Journal of Statistical Physics, Vol. 9, No. 4, 1973

On the Kullback Information Measure as a Basis for Information Theory: Comments on a Proposal by Hobson and Chang

Myron Tribus[1] and Richard Rossi[1]

Received July 19, 1973

Hobson and Chang recommend that the Kullback information measure replace the Shannon information measure as a basis for information theory. They cite several items in support of their proposal. The items are considered individually and it is shown that they do not in fact constitute sufficient reasons for accepting the Hobson/Chang proposal. It is concluded that the Shannon information measure should be retained as the basis of information theory.

KEY WORDS: Information; information theory; Shannon; Kullback; communication theory; Bayesian statistics.

1. INTRODUCTION

Hobson and Chang[(1)] have proposed that, "The Kullback [information] measure can, but the Shannon [information] measure cannot, form the basis of a consistent, general (i.e., extending to continuous sample spaces and nonconstant prior distributions) theory of information". In support of this proposal they cite the following:

[1] Xerox Corporation, Rochester, New York.

1. The Shannon information measure I_S is a special case of the Kullback information measure I_K.
2. The expressions for I_K in the discrete and continuous cases arise from the same Lebesgue–Stiltjes integral. I_S does not have this property.
3. I_S is not invariant under certain transformations, while I_K is.
4. Jaynes' principle of minimum prejudice "makes sense" when applied to the Kullback measure, but does not (in general) when applied to the Shannon measure.
5. I_S is additive over all two-step processes, while I_K is additive only when the data are additive.

It is the purpose of this note to document our disagreement with the Hobson–Chang proposal and to take issue with each of the supporting items cited above.

2. NOTATION

The mathematical concept of *uncertainty* is considered to have specific meaning only in respect to a well-defined *question*. A question is thus considered to be "well defined" only if it is possible to define an exhaustive set of possible answers. In Boolean symbols, if Q represents the question, $+$ represents "or", and A_i represents the ith member of the set of possible answers, then a "well-defined" question implies the ability to define all the symbols on the right-hand side of the equation

$$Q = A_1 + A_2 + \cdots + A_n \tag{1}$$

When a statistician says, "Define the sample space" (i.e., enumerate the A_i) the terminology used here translates to, "Pose a well-defined question." *Uncertainty* about a question refers to the inability to say which A_i is true. There is, of course, a higher level of uncertainty which might be associated with deciding what the *question* is, but this kind of uncertainty does not at this time lend itself to straightforward analysis.

Knowledge pertaining to Q is said to be deterministic if possession of that knowledge permits a person to state that one of the A_i is true and the rest false. Knowledge which pertains to the set of answers but does not tell which one is true is said to be "uncertain."

There exists a unique code for transmitting *uncertain knowledge* which guarantees against the *deliberate* introduction of inconsistency, ambiguity, ad hoc procedures, and lack of candor.[2] According to the mathematical derivation which leads to this result, *probability is an encoding of knowledge.* It is appropriate to speak of a "state of knowledge" and denote the state by

X. The encoding is symbolized by $p(A_i \mid X)$, where p is the probability measure assigned to A_i to represent knowledge X. If we write as a short-hand notation $p_i^k = p(A_i \mid X^k)$, then the set $P^k = (p_1^k, p_2^k, \ldots, p_i^k, \ldots)$ encodes the knowledge X^k pertaining to $Q = (A_1 + A_2 + \cdots + A_i + \cdots)$.

3. THE SHANNON AND KULLBACK MEASURES FOR DISCRETE VARIABLES

The concept "state of knowledge," represented by X in the foregoing, is distinct from the concept "information in a message." The distinction is critical to any comparison between the Kullback and Shannon measures. Since the connection between the Shannon definition of entropy as a measure of uncertainty and the Clausius definition of entropy (in terms of heat and absolute temperature) has already been shown,[3-6] arguments based on thermodynamic reasoning may be employed here without additional justification. The distinction between a state function and a path function is useful. (Energy is a state function; work is a path function.) Shannon's entropy is a state function which measures the *uncertainty* of X about Q. It depends on the encoding P, i.e., for discrete A_i

$$U_S = U_S(Q \mid X) = U_S(P) = -k \sum_i p_i \ln p_i \tag{2}$$

When a given problem is analyzed it is important to decide whether Q has changed (a new problem or a new system) or X has changed (a new state of knowledge).

The Kullback measure of *information* is a *path* function which depends upon Q and *two* states of knowledge X and X^0. If the states of knowledge X and X^0 are encoded by P and P^0, the Kullback measure is

$$I_K(Q \mid X, X^0) = \sum_i p_i \ln(p_i/p_i^0) \tag{3}$$

This function is always positive if the p's satisfy $\sum_i p_i = 1, 0 \leqslant p \leqslant 1$ (Ref. 2, p. 100).

The Shannon measure of information is taken as the difference in uncertainties, i.e.,

$$I_S(Q \mid X, X^0) = U_S(Q \mid X^0) - U_S(Q \mid X) \tag{4}$$

I_S is not always positive (Ref. 2, p. 116). From the definitions given in (3) and (4) it is seen that Shannon's measure of information is "reversible" but Kullback's is not, i.e.,

$$I_K(Q \mid X, X^0) + I_K(Q \mid X^0, X) \neq 0 \tag{5}$$

$$I_S(Q \mid X, X^0) + I_S(Q \mid X^0, X) = 0 \tag{6}$$

The fact that I_S is not antisymmetric in X and X^0 gives rise to irreversibility in the "mechanical" sense (i.e., cannot go backward) but it remains to be seen if change in the "forward" direction is non-isentropic (i.e., loses information).

In a test of the two measures, therefore, it is important to avoid inadvertently going "backward." The idea of "forward" and "backward" is related to the idea of learning a sequence of "truths." ("Truth" is an elusive concept; the context in which it is used here will become clearer.) If we begin with a state of knowledge X^0 and acquire a new datum D^1, the new state of knowledge is represented by

$$X^1 = X^0D^1 \tag{7}$$

(writing two Boolean symbols together as though multiplied implies the word "and" between them). If we learn another datum D^2, the new state of knowledge is

$$X^2 = X^0D^1D^2 \tag{8}$$

The order in which the symbols is written is taken as having no temporal significance, i.e.,

$$X^2 = X^0D^1D^2 = X^0D^2D^1 = D^1X^0D^2, \quad \text{etc.}$$

If we wish to introduce time, we must do so explicitly (Ref. 2, p. 8). If we say X^2 is "true," we mean that *none* of X^0, D^1, or D^2 is false. In particular, this means X^0, D^1, and D^2 do not contradict one another. The introduction of a new datum D^k which contradicts a previous state of knowledge is called "going backward" and is a form of inconsistency, for if probabilities encode knowledge, the encoding can be consistent only if the knowledge is consistent.

If $D^2 =$ "D^1 is false," then we are led to interpret $X^0D^1D^2$ as implying a return to X^0. The word "true" is here used in the very limited sense; "postulated as true" and "not in conflict with other postulates."

Shannon's measure has a simple interpretation (Ref. 2, p. 111). It indicates an expectation for how much there is to learn in going from state X to a deterministic state (for which $U_S = 0$). Since U_S is a state function, the change in entropy between any two states of knowledge about a given question is independent of the path.

Kullback's measure also has a simple interpretation (Ref. 2, p. 109). It indicates an "expectation" for what will be learned if one believes X and considers going to some other state X^0 (which one does not really believe).

4. THE KULLBACK AND SHANNON MEASURES IN JAYNES' FORMALISM

In the paper by Hobson and Chang it is suggested that the Kullback measure (plus an irrelevant constant) be used to replace Shannon's measure in Jaynes' "principle of minimum prejudice."[7] This principle is used to encode various kinds of knowledge:

> *JAYNES' PRINCIPLE OF MINIMUM PREJUDICE. Assign the set of probabilities which maximizes the entropy* U_S,
>
> $$U_S = -k \sum_i p_i \ln p_i$$
>
> *subject to what is known.*

The proposal by Hobson and Chang is to substitute $-I_k$ for U_S (in Ref. 1 a new function $U_k = I_k{}^m - I_k$ is proposed, but since $I_k{}^m$ does not depend on the set P, the constant is irrelevant mathematically).

In testing $-I_k$ as a criterion for choosing P, it is important not to use initial states of knowledge for which the nonzero probabilities are uniform. For, if $P_i{}^0 = 1/n$ for $i = 1, 2,..., n$, the Kullback information measure and the Shannon information measure differ only by $k \ln n$.

5. USE OF THE KULLBACK MEASURE IN JAYNES' FORMALISM

Consider that knowledge of state X^1 has been encoded in probability distribution P^1 $[=(p_i{}^1, p_2{}^1,..., p_i{}^1,...)]$. Let D^2 represent added information of the form

$$D^2 = \text{“}\sum_i p_i g_i = \langle g\rangle\text{”}$$

Let $X^2 = X^1D^2$. Using the Kullback measure, the problem of assigning P^2 becomes:

$$\text{maximize} \qquad -I_K = -k \sum_i p_i{}^2 \ln p_i{}^2 + k \sum_i p_i{}^2 \ln p_i{}^1$$

$$\text{subject to:} \qquad \sum_i p_i{}^2 = 1, \qquad \sum_i p_i{}^2 g_i = \langle g\rangle$$

The solution is

$$p_i{}^2 = p_i{}^1 C e^{-\lambda g_i}$$

where $C^{-1} = \sum_i p_i{}^1 e^{-\lambda g_i}$ and λ satisfies

$$\langle g\rangle = \left(\sum g_i p_i{}^1 e^{-\lambda g_i}\right) \Big/ \sum p_i{}^1 e^{-\lambda g_i}$$

The new probability distribution is a product of the prior distribution with a new distribution.

6. A SPECIFIC EXAMPLE

Suppose $Q = A_1 + A_2 + \cdots + A_6$,

A_i = "the value is i"

X^0 = "the labels on the A_i tell us nothing about the truth of A_i"

D^1 = "the mean value of $i = 4$"

D^2 = "the mean value of $i^2 = 17$"

From the Jaynes' principle, using Shannon's information measure, we find the following results (see Table I). When the Kullback information measure is used the same results occur for states of knowledge X^0, X^1, and X^2 because X^0 gives a "uniform prior". In these cases the Kullback and Shannon information measures differ only by a constant, so they yield the same probabilities. On the other hand, if we define $X^{3'}$ to be the same as X^3 but compute P^3 first by the path $X^0 \rightarrow X^1 \rightarrow X^3$ and second by the path $X^0 \rightarrow X^2 \rightarrow X^3$ using the Kullback measure, we arrive at two different answers. It is found that the resulting probability distribution satisfies the last constraint imposed but does not preserve previous constaints. Thus there is a loss of information.

Thus we find that the fourth and fifth of Hobson and Chang's supporting items lead to unsatisfactory results. The other objections are also of interest.

Item 1: The Shannon information measure I_S is a special case of the Kullback information measure I_K. As shown by Hobson and Chang, the change in information associated with acquiring a new datum is numerically the same whether one measures information by I_K or by I_S, provided the prior distribution is uniform. But this misses the point that the Shannon information is a state variable and the Kullback information is a path variable, to borrow terms from thermodynamics. That is, for a specified

Table I

State of knowledge	Probability distribution	$\langle i \rangle$	$\langle i^2 \rangle$
X^0	$p_i = 1/6$	3.50	15.17
$X^1 = X^0D^1$	$p_i = 0.0866e^{0.1746i}$	4.00	18.76
$X^2 = X^0D^2$	$p_i = 0.1376 \exp(0.0119i^2)$	3.75	17.00
$X^3 = X^0D^1D^2$	Shannon:		
	$p_i = 0.0001812 \exp(3.82i - 0.476i^2)$	4.00	17.00
	Kullback:		
	via $X^0 \rightarrow X^1 \rightarrow X^3$, $p_i = 0.1062 \exp(0.1746i - 0.01142i^2)$	3.77	17.00
	via $X^0 \rightarrow X^2 \rightarrow X^3$, $p_i = 0.0983 \exp(0.0869i + 0.01191i^2)$	4.00	18.83

change in state of knowledge the Shannon information is uniquely defined while the Kullback information depends on how the change takes place. If the extent of one's knowledge about the answer to a question depends only on the available data without regard to the temporal order in which the data were gathered, to express information in a way which depends on that order is inconsistent.

Item 2: I_S does not have a representation as a Lebesgue–Stiltjes integral which reduces to the usual expressions in both the discrete and continuous cases. Hobson and Chang are concerned because I_S (like the probability mass functions) behaves badly when passing smoothly between discrete and continuous spaces. As it happens, I_S can be represented in measure-theoretic terms in a form which does reduce to the usual expressions in the discrete and continuous cases. (See Wilks,[8] Section 12.1 for the construction of such an expression for a function essentially identical to the Shannon information.) The fact that the representation is not a Lebesgue–Stiltjes integral seems unimportant.

Item 3: I_K is and I_S is not invariant under a change of continuous variables. I_S can be made measure invariant by carefully going from the discrete to the continuous variable. The result is to introduce a singularity which is usually suppressed but should be retained when a change of variable is desired. There seems to be little use for the property.

Hobson and Chang attach significance to this phenomenon since there is an intuitive appeal to the idea that the information should be unchanged by a mere mathematical transformation. This misses the point that in performing the transformation one at the same time transforms the question.

To see this, consider the following experiment. Let there be two boxes—one red, one blue. Let it be known that the red box contains a collection of spheres all of whose diameters are less than 1 in. Given the box is very large, the appropriate prior probability on the diameter of a sphere drawn at random from the red box is uniformly distributed on zero to one.

Now suppose it is known of the blue box that it contains a collection of spheres, all of whose surface areas are less than pi inches. The prior probability assignable to the surface area of a sphere from the blue box is uniformly distributed on zero to pi.

The given information is the same in each case, yet by the way the question is asked the priors are incompatible. As a matter of fact, there is no "right" answer to the paradox.

The paradox arises in practical situations. See Ref. 2, p. 149 for an example from chemistry and Ref. 2, Chapter 10, where the problem arises in the relation between hazard rate and MTBF in reliability theory. The fact that I_K is constant under changes of variables is relevant only to the extent that the question in some sense remains invariant. In general, questions do not remain invariant under transformations, so why should the information?

The answer lies in a simple interpretation of I_S as the expected amount by which you will be "surprised" by the outcome of a discrete experiment. The extent of your surprise is measured by the negative log of the probability you had assigned to what actually happened. The Shannon information measure thus has an appealing intuitive basis. Since the information at hand about a question may indeed be a function of the manner in which the question is asked, the inconsistent priors merely reflect the inconsistency of our own thought processes.

REFERENCES

1. A. Hobson and B. Chang, *J. Stat. Phys.* **7**(4):301 (1973).
2. M. Tribus, *Rational Descriptions, Decisions and Designs*, Pergamon, New York (1969).
3. M. Tribus, *J. Appl. Mech.*, **1961** (March):1–8.
4. M. Tribus and R. Evans, *App. Mech. Rev.* **16**(10):765–769 (1963).
5. M. Tribus, P. Shannon, and R. Evans, *AIChE J.* **1966** (March):244–248.
6. M. Tribus, *Am. Scientist* **54**(2) (1966).
7. E. T. Jaynes, *Phys. Rev.* **106**:620–630 (1957); **108**:171–190 (1957).
8. S. Wilks, *Mathematical Statistics*, Wiley, New York (1962).