Journal of Statistical Physics, Vol. 7, No. 4, 1973

A Comparison of the Shannon and Kullback Information Measures

Arthur Hobson[1] and Bin-Kang Cheng[1]

Received October 13, 1972

Two widely used information measures are compared. It is shown that the Kullback measure, unlike the Shannon measure, provides the basis for a consistent theory of information which extends to continuous sample spaces and to nonconstant prior distributions. It is shown that the Kullback measure is a generalization of the Shannon measure, and that the Kullback measure has more reasonable additivity properties than does the Shannon measure. The results lend support to Jaynes's entropy maximization procedure.

KEY WORDS: Information theory; entropy; Shannon; Kullback; Jaynes.

1. INTRODUCTION

Information theory was founded in 1948 by Shannon[1] and Wiener,[2] who introduced the expression

$$U_s[P] = -\sum p_i \ln p_i \tag{1}$$

as a measure of the missing information, or uncertainty, in a probability distribution $P \equiv (p_1, p_2, ..., p_n)$. Expression (1) is widely used in communication theory[3] and is generally used to represent the physical entropy in statistical mechanics.[4] In 1951 Kullback[5] introduced

$$I_k[P : P^0] = \sum p_i \ln(p_i/p_i^0) \tag{2}$$

[1] Department of Physics, University of Arkansas, Fayetteville, Arkansas.

as a measure of the "information for discrimination" between two distributions P and P^0. Expression (2) is widely used in statistics[6] and has been used in statistical mechanics.[7]

The purpose of this paper is to compare the Shannon and Kullback information measures; we will find that the Kullback expression represents a generalization of the Shannon expression, and that the Shannon expression has several drawbacks not shared by the Kullback expression. The results will lend support to Jaynes's maximum entropy principle,[8–12] according to which the probability distribution representing any given data must maximize the uncertainty subject to the data.

In Sections 2 and 3 we will review the basic properties of the two information measures and show the precise relation between them. In Section 4 we will discuss the additivity properties of the two measures when information is given in two successive steps, and we will use Jaynes's principle to show that the Kullback measure (in contrast to the Shannon measure) is additive in precisely those cases for which the *data* are additive. We will discuss the results in Section 5.

2. BASIC PROPERTIES OF THE TWO MEASURES

Shannon's expression (1) represents the *uncertainty* about which outcome $(1, 2,..., n)$ will occur in one trial of a "random experiment" (i.e., one in which the outcome is not predictable) when predictions are based on the probability distribution P. In fact, Shannon[1] proved a uniqueness theorem according to which (1) is the *only* expression having the following intuitively reasonable properties (see also the elegant proofs of Feinstein and Khinchin[13]):

S-1. U_s is a continuous function of the p_i.

S-2. When $P = (n^{-1}, n^{-1},..., n^{-1})$, U_s is a monotonic increasing function of the integer n.

S-3. U_s is additive under decomposition of the sample space. By this we mean that, if the set (or "sample space") of possible outcomes $(1, 2,..., n)$ is divided into r groups, with probability q_k associated with the kth group, then the overall uncertainty $U_s[P]$ should be the *sum* of the uncertainty $U_s[q_1,..., q_r]$ about which group occurred, plus the weighted sum (with weighting factors q_k) of the uncertainties as to which outcome occurred within each group.

Kullback's expression $I_k[P : P^0]$ represents the *information gained* concerning the outcome of a random experiment when the probability distribution is changed from P^0 to P. In fact, it has been proven[10,14] that,

except for a positive multiplicative constant, (2) is the *only* expression having the following intuitively reasonable properties:

K-1. I_k is a continuous function of the p_i and p_i^0.

K-2. I_k does not depend on the manner in which the outcomes (1, 2,..., n) are labeled.

K-3. $I_k = 0$ when $P = P^0$.

K-4. When $P^0 = (n_0^{-1}, n_0^{-1},..., n_0^{-1})$ and $P = (n^{-1},..., n^{-1}, 0,..., 0)$ $(n \leqslant n_0)$ then I_k is an increasing function of the integer n_0 and a decreasing function of the integer n.

K-5. I_k is additive under decomposition of the sample space (see property S-3).

These five properties are fully as natural and intuitive (when applied to the concept of information gain) as S-1 through S-3 (when applied to the concept of uncertainty), but they are somewhat more complicated since the information *gain* must depend on *two* probability distributions. Postulates K-1, K-4, and K-5 are the precise analogs of S-1, S-2, and S-3; they are intuitively reasonable for the same reasons that the corresponding Shannon postulates are reasonable. The "extra" postulates K-2 and K-3 were needed in the uniqueness proof given in Ref. 14; perhaps a proof could be devised without requiring K-2 and K-3. At any rate, K-2 and K-3 are simple and reasonable properties; any expression which did *not* satisfy K-2 and K-3 could surely not be interpreted as information gain.

Difficulties arise when the Shannon uncertainty is extended to continuous sample spaces[9–11]: Equation (1) cannot be generalized to continuous sample spaces without arbitrarily "renormalizing" the uncertainty (i.e., throwing out an infinite contribution); furthermore the usual expression for U_s in the continuous case,

$$U_s[P] = -\int \rho(x) \ln \rho(x)\, dx \tag{3}$$

(obtained after renormalization) is not invariant under a change of variables, and is obviously incorrect dimensionally whenever x has dimensions. The Kullback information measure does not suffer from these difficulties: For continuous sample spaces, (2) may be generalized (without renormalization) to[2]

$$I_k[P : P^0] = \int \rho(x) \ln[\rho(x)/\rho^0(x)]\, dx \tag{4}$$

Equation (4) is invariant under a change of variables, and is dimensionally correct.

[2] More precisely, there exists a single Lebesgue–Stieltjes integral which reduces to (4) in the continuous case and (2) in the discrete case. See Ref. 6.

Jaynes[9] has obtained (4) from (1) [rather than from (2)] by introducing $\rho_0(x)$ as a "measure function" in making the transition from the discrete to the continuous case. Jaynes's procedure circumvents the invariance and dimensional difficulties associated with (3), but still contains the divergence difficulty noted above.

The difficulties exhibited by (1) represent inconsistencies in the theory of information; these inconsistencies will apparently be present in any theory which extends to continuous sample spaces and which is based on (1). Such inconsistencies should be tolerated only if no consistent theory is available. Fortunately, a consistent theory *is* available, starting from the Kullback information rather than from the Shannon uncertainty.

3. RELATION BETWEEN THE TWO MEASURES

If we take U_s to be the fundamental expression for missing information, then the information gained when the distribution changes from P^0 to P must be

$$I_s[P : P^0] = U_s[P^0] - U_s[P] = \sum p_i \ln p_i - \sum p_i^0 \ln p_i^0 \tag{5}$$

since this is the decrease in the missing information. If, on the other hand, we take the viewpoint that I_k is the correct expression for information gain, then the missing information in the distribution P must be

$$\begin{aligned} U_k[P : P^0, P^m] &= I_k[P^m : P^0] - I_k[P : P^0] \\ &= \sum p_i^m \ln(p_i^m/p_i^0) - \sum p_i \ln(p_i/p_i^0) \end{aligned} \tag{6}$$

where P^m is that distribution representing the maximum information consistent with the fundamental physical constraints of the random experiment.[3] The Kullback uncertainty (6) is the difference between the maximum obtainable information $I_k[P^m : P^0]$ and the actual information $I_k[P : P^0]$; hence, (6) gives the amount of information which is still missing, relative to P^0 and P^m.

For discrete sample spaces the precise relation between U_s and U_k is as follows: Let P^0 and P^m be given by $p_i^0 = n^{-1}$ and $p_i^m = \delta_{ij}$ ($i = 1,..., n$; j fixed); then (6) becomes

$$U_k = -\sum p_i \ln p_i = U_s \tag{7}$$

Thus U_k reduces to U_s in the special case of a constant prior distribution and a "zero or one" distribution of maximum information.

[3] For discrete sample spaces, P^m will usually be a "zero or one" function: $p_i^m = \delta_{ij}$ (j fixed).

For continuous sample spaces $a \leqslant x \leqslant b$, the Kullback uncertainty is given by the generalization of (6):

$$U_k[P : P^0, P^m] = \int \rho^m(x) \ln[\rho^m(x)/\rho^0(x)]\, dx - \int \rho(x) \ln[\rho(x)/\rho^0(x)]\, dx \quad (8)$$

The relation between U_s and U_k is as follows: Assume $\rho^0(x) = (b-a)^{-1}$ $(a \leqslant x \leqslant b)$ and $\rho^m(x) = L^{-1}\theta_R(x)$, where $\theta_R(x)$ is the characteristic function (zero-or-one step-function) of some Lebesgue-measurable region R having measure L; for this case (8) becomes

$$U_k = -\int \rho(x) \ln[L\rho(x)]\, dx \quad (9)$$

which is the same as (3) except for the factor L; this factor removes the dimensional and transformational difficulties associated with (3). Thus (except for the factor L) U_k reduces to U_s in the special case of a constant prior distribution and a "step-function" distribution of maximum information. Note that if $L \to 0$, then $U_k \to \infty$; i.e., it is "infinitely difficult" to pick a precise point out of a continuum, which seems reasonable.

We have already seen, in Section 2, that there is some reason for regarding the Kullback expression as more general and fundamental than the Shannon expression. If we accept this notion, then the above results show that *the Shannon uncertainty is a special case, valid only for constant prior distributions, of the Kullback uncertainty.*

4. ADDITIVITY OVER TWO-STEP PROCESSES

Suppose that the distribution is altered from P^0 to P^1, and then from P^1 to P^2. The Shannon information (5) obviously satisfies

$$I_s[P^2 : P^0] = I_s[P^2 : P^1] + I_s[P^1 : P^0] \quad (10)$$

for any P^0, P^1, P^2; i.e., I_s is "additive over every two-step process." The Kullback information, on the other hand, is *not* generally additive over two-step processes. This is easy to verify by choosing $P^2 = P^0$ and noting that $I_k[P^0 : P^0] = 0$ (see property K-3), whereas $I_k[P : P^0] \geqslant 0$ whenever $P \neq P^0$.[5]

But do we intuitively *want* the information to be additive over every two-step process? Consider, for example, the above-mentioned case in which $P^2 = P^0$, so that the two-step process is $P^0 \to P^1 \to P^0$, and the information in the first step is completely canceled out by the information in the second step. In this case we do learn something in each step, since the distribution is changed at each step, but the overall effect of the two-step process is to put

us back where we started, so that the overall information is zero. We do not expect the information to be additive in this case.

To further illustrate this point, consider another example. A die is thrown a single time. Consider the following four probability distributions over the six possible outcomes:

$$P^0 = (\tfrac{1}{6},\ldots,\tfrac{1}{6}), \qquad P^1 = (0, \tfrac{1}{3}, 0, \tfrac{1}{3}, 0, \tfrac{1}{3})$$

$$P^2 = (0, \tfrac{1}{5}, \tfrac{1}{5}, \tfrac{1}{5}, \tfrac{1}{5}, \tfrac{1}{5}), \qquad P^3 = (0, 0, 0, \tfrac{1}{2}, 0, \tfrac{1}{2})$$

These distributions correspond (via Laplace's principle,[15] or via Jaynes's principle—see below) to the following four data:

D^0: Any one of the six possible outcomes can occur.

D^1: The outcome is even.

D^2: The outcome is two or greater.

D^3: The outcome is four or six.

We do not expect the information to be additive in the process $P^0 \to P^1 \to P^2$, since the datum D^2 detracts from (rather than supplementing) the datum D^1 and hence there is less information in the overall result $P^0 \to P^2$ than there is in the individual steps $P^0 \to P^1$ and $P^1 \to P^2$. Numerically,

$$I_k[P^2 : P^0] = \ln(6/5)$$

while

$$I_k[P^2 : P^1] + I_k[P^1 : P^0] = \infty + \ln 2 = \infty$$

(note that there is an infinite amount of information in any statement which assigns a positive probability to an event whose prior probability was zero; this corresponds intuitively to the radical alteration in our "state of knowledge"). We do, on the other hand, expect the information to be additive over the process $P^0 \to P^1 \to P^3$, since the datum D^3 actually supplements D^1; more precisely, D^3 may be expressed in the form "D^1 is true and, furthermore, the outcome 'two' cannot occur." Numerically,

$$I_k[P^3 : P^0] = \ln 3, \qquad I_k[P^3 : P^1] + I_k[P^1 : P^0] = \ln(3/2) + \ln 2 = \ln 3$$

As expected intuitively, I_k is additive over the process $P^0 \to P^1 \to P^3$ in which the *data* are additive, but I_k is *not* additive over $P^0 \to P^1 \to P^2$. The Shannon information, on the other hand, is additive over both processes.

Thus, the additivity of the Shannon information over arbitrary processes is actually an undesirable property; the Kullback information, on the other hand, appears to be additive over precisely those processes for which the *data* are additive.

We will now generalize the above argument to arbitrary sample spaces and arbitrary two-step processes. We will prove that the Kullback information is additive in precisely those situations for which the *data* are additive. In order to show this, it is necessary to invoke the following basic relation between experimental data and probability distributions.[8–12]

Jaynes's maximum entropy principle: If data D are given concerning the outcome of a random experiment, then predictions about the outcome should be based on that distribution P which maximizes U_k subject to the restrictions imposed by D.

Comment. This is a generalization of Jaynes's original formulation, which stated that P should maximize U_s rather than U_k. Jaynes's original formulation is applicable if the prior distribution is constant, in which case maximization of U_s is equivalent to maximization of U_k. In case the prior distribution is *not* constant, then it is U_k rather than U_s which should be maximized. This may be demonstrated by means of an example: Let the sample space be $(0 \leqslant x \leqslant 1)$, let the prior distribution be $\rho^0(x) = 3x^2$, and let the data be $\langle x \rangle = 1/4$, where $\langle \cdots \rangle$ denotes an expectation value. Maximizing U_k[Eq. (8)] subject to the data, we get $\rho_k(x) = Ax^2e^{-\alpha x}$, where α and A are determined from $\langle x \rangle = 1/4$ and $\langle 1 \rangle = 1$ (normalization). Maximizing U_s [Eq. (3)] subject to the data, we get $\rho_s = Be^{-\beta x}$. The distribution ρ_k seems more reasonable than ρ_s, since ρ_k retains the effect of the original "weighting" $3x^2$ of the sample space, whereas ρ_s does not.

In case the reader is not convinced, the following transformation argument proves conclusively that maximization of U_s is inconsistent with the principles of probabilistic reasoning: Change variables to $y = x^3$. The transformed prior distribution is $\bar{\rho}^0(y) = 1$ $(0 \leqslant y \leqslant 1)$, the transformed data is $\langle y^{1/3} \rangle = 1/4$, and Jaynes's principle (applied to either U_s or U_k, since the two are equivalent when the prior distribution is constant) yields $\bar{\rho}_k(y) = \bar{\rho}_s(y) = C \exp(-\gamma y^{1/3})$. We now note that $\bar{\rho}_k(y)$ is just the transformed form (under $x \to y = x^3$) of $\rho_k(x) = Ax^2e^{-\alpha x}$, whereas $\bar{\rho}_s(y)$ is *not* the transformed form of $\rho_s(x)$. That is, if we choose to maximize U_s rather than U_k, then our maximization procedure does not have the proper transformation properties. Another way of stating this is to note that the following diagram is commutative if our maximization procedure means "maximize U_k" but not if our maximization procedure means "maximize U_s":

$$\begin{array}{ccc} \text{data in terms of } x & \xrightarrow{\text{maximization}} & \rho(x) \\ \Big\downarrow\, y = f(x) & & \Big\downarrow\, y = f(x) \\ \text{Data in terms of } y & \xrightarrow{\text{maximization}} & \bar{\rho}(y) \end{array}$$

The above transformation argument (which is not difficult to generalize to arbitrary data and arbitrary transformations) constitutes yet another reason for preferring the Kullback measure over the Shannon measure.

Additivity Theorem.[4] Let the distribution P^1 correspond (via Jaynes's principle) to the datum

$$\langle f_1(x) \rangle = F_1 \tag{11}$$

and let P^2 correspond to (11) supplemented by the new datum

$$\langle f_2(x) \rangle = F_2 \tag{12}$$

Then, for arbitrary P^0, the Kullback information is additive over the two-step process $P^0 \to P^1 \to P^2$:

$$I_k[P^2 : P^0] = I_k[P^2 : P^1] + I_k[P^1 : P^0] \tag{13}$$

Comment. It may seem that we are unduly restricting the significance of the theorem by assuming that the data are expressible as expectation values F_i of known functions $f_i(x)$. But to the authors' knowledge, the expectation value form is the only form in which data can be expressed in terms of a probability distribution. Note that data of the form "x is in the region R" can be expressed in the expectation value form: simply take $f(x) = \theta_R(x)$ and $F = 1$.

Proof of Theorem. The distribution P^1 must maximize (8) subject to (11). Since the first term on the right side of (8) is not varied, P^1 must minimize $\int \rho \ln(\rho/\rho^0)\, dx$ subject to (11) and fixed $\rho^0(x)$. The minimizing distribution is

$$\rho^1(x) = \rho^0(x)\, e^{-\alpha f_1(x)}/Z_1(\alpha) \tag{14}$$

where

$$Z_1(\alpha) = \int \rho^0(x)\, e^{-\alpha f_1(x)}\, dx$$

with α chosen to satisfy (11). The distribution P^2 minimizes $\int \rho \ln(\rho/\rho^0)\, dx$ subject to (11) and (12). The minimizing distribution is

$$\rho^2(x) = \rho^0(x)\, e^{-\beta f_1(x) - \gamma f_2(x)}/Z_2(\beta, \gamma) \tag{15}$$

where

$$Z_2(\beta, \gamma) = \int \rho^0(x)\, e^{-\beta f_1(x) - \gamma f_2(x)}\, dx$$

[4] The theorem is stated and proved for continuous sample spaces; it also holds for discrete spaces.

and where β and γ are chosen to satisfy (11) and (12). Using (4), (14), and (15), the right member of (13) becomes

$$\int \rho^2(-\beta f_1 - \gamma f_2 - \ln Z_2 + \alpha f_1 + \ln Z_1)\, dx + \int \rho^1(-\alpha f_1 - \ln Z_1)\, dx \quad (16)$$

while the left member of (13) becomes

$$\int \rho^2(-\beta f_1 - \gamma f_2 - \ln Z_2)\, dx \quad (17)$$

Using the relations

$$\int \rho^1\, dx = \int \rho^2\, dx = 1, \qquad \int f_1\rho^1\, dx = \int f_1\rho^2\, dx = F_1\,, \qquad \int f_2\rho^2\, dx = F_2$$

terms (16) and (17) both reduce to $-\beta F_1 - \gamma F_2 - \ln Z_2$.

5. DISCUSSION

We have compared the Kullback and Shannon information expressions and found that the two are equivalent whenever the prior distribution P^0 is constant and the distribution of maximum information P^m is δ_{ij} (discrete case) or $L^{-1}\theta_R(x)$ (continuous case), and we have shown that the Kullback expression is preferable whenever these conditions do *not* hold. Briefly, the reasons for preferring the Kullback measure are as follows.

1. I_k does not exhibit the divergence, transformational, and dimensional difficulties exhibited by I_s.
2. The uniqueness theorem for I_k implies that I_k is the *only* intuitively reasonable measure of information gain.
3. The maximization procedure for finding the distribution corresponding to given data makes sense when applied to the Kullback measure but not (in general) when applied to the Shannon measure.
4. I_k is additive over precisely those two-step processes for which the *data* are additive. On the other hand, I_s is additive even when the data are not additive.

It appears from these results that the Kullback measure can, but the Shannon measure cannot, form the basis of a consistent, general (i.e., extending to continuous sample spaces and nonconstant prior distributions) theory of information.

We have also provided further support for Jaynes's maximum entropy principle, since additive data lead to the expected additivity of the Kullback information only when the distributions are chosen in accordance with Jaynes's principle.

ACKNOWLEDGMENT

One of us (A. H.) would like to thank Professor Amnon Katz for pointing out that the Kullback information is not additive over arbitrary two-step processes; this remark provided the initial stimulus for this paper. We would also like to thank the National Science Foundation for providing research support.

REFERENCES

1. C. E. Shannon, *Bell System Tech. J.* **27**:379, 623 (1948); reprinted in C. E. Shannon and W. Weaver, *The Mathematical Theory of Communication*, Univ. of Illinois Press, Urbana, Ill. (1949).
2. N. Wiener, *Cybernetics*, Wiley, New York (1948).
3. F. M. Reza, *An Introduction to Information Theory*, McGraw-Hill, New York (1961), and references cited therein.
4. R. C. Tolman, *The Principles of Statistical Mechanics*, Oxford Univ. Press, London (1938).
5. S. Kullback, *Annals of Math. Statistics* **22**:79 (1951).
6. S. Kullback, *Information Theory and Statistics*, Wiley, New York (1951).
7. F. Schlögl, *Z. Physik* **249**:1 (1971), and references cited therein.
8. E. T. Jaynes, *Phys. Rev.* **106**:620 (1957); **108**:171 (1957).
9. E. T. Jaynes, in *Statistical Physics* (1962 Brandeis Lectures), ed. by K. W. Ford, Benjamin, New York (1963).
10. A. Hobson, *Concepts in Statistical Mechanics*, Gordon and Breach, New York (1971).
11. A. Katz, *Principles of Statistical Mechanics*, Freeman, San Francisco (1967).
12. R. Baierlein, *Atoms and Information Theory*, Freeman, San Francisco (1971).
13. A. Feinstein, *Foundations of Information Theory*, McGraw-Hill, New York (1968); A. I. Khinchin, *Mathematical Foundations of Information Theory*, Dover, New York (1957).
14. A. Hobson, *J. Stat. Phys.* **1**:383 (1969).
15. Pierre Simon de Laplace, *A Philosophical Essay on Probabilities*, Dover, New York (1951).