac{1}{2}x'Wx\, dx = \tfrac{1}{2}tr(VW)$, so that we found a maximum for the entropy under the condition that $tr(VW) = n$. The same maximum would be found under the condition $tr(VW) \leq n$, for suppose for the moment that inequality holds. The inequality between the geometric and the arithmetic mean implies that

$$|VW|^{1/n} \leq tr(VW)/n \tag{3.8}$$

so, in that case, we would have $|V| < |W|^{-1}$. But the maximization could be carried out with $W_1 = V^{-1}$ and we would find an entropy

$$H = \tfrac{1}{2}n + \tfrac{1}{2}n \ln(2\pi) - \tfrac{1}{2} \ln(|W_1|).$$

Therefore, any distribution, whether normal or not, for which $|V| < |W^{-1}|$, has an entropy smaller than (3.7).

The maximizing distribution must therefore have $|V| \geq |W|^{-1}$. Inequality is impossible because of (3.8). The only remaining case we have to investigate is that of a distribution different from (3.6) but with $|V| = |W|^{-1}$ and also normal, achieves the same maximum entropy. Then we have:

$$1 = |V|\ |W| = |VW| = |VW|^{1/n} \leq tr(VW)/n = 1,$$

e the last equality is the constraint. We must therefore have equality 3.8), which implies $V = W^{-1}$. This proves that (3.6) uniquely maximizes

3.3. THE COMPLEXITY OF A COVARIANCE MATRIX

Let $x_1,\ldots,x_n$ be jointly distributed random variables. In 1.3.2. we introduced the interaction $R(x_1,\ldots,x_n) = H(x_1) + \ldots + H(x_n) - H(x_1,\ldots,x_n)$ between them and this is also the sum of all interactions in a system with $x_1,\ldots,x_n$ as components, which we defined to be the complexity of that system. Complexity in its proper sense is defined in terms of entropies associated with a random vector. However, a covariance matrix does not uniquely determine the random vector for which this matrix is the covariance matrix. In order to be able to define the complexity of a covariance matrix V, we shall consider the normally distributed vector of which V is the covariance matrix because it has maximum entropy (see theorems 3.4 and 3.5). Then we find for the total amount of interaction

$$R(x_1,\ldots,x_n) = H(x_1) + \ldots + H(x_n) - H(x_1,\ldots,x_n)$$

$$= \sum_{i=1}^{n} (\tfrac{1}{2}\ln(2\pi) + \tfrac{1}{2}\ln(v_{ii}) + \tfrac{1}{2}) +$$

$$- \tfrac{1}{2}n\ln(2\pi) - \tfrac{1}{2}\ln|V| - \tfrac{1}{2}n =$$

$$= \tfrac{1}{2}\sum_{i=1}^{n}\ln(v_{ii}) - \tfrac{1}{2}\sum_{i=1}^{n}\ln(\lambda_i), \qquad \text{by (3.5).}$$

This cannot be used as an amount of complexity in the matrix because it depends on the coordinates. However, the maximum of this over all coordinate systems is only dependent on V and it may reasonably be interpreted as the complexity of V. To find the maximum of $R(x_1,\ldots,x_n)$ we must find the orthogonal transformation of V that maximizes $\ln(v_{11})+\ldots+\ln(v_{nn})$. Orthogonal transformations leave $v_{11} + \ldots + v_{nn}$ invariant and, as we assume, equal to 1. Under this condition, the inequality between the geometric and the arithmetic mean of $v_{11},\ldots,v_{nn}$ implies that the maximum is attained for $v_{11} = \ldots = v_{nn} = 1/n$. If there is an orthogonal transformation of V such that all diagonal elements are equal, we would find for the complexity:

$$C_1 = \tfrac{1}{2} \sum_{i=1}^{n} \ln(1/n) - \tfrac{1}{2} \sum_{i=1}^{n} \ln(\lambda_i) =$$

$$= -\tfrac{1}{2} \sum_{i=1}^{n} \ln(n\lambda_i).$$

In 3.4.2. we shall show that such a transformation indeed exists.

Note that the complexity thus defined vanishes if $\lambda_1 = \ldots = \lambda_n = 1/n$ and is positive otherwise: it may be regarded as a measure of the unequality among eigenvalues. In a neighbourhood of the point $\lambda_1 = \ldots = \lambda_n = 1/n$ the following series expansion converges:

$$C_1 = -\tfrac{1}{2} \sum_{i=1}^{n} \ln(n\lambda_i) =$$

$$= -\tfrac{1}{2} \sum_{i=1}^{n} ((n\lambda_i-1) + (n\lambda_i-1)^2/2 + O(n\lambda_i-1)^3)$$

(3.9)

$$= -\tfrac{1}{4} \sum_{i=1}^{n} (2n\lambda_i - 2 + n^2\lambda_i^2 - 2n\lambda_i + 1 + O(n\lambda_i-1)^3)$$

$$= -\frac{n^2}{4} \sum_{i=1}^{n} (\lambda_i^2 - 1/n^2) + O(\sum_{i=1}^{n} (n\lambda_i-1)^3)$$

The property of $-\frac{1}{2} \sum_{i=1}^{n} \ln(n\lambda_i)$ to vanish for $\lambda_1 = \ldots = \lambda_n = 1/n$ and be positive otherwise is shared by all functions of the form $= \sum_{i=1}^{m} (\phi(\lambda_i) - \phi(1/n))$ where ϕ is, like $-\ln$, a convex function (see Theorem 3.1). For the complexity of a covariance matrix we choose a function of this form which is convenient to use in inner-product space:

$$C = (1/n) \sum_{i=1}^{n} (\lambda_i^2 - 1/n^2). \qquad (3.10)$$

This is defined for any covariance matrix, irrespective of whether (3.9) converges. Because $||V||^2 = \sum_{i=1}^{n} \sum_{j=1}^{n} v_{ij}^2 = \sum_{i=1}^{n} \lambda_i^2$, the square of the Euclidean norm of V, is invariant under orthogonal transformation, we have

$$C = (1/n)\ ||V||^2 - 1/n^2 =$$

$$= (1/n) \sum_{i=1}^{n} (v_{ii}^2 - 1/n^2) + (2/n) \sum_{i=1}^{n} \sum_{j=i+1}^{n} v_{ij}^2. \tag{3.11}$$

This formula for C is convenient because no transformation of V is needed to compute it. Moreover, (3.10) coincides, apart from the constant factor, with the first two terms of the series expansion (3.9) which is the information-theoretic complexity of the random vector "naturally" (by Shannon's theorem 3.4) related to V. The factor 1/n is included because it makes C into the usual expression for the variance of a discrete random variable assuming the values $\lambda_1,\ldots,\lambda_n$ each with probability 1/n. The first term in (3.11) is the variance of the variances v_{ii} of the components.

3.4. REPRESENTATIONS OF COMPLEXITY

3.4.1. CHANGE OF REPRESENTATION BY PLANE ROTATION

In section 3.2 we saw that the success of data compression depends on the "unequality" of the eigenvalues. Suppose one wants to approximate an n-dimensional random vector by its perpendicular projection on a k-dimensional subspace in such a way that the sum of the variances of the k components is at least $k\alpha$. It is a consequence of Watanabe's theorem 3.2 that this is only possible if $\lambda_1 + \ldots + \lambda_k \geq k\alpha$. For much data compression to be possible, $k\alpha$ must be close to unity for a small k; hence, it is necessary that the eigenvalues be very unequal. The complexity C is important because it indicates whether such is the case. The information-theoretic notion of redundancy of a set of variables implies that the whole set can be closely approximated by a small subset, and this is just what a high value of C implies.

The expression (3.11) shows that both unequality among variances and the existence of non-zero covariances contribute to C. When V is in diagonal form, the sum-of-covariances term of C vanishes and the variance-of-variances term is maximal. Therefore, in the diagonal form there is a maximum amount of unequality among diagonal elements if unequality is measured by the sum of squares. Watanabe (theorem 3.2) showed that this is

also the case if unequality is measured by $\sum_{i=1}^{n} -v_{ii} \ln(v_{ii})$, the entropy. In fact, his proof can be used to show that this holds for any measure of unequality of the form $\sum_{i=1}^{n} -\phi(v_{ii})$, where ϕ is convex.

This suggests that a change of coordinate axes in general transforms some variance of variances into covariances or vice versa, leaving the sum equal. That this mechanism may be traced quite precisely is shown as follows. The fact that C vanishes when all eigenvalues are equal and is positive otherwise is not the only reason for regarding it as a measure of unequality. Equality is relation between pairs of numbers and therefore we prefer to write C as the sum of measures of unequality between pairs of eigenvalues:

$$C = (1/n) \sum_{i=1}^{n} (\lambda_i^2 - 1/n^2) =$$

$$= (1/n^2) \sum_{i=1}^{n} \sum_{j=i+1}^{n} (\lambda_i - \lambda_j)^2$$

and, similarly, (3.11) may be written as

$$(3.12) \qquad C = (1/n^2) \sum_{i=1}^{n} \sum_{j=i+1}^{n} ((v_{ii} - v_{jj})^2 + 2nv_{ij}^2).$$

Here, we see that any pair (x_i, x_j) of components can give two contributions to the complexity in V: a contribution due to the unequality between their variances and one due to a non-zero covariance between them. The complexity of V can, apparently, be split up into contributions due to pairs of components.

The relation between the two contributions can be described more precisely by showing that, for any pair of components x_i and x_j, a new set of coordinate axes may be found with respect to which only x_i and x_j are changed; change in their variances being compensated by a change in the covariance between them. The new set of coordinate axes is the same as the old except for the i-th and the j-th, which are rotated in the same plane through an angle ϕ. The resulting covariance matrix is $U = P'VP$ where P is the n×n matrix equal to the identity matrix except for the elements

$$(3.13) \qquad \begin{aligned} p_{ii} &= p_{jj} = \cos(\phi), \\ p_{ji} &= -p_{ij} = \sin(\phi). \end{aligned}$$

Rotations of this kind are called plane rotations and they are the elementary steps by which Jacobi's method for finding the eigenvalues of a symmetric real matrix proceeds. Theorem 3.6 may be obtained from the relevant formulas in [67], or directly as follows.

THEOREM 3.6

For $i = 1,\ldots,n$ and $j = 1,\ldots,n$ we have

$$v_{ii}^2 + 2v_{ij}^2 + v_{jj}^2 = u_{ii}^2 + 2u_{ij}^2 + u_{jj}^2.$$

PROOF. Let $W = VP$; then $U = P'W$. The columns of W are equal to those of V except for the i-th and j-th columns. The rows of W have the same Euclidean norm as those of V because P is orthogonal. Therefore:

$$w_{ki}^2 + w_{kj}^2 = v_{ki}^2 + v_{kj}^2 \quad \text{for } k = 1,\ldots,n. \tag{3.14}$$

Similarly, we find:

$$u_{ik}^2 + u_{jk}^2 = w_{ik}^2 + w_{jk}^2 \quad \text{for } k = 1,\ldots,n. \tag{3.15}$$

From (3.14):

$$w_{ii}^2 + w_{ij}^2 = v_{ii}^2 + v_{ij}^2$$
$$w_{ji}^2 + w_{jj}^2 = v_{ji}^2 + v_{jj}^2$$
$$\text{---------------------------------} +$$
$$w_{ii}^2 + w_{ij}^2 + w_{ji}^2 + w_{jj}^2 = v_{ii}^2 + v_{ij}^2 + v_{ji}^2 + v_{jj}^2.$$

Similarly, from (3.15):

$$w_{ii}^2 + w_{ji}^2 + w_{ij}^2 + w_{jj}^2 = u_{ii}^2 + u_{ij}^2 + u_{ji}^2 + u_{jj}^2.$$

Using the fact that V and U are symmetric, we obtain

$$v_{ii}^2 + 2v_{ij}^2 + v_{jj}^2 = u_{ii}^2 + 2u_{ij}^2 + u_{jj}^2, \tag{3.16}$$

which completes the proof.

Note that $v_{kk} = u_{kk}$ for $k \neq i$, and $k \neq j$, and $\sum_{k=1}^{n} v_{kk} = \sum_{k=1}^{n} u_{kk} = 1$. This implies that $u_{ii} + u_{jj} = v_{ii} + v_{jj}$, and suppose it is equal to 2μ. From (3.16):

$$v_{ii}^2 + v_{jj}^2 - 2\mu^2 + 2v_{ij}^2 = u_{ii}^2 + u_{jj}^2 - 2\mu^2 + 2u_{ij}^2$$

(3.17)

$$\tfrac{1}{2}(v_{ii}-v_{jj})^2 + 2v_{ij}^2 = \tfrac{1}{2}(u_{ii}-u_{jj})^2 + 2u_{ij}^2$$

On each side, the first term is a contribution to the variance-of-variances term in (3.12) and the second term is a contribution to the square-of-covariances term. The expression (3.17) shows that under a plane rotation the sum of the contributions is invariant.

3.4.2. A VARIATIONALLY EQUILIBRATED FORM OF A COVARIANCE MATRIX

We saw that the complexity of a covariance matrix has two additive components: one arising from the unequality of diagonal elements and the other from the sum of the squares of the covariances. In the diagonal form of a matrix all complexity is represented as unequality. In fact, one well-known method for diagonalizing a matrix (known as Jacobi's; see, for instance, [67]) uses successive plane rotations chosen such that $u_{ij} = 0$ in (3.16).

In 3.3, where we justified C as a measure of the complexity in V, we supposed the existence of a matrix orthogonally equivalent to V where all complexity is present in the form of covariances and, therefore, all diagonal elements are equal; because these are the variances of the components of a random vector that has V as its covariance matrix, we shall call this a *variationally equilibrated* covariance matrix. We can say that such a form is a "most undiagonal" form of V.

THEOREM 3.7

A covariance matrix V is orthogonally equivalent to a variationally equilibrated matrix.

PROOF. Let

$$C_1(Q) = (1/n^2) \sum_{i=1}^{n} \sum_{j=i+1}^{n} (m_{ii}-m_{jj})^2$$

and

$$C_2(Q) = (1/n) \sum_{i=1}^{n} \sum_{j=i+1}^{n} 2m_{ij}^2,$$

where m_{ij}, $i = 1,\ldots,n$, $j = 1,\ldots,n$, are the elements of $M = Q'VQ$ for some orthonormal matrix Q. Then, according to (3.12), we have $C = C_1(Q) + C_2(Q)$. Apparently, $C_1(Q) \geq 0$; we shall show that zero is also the greatest lower bound. Consider the following algorithm:

Take some $d_k > 0$ and repeat as often as possible the following step: Find an i and a j such that $(m_{ii}-m_{jj})^2 > d_k$ and subject M to a plane rotation such that m_{ii} becomes equal to m_{jj}, which is always possible. Call the resulting matrix M again.

The algorithm consists of a finite number of such steps, because at each of these C_1 decreases at least by d_k/n^2. The finite product of the corresponding plane rotations is an orthonormal matrix which we call Q_k. After the execution of the algorithm, $C_1 \leq n(n-1)\, d_k/(2n^2)$. Because this holds for all $d_k > 0$, the greatest lower bound of C_1 is 0.

It now remains to be shown that there exists an orthonormal matrix Q such that $C_1(Q) = 0$. Take any sequence $d_1, d_2, \ldots$ of positive numbers that has 0 as limit. Execute the algorithm for each. Consider the element $p_{ij}^{(k)}$ of the i-th row and j-th column in the matrix

$$P^{(k)} = \prod_{i=1}^{k} Q_i, \quad k = 1,2,\ldots .$$

It may be verified that $p_{ij}^{(1)}, p_{ij}^{(2)}, \ldots$ is a Cauchy sequence and that its limit is the corresponding element of a matrix Q such that $C_1(Q) = 0$. This completes the proof.

3.4.3. A RECURSIVELY DOUBLY SYMMETRIC FORM OF A COVARIANCE MATRIX

Theorem 3.7 showed the existence of a variationally equilibrated form of a covariance matrix by describing a, generally infinite, sequence of matrices of which it is the limit. Here we shall be concerned with a

special form of the matrix, which may be described as "recursively doubly symmetric" and which is also variationally equilibrated if n, the order of the matrix, is a power of 2. The algorithm that computes this form uses only a finite number of steps. We do not know whether there is an algorithm that yields a variationally equilibrated form after a finite number of steps for arbitrary order of the matrix.

Let the covariance matrix be called V, with elements v_{ij}, i,j = $= 1,\ldots,n$. The set of elements $(v_{11},\ldots,v_{nn})$ is called the diagonal; the set of elements $(v_{1n},v_{2,n-1},\ldots,v_{n1})$ is called the counter diagonal. A matrix is said to be of doubly diagonal form if non-zero elements only occur in the diagonal or in the counter diagonal. A matrix is said to be doubly symmetric if it is symmetric with respect to the diagonal and also with respect to the counter diagonal.

LEMMA 3.1

A covariance matrix V is orthogonally equivalent to a doubly diagonal doubly symmetric matrix.

PROOF. Because any covariance matrix is orthogonally equivalent to a diagonal matrix, it suffices to show that a diagonal matrix is orthogonally equivalent to a doubly diagonal doubly symmetric matrix. It is easy to see that, if P is the matrix satisfying condition (3.13) and j = n + 1 - i and if V is doubly diagonal, U = P'VP is also doubly diagonal. A diagonal matrix is doubly diagonal, by definition. The angle ϕ which occurs in the conditions (3.13) may be chosen such that for the elements of U, $u_{ii} = u_{n+1-i,n+1-i}$, i = 1,...,n, which implies double symmetry in a doubly diagonal matrix.

Transposition of a matrix M with elements m_{ij}, i,j = 1,...,n, means the reflection of M with respect to the diagonal. The result is denoted by M'; we have $m_{ij} = m'_{ji}$. The fact that a matrix is symmetric may be expressed as M = M'. We shall call *counterposition* the reflection of M with respect to the counter diagonal. The result is denoted by M^*; we have $m_{ij} = m^*_{n+1-j,n+1-i}$. The fact that a matrix is doubly symmetric may be expressed as $M^* = M$, M = M'. As may be verified, the following lemma is implied by

these definitions.

LEMMA 3.2

$M^{*'} = M'^{*}$ and $(M_1M_2)^* = M_2^*M_1^*$ for arbitrary square matrices M, M_1, and M_2.

Suppose that the order n of V is even; let m = n/2. Let V_{11} be its leading principal submatrix of order m. We say that V is *recursively doubly symmetric* if it is either of order 1 or else if it is of even order, and doubly symmetric, and such that V_{11} is recursively doubly symmetric.

THEOREM 3.8

A covariance matrix V of order $n = 2^k$ (where k is a non-negative integer) is orthogonally equivalent to a recursively doubly symmetric matrix.

PROOF. By Lemma 3.1 we may, without loss of generality, suppose V to be doubly symmetric. Let us proceed by induction on k. For k = 0 the theorem is trivially true. Suppose it is true for k-1, then this implies that V_{11} is orthogonally equivalent to a recursively doubly symmetric matrix; specifically, suppose this matrix to be $Q_{11}'V_{11}Q_{11}$, where Q_{11} is orthonormal. Then

$$Q = \begin{pmatrix} Q_{11} & 0 \\ 0 & Q_{11}^{*'} \end{pmatrix}$$

is also orthonormal, as we verify by evaluating

$$Q'Q = \begin{pmatrix} Q_{11}'Q_{11} & 0 \\ 0 & Q_{11}^{*}Q_{11}^{*'} \end{pmatrix}.$$

Here, $Q_{11}^{*}Q_{11}^{*'} = (Q_{11}'Q_{11})^* = I$, the identity matrix of order n/2. Therefore, V is orthogonally equivalent to

$$Q'VQ = \begin{pmatrix} Q_{11}'V_{11}Q_{11} & Q_{11}'V_{12}Q_{11}^{*'} \\ Q_{11}^{*}V_{12}'Q_{11} & Q_{11}^{*}V_{11}^{*}Q_{11}^{*'} \end{pmatrix},$$

which is doubly symmetric because V is. By the induction hypothesis, $Q'_{11}V_{11}Q_{11}$ and $Q^*_{11}V^*_{11}Q^{*\prime}_{11} = (Q'_{11}V_{11}Q_{11})^*$ are recursively doubly symmetric; hence, by definition, V is. This concludes the proof.

If V is doubly symmetric, $\mathrm{tr}(V_{22}) = \mathrm{tr}(V^*_{11}) = \mathrm{tr}(V_{11}) = \frac{1}{2}\,\mathrm{tr}(V)$. Therefore, if V is recursively doubly symmetric, all its main diagonal elements are equal. Thus, in the special case where n is a power of 2, one of the variationally equilibrated forms of the matrix is a recursively doubly symmetric form. If the order of V is not a power of 2, it is easy to see that it is also orthogonally equivalent to a recursively doubly symmetric matrix (if we would choose m, the order of V_{11}, such that $2m + 1 = n$ in the case n is odd). This fact, which may be of interest in itself, does not concern us here, because in that case the doubly symmetric form is not necessarily variationally equilibrated.

3.5. COMPLEXITY AND CONDITION NUMBER

For a covariance matrix V the *condition number* $k = \lambda_1/\lambda_n$. In numerical computation this is an important quantity because it measures the relative precision with which V defines the solution of the vector equation $Vx = b$. If k is large (then V is called *ill-conditioned*), small relative errors in the elements of V can lead to large relative errors in the solution x. For a systematic treatment including general matrices the reader is referred to [27].

The condition number of V is related to its complexity: if the complexity is large, λ_1 must be large and λ_n must be small and, therefore, V is ill-conditioned. The converse is not true. Suppose, for instance, that $\lambda_1 = \ldots = \lambda_{n-1}$ and λ_n very close to zero. Then V is ill-conditioned, while the complexity is only slightly more than its minimum value. In this case, small relative errors in the elements of V may cause large changes in x, but these lie approximately in a one-dimensional subspace; the projection of the solution on the eigenspace of $\lambda_1,\ldots,\lambda_{n-1}$ is very precisely determined by V.

Thus we see that V may be ill-conditioned with respect to the entire space, but well-conditioned with respect to a suitably chosen (n-1)-dimensional subspace. This is not possible if complexity is close to 1.

Apparently, high complexity is a more serious condition than a large k by itself implies. This property may make complexity a useful concept in numerical computation.

We shall now address ourselves to the task of giving a more precise relation between complexity and condition number. To this end, we shall derive a lower bound for λ_1 and an upper bound for λ_n. This gives a lower bound for the condition number in terms of complexity. By the same method we shall find an upper bound for λ_1 and, for covariance matrices with very small complexity, a positive lower bound for λ_n. These provide an upper bound for the condition number in terms of complexity.

THEOREM 3.9

Let V be a covariance matrix of order n with $tr(V) = 1$ and complexity C. Let h be the integer such that $h \leq 1/(nC+1/n) < h + 1$, α_0 the larger root of $nC + 1/n = h\alpha^2 + (1-h\alpha)^2$, and β_1 the smaller root of $nC + 1/n = (1 - (n-1)\beta)^2 + (n-1)\beta^2$. Then we have for k, the condition number of V:

$$k \geq \alpha_0/\beta_1.$$

If $nC < 1/(n-1) - 1/n$, then $k \leq \alpha_1/\alpha_0$, where α_1 and α_0 are, respectively, the larger and smaller roots of $nC + 1/n = \alpha^2 + (1-\alpha)^2/(n-1)$.

PROOF. To find the lower bound for λ_1 we shall determine the maximum value of $\lambda_1^2 + \ldots + \lambda_n^2$. Suppose that

$$\lambda_1 = \alpha \geq 1/n. \tag{3.18}$$

We also have

$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n \geq 0 \tag{3.19}$$

and

$$\lambda_1 + \lambda_2 + \ldots + \lambda_n = 1. \tag{3.20}$$

Let m be the integer such that $m \leq 1/\alpha < m + 1$. The case $m = n$ is easy to dispose of: $\lambda_1^2 + \ldots + \lambda_n^2$ can only assume the value $1/n$. Therefore, suppose that $n > m$ and put $\rho_1 = \ldots = \rho_m = \alpha, \rho_{m+1} = 1 - m\alpha$, and, if

$n > m + 1$, $\rho_{m+2} = \ldots = \rho_n = 0$.

It may be verified that for all $\lambda_1,\ldots,\lambda_n$ that satisfy (3.18, 19 and 20) $\rho_1 + \ldots + \rho_j \geq \lambda_1 + \ldots + \lambda_j$ for $j = 1,\ldots,n$ with equality for $j = n$. According to theorem 3.1

$$\rho_1^2 + \ldots + \rho_n^2 = m\alpha^2 + (1-m\alpha)^2 \geq$$

$$\geq \lambda_1^2 + \ldots + \lambda_n^2 = ||V||^2 = nC + 1/n.$$

Apparently, for $1/(m+1) < \alpha \leq 1/m$, the maximum value of $||V||^2$ is $m\alpha^2 + (1-m\alpha)^2$. For these values of α, $||V||^2$ is monotone increasing; $||V||^2 = 1/(m+1)$ for $\alpha = 1/(m+1)$ and $||V||^2 = 1/m$ for $\alpha = 1/m$. Thus, if we consider a certain value of $||V||^2$ as given and if h is the integer such that $h \leq 1/||V||^2 < h + 1$, we find that $\lambda_1 \geq \alpha_0$, where α_0 is the larger root of

$$nC + 1/n = h\alpha^2 + (1-h\alpha)^2.$$

To find the upper bound for λ_1, we shall determine the minimum value of $\lambda_1^2 + \ldots + \lambda_n^2$ under conditions (3.18, 19, and 20). Suppose that $\rho_1 = \alpha$, $\rho_2 = \ldots = \rho_n = (1-\alpha)/(n-1)$ for $1/n \leq \alpha \leq 1$, and $n > 1$. It may be verified that for $\lambda_1,\ldots,\lambda_n$ that satisfy (3.18, 19 and 20) $\rho_1 + \ldots + \rho_j \leq$ $\leq \lambda_1 + \ldots + \lambda_j$ for $j = 1,\ldots,n$ and with equality for $j = n$. According to theorem 3.1

$$\rho_1^2 + \ldots + \rho_n^2 = \alpha^2 + (1-\alpha)^2/(n-1) \leq$$

$$\leq \lambda_1^2 + \ldots + \lambda_n^2 = ||V||^2 = nC + 1/n.$$

This implies that, for a given $||V||^2$, $\lambda_1 \leq \alpha_1$, where α_1 is the larger root of

$$nC + 1/n = \alpha^2 + (1-\alpha)^2/(n-1).$$

In a similar way we derive bounds for λ_n. Suppose that

(3.21) $\lambda_n = \beta \leq 1/n.$

We also have

(3.22) $$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n \geq 0$$

and

(3.23) $$\lambda_1 + \ldots + \lambda_n = 1.$$

Under these conditions we shall find the maximum value of $\lambda_1^2 + \ldots + \lambda_n^2$ to derive an upper bound for λ_n. Put $\rho_1 = 1 - (n-1)\beta$, $\rho_2 = \ldots = \rho_n = \beta$. For all $\lambda_1,\ldots,\lambda_n$ that satisfy (3.21, 22, and 23) $\rho_1 + \ldots + \rho_j \geq \lambda_1 + \ldots + \lambda_j$ for $j = 1,\ldots,n$ with equality for $j = n$. According to theorem 3.1

$$\rho_1^2 + \ldots + \rho_n^2 = (1 - (n-1)\beta)^2 + (n-1)\beta^2 \geq$$

$$\geq \lambda_1^2 + \ldots + \lambda_n^2 = ||V||^2 = nC + 1/n.$$

This implies that, for a given $||V||^2$, $\lambda_n \leq \beta_1$, where β_1 is the smaller root of

$$nC + 1/n = (1 - (n-1)\beta)^2 + (n-1)\beta^2.$$

For the condition number k we now have $k = \lambda_1/\lambda_n \geq \alpha_0/\beta_1$. It may be possible to derive a higher lower bound for k using only $||V||^2$ and the fact that $tr(V) = 1$. It is not possible to improve the bounds α_0 and β_1 because, for each, we have indicated a class of matrices, namely those that have $\rho_1,\ldots,\rho_n$ as eigenvalues, for which these are attained. These classes are, in general, not the same; this leaves room for improving $k \geq \alpha_0/\beta_1$.

To obtain the lower bound for λ_n we find the minimum value of $\lambda_1^2 + \ldots + \lambda_n^2$ under the conditions (3.21, 22, and 23). Suppose that $\rho_1 = \ldots = \rho_{n-1} = (1-\beta)/(n-1)$, $\rho_n = \beta$. For all $\lambda_1,\ldots,\lambda_n$ that satisfy (3.21, 22, and 23) $\rho_1 + \ldots + \rho_j \leq \lambda_1 + \ldots + \lambda_j$ for $j = 1,\ldots,n$ with equality for $j = n$. According to theorem 3.1

$$\rho_1^2 + \ldots + \rho_n^2 = (1-\beta)^2/(n-1) + \beta^2 \leq$$

$$\leq \lambda_1^2 + \ldots + \lambda_n^2 = ||V||^2 = nC + 1/n.$$

This implies that, for a given $||V||^2$, $\lambda_n \geq \beta_0$, where β_0 is the smaller root of

$$nC + 1/n = (1-\beta)^2/(n-1) + \beta^2.$$

Only for $||V||^2 < 1/(n-1)$ do we have that $\beta_0 > 0$; otherwise the trivial lower bound 0 is at least as good. Thus, we find for the condition number $k = \lambda_1/\lambda_n \leq \alpha_1/\beta_0$ if $||V||^2 < 1/(n-1)$.

3.6. INTERACTION AND COMPUTATIONAL COMPLEXITY

3.6.1. INTRODUCTION

As far as we are aware, the only subject where a concept named "complexity" is being studied quantitatively, is the theory of automata. Here, the complexity of a sequence of symbols is defined to be the time required by a (suitably restricted) universal automaton to recognize that sequence (see, for instance, the relevant chapter in [37]). This definition of complexity as the difficulty of a computation seems to have little to do with ours, which is a difference of entropies. However, Kolmogorov [30] has proposed a basis for information theory in which the entropy of a sequence is defined to be its computational difficulty. The present section provides another, more indirect, link between complexity in terms of entropy and complexity in terms of length of computation.

Suppose we have a system of which the components are equations and where a measure of interaction is defined between disjoint sets of equations. Then, if the system is partitioned into two subsets, the interaction between these is, according to our definition in chapter 1, a contribution to the complexity of the system. We shall show that, for a particular type of system of equations, the interaction is related to the time required to compute the solution of the system, if a certain method of solution is used.

The particular type of system of equations and the particular method of solving it will be described below. We shall first sketch the result. Let each equation be associated with an unknown. The method of solution consists of the iteration of the following step. First, the equations in

one subset are used to obtain values for the variables associated with it; in parallel, the same is done for the other subset. In general, the approximation obtained by concatenating the solutions for the subsystems does not satisfy the whole system and, in that case, the step is repeated. For certain types of system, the successive approximations obtained by this method are guaranteed to converge to the solution. The lower the "rate of convergence", the more steps have to be done to obtain a result of sufficient accuracy and the greater is the computational complexity. We shall obtain an inequality that provides a lower bound for the rate of convergence in terms of interaction. In other words, we shall show that the better classification exists among the equations, the faster convergence must be, if the partitioning into subsystems is made according to the classification.

Let the system of equations be the linear equations contained in

$$Ax = b \tag{3.24}$$

where A is a positive definite symmetric real matrix with elements a_{ij} and where b and x are the known and unknown vectors respectively. A, b, and x are of order n. Let us consider the two subspaces R_1 and R_2 spanned by the first l and the last m ($l \leq m$, $l + m = n$) axes of the coordinate system with respect to which the equation has the representation (3.24). The corresponding partition of (3.24) is

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

The interaction between the first l and the last m scalar equations then depends on the elements of A_{12} and A_{21}.

According to Jacobi's method for block iteration (see, for instance, [16]), an initial approximation $x^{(0)}$ to x is chosen and a sequence $x^{(1)}, x^{(2)}, \ldots$ of successive approximations is constructed by solving in parallel

$$\begin{aligned} A_{11}\, x_1^{(i)} &= b_1 - A_{12}\, x_2^{(i-1)} \\ A_{22}\, x_2^{(i)} &= b_2 - A_{21}\, x_1^{(i-1)} . \end{aligned} \tag{3.25}$$

If the iteration converges, it converges faster the smaller the absolute values of the elements of A_{12} and A_{21} are. If all elements in A_{12} are zero, the iteration has converged when the second cycle is completed. In the sequel we show how it is possible to borrow a definition of interaction from information theory and to find a relation between such an amount of interaction and the speed of convergence in Jacobi's method for block iteration.

Let A be the covariance matrix of a random vector z. Then, in case z is normally distributed, the interaction (see 1.13 and 3.5) between the sets of components $(z_1,\ldots,z_l)$ and $(z_{l+1},\ldots,z_n)$ is:

$$R = -(1/2)\ln(|A|/(|A_{11}|.|A_{22}|)). \tag{3.26}$$

A well-known theorem (Beckenbach and Bellman [8]) states that

$$|A|/(|A_{11}|.|A_{22}|) \leq 1.$$

Their proof may also be used to show that equality obtains only if all elements of $A_{12} = A_{21}'$ vanish. We shall use the interaction R as the definition for our intuitively introduced concept of interaction. This implies that the less interaction there is, the closer the left hand side approaches 1. It is the purpose of the next two sections to find a relation between the speed of convergence of Jacobi's iteration and the interaction as defined above.

3.6.2. INTERACTION AND THE PERFORMANCE OF JACOBI'S ITERATION ACCORDING TO THE USUAL DEFINITION

The equations (3.25), that may be solved in parallel, may also be written as a single vector equation:

$$x^{(i)} = D^{-1}b + (I-D^{-1}A)x^{(i-1)}, \quad i = 1,2,\ldots, \tag{3.27}$$

where $D = \begin{pmatrix} A_{11} & 0 \\ 0 & A_{22} \end{pmatrix}$ and I is the identity matrix of order n. The solution to (3.27) must satisfy:

$$x = D^{-1}b + (I-D^{-1}A)x.$$

Hence, we must have for the error vector $e^{(i)} = x^{(i)} - x$:

$$(3.28) \qquad e^{(i)} = (I-D^{-1}A)\, e^{(i-1)} = (I-D^{-1}A)^{i}\, e^{(0)}, \quad i = 1,2,\ldots .$$

Suppose we want to approximate the eigenvector corresponding to the largest (in absolute value) eigenvalue ρ_1 of $(I-D^{-1}A)$, then the iteration (3.28) would correspond to the "power method" for finding successive approximations $e^{(i)}$ to this eigenvector. If ρ_1 is of unit geometric multiplicity, we have

$$(3.29) \qquad \lim_{i\to\infty} ||e^{(i)}||/||e^{(i-1)}|| = |\rho_1|.$$

The iteration diverges for $|\rho_1| > 1$, which cannot be the case for A symmetric and positive definite (see [16,19]). The iteration (3.27) converges faster the smaller $|\rho_1|$ is. We shall now derive an upper bound for $|\rho_1|$ in terms of the interaction (3.26).

The eigenvalues of $(I-D^{-1}A)$ are, by definition, the values of ρ for which there is a non-zero vector y that satisfies

$$(I-D^{-1}A)y = \rho y, \quad \text{or} \quad D(1-\rho)y = Ay.$$

This equation has a solution for non-zero y only if:

$$(3.30) \qquad |A - (1-\rho)D| = \begin{vmatrix} \rho A_{11} & A_{12} \\ A_{21} & \rho A_{22} \end{vmatrix} = \rho^{-1} \begin{vmatrix} \rho^2 A_{11} & \rho A_{12} \\ A_{21} & \rho A_{22} \end{vmatrix} =$$

$$= \rho^{m-1} \begin{vmatrix} \rho^2 A_{11} & A_{12} \\ A_{21} & A_{22} \end{vmatrix} = 0.$$

Because we supposed that $1 \leq m$, we find that this equation has m-1 roots equal to zero and the remaining 21 roots occur in pairs of opposite sign: $\pm\rho_1, \pm\rho_2, \ldots, \pm\rho_1$, where $|\rho_1| \geq |\rho_2| \geq \ldots \geq |\rho_1|$. These last roots must be such that there is a non-zero $y = y_1 + y_2$ ($y_1 \epsilon R_1$, $y_2 \epsilon R_2$) which

satisfies:

$$\rho^2 A_{11} y_1 + A_{12} y_2 = 0$$

$$A_{21} y_1 + A_{22} y_2 = 0$$

and

$$\rho^2 A_{11} y_1 + A_{12} y_2 = 0$$

$$A_{12} A_{22}^{-1} A_{21} y_1 + A_{12} y_2 = 0$$

which implies

$$\rho^2 y_1 = A_{11}^{-1} A_{12} A_{22}^{-1} A_{21} y_1$$

and

$$|A_{11}^{-1} A_{12} A_{22}^{-1} A_{21}| = \prod_{j=1}^{l} \rho_j^2.$$

Hence we have

$$|I - A_{11}^{-1} A_{12} A_{22}^{-1} A_{21}| = \prod_{j=1}^{l} (1-\rho_j^2)$$

but also

$$|A|/(|A_{11}|.|A_{22}|) = \left| \begin{pmatrix} A_{11}^{-1} & 0 \\ 0 & A_{22}^{-1} \end{pmatrix} \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} \right| =$$

$$= |I - A_{11}^{-1} A_{12} A_{22}^{-1} A_{21}|.$$

This leads to the result (see [19]):

$$|A|/(|A_{11}|.|A_{22}|) = \prod_{j=1}^{l} (1-\rho_j^2).$$

By means of the inequality between the geometric and the arithmetic mean we find:

$$1 - (|A|/(|A_{11}|.|A_{22}|))^{1/l} = 1 - \prod_{j=1}^{l} (1-\rho_j^2)^{1/l} \geq$$

(3.31)

$$\geq 1 - \sum_{j=1}^{l} (1-\rho_j^2)/l =$$

$$= \sum_{j=1}^{l} \rho_j^2/l \geq \rho_1^2/l.$$

The left-hand side is related to the interaction (3.26) and it provides an upper bound for $|\rho_1|$, which determines the asymptotic speed of convergence for Jacobi's iteration. The smaller the amount of interaction, the closer $|A|/(|A_{11}|.|A_{22}|)$ to 1, the smaller the upper bound for $|\rho_1|$, and the faster the convergence is.

In statistics, ρ_1 is well-known as a measure of the relatedness between the two groups of random variables $(z_1,\ldots,z_l)$ and $(z_{l+1},\ldots,z_n)$. There, Hotelling [26] posed and solved the problem of finding a linear combination (constrained to unit variance) of each group in such a way that their ordinary correlation coefficient has maximum absolute value. It turns out that this correlation, called the first canonical correlation, equals ρ_1.

3.6.3. INTERACTION AND THE PERFORMANCE OF JACOBI'S ITERATION ACCORDING TO ANOTHER DEFINITION

The asymptotic speed of convergence $|\rho_1|$ (see (3.29)) is not a satisfactory measure for the performance of the iteration because the number of iterations after which the actual rate of convergence approaches the asymptotic rate depends on the initial error vector $e^{(0)}$ and on $|\rho_2|/|\rho_1|$. In this section we shall first give a measure for the efficacy that does not have these drawbacks by considering *the amount* ΔI *of information about the unknown solution yielded by a cycle of the iteration*. Finally, we shall establish a lower bound for ΔI in terms of R.

The performance index of an algorithm should not depend on the value of the initial error vector if it works for many different values. But the algorithm may well converge faster for one value than for another. In his pioneering work in cybernetics, Wiener [66] was faced with a similar

dilemma. His problem was to design a filter that would optimally separate a message from noise. The difficulty was that one filter would be better for one message and another filter for another message. The solution adopted by Wiener was to consider not this or that particular message, but a set of possible messages, each with the probability with which it would occur in the environment in which the filter would have to operate, in short, an *ensemble* of messages.

Similarly, we shall consider the performance of Jacobi's iteration with respect to an ensemble of initial error vectors, that is, we shall assume that there exists an n-dimensional probability density function f_0 for $e^{(0)}$; the iteration (3.28) defines the successive density functions f_i of $e^{(i)}$, $i = 1,2,\ldots$. The assumed existence of f_0 is not a serious loss of generality, because the resulting performance index turns out to be independent of f_0; the assumption merely allows us to express the performance of the algorithm as the gain in information per cycle of the iteration, which is the uncertainty in $e^{(i-1)}$ minus the uncertainty in $e^{(i)}$. Intuitively, it is clear that the more a density function is concentrated in a small region of n-space, the less uncertainty there is in the vector distributed according to this density function. Therefore, a monotone increasing function of the generalized variance $|S_i|$ (the determinant of the covariance matrix S_i of f_i) of $e^{(i)}$ seems an appropriate measure for the uncertainty in $e^{(i)}$.

Note that $I - D^{-1}A$ is of rank at most 2l; we shall suppose that it equals 2l. Then, (3.28) implies that, for $i \geq 1$, $e^{(i)}$ is, with probability one, in a 2l-dimensional subspace. Let us now change to a basis such that its first 2l vectors span the range of $I - D^{-1}A$. In the sequel, we shall suppose that $i > 1$ and we shall replace I, D, A, and S_i by submatrices obtained from the first 2l rows and columns of the accordingly transformed matrices, that is, by that part that acts only within the range of $I - D^{-1}A$.

A convergent iteration causes the entire probability mass to be successively more concentrated in an arbitrarily small neighbourhood of the origin. The speed with which this happens is reflected in the ratio $|S_i|/|S_{i-1}|$, which is smaller than one in the case of convergence. We shall use $\ln|S_i|$ as the measure of uncertainty in $e^{(i)}$, so that the gain in in-

formation is expressed as a difference:

$$\Delta I_i = \ln|S_{i-1}| - \ln|S_i|.$$

With the definition of the covariance matrix S_i of $e^{(i)}$ and (3.28) we find (E denotes the mean):

$$S_i = E(e^{(i)}e^{(i)'}) = E((I-D^{-1}A)\, e^{(i-1)}\, e^{(i-1)'}(I-D^{-1}A)')$$

$$S_i = (I-D^{-1}A)\, S_{i-1}(I-D^{-1}A)'$$

$$\Delta I_i = \ln\left(\frac{|S_{i-1}|}{|S_i|}\right) = -2\ln|I-D^{-1}A| = -2\ln(\rho_1^2 \ldots \rho_1^2)$$

$$\Delta I = -2\ln(\rho_1^2 \ldots \rho_1^2).$$

This index for the performance of Jacobi's iteration is independent of i, independent of the initial error vector, and it is applicable from the first iteration onwards. Furthermore, it has the advantage that it may be interpreted to be the amount of information about the error vector (and, as the initial approximation $x^{(0)}$ is known, also about the solution x) gained in any one cycle of the iteration. Also, we find that the less interaction there is according to the definition (3.26), the greater ΔI. For $j = 1,\ldots,l$ we have $\rho_j^2 \leq 1$; hence

$$\prod_{j=1}^{l} \rho_j^2 \leq 1 - \prod_{j=1}^{l} (1-\rho_j^2) = 1 - |A|/(|A_{11}||A_{22}|) \tag{3.32}$$

$$\Delta I \geq -2\ln(1 - |A|/(|A_{11}||A_{22}|)).$$

3.6.4. CONCLUDING REMARK

We have derived a relationship between the amount of computation required to solve (3.24) and a contribution to complexity, namely, the interaction. Such a relationship may be useful in a theory for the amount of computational effort if we average the amount of interaction over all coordinate systems (or, equivalently, over all orthogonally equivalent forms

of A), because the amount of computing time may depend very much on the coordinate system: if it consists of eigenvectors, A is diagonal and the time required is proportional to n.

But, as we showed in 3.3, complexity is not only contributed to by elements not on the diagonal, but also by unequality among diagonal elements. It may happen that a matrix with much redundancy is diagonal; however, we are not interested in such an exception, but in the *average* over the set of all orthogonally equivalent forms of such a matrix.

It may well be possible to find a measure of such a set that increases with complexity. For instance, the set of matrices orthogonally equivalent to the identity matrix contains only this matrix. The measure of this set would be zero and the identity matrix is the only matrix for which the complexity vanishes. Averaged over the set of orthogonally equivalent forms, a matrix with much complexity would give much interaction and according to (3.31 and 32), this suggests more computational effort. Thus we hope to have reinforced the intuitively plausible relation between complexity and computational effort in solving a certain type of system of linear equations.

REFERENCES

[1] ALEXANDER, C., Notes on the synthesis of form, Harvard University Press, Cambridge, Massachusetts, 1967.

[2] ANDERSON, T.W., An introduction to multivariate analysis, Wiley, New York, 1958.

[3] ARBIB, M.A., Theories of abstract automata. Prentice-Hall, 1969.

[4] ASH, R., Information theory, Wiley, New York, 1965.

[5] ASHBY, W.R., An introduction to cybernetics, Chapman and Hall, London, 1956.

[6] ASHBY, W.R., The set theory of mechanism and homeostasis, L. von Bertalanffy and A. Rapoport (eds): General Systems, 9 (1964), 83-97.

[7] ASHBY, W.R., Measuring the informational exchange in a system, Cybernetica, 8 (1965), 5-22.

[8] BECKENBACH, E.F. & R. BELLMAN, Inequalities, Springer, Berlin, 1961.

[9] BERTALANFFY, L. von, General Systems, Braziller, New York, 1968.

[10] BOULTON, D.M. & C.S. WALLACE, A program for numerical classification, Computer Journal, 13 (1970), 63-9.

[11] BRAMS, S.J., Transaction flows in the international system. The American Political Science Review, 60 (1966), 880-98.

[12] BRAMS, S.J., Measuring the concentration of power in political systems, The American Political Science Review, 62 (1968), 461-75.

[13] BURR, E.J., Cluster sorting with mixed character types II: Fusion strategies, Australian Computer Journal 2 (1968), 98-103.

[14] CHERRY, C. (ed.), Information Theory, Butterworths, London, 1961.

[15] COLE, A.J. (ed.), Numerical taxonomy, Academic Press, London 1969.

[16] FADDEEV, D.K. & V.N. FADDEEVA, Computational methods of linear algebra, Freeman, San Francisco, 1963.

[17] GALLAGER, R.G., Information theory and reliable communication, Wiley, New York, 1968.

[18] GARNER, W.R. & W.J. McGILL, The relation between information and variance analyses, Psychometrika, 21, 219-28.

[19] GELFAND, I.M. & A.M. JAGLOM, Über die Berechnung der Menge an Information über eine zufällige Funktion, Arbeiten zur Informationstheorie, II. H. Grell (ed.), VEB Deutscher Verlag der Wissenschaften, Berlin, 1958.

[20] GOOD, I.J., Probability and the weighing of evidence, Griffon, London, 1950.

[21] GOOD, I.J., The estimation of probabilities, MIT Press, Cambridge, Massachusetts, 1965.

[22] GOOD, I.J., Some statistical methods in machine intelligence research, Virginia, Journal of Science, 19 (1968), 101-10.

[23] HALMOS, P., Finite-dimensional vector spaces, Van Nostrand, New York, 1958.

[24] HARDY, G.H., J.E. LITTLEWOOD, & G. PÓLYA, Inequalities, Cambridge University Press, 1934.

[25] HOTELLING, H., Analysis of a complex of statistical variables into principal components, Journal of Educational Psychology, 24 (1933), 417-41, 498-520.

[26] HOTELLING, H., The most predictable criterion, Journal of Educational Psychology, 26 (1935), 139-42.

[27] HOUSEHOLDER, A.S., The theory of matrices in numerical analysis, Blaisdell, New York, 1964.

[28] JARDINE, N. & R. SIBSON, A model for taxonomy, Mathematical Biosciences, 2 (1968), 465-82.

[29] KHINCHIN, A.I., Information theory, Dover, New York, 1957.

[30] KOLMOGOROV, A.N., A logical basis for information theory and probability theory, IEEE Transactions on Information Theory, vol. IT-14 (1968), 662-4.

[31] KULLBACK, S. Information theory and statistics, Dover, New York, 1968.

[32] LANCE, G.N. & W.T. WILLIAMS, Computer programs for hierarchical polythetic classification, Computer Journal, 9 (1966), 60-4.

[33] LANCE, G.N. & W.T. WILLIAMS, A general theory of classificatory sorting strategies I: Hierarchical systems. Computer Journal, 9 (1967), 373-80.

[34] LANCE, G.N. & W.T. WILLIAMS, A general theory of classificatory sorting strategies II: Clustering systems. Computer Journal, 10 (1967), 271-7.

[35] McGILL, W.J., Multivariable information transmission, Psychometrika, 19 (1954), 97-116.

[36] McLANE, S. & G. BIRKHOFF, Algebra, McMillan, New York, 1967.

[37] MOOD, A.M. & F.A. GRAYBILL, Introduction to the theory of statistics, McGraw-Hill, New York, 1965.

[38] MORRELL, A.J.H., Information Processing 68, North Holland, Amsterdam, 1969.

[39] NAGY, G., The state of the art in pattern recognition, Proc. IEEE, 56 (1968), 836-62.

[40] NEGROPONTE, N., The architecture machine, MIT Press, Cambridge, Mass., 1970.

[41] OKAMOTO, M., Optimality of principal components,MMultivariate Analysis II, P.R. Krishnaiah (ed.), Academic Press, New York, 1969.

[42] OKAMOTO, M. & M. KANAZAWA, Minimization of eigenvalues of a matrix and optimality of principal components, Annals of Mathematical Statistics, 39 (1968), 859-63.

[43] PEARSON, K., On lines and planes of closest fit to systems of points in space, Philosophical Magazine, 2 (1901), 559-72.

[44] PICARD, C., Théorie des questionnaires, Gauthiers-Villars, Paris, 1965.

[45] RAJSKI, C., Entropy and metric spaces, Information Theory, C. Cherry (ed.), Butterworths, London, 1961.

[46] RAO, C.R., Linear statistical inference and its applications, Wiley, New York, 1956.

[47] RESCIGNO, A. & G.A. MACCACARO, The information content of biological classifications. Paper in [14].

[48] RESTLE, F., A metric and an ordering on sets, Psychometrika, 24 (1959).

[49] ROGERS, D.J. & T.T. TANIMOTO, A computer program for classifying plants, Science, 132 (1960), 1115-8.

[50] ROSENBLATT, F., Principles of neurodynamics, Spartan Books, Washington D.C., 1962.

[51] SEBESTYEN, G., Decision-making processes in pattern recognition, McMillan, New York, 1962.

[52] SEGAL, S., Ecological notes on wall vegetation, W. Junk, The Hague, 1969.

[53] SHANNON, C.E., A mathematical theory of communication, Bell System Technical Journal, 27 (1948), 379-423, 623-56. Also in [54].

[54] SHANNON, C.E. & W. WEAVER, The mathematical theory of communication, University of Illinois Press, Urbana, Illinois, 1963.

[55] SIMON, H.A., The architecture of complexity, Proceedings of the American Philosophical Society, 106 (1962), 467-82. Also in [56].

[56] SIMON, H.A., The sciences of the artificial, MIT Press, Cambridge, Massachusetts, 1969.

[57] SOKAL, R.R. & P.H.A. SNEATH, Principles of numerical taxonomy, Freeman, San Francisco, 1963.

[58] TOU, J. & R.P. HEYDORN, Some approaches to optimum feature extraction, Computer and information sciences II (J. Tou, ed.), Academic Press, New York, 1967.

[59] TOU, J., Engineering principles of pattern recognition, J. Tou (ed.), Advances in information systems science, vol. 1, Plenum Press, New York, 1969.

[60] WATANABE, S., Information-theoretical analysis of multivariate correlation, IBM Journal of Research and Development, 4 (1960), 66-82.

[61] WATANABE, S., Karhunen-Loève expansion and factor analysis, Proc. 4th Prague conference on information theory (1965), 635-59.

[62] WATANABE, S., Object-predicate reciprocity and its application to pattern recognition, Paper in [38].

[63] WATANABE, S., Knowing and guessing, Wiley-Interscience, New York, 1969.

[64] WATANABE, S., Feature compression (to be published).

[65] WEYL, H., Philosophy of mathematics and natural science, Princeton University Press, 1949.

[66] WIENER, N., Cybernetics, MIT Press, Cambridge, Massachusetts, 1961.

[67] WILKINSON, J., The algebraic eigenvalue problem, Oxford University Press, London, 1965.

[68] WILLIAMS, W.T. & J.M. LAMBERT, Multivariate methods in plant ecology I, Journal of ecology, 47 (1959), 83-101.