qualify themselves, and changes of venue are sought and granted, all on behalf of this principle. The scientific investigator, on the other hand, is not under so urgent a mandate. An article by the eminent Dirac [1] was published in spite of its central theme that "a theory with mathematical beauty is more likely to be correct than an ugly one that fits some mathematical data." In fact, this patently biased statement was singled out for prominent display by the editors of *Physics Today*. Similarly, it is well known that Einstein resisted the tide of modern physics because he refused to believe that "God rolls dice with the universe." That Einstein's intuition is right or wrong is immaterial. (Indeed, it has recently been suggested by this author that he was right after all [2].) What matters is that the source of his conclusion rests in intuitive bias. In general, it is quite apparent from the literature that the rules of the game do not preclude bias on the part of the scientist. If anything, as Jensen learned, he may incur the displeasure of his colleagues if he fails to display sufficient bias [3]. Thus whereas the judicial system filters out bias entirely, the scientific system filters out only *unpopular* bias. Again pragmatics must be considered. The scientific researcher needs to be knowledgeable in his subject, and knowledge, according to the judicial premise, breeds bias. But still again, in its total intolerance of any bias, popular or unpopular, the judicial system remains more objective than the scientific system.

Summary and Conclusions

While the scientific and judicial fact-finding systems are analogous in many ways, the latter is more skeptical and objective. This is due, at least in part, to pragmatics. However, it would benefit scientists to emulate the judicial method to the extent possible in the areas of skepticism and objectivity. For example, it might be worthwhile, and certainly would be interesting, if basic scientific ideas were presented to an intelligent but unknowledgeable finder of fact under the judicial system. If, for example, the theory of the calculus were on trial before a nonmathematician, and were being defended by Dr. Function, a portion of the transcript might read as follows.

Prosecutor: Now, Dr. Function, you have stated that if a and b are real numbers such that $a + b = a$, then b must be zero.

Dr. Function: That is correct.

Prosecutor: You have also stated that if Δx is allowed to assume the value dx, then $f'(x) + \Delta x = f'(x)$.

Dr. Function: That is also correct.

Prosecutor: Then you admit that $dx = 0$.

Dr. Function: I didn't say that.

Prosecutor: But you did say that if $a + b = a$, then $b = 0$, and you also said that $f'(x) + dx = f'(x)$. So either $dx = 0$, or your first statement isn't true; if $a + b = a$, then b might be zero, and it might be something else called dx. Isn't that right?

Dr. Function: I suppose you could think of it that way.

Prosecutor: All right. Now, will you please tell the court how you know dx isn't zero.

Dr. Function: Well, of course, that's trivial. Dividing by zero is prohibited because it does not produce a unique result. Dividing by dx does produce a unique result. So, clearly, zero and dx are not the same.

Prosecutor: That nonunique result, that's bad, is it?

Dr. Function: I don't know what you mean by "bad." A nonunique result is a meaningless and arbitrary result. So we prohibit division by zero in order to avoid being meaningless and arbitrary.

Prosecutor: I see. Now, a moment ago you admitted that if $a + b = a$, then b might be zero, and it might be dx. I'm not a mathematician, but I do know that if $a + b = c$, then $a = c - b$. So I would point out that you have agreed that $a - a = 0$ or dx. I believe that's what you mean by a nonunique result, and so I must conclude that you're in favor of prohibiting the subtraction of a number from itself. Is that right?

Dr. Function: You're twisting my words. I never said that subtraction could produce a nonunique result.

Prosecutor: The record will show that you did. And while I have no doubt that the calculus works, I submit that you have no idea at all of why it works. No more questions.

Acknowledgment

The author thanks Mr. Robert H. Koenig for helpful discussions.

References

[1] P. A. M. Dirac, "Can equations of motion be used in high-energy physics?" *Phys. Today*, p. 30, Apr. 1970.
[2] A. D. Allen, "Physical reconciliations from the resolution of a Cantorian paradox," presented at the Suppl. Progr., Relativity and Gravitation Sec., 1972 Joint Annu. Meeting APS/AAPT.
[3] The Editors, *Atlantic Monthly*, p. 44, Sept. 1971.

Detecting Subsystems of a Complex System

ROGER C. CONANT

***Abstract*—In studies of complex systems one may be faced with a bewildering amount of data collected on variables of the system without knowing much about the way these variables interact. A measure for the strength of the pairwise interaction of variables in a dynamic system is proposed. In the common case in which the variables have a natural grouping into subsystems within which interaction is strong and between which interaction is weak, the pairwise interactions will usually suggest what the grouping is. Another measure, of interaction within and between subsystems, then serves to verify or discredit the grouping. This decomposition of the system into weakly connected subsystems allows the observer to more readily understand the system by lowering its apparent complexity and directing his attention to appropriate parts of it for more intense scrutiny.**

I. Introduction

A system is commonly thought of as "complex" if it is made up of a large number of parts interacting in a nonsimple way. The linguistic usage, implying as it does that complexity is an attribute of the system itself, obscures the fact that complexity is a relation between an observer and the thing observed, i.e., that the same system can be seen as of greater or lesser complexity by different observers or even by a single observer before and after he comes to "understand" it. This correspondence is motivated by the assumption that the apparent complexity of a system can be lowered by a judicious grouping of its variables into subsystems, as for example, when on the schematic diagram of a radio its subsystems (power supply, detector, amplifier, etc.) are outlined by dotted-line boxes which suggest a block diagram. Broadly put, the goal of this correspondence is to enable the observer of a complex system to see it as less complex by en-

Manuscript received September 1, 1971.
The author is with the Department of Information Engineering, University of Illinois, Chicago, Ill. 60680.

TABLE I

	Time														
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
X_1	1	1	3	3	3	2	2	1	1	2	2	1	1	2	2
X_2	1	1	2	2	1	1	1	2	2	2	1	1	2	2	2
X_3	1	3	3	3	3	3	2	2	1	3	2	2	1	3	2
X_4	1	2	2	1	1	1	2	2	2	1	2	2	2	2	1
X_5	1	1	2	1	1	1	1	2	1	2	2	1	2	1	2

abling him to interpret it, when possible, as nearly decomposable into a set of interrelated subsystems between which the interactions are weak but not necessarily negligible.

Not all complex systems have a hierarchic[1] structure allowing such a decomposition (into subsystems, sub-subsystems, etc.), but Simon in his classic paper [1] has set forth strong arguments for believing that hierarchic structure is the rule rather than the exception in physical, biological, social, symbolic, and many other types of systems. When faced with a complex system which one is trying to understand, then, it is reasonable to start by testing the hypothesis that it has a hierarchic structure and is nearly decomposable into subsystems within which the interaction of variables is relatively intense and between which the interaction is relatively weak, for if that is the case attention can be turned to the detailed workings of each subsystem. One of Simons' main theoretical findings is that [1] "in a nearly decomposable system, the short-run behavior of each of the component subsystems is approximately independent of the short-run behavior of the other components." This result, together with the assumption (nearly a corollary) that the short-run behavior of each of the parts within a subsystem is *not* approximately independent of all other parts in its subsystem, indicates that to form a tentative grouping of parts into subsystems we need a measure of the intensity with which the parts interact. The following sections explore such a measure as well as a measure of the degree of interaction within and between the groups, suggest how these measures can be used to produce and test hypotheses about subsystem groupings, and illustrate the technique with an example.

We suppose, then, that the investigator has before him a mass of data obtained by repeatedly sampling the values (not necessarily metric) of many variables of a complex system which he can at least partially observe but not necessarily control (an ecological system, for example); he would like guidance in searching for subsystems of the complex whole. Part of the data for a rather modest system might look like that given in Table I.

II. The Decomposition Technique

We assume a set of K primary variables, not necessarily metric, each of which has been observed once every "standard time increment" for N increments, giving a total of $K \cdot N$ observations. With each primary variable is associated a derived variable X_j, $1 \leq j \leq K$, whose values are taken to be the positive integers from 1 through M_j (finite) for notational convenience. If the primary variable is not metric, these integers represent its categories, and if the primary variable is metric, these integers represent its values, or ranges of its values if it is a continuous variable. Categories or values must be grouped or ranges quantized so as to make M_j reasonably small; that matter and limits on the "standard time increment" will be discussed in more detail further on.

The variables X_j may be grouped into sets; no confusion results if in this case we let S_j denote the set $\{X_{j1},X_{j2},\cdots,X_{jn_j}\}$ so grouped and also the vector variable $\langle X_{j1},X_{j2},\cdots,X_{jn_j}\rangle$ whose components comprise the set. When necessary, reference will be made to values at different times by superscripts or primes; S_j^k denotes S_j at the kth measurement, $\langle X_i,X_i'\rangle$ is a vector variable whose components are X_i measured at two successive time increments, and so on.

It is well known [2]–[4] that the *entropy* of X_j, denoted $H(X_j)$, is a reasonably good measure of the nonconstancy or variability of X_j. $H(X_j)$ is calculated from the observations on X_j by the following formula:

$$H(X_j) = \log_2 N - \frac{1}{N}\sum_{i=1}^{M_j} n_i \log_2 n_i \tag{1a}$$

in which n_i is the observed number of occurrences of the event $\{X_j = i\}$ and $\sum_{i=1}^{M_j} n_i = N$. If the events occur with definite probabilities p_i, then $H(X_j) \to -\sum p_i \log p_i$ as $N \to \infty$, and (1a) can be thought of as an empirical estimate of the true entropy; however, we do not need to assume the existence of the p_i's in what follows. All quantities discussed in this correspondence, as well as the relations "statistical independence" and "statistical dependence," are interpreted as estimations based on the observations over a finite time span.

The entropy of $S_j = \langle X_{j1},X_{j2},\cdots,X_{jn_j}\rangle$ is denoted by $H(S_j)$. It is a nonnegative measure of the total amount of activity or variability in the set S_j, and it is calculated by a formula similar to (1a):

$$H(S_j) = \log_2 N - \frac{1}{N}\sum_{i} n_i \log_2 n_i \tag{1b}$$

in which, however, n_i is interpreted as the number of occurrences of the ith possible value of the vector S_j. As before, $\sum n_i = N$. The entropy of the union of two sets S_i and S_j is denoted by $H(S_i,S_j)$, and so forth.

The observed *transmission* between S_i and S_j is denoted $T(S_i\colon S_j)$ and is defined as follows:

$$T(S_i\colon S_j) = H(S_i) + H(S_j) - H(S_i,S_j). \tag{2}$$

This is a nonnegative measure of the strength of the relation between S_i and S_j; it is zero if and only if S_i^k and S_j^k are statistically independent (when averaged over all k) [5, pp. 41–43], and it is a maximum (equal to $\min\{H(S_i),H(S_j)\}$) if and only if one vector variable is strictly dependent upon the other, that is, S_i^k is a function of S_j^k, for all $k \leq N$, or vice versa. The generalization of (2) is as follows:

$$T(S_1 : S_2 : \cdots : S_m) = \sum_{i=1}^{m} H(S_i) - H(S_1,S_2,\cdots,S_m). \tag{3}$$

This is a measure of the total constraint holding between, but not within, the sets S_1 through S_m.

Simultaneously measured variables are implied unless a contrary indication is given by primes or superscripts. Thus $T(S_i\colon S_j)$ is a measure of the strength of the relation between S_i and S_j when observed simultaneously; $T(X_i\colon X_j')$ measures the effect of X_i on X_j one time increment later, etc. With regard to $T(X_i\colon X_j')$, note that in observations over N time increments one would obtain N samples of X_i but only $N-1$ samples of X_j' and $\langle X_i,X_j'\rangle$. Since for statistical validity one must have an equal number of samples from each variable, the last sample of X_i

[1] *Hierarchic* does not imply subordination by an authority relation. Quoting Simon [1], "By a hierarchic system, or hierarchy, I mean a system that is composed of interrelated subsystems, each of the latter being, in turn, hierarchic in structure until we reach some lowest level of elementary subsystem."

would not be used, and all entropies would be calculated from $N - 1$ occurrences.

Entropies and transmissions have been used for some time as measures of variability and relatedness, respectively, and their properties are well known. The notation used in this correspondence is consistent with that of Ashby [6], who has developed many identities relevant to the calculations suggested here.

The usefulness of $T(\ )$ in the decomposition of complex systems is suggested by Simon's statement quoted earlier. Suppose a system is in fact "nearly decomposable" into subsystems $S_1,S_2,\cdots,S_m$ (implying a partition of all variables in the system); then one would expect that the constraint holding *between* the subsystems over a short time span would be weak compared to the constraint *within* them. If the time increment is chosen properly (on the order of the time constants of the variables and short compared with time constants of subsystem interactions), the constraint holding over one time increment *within* the jth subsystem $S_j = \{X_{j1},X_{j2},\cdots,X_{jn_j}\}$ will be measured reasonably well by T_{wj}, defined as

$$T_{wj} = T(X_{j1} : X_{j1}' : X_{j2} : X_{j2}' : \cdots : X_{jn_j} : X_{jn_j}')$$

since this transmission measures the nonindependence of all variables in the subsystem over the time increment. The strength of the relation *between* the ith and jth subsystems over one time increment is measured by T_{bij}, defined as

$$T_{bij} = T(\langle S_i,S_i'\rangle : \langle S_j,S_j'\rangle)$$

and the constraint between *all* subsystems over one time increment is measured by T_b:

$$T_b = T(\langle S_1,S_1'\rangle : \langle S_2 : S_2'\rangle : \cdots : \langle S_n,S_n'\rangle).$$

T_b is an upper bound for T_{bij} [6].

In a nearly decomposable system T_{bij} is small compared to $T_{wi} + T_{wj}$, for all i and j, and in addition T_b is small compared to $\sum_j T_{wj}$; the calculation of these transmissions thus allows verification of a proposed grouping of variables into subsystems.

The question remains, how does one infer a proposed grouping from the observations? A reasonable measure of the effect of X_i on X_j one time increment later is the normalized transmission t_{ij}:

$$t_{ij} = \frac{T(X_i : X_j')}{H(X_j')}.$$

$T(X_i : X_j')$ is strongly affected by the number of categories or quantum levels in X_i and X_j, and the normalization indicated largely eliminates that effect. The result, t_{ij}, is always between zero and unity, zero if and only if X_i and X_j' are statistically independent, and unity if and only if X_j' is strictly determined by X_i. Although odd situations can be contrived in which it works poorly, one reasonable procedure for generating a grouping of variables into subsystems is to calculate the t_{ij} for all i and j and then deduce, by starting with the largest value and working down, which variables most strongly affect which others. The grouping can then be checked with the verification procedure suggested in the previous paragraph.

An example will be given in the next section to illustrate the procedure, and in Section IV some of the strengths, weaknesses, and constraints upon the use of the technique will be discussed.

III. Example

A synchronous system of one binary random source and five finite state machines (each having three or fewer states and three or fewer inputs) was simulated, and a history of 1000 steps was generated. The first few values taken by the finite state machines are shown in Table I. From the history, frequency tables for all the pairs $\langle X_i,X_j'\rangle$ were compiled, and the t_{ij} were calculated. The results are shown in Table II. Table II indicates that the strongest effect is that of X_1 on X_3'; then X_4 on X_2', and so on. Fig. 1 shows the result of drawing arrows whose thickness represents t_{ij} between circles representing the variables down to t_{14}, which is the first link between the sets $S_1 = \{X_1,X_3\}$ and $S_2 = \{X_2,X_4,X_5\}$. The figure clearly suggests that the system is nearly decomposable into S_1 and S_2 as subsystems. To confirm this, frequency tables were compiled to enable calculation of

$$\begin{aligned} T_{w1} &= T(X_1 : X_1' : X_3 : X_3') \\ T_{w2} &= T(X_2 : X_2' : X_4 : X_4' : X_5 : X_5') \\ T_b &= T(\langle X_1,X_1',X_3,X_3'\rangle : \langle X_2,X_2',X_4,X_4',X_5,X_5'\rangle). \end{aligned}$$

The frequency tables required are those for X_1 through X_5 (the tables for X_1' through X_5' are assumed to be nearly identical to these) and for the vector variables $\langle S_1,S_1'\rangle$, $\langle S_2,S_2'\rangle$, and $\langle S_1,S_1',S_2,S_2'\rangle$. The entropies are calculated from these tables by (1) and the transmissions from the entropies by (3).

The results are as follows: $T_{w1} = 1.957$, $T_{w2} = 2.721$, and $T_b = 0.422$. Since $T_b \ll (T_{w1} + T_{w2})$, the calculations confirm that the system is indeed nearly decomposable into S_1 and S_2. By way of contrast, if we had chosen $S_1^* = \{X_1,X_2\}$ and $S_2^* = \{X_3,X_4,X_5\}$, a decomposition which Fig. 1 indicates is a poor grouping, the results would have been as follows: $T_{w1} = 0.168$, $T_{w2} = 1.966$, and $T_b = 2.967$. Clearly, the constraint between S_1^* and S_2^* would not have been small compared to the constraints within them.

A complete description of this system and an analysis of its behavior is given elsewhere [5, pp. 129–141]. Here it suffices to say that a detailed analysis supports the reasonableness of the decomposition which has been suggested and confirmed in this example.

TABLE II

	t_{ij}	X_1'	X_2'	X_3'	X_4'	X_5'
				X_j'		
	X_1	0.098	0.013	0.690	0.161	0.073
	X_2	0.002	0.023	0.002	0.145	0.012
X_i	X_3	0.109	0.012	0.353	0.044	0.017
	X_4	0.002	0.413	0.002	0.009	0.021
	X_5	0.000	0.186	0.002	0.259	0.195

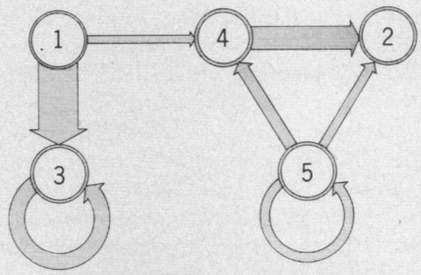

Fig. 1. Seven strongest pairwise relations, as indicated by arrows whose thickness is proportional to t_{ij}.

IV. Discussion

Of the three common measures of relatedness—the correlation coefficient r, the correlation ratio η, and the transmission T—the transmission is the most general since it does not require the values of the primary variables to be numerical or ordered. Attneave [2] has discussed the relation of T and r; the relation

is simple and monotonic for some common distributions. The T measure is nonnegative, which for our purposes here is an advantage, and the fact that T is zero if and *only* if the variables are unrelated is a clear advantage over the correlation measure, which, for example, would show $X_1 = \sin t$ and $X_2' = |\sin t|$ to be uncorrelated even though they are clearly related and would yield a high value of t_{12}.

All three measures require at least weak time invariance of the system under study, but one can hardly hope to analyze a system if that constraint is not satisfied.

It is important that the time increment be appropriate—small compared to the time constants of subsystem interactions but large enough to allow the variables a fair chance of changing values from one observation to the next. Too small an increment may result in entropies and therefore transmissions which are so small as to be statistically unreliable; too large an increment may result in insufficient data on the interactions of quickly changing variables. Within the acceptable range of increments, different subsystem groupings may appear with different choices of the increment. This is not necessarily a disadvantage of the technique; rather, it is useful information for the analysis of the system.

One danger in using t_{ij} as a measure of intervariable relation is that it measures only pairwise dependence and would fail if, for instance, X_k' were strongly dependent upon $\langle X_i, X_j \rangle$ but only weakly upon X_i and X_j individually. This situation is possible but presumably uncommon; if suspected, it can be looked into by calculating all the normalized transmissions $t_{ij,k}$ defined by

$$t_{ij,k} = \frac{T(\langle X_i, X_j \rangle : X_k')}{H(X_k')}$$

(and so forth for higher ordered dependencies if need be).

Since the techniques for decomposition and verification depend on relative rather than absolute values of transmissions, the units in which H and T are expressed need not be the conventional *bits*; thus logs to base 10 or base e may be used in (1) without affecting the results.

H and T as calculated from a finite number of observations tend to be biased estimates of the "true" entropies and transmissions when these exist by virtue of stable probabilities; H tends to be an underestimate and T an overestimate [2]. If the amount of data is too small, they and the technique using them naturally become statistically unreliable. In a discussion of the statistical reliability of H and T, Attneave [2] suggests that N, the number of observations, should at least be as large as df, the degrees of freedom of the frequency table; if X_i has M_i values and X_j has M_j, reliable calculations of t_{ij} would require $N > (M_i - 1)(M_j - 1)$, for example. Preferably N should be much larger than that minimum so that each event allowed by the laws of the system could be expected to occur at least a few times. To satisfy this requirement it is often convenient to combine categories and set quantum levels of the primary variables in a manner which keeps all the M_j relatively small and makes the various values of each variable relatively equiprobable.

At least in the systems to which this decomposition technique is being applied (mostly ecological systems), this requirement for large amounts of data seems not to be a serious problem in the calculation of t and therefore in the production of a grouping of variables into subsystems. It is a much more severe problem in the verification portion of the technique since the df of the frequency tables there (the largest being $df = \prod_{j=1}^{K} (M_j - 1)^2$) can create requirements for unreasonable amounts of data. In such cases one could be satisfied with verifying that $(T_{wi} + T_{wj}) \gg T_{bij}$, for all i and j, or one could reduce the data requirement by grouping categories or ranges of the primary variables (thus reducing the M_j and therefore the largest df), or one could verify the decomposition by inspection of the dynamics of the original system.

This technique lays no claim to infallibility, after all, and it is best viewed as an analytical tool allowing the observer to interact with the data, directing his attention to subsystems which he could then analyze in more detail. It is a tool to be used, like all others, with discretion and with an understanding of its limitations. Properly used, it can be a valuable aid to the understanding of complex systems.

References

[1] H. A. Simon, "The architecture of complexity," *Proc. Amer. Phil. Soc.*, vol. 106, pp. 467–482, Dec. 1962.
[2] F. Attneave, *Applications of Information Theory to Psychology*. New York: Holt, Rhinehart, and Winston, 1959.
[3] W. J. McGill, "Multivariate information transmission," *Psychometrika*, vol. 19, pp. 97–116, June 1954.
[4] W. R. Garner and W. J. McGill, "The relation between information and variance analyses," *Psychometrika*, vol. 21, pp. 219–228, Sept. 1956.
[5] R. C. Conant, "Information transfer in complex systems with applications to regulation," Ph.D. dissertation, Dep. Elec. Eng., Univ. of Illinois, Urbana, 1968.
[6] W. R. Ashby, "Measuring the internal informational exchange in a system," *Cybernetica*, vol. 1, pp. 5–22, 1965.

A Stopping Rule for Trainable One-Dimensional Threshold Learning

J. SKLANSKY AND H. R. RAMANUJAM

***Abstract*—A stopping rule is developed for a class of threshold learning processes (TLP's) that includes the training procedures occurring in certain pattern classifiers, psycho-physical and neural models of perception, and in stochastic approximation. The present correspondence is restricted to one-dimensional pattern spaces. Of the published work on stopping rules, all but Farrell [1] assume that the sequence of observations are independent and identically distributed. The TLP model discussed here gives rise to sequential samples from dependent nonidentical distributions. The stopping rule presented is a result of the technique developed here for obtaining a bounded length confidence interval for a parameter which varies from trial to trial in these distributions. Training is stopped when the interval falls within prespecified limits, thereby assuring a specified performance at any desired confidence level. The rule is illustrated by a numerical example. In the example, both variable and small fixed increment training are considered. An expression is also given for the limit to which the probability of acquiring the stopping criterion converges in probability.**

I. Introduction

We consider here a simple but nontrivial model of learning with a teacher, called the threshold learning process (TLP) [2]. The TLP is a trainable pattern recognizer or signal detector that undergoes a training phase followed by a working phase. It exhibits important features of learning, is representative of some actual natural and artificial learning systems, and can be easily simulated on the digital computer.

Manuscript received May 12, 1971. This work was supported by the U.S. Air Force Office of Scientific Research under Grant AFOSR 69-1813. This correspondence was presented at the IEEE Symposium on Adaptive Processes, University of Texas, Austin, December 7–9, 1970.

The authors are with the School of Engineering, University of California, Irvine, Calif. 92664.