On the Identification of Structures in Multivariate Data by the Spectral Analysis of Relations

Klaus Krippendorff
University of Pennsylvania, kkrippendorff@asc.upenn.edu

Follow this and additional works at: http://repository.upenn.edu/asc_papers

Part of the Communication Commons

Recommended Citation

This paper is posted at ScholarlyCommons. http://repository.upenn.edu/asc_papers/207
For more information, please contact repository@pobox.upenn.edu.
On the Identification of Structures in Multivariate Data by the Spectral Analysis of Relations

Abstract

Complex systems are typically manifest in multi-variate data. The analysis of such data is therefore an intrinsic effort of systems research.

In this paper, a distinction is made between the logic of many-valued relations and the quantification of the strength of these relations in data. The former is unrelated to data whereas the latter is invoked in two empirical tasks: the identification of structures in data and the construction of models and theories that reproduce or explain these data at least approximately so. A unified calculus is proposed to aid both tasks. It required a third generalization of Shannon's quantity of communication. The paper presents several algebraic identities of the quantity with entropies, transmissions and interactions.

These identities are intended to provide the basis of two separate identification strategies; decomposition and composition. These are exemplified. The paper concludes with pointing out several yet unresolved problems.

Disciplines
Communication | Social and Behavioral Sciences

This book chapter is available at ScholarlyCommons: http://repository.upenn.edu/asc_papers/207
ON THE IDENTIFICATION OF STRUCTURES
IN MULTIVARIATE DATA
BY THE SPECTRAL ANALYSIS OF RELATIONS

by

Klaus Krippendorff

The Annenberg School of Communications
University of Pennsylvania, Philadelphia

ABSTRACT

Complex systems are typically manifest in multi-variate data. The analysis of such data is therefore an intrinsic effort of systems research.

In this paper, a distinction is made between the logic of many-valued relations and the quantification of the strength of these relations in data. The former is unrelated to data whereas the latter is invoked in two empirical tasks: the identification of structures in data and the construction of models and theories that reproduce or explain these data at least approximately so. A unified calculus is proposed to aid both tasks. It required a third generalization of Shannon's quantity of communication. The paper presents several algebraic identities of the quantity with entropies, transmissions and interactions.

These identities are intended to provide the basis of two separate identification strategies: decomposition and composition. These are exemplified. The paper concludes with pointing out several yet unsolved problems.

BACKGROUND

I was acquainted with Ashby's work on constraint analysis of many-dimensional relations (1964) before I had the occasion to apply statistical techniques on social data. Prejudiced against reducing everything to binary relations, I soon became dissatisfied with the kind of relational atomism that is built into many of these techniques, notably correlation, multiple regression, factor analysis and many more. They ignore the kind of complexities so typical of systems in general, and of many phenomena of social communication in particular. This motivated my search for new ways of conceptualizing dependencies of higher ordainal so as to make them accessible to analysis.

The idea of a spectral analysis of relations had been in the back of my mind for some time before I finally presented the notion to the 1976 Society for General Systems Research meeting in Denver and subsequently, in the form of a paper, to the International Congress for Communication Sciences in Berlin (1976). The method allows multivariate data on a complex system with, say, m components to be analysed into \(2^m - m - 1\) possible interdependencies (the spectra). To assess how much these logically distinct interdependencies account for the whole, each is assigned a magnitude. And the algebraic sum of all these magnitudes add up to the total amount of information transmitted within the system. A year later, I came in contact with Kliir's (1976) work which took a rather different approach, focussing attention to the identification of structures of a certain type but within the same kind of data I was concerned with. The links between the spectral analysis of relations and the problem of structure identification became apparent in a paper of mine delivered at the Fourth European Congress for Cybernetics and Systems Research in Linz (1978). There I touched base with Broekstra who had applied information theoretical measures on Kliir's structure types (1976, 1977).

Naturally, the problems associated a spectral analysis of relations, with constraint analysis, and with the design of techniques for identifying structures are far from being all solved. While the results presented in this paper are not entirely unproblematic either, I hope that they are a step in a fruitful direction.

ON THE LOGIC OF MANY-VALUED RELATION

At least since Wiener's (1914) paper we conceive of a relation as a subset of a larger set. Typically, this larger set is a product of several component sets, constituting the relation's arguments. The "Gestalt switch" from conceiving a relation as a single link between two entities (John communicates with Mary, 7 is larger than 5, etc.) to its conception as a set of elements (e.g. \(<\text{John, Mary}> \cup \{7, 5\}\) ) whose values satisfy the relation. (X communicates with Y, X is larger than Y) made relations amenable to mathematical treatment and facilitated generalizations to complexities that natural language cannot easily, if at all, describe. It also allowed us to consider relations under the dual aspect of freedom and constraint which lies at the root of the notion of information: freedom (variety and uncertainty) in the sense that some and often many elements are equal in satisfying a relation, thus including alternatives, and constraint in the sense that some and often many combinatorially possible elements are excluded by not satisfying a relation. Graphically:

```
constraint
\[\text{freedom} \quad \text{variety}\]
\hspace{0.5cm} a distinction
```

A relation is seen as invoking a distinction within a domain of possible elements.

Extending these notions, many-valued relations are nothing more than subsets of the products of several sets or contained in multi-variable spaces. A logic of many-valued relations concerns the dependencies that do exist between relations, particularly the relations of different ordainal, and the ordering of sets of such relations. Inasmuch as data are collected on many variables they do constitute many-valued relations that need to be explained, accounted for and represented using the same logic.
Many analytical problems with multi-variate data and presumably the reason for avoiding many-valued relations as a source of explanation stem from the large computational efforts they require. In the number \(2^{(k^m)}\) of possible relations in an \(m\)-dimensional space of \(k\) alternatives in each, the dimensionality \(m\) appears as the exponent and signals explosive increases in the number of relations with only moderate increases in the number of variables. Ashby (1964) was probably the first to recognize the virtue of attempting to explain a complex relation in terms of the conjunction of several less complex relations. His "constraint analysis" in effect tests whether an \(m\)-valued relation can be explained in terms of:

- \(m\) unary relations or properties
- \(\frac{(m-1)}{2}\) binary relations
- \(\frac{m!}{(m-r)!r!}\) \(r\)-valued relations
- 1 \(m\)-valued relation (the data)

The outcome of Ashby's constraint analysis is a measure he calls "cylindrance" which is the smallest ordinality of the \(\frac{m!}{(m-r)!r!}\) \(r\)-valued relations that do account for the given \(m\)-valued data. Often an increase in the cylindrance improves the account only negligibly, giving the researcher an indication of the complexity he must, may or does not need to consider in analyzing his data. (The point made earlier is that many statistical techniques in the social sciences are arbitrarily set to a cylindrance of two and are therefore unable to recognize higher order dependencies if they exist. A peculiarity of such techniques is that without variability in ordi-nality, their own inadequacy becomes unrecognizable.)

The approach I am taking differs from Ashby's and conforms with Klir's (1976) insofar as it allows the set of relations that are tested for their accountabili-ty of data to have different ordinality. Also, following Klir I am concerned with the realization of components and links between them which would within a margin of tolerable error reproduce the original data. If, for example, a real system consists of two interdependent components:

\[
A \rightarrow B \rightarrow C \rightarrow D
\]

the four-valued data such a system can exhibit are less then maximally complex. They need not be described in terms of the all-variables-embracing quaternary relation. And because \(D\) is linked to \(A\) and \(B\) only via \(C\), three of the four tertiary relations between \(ACD\), \(BCD\) and \(ABD\) are also absent leaving simple \(\alpha(ABC)\) and \(g(CD)\). If the relation between \(ABC\) is not reducible without the loss of the above binary relations in this case, a realization of \(\alpha\) will have to include whatever additional dependencies might exist in the data between \(A\) and \(B\), \(A\) and \(C\) and \(B\) and \(C\). An account of data in terms of components of min-imal ordinality does not need to concern relations that are merely embedded.

One task of the logic of many-valued relations is to differentiate structures (such as depicted above) that would explain many-valued data and to show any logical ordering (>) between them. Avoiding formalism, I shall define these and explain the conventions to depict them as follows:

(a) An \(m\)-valued relation, i.e., a relation with \(m\) arguments is depicted by a box with \(m\) inputs or possible connections, each signed by a different argument.

\[
\begin{array}{c}
A \hspace{1cm} B \hspace{1cm} AB \hspace{1cm} ABC \hspace{1cm} \text{etc.}
\end{array}
\]

(b) Two many-valued relations are connected with each other if their arguments overlap, being neither included, equivalent nor empty. Connected relations are depicted by separate boxes with connecting lines representing shared values, e.g.:

\[
\begin{array}{c}
\alpha(ABC) \hspace{1cm} \beta(BCD) \hspace{1cm} \gamma(ACD)
\end{array}
\]

(c) One relation is embedded in another if all of the former's arguments are contained in the latter's. Embedded relations are obtained by projecting a relation in one space on a space of lower dimensionality. Embedded relations do not differentiate among structures and are not depicted, e.g.:

\[
\begin{array}{c}
R(AC) \cap R(BC) \cap R(CD) \quad \text{and} \quad R(ABC) \cap R(ACD) \quad \text{is} \quad R(ABC) \cap R(CD)
\end{array}
\]

(d) Two relations are equivalent if they share the same arguments. Being realizable by one component, equivalent relations are not differentiated, e.g.:

\[
\begin{array}{c}
A \quad B \quad \text{equals} \quad A \quad B
\end{array}
\]

(e) A structure is any collection of relations that are neither embedded in another, nor equivalent. In a lattice of possible relations, a structure represents a cut through this lattice, distinguishing the relations of a structure and its embedded relations from all others, e.g.:

\[
\begin{array}{c}
\text{structure}
\end{array}
\]

(f) Starting with the relation of the highest possible ordinality on its top, the original \(m\)-valued data, structures too form a lattice of descending complexity by removing one non-embedded relation after the other. The base of this lattice, the least complex structure, consists of an account of the data in terms of \(m\) unary relations, which regard the data as a mere aggregate of independent variables. (See Fig. 1 for example)

(g) Two structures are equal in complexity if they are obtainable from a third, more complex structure by the removal of an equal number of non-embedded relations. One struc-
ture is more complex than another structure, if the former is obtainable from a third by removing fewer non-embedded relations than required to obtain the latter.

In effect I am proposing a measure of structural complexity. It is zero for the least complex structure in which all variables are independent and otherwise counts the number of non-embedded relations that have to be removed to obtain this most simple case.

For data with four variables, for example, the most complex account involves the one quaternary relation corresponding to the original data with each unique quadruple of values regarded as significant. Removing this (a priori non-embedded) relation, the four tertiary relations now become non-embedded and constitute a somewhat simpler representation of the original data. The next two steps call for decisions regarding which of the tertiary relations can be removed without or with minimal loss. The fourth step towards a simpler structure provides the option of either removing one of the two remaining tertiary relations or the one binary relation that is now non-embedded, etc., until the structure is merely nominal and the data are explained in terms of four separate components. Ignoring, as in the above description, structures that differ merely in the permutation of variables, the lattice of structure types for four variables is depicted in Figure 1. The figure also shows the structures that Ashby's constraint analysis and Klir's structure identification algorithm would respectively identify.

A few comments might be in order. Klir's algorithm for structure identification proceeds from the most complex to the least complex structure. Ashby's goes the other way around. Since simple relations are more easily computable than complex ones, there are considerable computational advantages in starting with the least complex case. Although I ordered the lattice of possible structures by the number of non-embedded relations that have to be removed to obtain one structure from the other, the last section of this paper will show a way of accounting for such losses when the identification proceeds from the least to the most complex structure.

In the comparison of the three approaches it is apparent that Klir's structure types do not include structures with circular dependencies. Ashby, on the other hand includes only circular dependencies between components of the same ordinality. (It should be noted though that Ashby never considered his a structure identification procedure).

Before proceeding to the quantification of such structures, I should like to reiterate that a lattice of possible structure types is a logical device which can be considered separately from quantitative arguments and criteria that a researcher might employ to justify a simplification of his data. Ashby's set-theoretical operation \( p^{-1}_e p \) provides the basis for all non-probabilistic identification procedures. Klir's \( p_{o \rightarrow e} \) is an example of a probabilistic criterion.

In the following I will rely on information theoretical measures. All involve the logic of many-valued relations.

### A NEW MEASURE OF INTERDEPENDENCE

Based on the primary measure of entropy:

\[
H(X) = -\sum_{x \in X} p_x \log_2 p_x
\]

\[
H(A \ldots X) = -\sum_{a \in A} \ldots \sum_{x \in X} p_{a \ldots x} \log_2 p_{a \ldots x}
\]

information theory offers a variety of secondary measures, i.e., measures that are defined in terms of sums and differences. For example, the conditional entropy is:

\[
H_A(X) = H(XA) - H(X)
\]

The most important property of this function is that whenever the probabilities of joint events are the product of the probabilities of individual events, entropies are additive and the lack of additivity signifies interdependence. Thus, if \( X \) and \( Y \) are independent, \( H(XY) = H(X) + H(Y) \) with the difference \( H(X) + H(Y) - H(XY) \) being a measure if interdependence.

In developing information theory into a multivariate calculus, McGill (1954) and Ashby (1965, 1969) have generalized Shannon's (1949) initially binary notion of "information rate," "association," etc.
into two multivariate expressions: the amount of information transmission \( T \), and the amount of interaction \( Q \). Transmission has become a measure of the overall constraint within a given system, the difference between the maximum entropy that is to be expected when all variables were independent and the observed entropy. Interaction has become a measure of the uniqueness of the relationship within a set of variables with the effects of all embedded relations on this measure removed. These quantities are defined as (3) and (4) in Table 1.

In my original proposal for a spectral analysis of relations (1976) I suggested that the overall constraint implied by \( m \)-valued data be decomposed into the constraints implied by each of the \( 2^m - 1 \) embedded relations. For this purpose I developed the identity:

\[
T(AB) = Q(AB) \\
T(ABC) = Q(AB) + Q(AC) + Q(BC) + Q(ABC) \\
T(ABCD) = Q(AB) + Q(AC) + Q(AD) + Q(BC) + Q(BD) + Q(CD) + Q(ABCD) + Q(ACD) + Q(BCD) + Q(ABD) + Q(ABC) + Q(ACB) + Q(BCA) + Q(DAB) + Q(DAC) + Q(DBC) + Q(DCB) + Q(DABCD) \\
\text{etc.}
\]

It shows the total transmission \( T \) within a system to be the algebraic sum of the interactions \( Q \) within each subset of variables. Being more specific than transmission measures, interactions are indispensable for all efforts to account for the organization of multivariate data, specifically for identifying structures.

In applying the spectral analysis of relations to identify structures (1977) I encountered several oddities, however. For example, if one adds a dummy-variable to the list, one that merely copies the values of another variable already included, then, without increasing the entropy of the data, the transmission measure increases by the entropy of the copied variable. Transmission measures always increase with the addition of new variables and are thus, more so than others, an artefact of the researcher's description. Interaction measures, on the other hand, change signs by the addition of such a dummy-variable and can thus take negative values.

Previously, I had interpreted negative interactions as an indication of overdetermination by lower order relations. This seems not generally so. Broekstra responded (1978) with the suggestion that the sign of interactions simply be ignored which would bury the concept of overdetermination altogether. After further exploration, I now like to qualify my interpretation of the sign of unconditional interactions:

A \( Q \)-measure is indicative of over determination only if all \( Q \)-measures of immediately embedded relations are of opposite sign.

It should be noted that binary interactions are always positive. But for strictly linear relations (which are in fact prime examples of over-determination) all tertiary relations are negative, all quaternary relations again positive, etc. And when linear relations are one-to-one-to...to-one, all interactions are also equal in absolute value. The quantitative account for such a relation appears almost as an "undamped oscillation" passing through the whole lattice of possible relations, finding no clear focus on a single expression that could be said to represent the structure. Under these conditions, the sum of any two \( Q \)-measures of relations, of which one is immediately embedded in the other, is zero, suggesting that the "oscillation" can be represented by the more "stable" sum. But according to (8) below, this sum is a conditional \( Q \)-measure. Conditional \( Q \)-measures which also correct for spurious interactions thus seem to be much more important in structure identification than previously assumed.

Another and probably related problem is that unconditional interactions in (5) serve at least in part, to compensate for the way transmissions happen to account for the variety in data. In the binary case, the situation is still clear; Bivariate interaction measures are always positive and the portion of the entropy \( H(AB) \) that contributes to \( T(AB) \) is "counted" only once. But in the three-variable case:

<table>
<thead>
<tr>
<th>Information Transmission ( T(\ ) )</th>
<th>Interaction ( Q(\ ) )</th>
<th>Systematic Entropy ( S(\ ) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>(A)</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>(AB)</td>
<td>( H(A) + H(B) ) - ( H(AB) )</td>
<td>-</td>
</tr>
<tr>
<td>(ABC)</td>
<td>( H(A) + H(B) + H(C) ) - ( H(ABC) )</td>
<td>-</td>
</tr>
<tr>
<td>(ABCD)</td>
<td>( H(A) + H(B) + H(C) + H(D) ) - ( H(ABCD) )</td>
<td>-</td>
</tr>
<tr>
<td>(m variables)</td>
<td>( m ) terms ( H(1 ) variable) ( - H(m ) variables)</td>
<td>-</td>
</tr>
</tbody>
</table>

A Comparison of Transmissions, Interactions, and Systematic Entropies

Table 1
the entropy contributing to the tertiary interaction is also included in the three bivariate measures, so that Q(ABC) is seen as compensating for the double count by the tri-variate transmission T(ABC).

This can be seen in the following diagrams:

[Diagrams showing entropy contributions]

In fact, r-valued transmission measures contain the entropies that contribute to r-valued interactions r-1 times! This suggests, perhaps, that transmission measures are not the kind of measures that should be accounted for.

What I am proposing here is a third generalization of the bivariate notion of interdependence. Its definition is given as (6) in the third column of Table 1, which is constructed to make the difference between the three measures apparent.

By excluding the noise within each variable, the new measure assesses the amount of entropy that is responsible for the interdependencies among variables. It is the quantity which a theory or model attempts to account for or explain. In the tradition of distinguishing between systematic and random variation, it is a measure of the systematic variation encountered in data. For lack of a better term, I shall call it a measure of systematic entropy, S.

Only in the bivariate case are transmissions, interactions and systematic entropies equivalent: T(AB) = Q(AB) = S(AB). Although S assumes otherwise different values, it is always positive and bounded by: 0 < S(AB...Z) < H(AB...Z). Its upper limit is reached when the noise within individual variables is absent and the total entropy is devoted in full to the interdependencies among variables. It is zero when all variables are independent and the data exhibit no structure whatsoever. Thus, S is a third way of indicating the amount of structure in data. It is basically still an entropy measure and therefore the least abstract or weakest of the three measures of interdependence. I suggest that its "strength" lies in its interpretability as the next section may show.

For the following it must be noted that all conditional measures can be formulated by analogy to the definitions in Table 1; the variable(s) on which these measures are (are) conditional must occur in all subscripts of those entropies in terms of which they are defined. E.g. for three variables, conditional on a third:

\[ T_X(ABC) = H_X(A) + H_Y(B) + H_Z(C) - H_{X|Y|Z}(ABC) \]

\[ Q_X(ABC) = -H_X(A|BC) + H_X(AB) + H_X(AC) + H_X(BC) \]

\[ S_X(ABC) = H_X(AB) - H_{BCX}(A) - H_{ACX}(B) - H_{AXC}(C) \]

Generally:

\[ T_X(ABC...) = T(ABC...,X) \]

\[ Q_X(ABC...) = Q(ABC...,X) - Q(ABC...) \]

\[ S_X(ABC...) = S(ABC...,X) - S(ABC...,X) \]

Where the underlined variables are regarded as one compound variable.

QUANTITIES COMPONENTS HAVE TO REPRESENT

A model may be said to simulate and thereby represent certain essential features of the behavior of a designated portion of reality. Reality is manifest in data and the purpose of the analyzing structures in data is to guide the researcher in his attempt to construct models of appropriate composition and with minimal complexity. Although I should like to keep identification or analysis and realization or synthesis as separate problems, an integrated calculus is desirable so that solutions to these two problems do not conflict. This section develops that part of this calculus which is concerned with realization only.

Unfortunately, there is no universal agreement regarding what "minimal complexity" is and how it could be measured. The International Journal of General Systems recently devoted a whole issue to the subject (Vol.3, No.4, 1977). Obviously "complexity" cannot be abstracted entirely from the purpose of the situation in which it presents a problem. Since I am concerned here with the construction of models from components, my preference would be for a very simple measure: the number of states that the components of a model must possess in order to reproduce given data.

Happily, entropies can be interpreted as a kind of count. When a variable X consists of r equally likely events, this number is related to the entropy by: \( 2^r \cdot X = n \). Under optimum coding conditions \( 2^r \) is the number of alternatives required to reproduce data. Entropies merely express these alternatives by the exponent of two. But since entropies are averages, less optimal coding conditions tend to prevail in which case \( 2^r \) expresses the minimum number of alternatives a representation requires. In either case, I am suggesting that entropies serve as a good indication of the order of magnitudes of the number of alternatives required in realizing models.

Reemphasizing the point made earlier, I am not so sure whether transmission measures have suitable interpretation in model construction. Although transmissions are related to the number of alternatives theoretically possible but not actually used, they do not assess how much variety actually goes into a model. Transmission measures are typically large for linear models which are inherently simple to build and small for non-linear and generally more complex models. I am also suggesting that interaction measures are of limited value in realization problems. Once a component has been identified as no longer decomposable, interaction measures have served their purpose. A component must realize not only what is unique among its variables. It must represent all dependencies, unique and embedded, among them.

Modelling suggests a distinction between systematic entropy and noise. Noise is free variation. It cannot be accounted for by reference to variables other than itself. In a system with variables A, B, C, Z, random variation or noise is:

\[ H_{BC...,A}(A) + H_{AC...,B}(B) \ldots + H_{ABC,...,Z}(Z) \]

If this sum equals the total entropy \( H(AB...Z) \), the model has no structure and the data can be reproduced by independent components, it is the systematic entropy \( S(AB...Z) \) that a model needs to account for by its structure. In separating systematic entropy from noise, one commits oneself, albeit implicitly, to the conception of a model whose interdependencies are represented in (initially) one complex component to
which several independent noise generators attached:

\[
\begin{array}{c}
\text{S(AB...Z)} \\
\text{A} \\
\text{B} \\
\text{C} \\
\text{D} \\
\text{Z} \\
\end{array}
\]

Supposing now, that the structured part of a model is composed of two interdependent tertiary components:

A \[\rightarrow\] B \[\rightarrow\] C \[\rightarrow\] D

each component could be held respectively accountable for the quantities S(ABC) and S(BCD). However, such a separate realization would include two sources of redundancies. The first lies in the fact that variation is exaggerated by variables that are not part of the component, for the ABC-component and A for the BCD-component. Discounting this variation, the \( S_D(ABC) + S_A(BCD) \) may still be larger than necessary because the two components are in communication via B and C and thus share some entropy. This second source of redundancy appears as the shared quantity \( S_{AD}(BC) \) which may be realized in either of these components but not in both. \( S \), the total amount of systematic entropy the above model is able to account for thus is:

\[
S(ABC/BCD) = S_D(ABC) + S_A(BCD) - S_{AD}(BC)
\]

Generalizing the quantity of systematic entropy to any structure involving interdependent or independent components, one obtains:

\[
S(\text{a structure}) = \text{sum of all conditional terms for variables in each component} - \text{sum of all conditional terms for variables shared by any two components} + \text{sum of all conditional terms for variables shared by any three components} - \ldots \text{four} \ldots + \ldots \text{five} \ldots \text{etc.}
\]

**Two Strategies for the Identification of Structures**

Whereas realization problems seem to be more clearly tied to entropies, structures are justified only in terms of interactions. I also registered some doubts concerning the use of transmission measures in realization problems and suggested that conditional interaction measures prevent spurious interdependencies from entering an account. With this in mind let me now present a new accounting equation for the spectral analysis of relations that seems particularly suited for structure identification problems and devote the remainder of this paper to its use. This equation decomposes the total entropy in data into the interaction measures within all \( 2^{m-1} \) subsets of \( m \) variables, each being conditional on its complement within \( m \) variables:

\[
\begin{align*}
H(A) &= -Q(A) \\
H(AB) &= -Q_B(A) - Q_A(B) + Q(AB) \\
H(ABC) &= -Q_{BC}(A) - Q_{AC}(B) - Q_{AB}(C) + Q_B(AC) + Q_A(BC) - Q(ABC) \\
\text{Generally,} \\
H(m \text{'v'les}) &= -m \text{ terms } Q_{m-1} \text{'v'les} \text{ (1st order)} + \frac{m(m-1)}{2} \text{ terms } Q_{m-2} \text{'v'les} \text{ (2nd order)} - \frac{m(m-1)(m-2)}{6} \text{ terms } Q_{m-3} \text{'v'les} \text{ (3rd order)} + \ldots \\
&\quad \ldots \\
&\quad \pm Q(m\text{-th order})
\end{align*}
\]

The signs of these conditional Q-measures (positive for even ordinalities, negative otherwise) do imply nothing about the sign of their value. In fact, the value of \(-Q_X(1\text{st order})\), being interpretable as conditional entropies are always positive, the values of \(-Q_X(2\text{nd order})\), being interpretable as conditional transmissions, are also always positive whereas the values of \(-Q_X(3\text{rd order})\) may take positive or negative values. In strictly linear and deterministic dependencies in data, all conditional Q-measures are zero and the total entropy equals the highest order (unconditional) interaction and is positive, regardless of whether this relation is of even or uneven ordinality. It follows from (6) and (11) that the systematic entropy can also be expressed in terms of conditional interactions.

\[
\begin{align*}
S(AB) &= Q(AB) \\
S(ABC) &= Q_C(AB) + Q_B(AC) + Q_A(BC) - Q(ABC) \\
S(ABCD) &= Q_{CD}(AB) + Q_{BD}(AC) + Q_{BE}(AD) + Q_{AD}(BC) + Q_{AC}(BD) + Q_{AB}(CD) - Q_B(ABC) - Q_C(ABD) - Q_B(ACD) - Q_A(BCD) + Q(ABCD)
\end{align*}
\]

etc.

I do not want to formally prove the identity but I will demonstrate, by means of a Venn diagram how the entropy is enumerated in (11). The shaded area represents the systematic entropy \( S \) and thus demonstrates (12) as well:

![Venn diagram](image-url)
Apparently, one can now express the total entropy in given data as the algebraic sum of the noise, the systematic entropy within a model, and the systematic entropy that the model with its particular structure is unable to represent. To use the previous example of two interdependent tertiary components within four values:

\[
H(ABCD) = H_{BCD}(A) + H_{ACD}(B) + H_{ABD}(C) + H_{ABC}(D)
\]

\[
S(ABCD) = Q(BCD) - Q_{AB}(ACD) - Q_{AB}(ABC)
\]

More generally, when representing all possible conditional Q-measures as a lattice with the one mth order (unconditional) Q-measure on the top and the m noise terms on the bottom, the cut that any m-valued structure defines in this lattice divides these Q-terms into those contained in the model and those excluded from it. The sum of all Q-measures above this cut equals the systematic entropy the model cannot represent, and the sum of all Q-measures below this cut and above the noise equals the systematic entropy the structure of a model can represent:

systematic entropy lost

systematic entropy representable in model

With (11) and (12) we have thus accomplished a complete partition of the observed entropy into three additive quantities:

\[
\text{Entropy} = \text{Noise} + \text{Structure} + \text{Structure in Model lost}
\]

This complete partition now enables us to formulate two strategies for the identification of structures: decomposition and composition. The decomposition strategy works from the top down the lattice, systematically removing non-embedded relations whose conditional Q-measures are zero or negligibly small in absolute value, controlling that the cumulative losses do not exceed specified limits. The composition strategy works from the bottom up the lattice, systematically adding relations that cover all of its embedded relations adding conditional Q-measures to the systematic entropy a model can represent in the order of their absolute value until the total systematic entropy is reasonably approximated. As discussed earlier, Klir's is a decomposition strategy while Ashby's is a composition strategy though neither was formulated to consider all possible structures. Figure 2 exemplifies the two strategies within four variables. The decomposition strategy removes interaction terms that are insignificantly small and adds them to the losses while the composition strategy adds interaction terms that are large in absolute value.

Systematic Entropies Accountable by Structure

\[
S(ABCD)
\]

+ Composition gains

+ Decomposition losses

0 Quantities for Composition and Decomposition in a Lattice of Structure Types for 4 Variables

Figure 2
ue to the systematic entropy. With decision criteria clearly spelled out, they should meet at the same structure.

Let me now demonstrate the procedures by two examples. The first is hypothetical and published as Case 1 by Broekstra (1979:23). With variables denoted by 1, 2, 3, and 4, the example is constructed so that $p_{1234} = \frac{p_{134} p_{23}}{p_3}$. The frequencies with subscripts denoting a vector in this four-dimensional space and a complete account in terms of conditional Q-mesures to express (11) and (12) are as follows:

$$
\begin{align*}
\phi_{1234} & = 0.4512 \\
\phi_{134} & = 0.8750 \\
\phi_{124} & = 0.1722 \\
\phi_{123} & = 0.3444 \\
\phi_{34} & = 0.0000 \\
\phi_{24} & = 0.2054 \\
\phi_{23} & = 0.4108 \\
\phi_{14} & = 0.0778 \\
\phi_{13} & = 0.0000 \\
\phi_{12} & = 0.1312 \\
\phi_{1} & = 0.0321 \\
\phi_{2} & = 0.0000 \\
\phi_{3} & = 0.0790 \\
\phi_{4} & = 0.0244 \\
\phi_{0} & = 0.0206
\end{align*}
$$

\[
R(1234) = 2.6249 \text{ bits}
\]

The first observation these quantities suggest is that the data contains more noise than structure with variable 2 being the largest source. Concentrating on the systematic entropy, the following steps can be computed either by decomposition or, as in this case, by composition, adding or subtracting the largest absolute quantities at each step:

<table>
<thead>
<tr>
<th>Step Components</th>
<th>Retained S-entropy</th>
<th>Lost S-entropy</th>
<th>Structure</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>-0.0000</td>
<td>0.7821</td>
<td>none</td>
</tr>
<tr>
<td>1</td>
<td>0.4108</td>
<td>0.3713</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>0.6162</td>
<td>0.1659</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>0.7474</td>
<td>0.0347</td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td>0.8252</td>
<td>0.0431</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>0.7462</td>
<td>0.0359</td>
<td>2</td>
</tr>
<tr>
<td>6</td>
<td>0.7783</td>
<td>0.0038</td>
<td>2</td>
</tr>
<tr>
<td>7</td>
<td>0.8027</td>
<td>0.0206</td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td>0.7821</td>
<td>0.0000</td>
<td>2</td>
</tr>
</tbody>
</table>

It is apparent that variable 2 becomes part of structure only at the 4th step. The negative losses at this step make it also apparent that the four binary components over-determine the data. This is corrected only after the negative quantities of interaction are added to the account at the 5th step. The approximation at the 6th step is probably the best approximation one can get. Indeed $Q_{12}(24)=0$ which suggests that the relation between 2 and 4, if it exists is mediated by 1 and 3.

This example is instructive because it makes the difference between accounting for transmissions and accounting for systematic entropies apparent. I indicated above how Broekstra constructed the data. It is therefore not surprising that, accounting for transmissions, $T(1234)=T(134)+T(23)$. This is quite convincing and indeed leads to questions why under the same conditions $S(134/23)<S(1234)$. The answer lies in the fact that $T$ ignores the possibility that interdependencies within the data are unequally associated with the values of the connecting links while $S$ is sensitive to this inequality. In the example, the values of variable 3, occurring with rather unequal probability, are also rather unequal in carrying the interdependencies among the remainder of the variables. Graphically:

<table>
<thead>
<tr>
<th>Variables:</th>
</tr>
</thead>
<tbody>
<tr>
<td>14 3 2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Values:</th>
</tr>
</thead>
<tbody>
<tr>
<td>11</td>
</tr>
<tr>
<td>00</td>
</tr>
<tr>
<td>01</td>
</tr>
<tr>
<td>10</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>n, 1 = 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>01</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>n, 0 = 14</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
</tr>
</tbody>
</table>

It follows that higher-order interactions are non-zero. The question thus remains whether such non-zero interactions are important. Broekstra's analysis ignores them and furthermore suggests that the structures 134/23, 134/234, 134/123 and of course 1234 offer equivalently perfect accounts of the data. I am not implying that one method is better than the other but I am suggesting that the differences need to be examined in view of what either method includes or omits.

The second example is taken from Lazarsfeld (1974) who presented NCB data on exposure to violent TV programming, E, and aggressive behavior, A both at different points in time. Time 2

<table>
<thead>
<tr>
<th>E</th>
<th>A</th>
</tr>
</thead>
<tbody>
<tr>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>+</td>
<td>+</td>
</tr>
</tbody>
</table>

Time 1

<table>
<thead>
<tr>
<th>E</th>
<th>A</th>
</tr>
</thead>
<tbody>
<tr>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>+</td>
<td>+</td>
</tr>
</tbody>
</table>

Let me discuss the results without listing all the Q-mesures. As it is quite usual for social science data, there is an abundance of noise. 2.7444 out of 3.3683 bits or 81% of the total entropy is noise, the remainder, .6239 bits or 19% contributes to the structure. The strongest autocorrelation exists within aggression. $S_{EE'}(AA')=0.3986$ or 64% of the systematic entropy: someone exhibiting aggressive behavior is likely to continue to do so. Adding the next largest relation, $S_{AA'}(EE')$, exposure to violent programming is likely to continue in time, now accounts for $S(EE'/AA')=0.5909$ or 95% of the systematic entropy. At this point of the analysis viewing habits and aggressive behavior seem independent but the third largest component is between exposure to violent programming and subsequent aggressive behavior. With the
addition of $S_{EA}^P (EA')$, exposure to TV violence contributes to overt aggression, $S(EE'/EA'/AA')=.6051$ or 97% of the systematic entropy. Thus a theory, model or explanation of the data in terms of the three binary components:

prior exposure to TV violence .1929 current exposure to TV violence .0140
prior aggressive behavior .3986 current aggressive behavior .0079

can account for 97% of the variation other than noise. And, continuing the composition process, a theory, model or explanation of these data in terms of the two tertiary and one binary components:

prior exposure to TV violence .1999 current exposure to TV violence .0079
prior aggressive behavior .4128 current aggressive behavior .0089

can account for 98% of the systematic entropy. The increase in explanatory power being negligible, a model with components of lower ordinality might be preferable in this case.

REMAINING PROBLEMS

There is no doubt in my mind that the way the lattice of possible structures is constructed provides a reasonable if not the only logical basis for the identification of structures in multi-variate data and for the realization of a model's components. I am not quite so certain, however, about the quantitative aspect of my proposal. For the identification of structures, I would consider my new accounting equations (11) and (12) an important improvement over the earlier (5), but the procedures for composition and decomposition outlined here need further tests. Let me mention a few remaining problems briefly:

(a) Most important ranks the validity of the results. This refers specifically to the choice of the entropy measure as a meaningful quantity to be accounted for. Some further clear thinking and practical applications are required to uphold the thesis that the number of states a model with several components requires is indeed correlated with the S-measure. I believe I gave several arguments that transmissions do not indicate degrees of difficulty in model construction. Whether systematic entropies do what I hope they will needs to be seen.

(b) Although I have followed the tradition of accounting for some quantity that I believe to be empirically meaningful and I am convinced that some such quantity is essential in designing suitable algorithms, I wonder if the lattice of possible structures cannot be approached from a different way. Taking again Broekstra's example, one finds three conditional Q-measures to be zero. $Q_4(24)=0$ implies that any 3rd order dependency, if it exists, between 1, 2 and 4 is channeled only through 3 and is spurious otherwise. This reduces the number of structures to be considered tremendously, to 7 in fact. Similarly, $Q_{13}(24)=0$ and $Q_{34}(12)=0$ and because $Q(24)=.0038$ and $Q(12)=.0115$, the two relations are once again spurious. This eliminates all structures containing components 12, 24, 123, and 234 leaving the structure 134/23 as the lone survivor. The process is non-sequential and requires more work before it can be formulated into a working algorithm.

(c) There remains the puzzling problem of deciding when an arrangement of components overdetermines the data. I gave a rule for absolute Q-measures above and although the examples suggest that it also applies for the conditional Q-measures of (11), the rationale must be developed more clearly, this is particularly important when an algorithm is used for structure identification. Overdeterminism is as much of an error as structural losses.

(d) The comparability of the quantities of interaction across different ordinalities is another problem. Here quantities have different ranges (different maximum and minimum values) which are ignored under addition. I am therefore not entirely convinced whether this is justifiable.

(e) Related to the foregoing is the problem to decide what is much and what is little, how large losses may become to be no longer tolerable, etc. I suppose this is to a large extent a matter of the practical significance of these quantities and is related to what the resulting structure is intended to accomplish, how costly the errors are. Practical uses of composition or decomposition procedures result require that the risks of making wrong identifications be operationalized in terms of the quantities accounted for.

(f) Finally, there remains the problem of the statistical significance of the decisions any structure identification procedure implicitly makes. Without adequate statistical tests on hand we do not know whether we quibble over distinctions that are mere artifacts of inadequate sample sizes. And the fact that statistical tests are available for other quantities, for transmissions, for example, is no consolation when these measures assess something other than what is needed here.

With these (and more) problems still remaining one wonders how big a step forward this paper was able to take. I for myself am happy with small ones, as long as they lead to testable results.

REFERENCES

Ashby, W.R.

Ashby, W.R.

Ashby, W.R.

Broekstra, G.

Broekstra, G.
Broekstra, G.
1978 Personal communication.

Broekstra, G.


Klir, G.J.

Klir, G.J.

Krippendorff, K.

Krippendorff, K.

Lazarsfeld, P.F.

McGill, W.J.

Shannon, C.E. and Weaver, W.

Wiener, N.

Author: Klaus Krippendorff
The Annenberg School of Communications
University of Pennsylvania
Philadelphia, PA 19104