Problems of Control and Information Theory, Vol. 18 (1), pp. 3—16 (1989)

MULTIINFORMATION AND THE PROBLEM OF CHARACTERIZATION OF CONDITIONAL INDEPENDENCE RELATIONS

M. STUDENÝ

(Prague)

(Received January 10, 1988)

Certain algebraic relation between multiinformation and conditional mutual information is established. It is shown to be applicable to the problem of characterization of conditional independence relations arising in connection with probabilistic expert systems. More concretely, a new axiom of these relations is derived. Some auxiliary results have their own significance: the characterization of marginally continuous measures in Proposition 1 and the information theoretical significance of the conditional product of measures mentioned in Consequence 3.

Introduction

The main concept of this paper is a certain generalization of the concept of mutual information, namely the so-called *multiinformation.* Simply, multiinformation is the relative entropy of the simultaneous distribution of a finite collection of random variables with respect to the product of the distributions of individual random variables. It is nonnegative and vanishes iff the corresponding random variables are independent. So, similarly as the mutual information which can serve as a measure of dependence of two random variables (see [13]), multiinformation enables us to characterize the level of dependence of more than two random variables. From this point of view it was studied by Perez in [8].

There are several papers belonging to information theory which indirectly handle multiinformation. For example, in [1] the studied algorithm IPFP converges to such probability measure which minimizes multiinformation in some given family of measures having prescribed marginals. As statistical properties of multiinformation are concerned, they are investigated in [12].

In this paper we want to show that multiinformation is also useful in apparently remote spheres. Namely, its certain "algebraic" properties can be applied to the problem of characterization of *conditional independence relations* (we shall use the abbreviation CIR here). This problem arises in connection with probabilistic expert systems, i.e. expert systems based on principles of probability theory.

Akadémiai Kiadó, Budapest
Pergamon Press, Oxford

The first section contains the definitions of the basic concepts and recalls some facts used later. Note that we take multiinformation as a characteristic of a probability measure; those who prefer to speak about random variables can regard the probability measure as the distribution of the corresponding variables. Moreover, we subjoin a proposition which establish an interesting equivalence connection between marginally continuous measures and measures that can be formed by a dominated kernel.

In the second section the conditional product of measures is defined and some facts about it are mentioned.

The third section deals with the concept of conditional mutual information which is defined by means of the concept of conditional product of measures. In Consequence 1 the fundamental formula for the conditional mutual information is given.

The fourth section considers both multiinformation and conditional mutual information as a set function on subsets of the index set. An important algebraic connection between them is established there.

Finally, the mentioned connection is applied in the last section. The problem of characterization of conditional independence relations (CIR's) is formulated there and it is shown how it is possible to utilize multiinformation.

1. Basic definitions, auxiliary concepts and results

Given measurable spaces $(X, \mathscr{X})$, $(Y, \mathscr{Y})$ and a probability measure R on $(X \times Y, \mathscr{X} \times \mathscr{Y})$ *the marginal measure* (or simply the *marginal*) *of* R *on* $(X, \mathscr{X})$ is defined by

$$R^X(A) = R(A \times Y), \qquad A \in \mathscr{X}.$$

We denote it by the symbol of the original measure having as upper index the symbol of the corresponding space.

Let us suppose that two probability measures P and Q on a measurable space $(X, \mathscr{X})$ are given. In case $P \ll Q$ we take some function $f\colon X \to \langle 0, \infty)$ (it means that f is defined everywhere on X and has all values finite and nonnegative) which is a version of the Radon–Nikodym derivative dP/dQ and define the *relative entropy of* P *w. r. to* Q (we use the abbreviation w. r. instead of "with respect") as the integral:

$$H(P, Q) = \int_{x \in X} \ln f(x) dP(x).$$

Since $P\{x \in X;\ f(x) = 0\} = 0$, it is not essential what is $\ln(0)$. Evidently, the value of $H(P, Q)$ does not depend on the choice of a version of dP/dQ. In case $P \not\ll Q$ we put $H(P, Q) = \infty$. In this paper we denote relative entropy by the letter H.

Relative entropy is always nonnegative and vanishes iff $P=Q$. Moreover, if $\mathscr{Y}$ is a sub-σ-algebra of $\mathscr{X}$ and $\tilde{P}$ or $\tilde{Q}$ is the restriction of P or Q on $\mathscr{Y}$ (respectively), then $H(\tilde{P}, \tilde{Q}) \leqq H(P, Q)$. Especially, it follows for every pair of probability measures P, Q on a product $(X \times Y, \mathscr{X} \times \mathscr{Y})$:

$$H(P^X, Q^X) \leqq H(P, Q). \tag{1}$$

These basic properties are well known, see e.g. [10] or [9].

If P is a probability measure on a product $(X \times Y, \mathscr{X} \times \mathscr{Y})$ then the *mutual information between X and Y* is defined as the relative entropy of P w. r. to $P^X \times P^Y$.

Analogously, given a finite nonempty collection of measurable spaces $(X_i, \mathscr{X}_i)$, $i \in N$ and a probability measure P on $\left(\prod_{i \in N} X_i, \prod_{i \in N} \mathscr{X}_i\right)$ we define the *multiinformation of P* as the relative entropy of P w. r. to the product of its one-dimensional marginals:

$$M(P) = H\left(P, \prod_{i \in N} P^{X_i}\right).$$

In this paper multiinformation is denoted by the letter M.

In this paragraph $(X, \mathscr{X})$, $(Y, \mathscr{Y})$ are measurable spaces and R is a probability measure on $(X \times Y, \mathscr{X} \times \mathscr{Y})$. By a *representative of conditional probability on* $(Y, \mathscr{Y})$ *w. r. to* $(X, \mathscr{X})$ we shall understand every mapping $K: \mathscr{Y} \times X \to \langle 0, 1 \rangle$ such that for each $B \in \mathscr{Y}$ the function $x \mapsto K(B|x)$ is a variant of conditional probability of the set B given the σ-algebra $\mathscr{X}$, i.e. $\mathscr{X}$-measurable function satisfying:

$$\int_{x \in A} K(B|x)\, dR^X(x) = R(A \times B) \qquad \text{for each} \quad A \in \mathscr{X}. \tag{2}$$

We shall use the abbreviation c. p. instead of "conditional probability". Note that (2) can be formulated equivalently as follows:

$$\int_{x \in X} g(x) \cdot K(B|x)\, dR^X(x) = \int_{(x,y) \in X \times B} g(x)\, dR(x, y) \tag{3}$$
$$\text{for each} \quad g: X \to \langle 0, 1 \rangle \quad \mathscr{X}\text{-measurable}.$$

The existence of a representative of c. p. is a trivial consequence of the Radon–Nikodym theorem. Indeed, for each $B \in \mathscr{Y}$ the function $A \mapsto R(A \times B)$ is a measure on $(X, \mathscr{X})$ which is absolutely continuous w. r. to R^X. Evidently, representatives of c. p. are determined uniquely in the framework of this equivalence:

$$K \simeq K' \quad \text{iff} \quad K(B|x) = K'(B|x) \quad \text{for} \quad R^X\text{-a.e.} \quad x \in X \quad \text{for every} \quad B \in \mathscr{Y}.$$

We shall use the symbol $R_{Y|X}$ to denote an arbitrary representative of c. p., i.e. the symbol of the original measure having as lower index the separate symbols of the respective spaces.

A representative K of c. p. on $(Y, \mathscr{Y})$ w. r. to $(X, \mathscr{X})$ is called *regular* iff for each $x \in X$ the function $B \mapsto K(B|x)$ is a probability measure on $(Y, \mathscr{Y})$. In case there exists a regular representative of c. p. on $(Y, \mathscr{Y})$ w. r. to $(X, \mathscr{X})$ we shall say that c. p. on $(Y, \mathscr{Y})$ w. r. to $(X, \mathscr{X})$ is regular.

In this paragraph we suppose measurable spaces $(X, \mathscr{X})$ and $(Y, \mathscr{Y})$ are given. By stochastic (or Markov) *kernel from* $(X, \mathscr{X})$ *to* $(Y, \mathscr{Y})$ we understand a collection $\mathscr{P} = \{P_x;\, x \in X\}$ of probability measures on $(Y, \mathscr{Y})$ such that for each $B \in \mathscr{Y}$ the function $x \mapsto P_x(B)$ is $\mathscr{X}$-measurable. This concept is also known as a crossing probability or as a channel (in information theory, especially).

We shall say that a kernel $\mathscr{P} = \{P_x;\, x \in X\}$ is *dominated* iff there exists a probability measure τ on $(Y, \mathscr{Y})$ such that for each $x \in X$ $P_x \ll \tau$.

Given a kernel $\mathscr{P} = \{P_x;\, x \in X\}$ from $(X, \mathscr{X})$ to $(Y, \mathscr{Y})$ and a probability measure Q on $(X, \mathscr{X})$ we can define a probability measure $Q * \mathscr{P}$ on $(X \times Y, \mathscr{X} \times \mathscr{Y})$ by:

$$Q * \mathscr{P}(A \times B) = \int_{x \in A} P_x(B)\, dQ(x) \qquad A \in \mathscr{X},\ B \in \mathscr{Y} \tag{4}$$

and by the standard extension argument (see [4], III.2.1). We shall say that *Q and $\mathscr{P}$ form the measure $Q * \mathscr{P}$*. Note that (4) can be extended as follows:

$$Q * \mathscr{P}(C) = \int_{x \in X} P_x(C_x)\, dQ(x) \qquad C \in \mathscr{X} \times \mathscr{Y}$$

where $C_x = \{y \in Y;\, (x, y) \in C\}$. Especially:

$$Q * \mathscr{P}(C) = 0 \quad \text{iff} \quad P_x(C_x) = 0 \quad \text{for} \quad Q\text{-a.e.} \quad x \in X. \tag{5}$$

Remark 1. a) Let us point out an interesting connection. Supposing that R is a probability measure on $(X \times Y, \mathscr{X} \times \mathscr{Y})$, we can easily derive that $R = Q * \mathscr{P}$ for some probability measure Q on $(X, \mathscr{X})$ and some kernel $\mathscr{P}$ from $(X, \mathscr{X})$ to $(Y, \mathscr{Y})$ iff the c. p. on $(Y, \mathscr{Y})$ w. r. to $(X, \mathscr{X})$ is regular.

b) Further, we note the known fact that the assumption saying that Y is a separable complete metric space and $\mathscr{Y}$ is the σ-algebra of its Borel subsets suffices for regularity of c. p. on $(Y, \mathscr{Y})$ w. r. to $(X, \mathscr{X})$ (see [4], it follows from the consequence of V.4.4). Especially, it holds for finite Y with $\mathscr{Y} = \exp Y$.

If a probability measure P on a product $(X \times Y, \mathscr{X} \times \mathscr{Y})$ satisfies $P \ll P^X \times P^Y$, then it is called *marginally continuous*. Evidently, this condition is necessary for finiteness of the mutual information between X and Y. The following lemma leads to some characterization of marginally continuous measures in Proposition 1.

Lemma 1. Let $(X, \mathscr{X})$, $(Y, \mathscr{Y})$ be measurable spaces, λ a probability measure on $(X, \mathscr{X})$, τ on $(Y, \mathscr{Y})$ and μ on $(X \times Y, \mathscr{X} \times \mathscr{Y})$. Then $\mu \ll \lambda \times \tau$ iff $\mu^X \ll \lambda$ and there exists a kernel $\mathscr{P} = \{P_x; x \in X\}$ from $(X, \mathscr{X})$ to $(Y, \mathscr{Y})$ such that $\mu = (\mu^X) * \mathscr{P}$ and for each $x \in X$ it holds $P_x \ll \tau$.

Proof. a) In case $\mu \ll \lambda \times \tau$ we can take such a version $L\colon X \times Y \to \langle 0, \infty)$ of $d\mu/d(\lambda \times \tau)$ that

$$l(x) = \int_{y \in Y} L(x, y)\, d\tau(y) < \infty \qquad x \in X .$$

Evidently, l is a version of $d(\mu^X)/d\lambda$. We define $k(x, y) = l^{-1}(x) \cdot L(x, y)$ if $l(x) > 0$ and $k(x, y) = 1$, otherwise. Finally, we put:

$$P_x(B) = \int_{y \in B} k(x, y)\, d\tau(y) \qquad x \in X, B \in \mathscr{Y} .$$

It makes no problem to verify that $\mathscr{P} = \{P_x;\ x \in X\}$ is the desired kernel.

b) The sufficiency can be seen using (5). For $C \in \mathscr{X} \times \mathscr{Y}$ the relation $(\lambda \times \tau)(C) = 0$ implies $\tau(C_x) = 0$ for λ-a.e. $x \in X$. So $P_x(C_x) = 0$ for μ^X-a.e. $x \in X$, i.e. $(\mu^X) * \mathscr{P}(C) = 0$. ■

Proposition 1. Let $(X, \mathscr{X})$, $(Y, \mathscr{Y})$ be measurable spaces. Then the following conditions on a probability measure μ on $(X \times Y, \mathscr{X} \times \mathscr{Y})$ are equivalent:

(a) μ is marginally continuous

(b) there exist a probability measure λ on $(X, \mathscr{X})$ and a probability measure τ on $(Y, \mathscr{Y})$ such that $\mu \ll \lambda \times \tau$

(c) μ can be formed by a dominated kernel from $(X, \mathscr{X})$ to $(Y, \mathscr{Y})$.

Proof. Directly from Lemma 1 we conclude that (b) implies (c). Conversely, if $\mu = Q * \mathscr{P}$ where Q is a measure on $(X, \mathscr{X})$ and $\mathscr{P}$ is the mentioned kernel, then necessarily $Q = \mu^X$. So, we can take $\lambda = \mu^X$ in Lemma 1 to show that (c) implies (b). In fact we have just proved that $\mu \ll \lambda \times \tau$ implies $\mu \ll \mu^X \times \tau$. Replacing of $(X, \mathscr{X})$ by $(Y, \mathscr{Y})$ we get that $\mu \ll \lambda \times \tau$ implies $\mu \ll \lambda \times \mu^Y$. So, let us take $\lambda = \mu^X$ here and see that (b) implies (a). The converse is trivial. ■

Note that Proposition 1 yields a sufficient condition for regularity of c. p., which is not of topological nature (see Remark 1).

2. Conditional product of measures

Definition 1. Let $(X, \mathscr{X})$, $(Y, \mathscr{Y})$, $(Z, \mathscr{Z})$ be measurable spaces and P a probability measure on $(X \times Y \times Z, \mathscr{X} \times \mathscr{Y} \times \mathscr{Z})$. We shall say that P is a *conditional product on $X \times Y$ under condition Z* iff it holds

$$\left.\begin{aligned} P_{X \times Y|Z}(A \times B|z) = P_{X|Z}(A|z) \cdot P_{Y|Z}(B|z) \quad &\text{for} \quad P^Z\text{-a.e.} \quad z \in Z \\ \text{for each} \quad A \in \mathscr{X}, \quad &B \in \mathscr{Y} . \end{aligned}\right\} \tag{6}$$

Naturally, we write $P_{X|Z}$ instead of $(P^{X\times Z})_{X|Z}$. Evidently, the validity of (6) does not depend on the choice of representatives of c. p. Further, it is easy to see that (6) is equivalent to:

$$\left.\begin{aligned} P(A\times B\times C) = \int_{z\in C} P_{X|Z}(A|z)\cdot P_{Y|Z}(B|z)\,dP^Z(z) \\ \text{for each}\quad A\in\mathscr{X},\ B\in\mathscr{Y},\ C\in\mathscr{Z}. \end{aligned}\right\} \tag{7}$$

We use this terminology in order not to impair analogy with the "unconditional" case: a probability measure R on $(X\times Y, \mathscr{X}\times\mathscr{Y})$ is the product of its marginals R^X and R^Y iff in the probability space $(X\times Y, \mathscr{X}\times\mathscr{Y}, R)$ the σ-algebras $\mathscr{X}\times\mathscr{Y}'$ and $\mathscr{X}'\times\mathscr{Y}$ are independent ($\mathscr{X}'$, $\mathscr{Y}'$, $\mathscr{Z}'$ are respectively trivial σ-algebras on $X\times Y\times Z$). Analogously, (6) means that in the probability space $(X\times Y\times Z, \mathscr{X}\times\mathscr{Y}\times\mathscr{Z}, P)$ the σ-algebras $\mathscr{X}\times\mathscr{Y}'\times\mathscr{Z}'$ and $\mathscr{X}'\times\mathscr{Y}\times\mathscr{Z}'$ are conditionally independent given the σ-algebra $\mathscr{X}'\times\mathscr{Y}'\times\mathscr{Z}$ (see [5], chapter VII, § 25.3).

Remark 2. The usual "unconditional" product of measures can be viewed as a special case of the conditional product. Indeed, supposing that $\mathscr{Z}$ is the trivial σ-algebra on Z, a measure P on $(X\times Y\times Z, \mathscr{X}\times\mathscr{Y}\times\mathscr{Z})$ is a conditional product on $X\times Y$ under condition Z iff $P^{X\times Y} = P^X\times P^Y$.

Definition 2. Let $(X, \mathscr{X})$, $(Y, \mathscr{Y})$, $(Z, \mathscr{Z})$ be measurable spaces, $Q_{X\times Z}$ and $Q_{Y\times Z}$ be consonant probability measures, respectively, on $X\times Z$ and on $Y\times Z$, i.e. $(Q_{X\times Z})^Z = (Q_{Y\times Z})^Z$.

In case there exists a measure P on $(X\times Y\times Z, \mathscr{X}\times\mathscr{Y}\times\mathscr{Z})$ having $Q_{X\times Z}$ and $Q_{Y\times Z}$ as marginals which is moreover a conditional product on $X\times Y$ under condition Z, we shall call it the *conditional product of* $Q_{X\times Z}$ *and* $Q_{Y\times Z}$.

Proposition 2. Under assumptions of Definition 2 it holds:

a) The conditional product of $Q_{X\times Z}$ and $Q_{Y\times Z}$ is determined uniquely.

b) Supposing that $Q_{X\times Z}$ has regular c. p. on $(X, \mathscr{x})$ w. r. to $(Z, \mathscr{z})$ or that $Q_{Y\times Z}$ has regular c. p. on $(Y, \mathscr{y})$ w. r. to $(Z, \mathscr{z})$) there exists the conditional product of $Q_{X\times Z}$ and $Q_{Y\times Z}$.

Since measures having the same marginals on $X\times Z$ have the same set of representatives of c. p. on $(X, \mathscr{X})$ w. r. to $(Z, \mathscr{Z})$, we can show the first part of Proposition 2 using (7). For the proof of the second part we refer to the translator's remarks to chapter 3 of [10].

Combining Propositions 1 and 2b we see the known fact mentioned in [10] (p. 56), namely: supposing that $Q_{X\times Z}$ (or $Q_{Y\times Z}$) in Definition 2 is marginally continuous, there exists the conditional product of $Q_{X\times Z}$ and $Q_{Y\times Z}$.

Nevertheless, under assumptions of Definition 2 the conditional product of $Q_{X \times Z}$ and $Q_{Y \times Z}$ may not exist, moreover it holds:

Proposition 3. There exist measurable spaces $(X, \mathscr{X})$, $(Y, \mathscr{Y})$, $(Z, \mathscr{Z})$ and a probability measure P on $(X \times Y \times Z, \mathscr{X} \times \mathscr{Y} \times \mathscr{Z})$ such that the conditional product of $P^{X \times Z}$ and $P^{Y \times Z}$ does not exist.

For the proof we refer to [11], where the desired example is constructed.

3. Conditional mutual information

Definition 3. Let $(X, \mathscr{X})$, $Y, \mathscr{Y})$, $(Z, \mathscr{Z})$ be measurable spaces and P a probability measure on $(X \times Y \times Z, \mathscr{X} \times \mathscr{Y} \times \mathscr{Z})$. In case there exists the conditional product of $P^{X \times Z}$ and $P^{Y \times Z}$ (denoted by $\bar{P}$), we put:

$$C(X; Y|Z) = H(P, \bar{P}).$$

In the opposite case we put $C(X; Y|Z) = \infty$. The number $C(X; Y|Z)$ we shall call the *conditional mutual information between X and Y under condition Z.*

The following lemma is a trivial consequence of the basic properties of the relative entropy:

Lemma 2. Under assumptions of Definition 3 it holds $C(X; Y|Z) \geqq 0$. Moreover, $C(X; Y|Z) = 0$ iff P is a conditional product on $X \times Y$ under condition Z.

The well-known notion of mutual information can be viewed as a special case of conditional mutual information, if we take $\mathscr{Z}$ as the trivial σ-algebra on Z (cf. Remark 2). Indeed, it must hold

$$P = P^{X \times Y} \times P^Z \quad \text{and} \quad \bar{P} = P^X \times P^Y \times P^Z \quad \text{and} \quad H(P, \bar{P}) = H(P^{X \times Y}, P^X \times P^Y).$$

The following lemma we need for the proof of the fundamental formula (10) in Consequence 1:

Lemma 3. Under assumptions of Definition 3 we denote $R = P^Y \times P^{X \times Z}$.

a) If $P^{Y \times Z} \ll P^Y \times P^Z$, then there exists the conditional product $\bar{P}$ of $P^{X \times Z}$ and $P^{Y \times Z}$. Moreover, $\bar{P} \ll R$ and there exists a function k: $Y \times Z \to \langle 0, \infty)$ which is a version of $d(P^{Y \times Z})/d(P^Y \times P^Z)$ and viewed as a function on $X \times Y \times Z$ a version of $d\bar{P}/dR$.

b) The following two conditions are equivalent:

$$P \ll R, \tag{8}$$

$$\left.\begin{array}{l} P^{Y \times Z} \ll P^Y \times P^Z \text{ and there exists the conditional product} \\ \bar{P} \text{ of } P^{X \times Z} \text{ and } P^{Y \times Z} \text{ which, moreover, satisfies } P \ll \bar{P}. \end{array}\right\} \tag{9}$$

We shall not prove this lemma. The proof can be found in [2] (pp. 42–44), but with the proviso that one must be careful whether the conditional product of measures exists. Namely, in the mentioned paper there is an erroneous consideration leading to the conclusion that the existence of P suffices for the existence of the conditional product of $P^{X\times Z}$ and $P^{Y\times Z}$ (more exactly, the set function (2.7.7) is not countably additive). It was said in Proposition 3 that the mentioned conclusion is wrong.

In this paper we extended the definition of conditional mutual information in order to preserve the general validity of relation (16) mentioned below.

Consequence 1. Under assumption of Definition 3 it holds

$$H(P, P^{Y}\times P^{X\times Z}) = C(X;Y|Z) + H(P^{Y\times Z}, P^{Y}\times P^{Z}). \tag{10}$$

Proof. If (8) does not hold, then according to Lemma 3b both sides of (10) are infinite. In case (8) holds, we use Lemma 3a and fix the function $k\colon Y\times Z\to\langle 0,\infty)$ mentioned there. Further, according to Lemma 3b we may consider some version $l\colon X\times Y\times Z\to\langle 0,\infty)$ of $dP/d\bar{P}$. So, k being considered as a function on $X\times Y\times Z$, the product $k\cdot l$ is a version of dP/dR. Finally, integrating the identity

$$\ln(k\cdot l) = \ln(k) + \ln(l) \qquad (\text{where } \ln 0 = -\infty)$$

with respect to P, we get (10). ■

4. Multiinformation viewed as a set function

In the remaining two sections we shall consider the following situation.

$$\text{(S)}\quad \left\{\begin{array}{l} \text{A finite nonempty collection of measurable spaces} \\ (X_i, \mathscr{X}_i),\ i\in N \text{ is given. If } A\subset N \text{ is nonempty, we shall} \\ \text{write } (X_A, {}_A) \text{ instead of } \left(\prod\limits_{i\in A} X_i,\ \prod\limits_{i\in A}\mathscr{X}_i\right). \\ \text{Further, a probability measure } P \text{ on } (X_N, \mathscr{X}_N) \text{ is given.} \\ \text{For the sake of brevity, the marginal of } P \text{ on } (X_A, \mathscr{X}_A) \text{ will be} \\ \text{denoted by } P^A. \end{array}\right.$$

Definition 4. Assuming (S), we define for nonempty $A\subset N$:

$$I_m[A] = M(P^A).$$

Moreover, for empty A we put $I_m[\emptyset]=0$.

From basic properties of relative entropy we easily conclude that assuming (S) the function $I_m : \exp N \to \langle 0, \infty \rangle$ satisfies:

$$A \subset B \quad \text{implies} \quad I_m[A] \leqq I_m[B] \tag{11}$$

$$\text{if card } A \leqq 1 \qquad \text{then} \quad I_m[A] = 0. \tag{12}$$

Definition 5. Assuming (S), we define for every ordered triplet $\langle A, B, C \rangle$ of disjoint subsets of N the number $I_c[A; B|C] \in \langle 0, \infty \rangle$. If all the sets A, B, C are nonempty, then we define it as the conditional mutual information between X_A and X_B under condition X_C (logically it is computed from $P^{A \cup B \cup C}$), i.e.

$$I_c[A; B|C] = C(X_A; X_B | X_C).$$

For empty C and nonempty A, B we define $I_c[A; B|\emptyset]$ as the mutual information between X_A and X_B, i.e.

$$I_c[A; B|\emptyset] = H(P^{A \cup B}, P^A \times P^B).$$

Finally, in case that A or B is empty we put:

$$I_c[\emptyset; B|C] = 0 \quad \text{and} \quad I_c[A; \emptyset|C] = 0.$$

Lemma 4. Assuming (S), the function I_c satisfies (A, B, C are supposed to be disjoint):

$$I_c[A; B|C] = I_c[B; A|C] \tag{13}$$

$$0 \leqq I_c[A; B|C] \tag{14}$$

$$I_c[A; B \cup C|\emptyset] = I_c[A; B|C] + I_c[A; C|\emptyset] \tag{15}$$

$$\text{if} \quad A' \subset A, \; B' \subset B, \quad \text{then} \quad I_c[A', B'|C] \leqq I_c[A; B|C]. \tag{16}$$

Proof. (13) and (14) are easy consequences of the definition; (15) follows directly from (10) and (13). (16) is trivial in case $I_c[A; B|C] = \infty$. In the opposite case there exists the conditional product of $P^{A \cup C}$ and $P^{B \cup C}$. It makes no problem to verify that its restriction onto $X_{A' \cup B' \cup C}$ is the conditional product of $P^{A' \cup C}$ and $P^{B' \cup C}$. So (16) follows from (1). ■

The substantial relation between I_m and I_c is established by the following statement.

Proposition 4. Assuming (S), it holds for every $D, E \subset N$ (not necessarily disjoint):

$$I_m[D \cup E] + I_m[D \cap E] = I_m[D] + I_m[E] + I_c[E \backslash D; D \backslash E | D \cap E]. \tag{17}$$

Proof. a) First we prove (17) for disjoint D and E. So, if D and E are nonempty (otherwise trivial), then we denote $Q_A = \prod_{i \in A} P^{\{i\}}$ for nonempty $A \subset N$. In case $P^{D \cup E} \not\ll P^D \times P^E$ it is, according to Proposition 1, $P^{D \cup E} \not\ll Q_D \times Q_E = Q_{D \cup E}$. So, both $I_m[D \cup E]$ and $I_c[E \setminus D; D \setminus E | D \cap E]$ are infinite and (17) holds. Analogously we proceed in case $P^D \not\ll Q_D$ or $P^E \not\ll Q_E$ (using (11)). So, we can suppose $P^D \ll Q_D$, $P^E \ll Q_E$ and $P^{D \cup E} \ll P^D \times P^E$. We take a version $f: X_{D \cup E} \to \langle 0, \infty)$ of $d(P^{D \cup E})/d(P^D \times P^E)$, a version $h: X_D \to \langle 0, \infty)$ of dP^D/dQ_D and a version $g: X_E \to \langle 0, \infty)$ of dP^E/dQ_E. The proof we conclude similarly as the proof of Consequence 1.

b) Now we suppose arbitrary D, E. According to part a) we see:

$$I_m[D \cup E] = I_m[E \setminus D] + I_m[D] + I_c[E \setminus D; D | \emptyset]$$

$$I_m[E] = I_m[E \setminus D] + I_m[D \cap E] + I_c[E \setminus D; D \cap E | \emptyset].$$

So, for the proof of (17) it suffices to prove the identity:

$$I_c[E \setminus D; D | \emptyset] = I_c[E \setminus D; D \setminus E | D \cap E] + I_c[E \setminus D; D \cap E | \emptyset].$$

We simply put $A = E \setminus D$, $B = D \setminus E$, $C = D \cap E$ in (15). ■

Consequence 2. Assuming (S), the function I_m: $\exp N \to \langle 0, \infty \rangle$ is convex (or supermodular), i.e. it holds:

$$I_m[D \cup E] + I_m[D \cap E] \geqq I_m[D] + I_m[E] \qquad \text{for each} \quad D, E \subset N. \tag{18}$$

Proof. (14) implies $I_c[E \setminus D; D \setminus E | D \cap E] \geqq 0$. We add $I_m[D] + I_m(E)$ to both sides and use (17). ■

So, Consequence 2 leads to the following question.

Problem 1. We know that, assuming (S), function I_m satisfies (12) and (18) ((11) follows from them). Can it be conversed? More precisely, whether these conditions on a function $I: \exp N \to \langle 0, \infty \rangle$ suffice for the existence of measurable spaces and probability measure described in (S) such that $I = I_m$.

The last consequence shows some information-theoretical significance of the conditional product of measures.

Consequence 3. Let $(X_i, \mathcal{X}_i)$, $i \in N$ be measurable spaces and $\{A, B, C\}$ some decomposition of N (finite, nonempty sets). Let μ and τ be consonant probability measures, μ on $(X_{A \cup C}, \mathcal{X}_{A \cup C})$, τ on $(X_{B \cup C}, \mathcal{X}_{B \cup C})$. Further, we denote

$$\Phi = \{P;\ P \text{ is a probability measure on } X_N,\ M(P) < \infty,\ P^{A \cup C} = \mu,\ P^{B \cup C} = \tau\}.$$

Then a) $\Phi \neq \emptyset$ iff $M(\mu) < \infty$ and $M(\tau) < \infty$.

b) Supposing $\Phi \neq \emptyset$, there exists the conditional product of μ and τ and minimizes the multiinformation on Φ.

Proof. If $\Phi \neq \emptyset$, then $M(\mu)$ and $M(\tau)$ are finite according to (11). Conversely, let $M(\mu)$ and $M(\tau)$ be finite. We deduce that $\mu \ll \prod_{i \in A} \mu^{\{i\}} \times \prod_{i \in C} \mu^{\{i\}}$ and by Proposition 1 and Remark 1a we see that c. p. on $(X_A, \mathscr{X}_A)$ w. r. to $(X_C, \mathscr{X}_C)$ is regular. So, Proposition 2 yields the existence of the conditional product P of μ and τ. Evidently, for this measure $I_c[A; B|C] = 0$ and, according to Proposition 4, it is $M(P) = M(\mu) + M(\tau) - M(\mu^C) < \infty$, so $P \in \Phi$ and $\Phi \neq \emptyset$. Moreover, by (17) and (14) applied to another $Q \in \Phi$ we deduce $M(Q) \geqq M(\mu) + M(\tau) - M(\mu^C) = M(P)$. ■

5. Application to the problem of characterization of CIR's

Definition 6. Assuming (S), we define a ternary relation $I(.,.|.)$ having as the domain all ordered triplets $\langle A, B, C\rangle$ of mutually disjoint subsets of N. If both A and B is nonempty, then $I(A; B|C)$ holds iff $P^{A \cup B \cup C}$ is the conditional product of $P^{A \cup C}$ and $P^{B \cup C}$ (for empty C it means $P^{A \cup B} = P^A \times P^B$). If A or B is empty, then we postulate that $I(A; B|C)$ holds. We shall call this relation the *conditional independence relation* corresponding to P and shall use the abbreviation CIR.

Note that CIR determines the conditional dependence relation D as its complementary relation (i.e. $D(A; B|C)$ holds iff $I(A; B|C)$ does not hold). Now, what is the problem of characterization of CIR's?

Problem 2. Let N be nonempty finite set. The problem is to find all independent properties (axioms) of a ternary relation I (defined on all ordered triplets of disjoint subsets of N) which together yield a necessary and sufficient condition for the existence of finite spaces X_i, $i \in N$ and of a probability measure P on $\prod_{i \in N} X_i$ such that I coincides with the CIR corresponding to P.

In this form the CIR was introduced by Pearl in [6] and his previous papers. But restricts to strictly positive measures. In the mentioned paper five properties of CIR's are formulated. The first one is the axiom of symmetry:

$$I(A; B|C) \Leftrightarrow I(B; A|C). \tag{A.1}$$

Three other axioms can be integrated into the following one:

$$I(A; B \cup C|D) \Leftrightarrow [I(A; B|C \cup D) \wedge I(A; C|D)]. \tag{A.2}$$

These two axioms hold without the assumption of strict positivity of the measure, while the last property:

$$[I(A; B|C\cup D) \wedge I(A; C|B\cup D)] \Rightarrow I(A; B\cup C|D) \tag{19}$$

does not so (i.e. it is not relevant to Problem 2).

Pearl expressed the completeness conjecture, i.e. (A.1), (A.2), (19) is the solution of Problem 2 modified by the demand that P must be strictly positive. The rest of Pearl's paper is concerned with graphical representations of probabilistic knowledge that are possible owing to (A.1), (A.2), (19).

The desired solution of Problem 2 seems to be significant in the theory of probabilistic expert systems. Let us mention the intensional expert system INES (see [7]). According to this approach, the knowledge base of an expert system is modelled by a multidimensional probability measure, while partial knowledges obtained from experts are described by means of less-dimensional probability measures which should be marginals of the mentioned multidimensional one. For capacity reasons it is usually impossible to store the multidimensional measure in the memory of a computer. This imperfection is solved by the help of so-called DSS's (dependence structure simplifications). These multidimensional measures are "formed successively as conditional products of given less-dimensional measures". So, we have to store only those in the memory. The choice of the DSS (i.e. of the order of making conditional products) is made from a certain information-theoretical point of view.

The solution of Problem 2 would make possible some improvement. Since the notion of conditional independence (or dependence) is easy to interpret we would be able to determine the proper structure of dependences and independences directly by asking experts. By means of the solution of Problem 2 we would be able to decide whether the statements of various experts are contradictory or whether there exists a probabilistic model having the requisite dependences and independences (i.e. there exists a CIR having prescribed dependences and independences).

Now, how to use the multiinformation? From Lemma 2 and Definitions 5, 6 it is easy to see:

Proposition 5. Assuming (S), it holds for disjoint A, B, $C\subset N$:

$$I(A; B|C) \text{ holds} \Leftrightarrow I_c[A; B|C]=0. \tag{20}$$

Further, according to (17) we can express $I_c[A; B|C]$ by means of the function I_m (in Problem 2 X_i are finite, so I_m is finite). So, by this procedure we verify for disjoint A, B, C, D:

$$\left.\begin{aligned} &I_c[A; B|C\cup D]+I_c[C; D|A]+I_c[C; D|B]+I_c[A; B|\emptyset]= \\ &=I_c[C; D|A\cup B]+I_c[A; B|C]+I_c[A; B|D]+I_c[C; D|\emptyset]. \end{aligned}\right\} \tag{21}$$

Finally, from Proposition 5 we easily derive using (14):

$$\left.\begin{aligned}&[I(A;B|C\cup D] \wedge I(C;D|A) \wedge I(C;D|B) \wedge I(A;B|\emptyset)]\Leftrightarrow\\ &\Leftrightarrow[I(C;D|A\cup B) \wedge I(A;B|C) \wedge I(A;B|D) \wedge I(C;D|\emptyset)].\end{aligned}\right\} \quad (A.3)$$

Example. We can take $N=\{a, b, c, d\}$ and construct a certain ternary relation as follows:

1. $I(a, b|cd)$, $I(c, d|a)$, $I(c, d|b)$, $I(a, b|\emptyset)$ and symmetric independences hold
2. $I(A; B|C)$ for empty A or B holds
3. no other independence holds.

The desired relation satisfies (A.1), (A.2), (19) but not (A.3).

So, using algebraic properties of multiinformation a new axiom (A.3) of CIR's was derived and Pearl's completeness conjecture was disproved. Note that (A.1), (A.2) can be derived similarly. Perhaps, it is possible to derive further axioms of CIR's analogously.

Nevertheless, I do not know the complete solution of Problem 2. I would like to ask readers for help. If somebody knows something relevant to this problem (maybe the solution is known since, for example, the theory of Markov fields meets with similar problems), I would like him (or her) to send me a reference or a reprint or any information. The similar wish concerns Problem 1.

References

1. *Csiszár, I.*, *I*-divergence geometry of probability distributions and minimization problems. Ann. Probab. **3** (1975), pp. 146–158.
2. *Добрушин, Р. Л.*, Общая формулировка основной теоремы Шеннона в теории информации. Успехи матем. наук. **14** (1959), *6*, с. 3–104. (In Russian, translation: Ann. Math. Soc. Translation **33,** *2*, pp. 323–438).
3. *Halmos, P. R.*, Measure Theory. D. v. Nostrand Comp. Inc., New York 1950.
4. *Neveu, J.*, Bases mathematiques du calcul des probabilités. Masson et Cie, Paris 1964 (in French, Russian translation: Мир, Москва 1969).
5. *Loéve, M.*, Probability Theory. D. v. Nostrand Comp. Inc., New York 1960.
6. *Pearl, J.*, Markov and Bayes networks. Technical Report CSD 860024, R-46-I, October 1986, University of California, Los Angeles.
7. *Perez, A., Jiroušek, R.*, Constructing an intensional expert system (INES). Medical Decision Making: diagnostic strategies and expert systems, North Holland 1985, pp. 307–315.
8. *Perez, A.*, ε-admissible simplifications of the dependence structure of a set random variables. Kybernetika **13** (1977), pp. 439–449.
9. *Perez, A.*, Notions généralisées d'incertitude, d'entropie et d'information du point de vue de la théorie des martingales. Trans. of the 1-st Prague Conference on Information Theory, Statistical Decision Function, Random Processes (1956), Prague 1957, pp. 183–208 (in French).

10. *Pinsker, M. S.*, Information and Information Stability of Random Variables and Processes. Translation from Russian by A. Feinstein, Holden-Day, San Francisco 1964. (The original: Информация и информационная устройчивость случайных величин и процессов, Академия наук СССР, Москва 1960).
11. *Сазонов, В. В.*, Ответ на один вопрос Р. Л. Добрушина. Теория вероятностей и ее применение **9** (1964), *1*, с. 180–181.
12. *Studený, M.*, Asymptotic behaviour of empirical multiinformation. Kybernetika **23** (1987), *2*, pp. 124–135.
13. *Zvárová, J.*, Informační míry statistické závislosti a výběrové vlastnosti zobecněné entropie řádu α. Thesis, Prague 1973 (in Czech).

Мультиинформация и проблема характеризации отношений условной независимости

М. СТУДЕНЫ

(Прага)

Установлено некоторое алгебраическое соотношение между мультиинформацией и условной информацией. Показано, что это соотношение применимо к проблеме характеризации отношений условной независимости, которая возникает в связи с вероятностными экспертными системами. Более конкретно выведена новая аксиома для этих отношений. Некоторые подготовительные результаты имеют самостоятельное значение: характеризация маргинально-непрерывных мер в теореме 1 и информационно-теоретическое значение условного произведения мер, упомянутое в следствии 3.

M. Studený
Ústav teorie informace a automatizace ČSAV
(Institute of Information Theory and Automation)
Czechoslovak Academy of Sciences
Pod vodárenskou věží 4, 182 08 Praha 8
Czechoslovakia