ON CHARACTERIZATION OF THE KULLBACK-LEIBLER MEAN INFORMATION FOR CONTINUOUS PROBABILITY DISTRIBUTIONS

BY SADAO IKEDA

(Received Jan. 26, 1962)

1. Introduction

Let (R, S, m) be a σ-finite measure space, and X and X' be two probability distributions defined over R, which are absolutely continuous with respect to m. Let $f(x)$ and $g(x)$ denote generalized probability density functions (gpdf.) of X and X' respectively. Further let $D(X)$ denote the carrier of a gpdf., $f(x)$, of X (The gpdf. of X is not always uniquely determined).

As is well-known, the Kullback-Leibler mean information for discrimination is given by

$$I(X: X')=\int_R f(x) \log \frac{f(x)}{g(x)} dm, \tag{1}$$

and we have $I(X: X')\geqq 0$, where equality holds if and only if $f(x)=g(x)$ (a.e.m) on R. If $m(D(X)-D(X'))>0$, then $I(X: X')=\infty$ [1], [2], [3]. Various properties of this information measure have been listed in [1], and some additional properties of convergence will be seen in [4].

In order to clearify the role of this information measure in various statistical applications, it would be desirable to characterize it by means of some concepts familiar to the theory of statistical inference. As the first step to work out this consideration, we shall consider, in the present paper, the characterization problem in a certain conventional manner.

Now, we introduce some notations necessary to our present discussions (The reader may refer to [5]). Let (R, S, m) be a σ-finite measure space belonging to a set $\mathscr{T}$, of σ-finite measure spaces consisting of all product spaces of a certain fixed σ-finite measure space (R_0, S_0, m_0). Denote by $V(R, S, m)$ the family of all probability distributions which are absolutely continuous with respect to m, and the totality of such distributions for all (R, S, m) belonging to $\mathscr{T}$ will be denoted by $V(\mathscr{T})$. We shall assume that, for any pair of probability distributions (X, Y) both belonging to $V(\mathscr{T})$, the conditional distribution of Y always exists under the condition that X is given.

Our problem of characterization may be stated in the following way; For any pair (X, X') of probability distributions both belonging to any

$V(R, S, m)$, for which we shall always assume that $D(X)\subseteqq D(X')$ up to a set of measure(m) zero, consider a non-negative, extended real-valued function $I(X: X')$, depending only on their gpdf.'s. We shall seek for the conditions, among those which are satisfied by the right-hand member of (1), which give the function $I(X: X')$ the expression (1).

Put, for a gpdf. $g(x)$ of X',

$$\mu(E)=\int_E g(x)dm, \qquad \text{for all } E \text{ in } S,$$

and let $f(x)$ be a gpdf. of X, $\lambda(x)=f(x)/g(x)$. Then in our present situation that $D(X)\subseteqq D(X')$ (m), $\lambda(x)$ determines a gpdf. of a certain probability distribution defined on (R, S, μ). In such a situation, the Kullback-Leibler mean information (1) has the expression

$$I(X: X')=\int_R \lambda(x)\log\lambda(x)d\mu. \tag{2}$$

We shall assume, throughout the present paper, that it holds that

$$M_\mu(S)=\{\mu(E)\,;\, E\in S\}=[0, 1] \tag{3}$$

for the second member X' of any pair of two probability distributions, (X, X'). It has been shown that some conditions characterize the Shannon-Wiener information measure for continuous probability distributions defined on the finite measure space (R, S, μ) [5], which is helpful to our present characterization procedure as will be seen in the following section.

2. Characterization

As before, we consider an extended real-valued function, $I(X: Y)$, defined for any probability distributions X and Y belonging to any member, $V(R, S, m)$, such that $(R, S, m)\in\mathscr{T}$. Here, it is assumed that the restriction (3) is always imposed on the distribution Y. First, some assumptions will be set on $I(X: Y)$.

Assumption I. Let X_1 and Y_1 be in $V(R_1, S_1, m_1)$ with gpdf.'s $f_1(x)$ and $g_1(x)$, respectively, and let X_2 and Y_2 be in $V(R_2, S_2, m_2)$ with gpdf.'s $f_2(x)$ and $g_2(x)$. Put $d\mu_1=g_1dm_1$ and $d\mu_2=g_2dm_2$. For the probability ratios, $\lambda_1(x_1)=f_1(x_1)/g_1(x_1)$ and $\lambda_2(x_2)=f_2(x_2)/g_2(x_2)$, and for E_1 in S_1 and E_2 in S_2 with $\mu_1(E_1)=v_1$, and $\mu_2(E_2)=v_2$, suppose that

$$\lambda_1(x_1)=\begin{cases}1/v_1, & \text{if } x_1\in E_1,\\ 0, & \text{otherwise,}\end{cases}$$

and

$$\lambda_2(x_2)=\begin{cases}1/v_2, & \text{if } x_2\in E_2,\\ 0, & \text{otherwise,}\end{cases}$$

almost everywhere with respect to m_1 and m_2, respectively. Then, the condition $v_1>$, or $=v_2$ implies respectively

$$I(X_1: Y_1)<, \quad \text{or } =I(X_2: Y_2).$$

Assumption II. Let (R_1, S_1, m_1) and (R_2, S_2, m_2) be any members of $\mathscr{T}$, and let X_1 and Y_1 be in $V(R_1, S_1, m_1)$, and let X_2 and Y_2 be in $V(R_2, S_2, m_2)$. Denote by $X=(X_1, X_2)$ and $Y=(Y_1, Y_2)$ the joint distributions. They are both the members of $V(R, S, m)=V(R_1\times R_2, S_1\times S_2, m_1\times m_2)$. Furthermore, let $f(x)=f_1(x_1)f_2(x_2|x_1)$ and $g(x)=g_1(x_1)g_2(x_2|x_1)$ be the gpdf.'s of X and Y, respectively, where $x=(x_1, x_2)$, $f_1(x_1)$ and $g_1(x_1)$ are the gpdf.'s of X_1 and Y_1, and $f_2(x_2|x_1)$ and $g_2(x_2|x_1)$ denote the gpdf.'s of the conditional probability distributions $X_2(x_1)$ and $Y_2(x_1)$ given $X_1=x_1$ and $Y_1=x_1$. Then, it holds that

$$I(X: Y)=I(X_1: Y_1)+I_{X_1}(X_2: Y_2), \tag{4}$$

where

$$I_{X_1}(X_2: Y_2)=\int_{R_1} I(X_2(x_1): Y_2(x_1))f_1(x_1)dm_1.$$

Assumption III. Let X, $\{X_i\}$, $(i=1, 2, \cdots)$ and Y be the members of $V(R, S, m)$, with gpdf.'s $f(x)$, $\{f_i(x)\}$, $(i=1, 2, \cdots)$ and $g(x)$, respectively. Put $\lambda(x)=f(x)/g(x)$, $\lambda_i(x)=f_i(x)/g(x)$, $(i=1, 2, \cdots)$, and $d\mu=gdm$. Then the conditions

$$D(X_i)\subseteqq D(X)(m),\ \mu(D(X)-D(X_i))\to 0, \qquad (i\to\infty),$$

and

$$d_i(X, X_i)\equiv \operatorname*{ess.\,sup}_{x\in D(X_i)}|\lambda(x)-\lambda_i(x)|\to 0, \qquad (i\to\infty),$$

imply that

$$I(X_i: Y)\to I(X: Y), \qquad (i\to\infty).$$

Assumption I is concerned, in statistical terminology, with the truncation of probability distributions, and assumption II is the so called additive property.

We now prove the following theorem, which is the purpose of this paper.

THEOREM. *Under the assumptions I, II and III, the function* $I(X: Y)$ *can be expressed as*

$$I(X: Y)=\int_R f(x)\log\frac{f(x)}{g(x)}dm, \tag{5}$$

up to a multiplicative positive constant, depending only on $\mathscr{T}$.

PROOF. The proof may be partitioned into three parts, as that of Theorem 2 of [5].

First, it will be shown that, if the ratio $\lambda(x)=f(x)/g(x)$ is constant (m) on a subset of R, E, with $\mu(E)=\int_E g(x)dm=v$, and is equal to zero elsewhere, then it holds that

$$I(X: Y)=c\log\frac{1}{v}, \qquad (c>0). \tag{6}$$

Indeed, Assumption I states that, for any X and Y belonging to $V(R, S, m)$, such that X is a truncation of Y, the function $I(X: Y)$ deponds only on $\mu(E)$ for any basic measure space (R, S, m), where $E=D(X)$.

To show (6), we consider two mutually independent sets of probability distributions, $\{X_i\}$, $(i=1, 2, \cdots, k)$, and $\{Y_i\}$, $(i=1, 2, \cdots, k)$, belonging to $V(R, S, m)$, (i.e., X_i's and Y_j's are mutually independent), where we assume that the gpdf.'s of X_i's are all equal(m) to $f(x)$ and those of Y_j's to $g(x)$, and $f(x)$ and $g(x)$ are such that

$$\lambda(x)=f(x)/g(x)=\begin{cases}1/v, & \text{if } x\in E,\\ 0, & \text{otherwise,}\end{cases}$$

for a set E of S with $\mu(E)=v$. Then, by assumption I, it holds that

$$I(X_1: Y_1)=I(X_2: Y_2)=\cdots=I(X_k: Y_k)(=I(v), \text{ say}),$$

and it is easy to see that Assumption II leads to the following functional equation

$$I(v^k)=kI(v)$$

for all positive integer k and $0<v\leqq 1$. According to Assumption I, the solution of this equation may be given by (6).

Second, we consider the case where the probability ratio $\lambda(x)=f(x)/g(x)$ for gpdf.'s $f(x)$ and $g(x)$ of X and Y is a simple function. Let $z=\{A_i\}$, $(i=1, 2, \ldots, k)$ be an m-partition of a subset E of R, with $\mu(E)=v(>0)$. Suppose that

$$\lambda(x)=\begin{cases}p_i/v_i, & \text{on } A_i, \quad (i=1, 2, \cdots, k),\\ 0, & \text{elsewhere, (a.e.m.),}\end{cases} \tag{7}$$

where $v_i=\mu(A_i)$, $p_i>0$, $(i=1, 2, \ldots, k)$, and $\sum_{i=1}^{k} p_i=1$. We can choose a

set of k mutually disjoint subsets of R, $Z'=\{A_i'\}$, $(i=1, 2, \cdots, k)$, such that $\mu(A_i')=p_i/\lambda v_i$, $(i=1, 2, \cdots, k)$, where $\lambda=\sum_{i=1}^{k} p_i/v_i(\geqq 1)$. Let X' be a member of $V(R, S, m)$ such that, if X falls in A_i, then X' is distributed over A_i' with gpdf. $f(x'|x)$ which is given by

$$(8) \qquad f(x'|x)=\frac{\lambda v_i}{p_i}g(x'), \qquad x\in A_i \text{ and } x'\in A_i', \qquad (i=1, 2, \cdots, k).$$

On the other hand, let Y' be in $V(R, S, m)$, and be distributed independently of Y, with the same (m) gpdf. as $g(x)$. Then, the joint distribution of (Y, Y') has the gpdf. defined by

$$g(x, x')=g(x)g(x')$$

on the product space $(R\times R, S\times S, m\times m)$, while that of (X, X') has gpdf.

$$f(x, x')=f(x)f(x'|x)=\begin{cases}\lambda g(x, x'), & \text{if } (x, x')\in F,\\ 0, & \text{elsewhere,}\end{cases}$$

where $F=\sum_{i=1}^{k}(A_i\times A_i')$, for which it is seen that

$$\mu\times\mu(F)=\int_F g(x, x')d(m\times m)=\frac{1}{\lambda}.$$

The probability ratio of $f(x, x')$ to $g(x, x')$, now, becomes

$$(9) \qquad \lambda(x, x')=\frac{f(x, x')}{g(x, x')}=\begin{cases}\lambda, & \text{if } (x, x')\in F,\\ 0, & \text{otherwise.}\end{cases}$$

By Assumption II and (6), we can obtain the following

$$(10) \qquad c\log\lambda=I(X: Y)+c'\sum_{i=1}^{k} p_i\log\frac{\lambda v_i}{p_i}.$$

Considering, in particular, the case where $v_i=p_i$, $(i=1, 2, \ldots, k)$, and $k>1$, in the above definitions (7) and (8), we have $I(X: Y)=0$, and $\lambda=k$, from which we get $c=c'$. Then, the equality (10) turns out to be the following

$$(11) \qquad I(X: Y)=\sum_{i=1}^{k} p_i\log\frac{p_i}{v_i}.$$

Finally, we shall examine the general case. Let E be the carrier $D(X)$ of gpdf. $f(x)$ of X. Then the carrier of $\lambda(x)=f(x)/g(x)$ is identical (m) with E. As was noted in the first section, $\lambda(x)$ is the gpdf. of

a certain probability distribution defined over the measure space (R, S, μ). In the analogous manner to that of [5], we take for each positive integer n a truncated function of $\lambda(x)$ such as

$$\lambda^{(n)}(x)=\frac{\varphi_n(x)}{\alpha_n},$$

where

$$\varphi_n(x)=\begin{cases}\lambda(x), & \text{if } 1/n\leqq\lambda(x)\leqq n,\\ 0, & \text{otherwise,}\end{cases}$$

and

$$\alpha_n=\int_E \varphi_n(x)d\mu.$$

Then, for each n, there exists a probability distribution $X^{(n)}$ belonging to $V(R, S, m)$, whose gpdf. is identical(m) with $f^{(n)}(x)=\lambda^{(n)}(x)g(x)$. From the definition of $\lambda^{(n)}(x)$, it will easily be seen that

$$(12)\qquad \begin{cases}\mu(D(X)-DX^{(n)})\to 0, & (n\to\infty),\\ \quad \text{and}\\ d_n(X, X^{(n)})=\underset{x\in D(X^{(n)})}{\text{ess. sup}}|\lambda(x)-\lambda^{(n)}(x)|\to 0, & (n\to\infty).\end{cases}$$

For each $X^{(n)}$, we can choose a sequence of probability distributions, $\{X_i^{(n)}\}$ $(i=1, 2, \cdots)$ in $V(R, S, m)$, with gpdf.'s $\{f_i^{(n)}(x)\}$ $(i=1, 2, \cdots)$ such that

$$\lambda_i^{(n)}(x)=\frac{f_i^{(n)}(x)}{g(x)}=\begin{cases}p_{ij}^{(n)}/v_{ij}^{(n)}, & \text{on } A_{ij}^{(n)}, \quad (j=1, 2, \cdots, k_{in}),\\ 0, & \text{elsewhere,}\end{cases}$$

where, for each i, $\{A_{ij}^{(n)}\}$ $(j=1, 2, \ldots, k_{in})$ is a μ-partition of $D(X^{(n)})$ with $v_{ij}^{(n)}=\mu(A_{ij}^{(n)})$, $p_{ij}^{(n)}>0$ and $\sum_{j=1}^{k_{in}} p_{ij}^{(n)}=1$, for which it holds that

$$(13)\qquad d_n(X^{(n)}, X_i^{(n)})=\underset{x\in D(X^{(n)})}{\text{ess. sup}}|\lambda^{(n)}(x)-\lambda_i^{(n)}(x)|\to 0, \qquad (i\to\infty).$$

(As for the construction of this sequence, see the proof of theorem 2.1 of [6].) From (11), it follows that

$$(14)\qquad I(X_i^{(n)}: Y)=\sum_{j=1}^{k_{in}} p_{ij}^{(n)}\log\frac{p_{ij}^{(n)}}{v_{ij}^{(n)}},$$

up to a multiplicative positive constant. Since, by Theorem 1 of [5] and (13), the right-hand member of (14) converges to the right-hand member of the following expression, Assumption III and (13) imply that

$$I(X^{(n)}: Y)=c\int_R f^{(n)}(x)\log\frac{f^{(n)}(x)}{g(x)}dm, \qquad (c>0). \tag{15}$$

According to Theorem 1 of [5] and (12) the right-hand member of (15) converges to the right-hand member of (16) below, hence, by Assumption III and (12) it holds that

$$I(X: Y)=c\int_R f(x)\log\frac{f(x)}{g(x)}dm, \qquad (c>0), \tag{16}$$

which completes the proof of our theorem, since Assumption I implies that c is a constant depending only on $\mathscr{T}$.

DEPT. OF MATH., COLLEGE OF SCI. AND ENG., NIHON UNIV.

REFERENCES

[1] S. Kullback, *Information Theory and Statistics*, John Wiley and Sons, 1959.
[2] S. Ikeda, "An application of the discrimination information measure to the theory of testing hypotheses, Part II," *Ann. Inst. Stat. Math.*, Vol. 13 (1961), pp. 61-89.
[3] S. Ikeda, "Necessary conditions for the convergence of Kullback-Leibler's mean information," to be published soon.
[4] S. Ikeda, "A remark on the converagence of Kullback-Leibler's mean information," *Ann. Inst. Stat. Math.*, Vol. 12 (1960), pp. 81-88.
[5] S. Ikeda, "A note on the characterization of Shannon-Wiener's measure of information," *Ann. Inst. Stat. Math.*, Vol. 13 (1962), pp. 259-266.
[6] S. Ikeda, "Continuity and characterization of Shannon-Wiener information measure for continuous probability distributions," *Ann. Inst. Stat. Math.*, Vol. 11 (1959), pp. 131-144.