THE MULTIINFORMATION FUNCTION AS A TOOL FOR MEASURING STOCHASTIC DEPENDENCE

M. STUDENÝ AND J. VEJNAROVÁ
Institute of Information Theory and Automation
Academy of Sciences of Czech Republic
Pod vodárenskou věží 4, 182 08 Prague
AND
Laboratory of Intelligent Systems
University of Economics
Ekonomická 957, 148 00 Prague
Czech Republic

Abstract. Given a collection of random variables $[\xi_i]_{i\in N}$ where N is a finite nonempty set, the corresponding *multiinformation function* ascribes the relative entropy of the joint distribution of $[\xi_i]_{i\in A}$ with respect to the product of distributions of individual random variables ξ_i for $i \in A$ to every subset $A \subset N$. We argue it is a useful tool for problems concerning stochastic (conditional) dependence and independence (at least in discrete case).

First, it makes possible to express the *conditional mutual information* between $[\xi_i]_{i\in A}$ and $[\xi_i]_{i\in B}$ given $[\xi_i]_{i\in C}$ (for every disjoint $A, B, C \subset N$) which can be considered as a good measure of conditional stochastic dependence. Second, one can introduce reasonable measures of dependence of level r among variables $[\xi_i]_{i\in A}$ (where $A \subset N$, $1 \leq r < \operatorname{card} A$) which are expressible by means of the multiinformation function. Third, it enables one to derive theoretical results on (nonexistence of an) axiomatic characterization of stochastic conditional independence models.

1. Introduction

Information theory provides a good measure of stochastic dependence between two random variables, namely the *mutual information* [7, 3]. It is always nonnegative and vanishes iff the corresponding two random variables are stochastically independent. On the other hand it achieves its maximal value iff one random variable is a function of the other variable [28].

Perez [15] wanted also to express numerically the degree of stochastic dependence among any finite number of random variables and proposed a numerical characteristic called 'dependence tightness'. Later he changed the terminology, started to call that characteristic systematically *multiinformation* and encouraged research on asymptotic properties of an estimator of multiinformation [18]. Note that multiinformation somehow appeared in earlier information-theoretical papers. For example, Watanabe [24] called it 'total correlation' and Csiszár [2] showed that the IPFP procedure converges to the probability distribution minimizing multiinformation within the considered family of distributions having prescribed marginals.

Further prospects occur when one considers multiinformation as a set function. That means if $[\xi_i]_{i\in N}$ is a collection of random variables indexed by a finite set N then the *multiinformation function* (corresponding to $[\xi_i]_{i\in N}$) assigns the multiinformation of the subcollection $[\xi_i]_{i\in A}$ to every $A \subset N$. Such a function was mentioned already in sixties by Watanabe [25] under name 'total cohesion function'. Some pleasant properties of the multiinformation function were utilized by Perez [15] in probabilistic decision-making. Malvestuto named the multiinformation function 'entaxy' and applied it in the theory of relational databases [9]. The multiinformation function plays an important role in the problem of finding 'optimal dependence structure simplification' solved in thesis [21], too. Finally, it has appeared to be a very useful tool for studying of formal properties of conditional independence.

The first author in modern statistics to deal with those formal properties of conditional independence was probably Dawid [5]. He characterized certain statistical concepts (e.g. the concept of sufficient statistics) in terms of generalized stochastic conditional independence. Spohn [17] studied stochastic conditional independence from the viewpoint of philosophical logic and formulated the same properties as Dawid. The importance of conditional independence in probabilistic reasoning was explicitly discerned and highlighted by Pearl and Paz [13]. They interpreted Dawid's formal properties in terms of axioms for irrelevance models and formulated a natural conjecture that these properties characterize stochastic conditional independence models. This conjecture was refuted in [19] by substantial use of the multiinformation function and this result was later strengthened by showing that stochastic conditional independence models cannot be characterized by a finite number of formal properties of that type [20].

However, as we have already mentioned, the original prospect of multiinformation was to express quantitatively the strength of dependence among random variables. An abstract view on measures of dependence was brought by Rényi [16] who formulated a few reasonable requirements on measures of dependence of two real-valued random variables. Zvárová [28] studied in more detail information-theoretical measures of dependence including mutual information. The idea of measuring dependence appeared also in nonprobabilistic calculi for dealing with uncertainty in artificial intelligence [22, 23].

This article is basically an overview paper, but it brings several minor new results which (as we hope) support our claims about the usefulness of the multiinformation function. The basic fact here is that the multiinformation function is related to *conditional mutual information*. In the first part of the paper we show that the conditional mutual information complies with several reasonable requirements (analogous to Rényi's conditions) which should be satisfied by a measure of degree of stochastic conditional dependence.

The second part of the paper responds to an interesting idea brought by Naftali Tishby and Joachim Buhmann in Erice during the workshop. Is it possible to decompose multiinformation (which is considered to be a measure of global dependence) into level-specific measures of dependence among variables? That means one would like to measure the strength of interactions of the 'first level' by a special measure of *pairwise dependence*, and similarly for interactions of 'higher levels'. We show that the multiinformation can indeed be viewed as a sum of such level-specific measures of dependence. Nevertheless, we have found recently that such a formula is not completely new: similar level-specific measures of dependence were already considered by Han [8].

Finally, in the third part of the paper, as an example of theoretical use of the multiinformation function we recall the results about nonexistence of an axiomatic characterization of conditional independence models. Unlike the original paper [20] we present a long didactive proof emphasizing the essential steps.

Note that all results of the paper are formulated for random variables taking a finite number of values although the multiinformation function can be used also in the case of continuous variables. The reason is that we wish to present really elementary proofs which are not complicated by measure-theoretical technicalities.

2. Basic concepts

We recall well-known information-theoretical concepts in this section; the most of them can be found in textbooks, e.g. [3]. The reader who is familiar with information theory can skip the section.

Throughout the paper N denotes a finite nonempty set of factors or shortly a *factor set*. In the sequel, whenever $A, B \subset N$ the juxtaposition AB will be used to shorten the notation for the set union $A \cup B$ and for any $i \in N$, the singleton will be sometimes denoted by i instead of $\{i\}$.

2.1. DISCRETE PROBABILITY DISTRIBUTIONS

The factors should correspond to discrete random variables. A discrete random variable ξ_i corresponding to a factor $i \in N$ has to take values in a nonempty finite set $\mathbf{X}_i$ called the *frame* for i. Under situation when a fixed frame $\mathbf{X}_i$ is assigned to every factor $i \in N$ and $\emptyset \neq A \subset N$ the symbol $\mathbf{X}_A$ denotes the Cartesian product $\prod_{i \in A} \mathbf{X}_i$, that is the frame for A. Whenever $\emptyset \neq B \subset A \subset N$ and $x \in \mathbf{X}_A$, then its coordinate projection to $\mathbf{X}_B$ will be denoted by x_B.

By a *probability distribution on* a nonempty finite set $\mathbf{Y}$ we understand every nonnegative real function P on $\mathbf{Y}$ with $\sum\{P(y)\,;\, y \in \mathbf{Y}\} = 1$. By a (discrete) *probability distribution over* a factor set N is understood any probability distribution on $\mathbf{X}_N$ where $\{\mathbf{X}_i\,;\, i \in N\}$ is an arbitrary collection of frames. Or equivalently, any particular joint distribution of a discrete random vector $[\xi_i]_{i \in N}$.

Having $\emptyset \neq A \subset N$ and a probability distribution P over N its *marginal distribution* P^A is a probability distribution over A defined as follows:

$$P^A(a) \;=\; \sum\{P(a,b)\,;\, b \in \mathbf{X}_{N \setminus A}\} \quad \text{for every } a \in \mathbf{X}_A\,.$$

It describes the distribution of the random subvector $[\xi_i]_{i \in A}$. In the sequel we accept a natural convention $P^{\emptyset} \equiv 1$.

Having $\emptyset \neq B \subset N$ and $b \in \mathbf{X}_B$ such that $P^B(b) > 0$ the *conditional distribution* $P^{|\,b}$ is a probability distribution over $N \setminus B$ defined by:

$$P^{|\,b}(a) \;=\; \frac{P(a,b)}{P^B(b)} \quad \text{for every } a \in \mathbf{X}_{N \setminus B}\,.$$

It describes the (conditional) distribution of $[\xi_i]_{i \in N \setminus B}$ under the condition $[\xi_i]_{i \in B} \equiv b$. For disjoint $A, B \subset N$ one can use the symbol $P^{A|\,b}$ to denote $(P^{AB})^{|\,b} = (P^{|\,b})^A$.

Every mapping between frames induces a transformation of distributions. Supposing P is a probability distribution on $\mathbf{Y}$ and $f : \mathbf{Y} \to \mathbf{Z}$ is a mapping into a nonempty finite set $\mathbf{Z}$, the formula

$$Q(z) \;=\; \sum\{P(y)\,;\, y \in \mathbf{Y} \;\&\; f(y) = z\} \quad \text{for every } z \in \mathbf{Z},$$

defines a probability distribution on $\mathbf{Z}$. In such a case we say that Q is an *image* of P (by f). Provided P is the distribution of a random vector ξ, Q is the distribution of the transformed vector $f(\xi)$.

Supposing $A, B \subset N$ are disjoint, P is a distribution over N we say that A is *functionally dependent* on B with respect to P and write $B \to A\,(P)$ if there exists a mapping $f : \mathbf{X}_B \to \mathbf{X}_A$ such that

$$\begin{array}{ll} P^{AB}(a,b) = P^B(b) & \text{for } a = f(b),\ b \in \mathbf{X}_B\,, \\ P^{AB}(a,b) = 0 & \text{for remaining } a \in \mathbf{X}_A,\ b \in \mathbf{X}_B\,. \end{array}$$

It reflects the situation when P is the distribution of a random vector $[\xi_i]_{i\in N}$ whose random subvector $[\xi_i]_{i\in A}$ is a deterministic function of another random subvector $[\xi_i]_{i\in B}$. Note that the function f is uniquely determined on the set $\{b \in \mathbf{X}_B\,;\ P^B(b) > 0\}$, outside that set it can take arbitrary values.

Supposing $A, B, C \subset N$ are disjoint and P is a distribution over N we say that A is *conditionally independent* of B given C with respect to P and write $A \perp\!\!\!\perp B|C\,(P)$ if the equality

$$P^{ABC}(a,b,c) \cdot P^C(c) = P^{AC}(a,c) \cdot P^{BC}(b,c)$$

holds for every $a \in \mathbf{X}_A$, $b \in \mathbf{X}_B$, $c \in \mathbf{X}_C$. It describes the situation when P is the distribution of a random vector $[\xi_i]_{i\in N}$ and in every situation when the values of $[\xi_i]_{i\in C}$ are known the values of $[\xi_i]_{i\in A}$ and $[\xi_i]_{i\in B}$ are completely unrelated (from a stochastic point of view).

2.2. INDEPENDENCY MODELS

The symbol $\mathcal{T}(N)$ will denote the collection of ordered triplets $\langle A, B|C\rangle$ of pairwise disjoint subsets of a factor set N, where $A \neq \emptyset \neq B$. These triplets will serve for identification of conditional independence statements within the factor set N.

In general, an *independency model* over N is a subset of the class $\mathcal{T}(N)$. Supposing $\langle A, B|C\rangle \in \mathcal{T}(N)$, its symmetric image is the triplet $\langle B, A|C\rangle \in \mathcal{T}(N)$. The *symmetric closure* of an independency model $\mathcal{I} \subset \mathcal{T}(N)$ is the class of triplets in $\mathcal{I}$ and their symmetric images.

The independency model *induced by a probability distribution* P over N consists just of those triplets $\langle A, B|C\rangle \in \mathcal{T}(N)$ such that $A \perp\!\!\!\perp B|C\,(P)$. *A probabilistic independency model* (over N) is an independency model induced by some probability distribution over N.

Lemma 2.1 *Supposing $\mathcal{I}, \mathcal{J} \subset \mathcal{T}(N)$ are probabilistic independency models the class $\mathcal{I} \cap \mathcal{J}$ is also a probabilistic independency model.*

Proof: Let P be a probability distribution on $\mathbf{X}_N$ inducing $\mathcal{I}$ and Q be a probability distribution on $\mathbf{Y}_N \equiv \prod_{i\in N} \mathbf{Y}_i$ inducing $\mathcal{J}$. Put $\mathbf{Z}_i = \mathbf{X}_i \times \mathbf{Y}_i$ for every $i \in N$ and define

$$R([x_i, y_i]_{i\in N}) \;=\; P([x_i]_{i\in N}) \cdot Q([y_i]_{i\in N}) \quad \text{for } [x_i, y_i]_{i\in N} \in \mathbf{Z}_N\,.$$

It is easy to verify that for every $\langle A,B|C\rangle \in \mathcal{T}(N)$ one has $A \perp\!\!\!\perp B|C\,(R)$ iff $[\,A \perp\!\!\!\perp B|C\,(P)\ \&\ A \perp\!\!\!\perp B|C\,(Q)\,]$. $\square$

2.3. RELATIVE ENTROPY

Supposing Q and R are probability distributions on a nonempty finite set $\mathbf{Y}$ we say that Q is *absolutely continuous* with respect to R iff $R(y) = 0$ implies $Q(y) = 0$ for every $y \in \mathbf{Y}$. In that case we can define the *relative entropy* of Q with respect to R as

$$\mathcal{H}(Q|R) = \sum \{\, Q(y) \cdot \ln \frac{Q(y)}{R(y)}\ ;\ y \in \mathbf{Y}\ \&\ Q(y) > 0\,\}\,.$$

Lemma 2.2 *Suppose that Q and R are probability distributions on a nonempty finite set $\mathbf{Y}$ such that Q is absolutely continuous with respect to R. Then*
(a) $\mathcal{H}(Q|R) \geq 0$,
(b) $\mathcal{H}(Q|R) = 0$ iff $Q = R$.

Proof: Consider the real function φ on the interval $[0,\infty)$ defined by
$\varphi(z) = z \cdot \ln z$ for $z > 0$, $\varphi(0) = 0$,
and the function $h : \mathbf{Y} \to [0,\infty)$ defined by
$h(y) = Q(y)/R(y)$ if $R(y) > 0$, $h(y) = 0$ otherwise.
Since φ is a continuous strictly convex function, one can use the well-known Jensen's inequality [3] with respect to R and write:

$$0 = \varphi(1) = \varphi(\sum_{y\in\mathbf{Y}} h(y) \cdot R(y)) \leq \sum_{y\in\mathbf{Y}} \varphi(h(y)) \cdot R(y) = \mathcal{H}(Q|R).$$

Owing to strict convexity of φ the equality holds iff h is constant on the set $\{y \in \mathbf{Y}; R(y) > 0\}$. That means $h \equiv 1$ there, i.e. $Q = R$. $\square$

Supposing that $\langle A,B|C\rangle \in \mathcal{T}(N)$ and P is a probability distribution over N the formula

$$\begin{array}{rcll} \hat{P}(x) & = & \frac{P^{AC}(x_{AC})\cdot P^{BC}(x_{BC})}{P^{C}(x_C)} & \text{for } x \in \mathbf{X}_{ABC} \text{ with } P^C(x_C) > 0, \\ \hat{P}(x) & = & 0 & \text{for remaining } x \in \mathbf{X}_{ABC}\,. \end{array} \tag{1}$$

defines a probability distribution on $\mathbf{X}_{ABC}$. Evidently, P^{ABC} is absolutely continuous with respect to $\hat{P}$. The *conditional mutual information* between A and B given C with respect to P, denoted by $I(A;B|C\,\|\,P)$ is the relative entropy of P^{ABC} with respect to $\hat{P}$. In case that P is known from the context we write just $I(A;B|C)$.

Consequence 2.1 *Supposing that $\langle A,B|C\rangle \in \mathcal{T}(N)$ and P is a probability distribution over N one has*
(a) $I(A;B|C\,\|\,P) \geq 0$,
(b) $I(A;B|C\,\|\,P) = 0$ *iff* $A \perp\!\!\!\perp B|C\,(P)$.

Proof: Owing to Lemma 2.2 it suffices to realize that $P^{ABC} = \hat{P}$ means nothing but the corresponding conditional independence statement. □

2.4. MULTIINFORMATION FUNCTION

The *multiinformation function* induced by a probability distribution P (over a factor set N) is a real function on the power set of N defined as follows:

$$M(D\,\|\,P) = \mathcal{H}(P^D|\prod_{i\in D} P^{\{i\}}) \quad \text{for } \emptyset \neq D \subset N, \quad M(\emptyset\,\|\,P) = 0\,.$$

We again omit the symbol of P when the probability distribution is clear from the context. It follows from Lemma 2.2(b) that $M(D) = 0$ whenever $\operatorname{card} D = 1$.

Lemma 2.3 *Let $\langle A,B|C\rangle \in \mathcal{T}(N)$ and P be a probability distribution over N. Then*

$$I(A;B|C) = M(ABC) + M(C) - M(AC) - M(BC)\,. \tag{2}$$

Proof: Let us write $\mathcal{H}(P^{ABC}|\hat{P})$ as

$$\sum \{P^{ABC}(x)\cdot \ln \frac{P^{ABC}(x)\cdot P^{C}(x_C)}{P^{AC}(x_{AC})\cdot P^{BC}(x_{BC})}\,;\; x \in \mathbf{X}_{ABC}\ \&\ P^{ABC}(x) > 0\}\,.$$

Now we can artificially multiply both the numerator and the denominator of the ratio in the argument of the logarithm by a special product $\prod_{i\in A} P^{\{i\}}(x_i)\cdot\prod_{i\in B} P^{\{i\}}(x_i)\cdot\prod_{i\in C} P^{\{i\}}(x_i)\cdot\prod_{i\in C} P^{\{i\}}(x_i)$ which is always strictly positive for any considered configuration x. Using well-known properties of logarithm one can write it as a sum of four terms:

$$\begin{aligned}
&\sum\{P^{ABC}(x)\cdot\ln\frac{P^{ABC}(x)}{\prod_{i\in ABC}P^{\{i\}}(x_i)}\,;\; x\in\mathbf{X}_{ABC}\ \&\ P^{ABC}(x)>0\}\\
+\;&\sum\{P^{ABC}(x)\cdot\ln\frac{P^{C}(x_C)}{\prod_{i\in C}P^{\{i\}}(x_i)}\,;\; x\in\mathbf{X}_{ABC}\ \&\ P^{ABC}(x)>0\}\\
-\;&\sum\{P^{ABC}(x)\cdot\ln\frac{P^{AC}(x_{AC})}{\prod_{i\in AC}P^{\{i\}}(x_i)}\,;\; x\in\mathbf{X}_{ABC}\ \&\ P^{ABC}(x)>0\}\\
-\;&\sum\{P^{ABC}(x)\cdot\ln\frac{P^{BC}(x_{BC})}{\prod_{i\in BC}P^{\{i\}}(x_i)}\,;\; x\in\mathbf{X}_{ABC}\ \&\ P^{ABC}(x)>0\}.
\end{aligned}$$

The first term is nothing but the value of the multiinformation function for ABC. To see that the second term is $M(C)$ one can sum there in groups of configurations for which the corresponding logarithm has the same value, that is groups of xs having the same projection to C:

$$\sum_{\substack{x_C \in \mathbf{X}_C \\ P^C(x_C)>0}} \; \sum_{\substack{y \in \mathbf{X}_{AB} \\ P^{ABC}(y,x_C)>0}} P^{ABC}(y,x_C) \cdot \ln \frac{P^C(x_C)}{\prod_{i\in C} P^{\{i\}}(x_i)} =$$

$$= \sum_{\substack{x_C \in \mathbf{X}_C \\ P^C(x_C)>0}} \ln \frac{P^C(x_C)}{\prod_{i\in C} P^{\{i\}}(x_i)} \cdot \sum_{\substack{y \in \mathbf{X}_{AB} \\ P^{ABC}(y,x_C)>0}} P^{ABC}(y,x_C) =$$

$$= \sum_{\substack{x_C \in \mathbf{X}_C \\ P^C(x_C)>0}} \ln \frac{P^C(x_C)}{\prod_{i\in C} P^{\{i\}}(x_i)} \cdot P^C(x_C)\,.$$

Similarly for the other two terms. $\square$

2.5. ENTROPY AND CONDITIONAL ENTROPY

If Q is a discrete probability distribution on a nonempty finite set $\mathbf{Y}$ the *entropy* of Q is defined by the formula

$$\mathcal{H}(Q) = \sum \{\, Q(y) \cdot \ln \frac{1}{Q(y)} \;;\; y \in \mathbf{Y} \;\&\; Q(y) > 0 \,\}\,.$$

Lemma 2.4 *Suppose that Q is a discrete probability distribution on a nonempty finite set $\mathbf{Y}$. Then*
(a) $\mathcal{H}(Q) \geq 0$,
(b) $\mathcal{H}(Q) = 0$ *iff there exists $y \in \mathbf{Y}$ such that $Q(y) = 1$.*

Proof: Since logarithm is an increasing real function one has $\ln Q(y)^{-1} \geq 0$ for every $y \in \mathbf{Y}$ with $Q(y) > 0$. Hence $Q(y) \cdot \ln Q(y)^{-1} \geq 0$ for every such y; the equality occurs here only if $Q(y) = 1$. It gives both (a) and (b). $\square$

The *entropic function* induced by a probability distribution P over a factor set N is a real function on the power set of N defined as follows:

$$H(D \,\|\, P) = \mathcal{H}(P^D) \quad \text{for } \emptyset \neq D \subset N, \qquad H(\emptyset \,\|\, P) = 0\,.$$

We will often omit the symbol of P when it is clear from the context. By using the same procedure as in the proof of Lemma 2.3 it is not difficult to see that

$$M(D) = -H(D) + \sum_{i \in D} H(\{i\}) \quad \text{for every } D \subset N.$$

Hence, using the formula (2) from Lemma 2.3 one derives

$$I(A;B|C) = -H(ABC) - H(C) + H(AC) + H(BC)\,. \tag{3}$$

Supposing $A, B \subset N$ are disjoint the *conditional entropy* of A given B is defined as a simple difference

$$H(A|B) = H(AB) - H(B)\,.$$

We use the symbol $H(A|B \,\|\, P)$ to indicate the corresponding probability distribution P.

Lemma 2.5 *Let P be a probability distribution over N, $A, B \subset N$ are disjoint. Then*

$$H(A|B \,\|\, P) \;=\; \sum \{\, P^B(b) \cdot H(A \,\|\, P^{|\,b}) \;;\; b \in \mathbf{X}_B \;\&\; P^B(b) > 0 \,\}\,. \tag{4}$$

Proof: One can easily see using the method used in the proof of Lemma 2.3 that the expression

$$\sum \{\, P^{AB}(ab) \cdot \ln \frac{P^B(b)}{P^{AB}(ab)} \;;\; a \in \mathbf{X}_A \;\&\; b \in \mathbf{X}_B \;\&\; P^{AB}(ab) > 0 \,\}$$

is nothing but $H(A|B \,\|\, P)$. On the other hand, one can utilize the definition of $P^{A|\,b}$ and write it in the form

$$\sum_{\substack{b \in \mathbf{X}_B \\ P^B(b) > 0}} P^B(b) \cdot \sum_{\substack{a \in \mathbf{X}_A \\ P^{A|\,b}(a) > 0}} P^{A|\,b}(a) \cdot \ln \frac{1}{P^{A|\,b}(a)}\,,$$

which gives the expression from (4). $\square$

3. Measure of conditional stochastic dependence

In this section we give several arguments why conditional mutual information should be considered as a suitable quantitative measure of degree of conditional stochastic dependence.

To motivate this topic let us consider the following specific task. Suppose that ξ_A, ξ_B, ξ_C are discrete random vectors and the joint distributions of ξ_{AC} and ξ_{BC} are already known (fixed or prescribed). What are then possible values for the conditional mutual information $I(A;B|C)$? By Consequence 2.1 zero is a lower bound for those values, and it is the precise bound since one can always find a distribution having prescribed marginals for AC and BC such that $I(A;B|C)=0$ (namely the 'conditional product' $\hat{P}$ given by the formula (1)).

3.1. MAXIMAL DEGREE OF CONDITIONAL DEPENDENCE

But one can also find an upper bound.

Lemma 3.1 *Let $\langle A,B|C\rangle \in \mathcal{T}(N)$ and P be a probability distribution over N. Then*

$$I(A;B|C) \le \min\{\, H(A|C)\,,\, H(B|C)\,\}\,.$$

Proof: It follows from (3) with help of the definition of conditional entropy

$$I(A;B|C) = H(A|C) - H(A|BC)\,.$$

Moreover, $0 \le H(A|BC)$ follows from (4) with Lemma 2.4(a). This implies $I(A;B|C) \le H(A|C)$, the other estimate with $H(B|C)$ is analogous. □

The following proposition generalizes an analogous result obtained in the unconditional case by Zvárová ([28], Theorem 5) and loosely corresponds to the condition E) mentioned by Rényi [16].

Proposition 3.1 *Supposing $\langle A,B|C\rangle \in \mathcal{T}(N)$ and P is a probability distribution over N one has*

$$I(A;B|C\,\|\,P) = H(A|C\,\|\,P) \quad \textit{iff} \quad BC \to A\,(P)\,.$$

Proof: By the formula mentioned in the proof of Lemma 3.1 the considered equality occurs just in case $H(A|BC\,\|\,P)=0$. Owing to the formula (4) and Lemma 2.4(a) this is equivalent to the requirement $H(A\,\|\,P^{|\,bc})=0$ for every $(b,c)\in \mathbf{X}_{BC}$ with $P^{BC}(b,c)>0$. By Lemma 2.4(b) it means just that for every such a couple $(b,c)\in\mathbf{X}_{BC}$ there exists $a\in\mathbf{X}_A$ with $P^{A|\,bc}(a)=1$. Of course, this $a\in\mathbf{X}_A$ is uniquely determined. This enables us to define the required function from $\mathbf{X}_{BC}$ to $\mathbf{X}_A$. □

A natural question arises how tight is the upper bound for $I(A;B|C)$ from Lemma 3.1. More exactly, we ask whether one can always find a distribution having prescribed marginals for AC and BC with $I(A;B|C)=\min\{H(A|C),H(B|C)\}$. In general, the answer is negative as shown by the following example.

Example 3.1 Let us put $\mathbf{X}_A = \mathbf{X}_B = \mathbf{X}_C = \{0,1\}$ and define P_{AC} and P_{BC} as follows

$$P_{AC}(0,0) = \tfrac{1}{3}\,,\ P_{AC}(0,1) = P_{AC}(1,1) = \tfrac{1}{4}\,,\ P_{AC}(1,0) = \tfrac{1}{6}\,,$$
$$P_{BC}(0,0) = P_{BC}(0,1) = P_{BC}(1,0) = P_{BC}(1,1) = \tfrac{1}{4}\,.$$

Since $(P_{AC})^C = (P_{BC})^C$ there exists a distribution on $\mathbf{X}_{ABC}$ having them as marginals. In fact, any such distribution P can be expressed as follows

$$\begin{aligned}
P(0,0,0) &= \alpha, \\
P(0,0,1) &= \beta, \\
P(0,1,0) &= \tfrac{1}{3} - \alpha, \\
P(0,1,1) &= \tfrac{1}{4} - \beta, \\
P(1,0,0) &= \tfrac{1}{4} - \alpha, \\
P(1,0,1) &= \tfrac{1}{4} - \beta, \\
P(1,1,0) &= \alpha - \tfrac{1}{12}, \\
P(1,1,1) &= \beta,
\end{aligned}$$

where $\alpha \in [\frac{1}{12}, \frac{1}{4}], \beta \in [0, \frac{1}{4}]$. It is easy to show that $H(A|C) < H(B|C)$. On the other hand, for every parameter α either $P(0,0,0)$ and $P(1,0,0)$ are simultaneously nonzero or $P(0,1,0)$ and $P(1,1,0)$ are simultaneously nonzero. Therefore A is not functionally dependent on BC with respect to P and by Proposition 3.1 the upper bound $H(A|C)$ is not achieved. $\diamond$

However, the upper bound given in Lemma 3.1 can be precise for specific prescribed marginals. Let us provide a general example.

Example 3.2 Suppose that P_{BC} is given, consider an arbitrary function $g : \mathbf{X}_B \to \mathbf{X}_A$ and define P_{AC} by the formula

$$P_{AC}(a,c) = \sum \{\, P_{BC}(b,c)\ ;\ b \in \mathbf{X}_B\ \&\ g(b) = a \,\} \quad \text{for } a \in \mathbf{X}_A,\ c \in \mathbf{X}_C\,.$$

Well, one can always find a distribution P over ABC having such a couple of distributions P_{AC}, P_{BC} as marginals and satisfying $I(A;B|C \,\|\, P) = H(A|C \,\|\, P)$. Indeed, define P over ABC as follows:

$$\begin{array}{ll}
P(a,b,c) = P_{BC}(b,c) & \text{if } g(b) = a, \\
P(a,b,c) = 0 & \text{otherwise.}
\end{array}$$

This ensures that $BC \to A\,(P)$, then use Proposition 3.1. $\diamond$

3.2. MUTUAL COMPARISON OF DEPENDENCE DEGREES

A natural intuitive requirement on a quantitative characteristic of degree of dependence is that a higher degree of dependence among variables should be reflected by a higher value of that characteristic. Previous results on conditional mutual information are in agreement with this wish: its minimal value characterizes independence, while its maximal values more or less corresponds to the maximal degree of dependence.

Well, what about the behavior 'between' these 'extreme' cases? On can imagine two 'comparable' nonextreme cases when one case represents evidently a higher degree of dependence among variables than the other case. For example, let us consider two random vectors ξ_{AB} resp. η_{AB} (take $C = \emptyset$) having distributions P_{AB} resp. Q_{AB} depicted by the following diagrams.

P_{AB}

0	$\frac{1}{7}$	$\frac{1}{7}$
$\frac{1}{7}$	$\frac{1}{7}$	$\frac{1}{7}$
$\frac{1}{7}$	$\frac{1}{7}$	0

Q_{AB}

0	0	$\frac{2}{7}$
$\frac{1}{7}$	$\frac{2}{7}$	0
$\frac{1}{7}$	$\frac{1}{7}$	0

Clearly, $(P_{AB})^A = (Q_{AB})^A$ and $(P_{AB})^B = (Q_{AB})^B$. But intuitively, Q_{AB} expresses a higher degree of stochastic dependence between $\eta_A = \xi_A$ and $\eta_B = \xi_B$ than P_{AB}. The distribution Q_{AB} is more 'concetrated' than P_{AB}: Q_{AB} is an image of P_{AB}. Therefore, we can anticipate $I(A; B|\emptyset \,\|\, P) \leq I(A; B|\emptyset \,\|\, Q)$, which is indeed the case.

The following proposition says that conditional mutual information has the desired property. Note that the property is not derivable from other properties of measures of dependence mentioned either by Rényi [16] or by Zvárová [28] (in the unconditional case).

Proposition 3.2 *Suppose that $\langle A, B|C\rangle \in \mathcal{T}(N)$ and P, Q are probability distributions over N such that $P^{AC} = Q^{AC}$, $P^{BC} = Q^{BC}$ and Q^{ABC} is an image of P^{ABC}. Then*

$$I(A; B|C \,\|\, P) \leq I(A; B|C \,\|\, Q)\,.$$

Proof: Let us write P instead of P^{ABC} throughout the proof and similarly for Q. Suppose that Q is an image of P by $f : \mathbf{X}_{ABC} \to \mathbf{X}_{ABC}$. For every

$x \in \mathbf{X}_{ABC}$ with $Q(x) > 0$ put $T = \{y \in \mathbf{X}_{ABC}\,;\, f(y) = x \ \&\ P(y) > 0\}$ and write (owing to the fact that the logarithm is an increasing function):

$$\sum_{y \in T} P(y) \cdot \ln P(y) \leq \sum_{y \in T} P(y) \cdot \ln \left(\sum_{z \in T} P(z) \right) = Q(x) \cdot \ln Q(x)\,.$$

We can sum it over all such xs and derive

$$\sum_{\substack{y \in \mathbf{X}_{ABC} \\ P(y)>0}} P(y) \cdot \ln P(y) \leq \sum_{\substack{x \in \mathbf{X}_{ABC} \\ Q(x)>0}} Q(x) \cdot \ln Q(x)\,.$$

Hence

$$-H(ABC \,\|\, P) \leq -H(ABC \,\|\, Q)\,.$$

Owing to the assumptions $P^{AC} = Q^{AC}$, $P^{BC} = Q^{BC}$ one has $H(AC \,\|\, P) = H(AC \,\|\, Q), H(BC \,\|\, P) = H(BC \,\|\, Q)$ and $H(C \,\|\, P) = H(C \,\|\, Q)$. The formula (3) then gives the desired claim. $\square$

Nevertheless, the mentioned inequality from Proposition 3.2 may not hold when the assumption that marginals for AC and BC coincide is released, as demonstrated by the following example.

Example 3.3 Take $C = \emptyset$ and consider the distributions P_{AB} and Q_{AB} depicted by the following diagrams:

P_{AB}

$\frac{3}{8}$	$\frac{1}{8}$
$\frac{1}{8}$	$\frac{3}{8}$

Q_{AB}

0	$\frac{1}{2}$
$\frac{1}{8}$	$\frac{3}{8}$

Evidently, Q_{AB} is an image of P_{AB}, but $I(A; B|\emptyset \,\|\, P) > I(A; B|\emptyset \,\|\, Q)$. $\diamond$

Remark One can imagine more general transformations of distributions: instead of 'functional' transformations introduced in subsection 2.1 one can consider transformations by Markov kernels. However, Proposition 3.2 cannot be generalized to such a case. In fact, the distribution P_{AB} from the motivation example starting this subsection can be obtained from Q_{AB} by an 'inverse' transformation realized by a Markov kernel.

Rényi's condition F) in [16] states that a one-to-one transformation of a random variable does not change the value of a measure of dependence. Similarly, Zvárová [28] requires that restrictions to sub-σ-algebras (which somehow correspond to separate simplifying transformations of variables) decrease the value of the measure of dependence.

The above mentioned requirements can be generalized to the 'conditional' case as shown in the following proposition. Note that the assumption of the proposition means (under the situation when P is the distribution of a random vector $[\xi_i]_{i\in N}$) simply that the random subvector $[\xi_i]_{i\in A}$ is transformed while the other variables ξ_i, $i \in BC$ are preserved.

Proposition 3.3 *Let $\langle A,B|C\rangle$, $\langle D,B|C\rangle \in \mathcal{T}(N)$, P,Q be probability distributions over N. Suppose that there exists a mapping $g : \mathbf{X}_A \to \mathbf{X}_D$ such that Q^{DBC} is an image of P^{ABC} by the mapping $f : \mathbf{X}_{ABC} \to \mathbf{X}_{DBC}$ defined by*

$$f(a,b,c) = [\,g(a),b,c\,] \quad \text{for } a \in \mathbf{X}_A\,, \;\; (b,c) \in \mathbf{X}_{BC}\,.$$

Then

$$I(A;B|C\,\|\,P) \geq I(D;B|C\,\|\,Q)\,.$$

Proof: Throughout the proof we write P instead of P^{ABC} and Q instead of Q^{DBC}. Let us denote by $\mathbf{Y}$ the class of all $(c,d) \in \mathbf{X}_{CD}$ such that $P(g_{-1}(d) \times \mathbf{X}_B \times \{c\}) > 0$ where $g_{-1}(d) = \{a \in \mathbf{X}_A; g(a) = d\}$. For every $(c,d) \in \mathbf{Y}$ introduce a probability distribution R_{cd} on $g_{-1}(d) \times \mathbf{X}_B$ by the formula:

$$R_{cd}(a,b) = \frac{P(a,b,c)}{P(g_{-1}(d) \times \mathbf{X}_B \times \{c\})} \qquad \text{for } a \in g_{-1}(d),\, b \in \mathbf{X}_B\,.$$

It can be formally considered as a distribution on $\mathbf{X}_A \times \mathbf{X}_B$. Thus, by Consequence 2.1(a) we have

$$0 \leq I(A;B|\emptyset\,\|\,R_{cd}) \quad \text{for every } (c,d) \in \mathbf{Y}\,.$$

One can multiply this inequality by $P(g_{-1}(d) \times \mathbf{X}_B \times \{c\})$, sum over $\mathbf{Y}$ and obtain by simple cancellation of $P(g_{-1}(d) \times \mathbf{X}_B \times \{c\})$:

$$0 \leq \sum_{(c,d)\in \mathbf{Y}} \; \sum_{\substack{(a,b)\in g_{-1}(d)\times \mathbf{X}_B \\ P(abc)>0}} P(abc) \cdot \ln \frac{P(abc) \cdot P(g_{-1}(d) \times \mathbf{X}_B \times \{c\})}{P(\{a\} \times \mathbf{X}_B \times \{c\}) \cdot P(g_{-1}(d) \times \{b\} \times \{c\})}\,.$$

One can apply basic properties of the logarithm and write the right-hand side of the obtained inequality as a sum of four terms (as in the proof of Lemma 2.3). We leave it to the reader to verify that each of these terms is certain entropy (possibly with the minus sign). We just give hints indicating formally the way of summation.

$$\sum_c \sum_d \sum_b \sum_a P(abc) \cdot \ln P(abc) = \sum_c \sum_b \sum_d \sum_a \ldots = \sum_c \sum_b \sum_a \ldots = -H(ABC \,\|\, P)$$
$$-\sum_c \sum_d \sum_a \sum_b P(abc) \cdot \ln P(\{a\} \times \mathbf{X}_B \times \{c\}) = \sum_c \sum_d \sum_a \ldots = \sum_c \sum_a \ldots = H(AC \,\|\, P)$$
$$-\sum_c \sum_d \sum_b \sum_a P(abc) \cdot \ln P(g_{-1}(d) \times \{b\} \times \{c\}) = \sum_c \sum_d \sum_b \ldots = H(DBC \,\|\, Q)$$
$$\sum_c \sum_d \sum_a \sum_b P(abc) \cdot \ln P(g_{-1}(d) \times \mathbf{X}_B \times \{c\}) = \sum_c \sum_d \ldots = -H(DC \,\|\, Q)$$

Thus, one can derive:

$$0 \le -H(ABC \,\|\, P) + H(AC \,\|\, P) + H(DBC \,\|\, Q) - H(DC \,\|\, Q)\,. \qquad (5)$$

Since $P^{BC} = Q^{BC}$ one also has

$$0 = H(BC \,\|\, P) - H(BC \,\|\, Q) - H(C \,\|\, P) + H(C \,\|\, Q)\,. \qquad (6)$$

Hence by summing (5) and (6) and using the formula (3)

$$0 \le I(A; B|C \,\|\, P) - I(D; B|C \,\|\, Q)\,,$$

which concludes the proof. □

If g is a one-to one mapping, one can apply Proposition 3.3 both to g and g^{-1}, from which the following consequence immediately follows (it corresponds exactly to the Rényi's requirement F)).

Consequence 3.1 *Supposing the mapping g in Proposition 3.3 is a one-to-one mapping one has*

$$I(A; B|C \,\|\, P) = I(D; B|C \,\|\, Q)\,.$$

Nevertheless, Proposition 3.3 cannot be strengthened to transformations involving variables in C (more exactly transformations of the subvector $[\xi_i]_{i \in AC}$), as the following example shows.

Example 3.4 Let us put $\mathbf{X}_A = \mathbf{X}_B = \mathbf{X}_C = \mathbf{X}_D = \{0, 1\}$, $\mathbf{X}_E = \{0\}$ and define a distribution P on $\mathbf{X}_{ABC}$ as follows

$$P(0,0,0) = P(1,0,0) = P(0,1,1) = P(1,1,1) = \frac{1}{4}\,,$$

where the remaining values of P zero. Since $A \perp\!\!\!\perp B|C\,(P)$ one has by Consequence 2.1(b) $I(A;B|C\,\|\,P)=0$. Let us consider a mapping $g:\mathbf{X}_{AC}\to\mathbf{X}_{DE}$ defined by

$$g(0,0)=g(1,0)=(0,0)\qquad g(0,1)=g(1,1)=(1,0)\,.$$

Then the image of P by the mapping $f:\mathbf{X}_{ABC}\to\mathbf{X}_{DBE}$ defined by

$$f(a,b,c)=[\,g(a,c),b\,]\quad \text{for } (a,c)\in\mathbf{X}_{AC}\,,\ b\in\mathbf{X}_{B}\,,$$

is the following distribution Q on $\mathbf{X}_{DBE}$:

$$Q(0,0,0)=Q(1,1,0)=\frac{1}{2}\,,\quad Q(0,1,0)=Q(1,0,0)=0\,.$$

Evidently $I(D;B|E\,\|\,Q)=\ln 2\,.$ ◇

4. Different levels of stochastic dependence

Let us start this section with motivation. Quite common 'philosophical' point of view on stochastic dependence is the following one. Global strength of dependence among variables $[\xi_i]_{i\in N}$ is considered as a result of various *interactions* among factors in N.

For example, in hierarchical log-linear models for contingency tables [4] one can distinguish the first-order interactions, i.e. interactions of pairs of factors, the second-order interactions, i.e. interactions of triplets of factors, etc. In substance, the first-order interactions correspond to pairwise dependence relationships, i.e. to (unconditional) dependences between ξ_i and ξ_j for $i,j\in N$, $i\neq j$. Similarly, one can (very loosely) imagine that the second-order interactions correspond to conditional dependences with one conditioning variable, i.e. to conditional dependences between ξ_i and ξ_j given ξ_k where $i,j,k\in N$ are distinct. An analogous principle holds for higher-order interactions. Note that we have used the example with log-linear models just for motivation – to illustrate informally the aim of this section. In fact, one can interpret only special hierarchical log-linear models in terms of conditional (in)dependence.

Well, it leads to the idea to distinguish different 'levels' of stochastic dependence. Thus, the first level could 'involve' pairwise (unconditional) dependences. The second level could correspond to pairwise conditional dependences between two variables given a third one, the third level to pairwise conditional dependences given a couple of variables, etc. Let us give a simple example of a probability distribution which exhibits different behavior for different levels. The following construction will be used in the next section, too.

Construction A Supposing $A \subset N$, $\mathrm{card}\, A \geq 2$, there exists a probability distribution P over N such that

$$\begin{array}{ll} M(B \,\|\, P) = \ln 2 & \text{whenever } A \subset B \subset N\,, \\ M(B \,\|\, P) = 0 & \text{otherwise.} \end{array}$$

Proof: Let us put $\mathbf{X}_i = \{0,1\}$ for $i \in A$, $\mathbf{X}_i = \{0\}$ for $i \in N \setminus A$. Define P on $\mathbf{X}_N$ as follows

$$\begin{array}{ll} P([x_i]_{i\in N}) = 2^{1-\mathrm{card}\, A} & \text{whenever } \sum_{i\in N} x_i \text{ is even,} \\ P([x_i]_{i\in N}) = 0 & \text{otherwise.} \end{array}$$

□

The distribution P from Construction A exhibits only the highest-level dependences within the factor set A. Indeed, for every couple $i, j \in A$, $i \neq j$, one can easily verify (by Consequence 2.1 and Lemma 2.3) that i is conditionally independent of j given any proper subset C of $A \setminus \{i,j\}$ (with respect to P) but i is *not* conditionally independent of j given $A \setminus \{i,j\}$. Or equivalently, supposing $[\xi_i]_{i\in N}$ has the distribution P, the variables $[\xi_i]_{i\in A}$ are 'collectively dependent' although the variables $[\xi_i]_{i\in D}$, where D is arbitrary proper subset of A, are 'completely independent'. Such distributions are called in [26] *pseudo-independent models*. The main conclusion of [26] is that in the case of such an underlying model standard algorithms for learning Bayesian network approximations fail to find a suitable network. This maybe justifies a wish to measure the strength of each level of dependence separately. Good quantitative level-specific measures of dependence may help one to recognize whether a considered distribution is similar to the fearful pseudo-independent model. They can provide a good theoretical basis for necessary statistical tests.

Thus, we wish to have an analogue of the above mentioned classification of interactions by order in log-linear models together with the possibility to express numerically the degree of dependence for each level.

4.1. LEVEL-SPECIFIC MEASURES OF DEPENDENCE

In the previous section we argued that the conditional mutual information $I(A; B \mid C)$ is a good measure of stochastic conditional dependence between $[\xi_i]_{i\in A}$ and $[\xi_j]_{j\in B}$ given $[\xi_k]_{k\in C}$ where $A, B, C \subset N$ are pairwise disjoint subsets of N. In the special case, when A and B are singletons, we will get a measure $I(i; j \mid K)$ of conditional dependence between ξ_i and ξ_j given

$[\xi_k]_{k \in K}$, where $K \subset N \setminus \{i, j\}$. It leads directly to our proposal how to measure the degree of dependence for a specific level.

Suppose that P is a probability distribution over N, $A \subset N$ with $\mathrm{card}\, A \geq 2$. Then for each $r = 1, \ldots, \mathrm{card}\, A - 1$ we put:

$$\Delta(r, A \,\|\, P) = \sum \{ I(a; b | K \,\|\, P) \, ; \, \{a, b\} \subset A, \; K \subset A \setminus \{a, b\}, \; \mathrm{card}\, K = r - 1 \}.$$

If the distribution P is known from the context, we write $\Delta(r, A)$ instead of $\Delta(r, A \,\|\, P)$. Moreover, we will occasionally write just $\Delta(r)$ as a shorthand for $\Delta(r, N)$. We regard the introduced number as a basis of a measure of dependence of level r among factors from A. Consequence 2.1 directly implies:

Proposition 4.1 *Let P be a probability distribution over N, $A \subset N$, $\mathrm{card}\, A \geq 2$, $1 \leq r \leq \mathrm{card}\, A - 1$. Then*
(a) $\Delta(r, A \,\|\, P) \geq 0$,
(b) $\Delta(r, A \,\|\, P) = 0$ *iff* $[\forall \, \langle a, b | K \rangle \in \mathcal{T}(A) \; \mathrm{card}\, K = r - 1 \quad a \perp\!\!\!\perp b | K \, (P)]$.

So, the number $\Delta(r)$ is nonnegative and vanishes just in case when there are no stochastic dependences of level r. Especially, $\Delta(1)$ can be regarded as a measure of degree of pairwise unconditional dependence. The reader can ask whether there are different measures of the strength of level-specific interactions. Of course, one can find many such information-theoretical measures. However, if one is interested only in symmetric measures (i.e. measures whose values are not changed by a permutation of variables) based on entropy, then (in our opinion) the corresponding measure must be nothing but a multiple of $\Delta(r)$. We base our conjecture on the result of Han [8]: he introduced certain level-specific measures which are positive multiples of $\Delta(r)$ and proved that every entropy-based measure of multivariate 'symmetric' correlation is a linear combination of his measures with nonnegative coefficients.

Of course, owing to Lemma 2.3 the number $\Delta(r)$ can be expressed by means of the multiinformation function. To get a neat formula we introduce a provisional notation for sums of the multiinformation function over sets of the same cardinality. We denote for every $A \subset N$, $\mathrm{card}\, A \geq 2$:

$$\sigma(i, A) = \sum \{ M(D \,\|\, P) \, ; \, D \subset A, \; \mathrm{card}\, D = i \} \quad \text{for } i = 0, ..., \mathrm{card}\, A \, .$$

Of course $\sigma(i)$ will be a shorthand for $\sigma(i, N)$. Let us mention that $\sigma(0) = \sigma(1) = 0$.

Lemma 4.1 *For every $r = 1, \ldots, n - 1$ (where $n = \mathrm{card}\, N \geq 2$)*

$$\Delta(r) = \binom{r+1}{2} \cdot \sigma(r+1) - r \cdot (n - r) \cdot \sigma(r) + \binom{n - r + 1}{2} \cdot \sigma(r - 1).$$

Proof: Let us fix $1 \leq r \leq n-1$ and write by Lemma 2.3

$$2\Delta(r) = \sum_{\langle a,b|K\rangle \in \mathcal{L}} \{\, M(abK) + M(K) - M(aK) - M(bK)\,\}\,, \qquad (7)$$

where $\mathcal{L}$ is the class of all $\langle a,b|K\rangle \in \mathcal{T}(N)$ where a, b are singletons and $\mathrm{card}\, K = r-1$. Note that in $\mathcal{L}$ the triplets $\langle a,b|K\rangle$ and $\langle b,a|K\rangle$ are distinguished: hence the term $2\Delta(r)$ in (7). Evidently, the sum contains only the terms $M(D)$ such that $r-1 \leq \mathrm{card}\, D \leq r+1$, and one can write

$$\Delta(r) = \sum \{\, k(D)\cdot M(D)\,;\; D \subset N,\; r-1 \leq \mathrm{card}\, D \leq r+1\,\}\,,$$

where $k(D)$ are suitable coefficients. However, since every permutation π of factors in N transforms $\langle a,b|K\rangle \in \mathcal{L}$ into $\langle \pi(a), \pi(b)|\pi(K)\rangle \in \mathcal{L}$ the coefficient $k(D)$ depends only on $\mathrm{card}\, D$. Thus, if one divides the number of overall occurrences of terms $M(E)$ with $\mathrm{card}\, E = \mathrm{card}\, D$ in (7) by the number of sets E with $\mathrm{card}\, E = \mathrm{card}\, D$, the absolute value of $2k(D)$ is obtained. Since $\mathrm{card}\, \mathcal{L} = n\cdot(n-1)\cdot\binom{n-2}{r-1}$ one can obtain for $\mathrm{card}\, D = r+1$ that $k(D) = \frac{1}{2}\cdot n(n-1)\binom{n-2}{r-1}/\binom{n}{r+1} = \binom{r+1}{2}$. Similarly, in case $\mathrm{card}\, D = r-1$ one has $k(D) = \frac{1}{2}\cdot n(n-1)\binom{n-2}{r-1}/\binom{n}{r-1} = \binom{n-r+1}{2}$. Finally, in case $\mathrm{card}\, D = r$ one derives $-k(D) = \frac{1}{2}\cdot 2n(n-1)\binom{n-2}{r-1}/\binom{n}{r} = r(n-r)$. To get the desired formula it suffices to utilize the definitions of $\sigma(r-1)$, $\sigma(r)$, $\sigma(r+1)$. $\square$

Lemma 4.1 provides a neat formula for $\Delta(r)$, but in the case when a great number of conditional independence statements is known to hold, the definition formula is better from the computational complexity viewpoint.

4.2. DECOMPOSITION OF MULTIINFORMATION

Thus, for a factor set N, $\mathrm{card}\, N \geq 2$, the number $M(N)$ quantifies global dependence among factors in N and the numbers $\Delta(r, N)$ quantify level-specific dependences. So, one expects that the multiinformation is at least a weighted sum of these numbers. This is indeed the case, but as the reader can expect, the coefficients depend on $\mathrm{card}\, N$.

For every $n \geq 2$ and $r \in \{1, \ldots, n-1\}$ we put

$$\beta(r,n) = 2\cdot r^{-1}\cdot \binom{n}{r}^{-1}.$$

Evidently, $\beta(r,n)$ is always a strictly positive rational number.

Proposition 4.2 *Let P be a probability distribution over N,* $\mathrm{card}\, N \geq 2$. *Then*

$$M(N\,\|\,P) = \sum_{r=1}^{n-1} \beta(r,n)\cdot \Delta(r, N\,\|\,P)\,.$$

Proof: Using Lemma 4.1 we write (note that the superfluous symbol of P is omitted throughout the proof and $\beta(r)$ is used instead of $\beta(r,n)$)

$$\sum_{r=1}^{n-1} \beta(r)\cdot\Delta(r) = \sum_{r=1}^{n-1} \beta(r)\cdot\binom{r+1}{2}\cdot\sigma(r+1)$$
$$-\sum_{r=1}^{n-1} \beta(r)\cdot r\cdot(n-r)\cdot\sigma(r) + \sum_{r=1}^{n-1} \beta(r)\cdot\binom{n-r+1}{2}\cdot\sigma(r-1)\,.$$

Let us rewrite it into a more convenient form:

$$\sum_{j=2}^{n} \beta(j-1)\cdot\binom{j}{2}\cdot\sigma(j) - \sum_{j=1}^{n-1} \beta(j)\cdot j\cdot(n-j)\cdot\sigma(j) + \sum_{j=0}^{n-2} \beta(j+1)\cdot\binom{n-j}{2}\cdot\sigma(j).$$

It is, in fact, $\sum_{j=0}^{n} l(j)\cdot\sigma(j)$, where $l(j)$ are suitable coefficients. Thus,
$l(n) = \beta(n-1)\cdot\binom{n}{2} = 1$,
$l(n-1) = \beta(n-2)\cdot\binom{n-1}{2} - \beta(n-1)\cdot(n-1) = \frac{2}{n} - \frac{2}{n} = 0$,
and moreover, for every $2 \leq j \leq n-2$ one can write
$l(j) = \beta(j-1)\cdot\binom{j}{2} - \beta(j)\cdot j\cdot(n-j) + \beta(j+1)\cdot\binom{n-j}{2} =$
$= \binom{n}{j}^{-1}\cdot\{(n-j+1) - 2(n-j) + (n-j-1)\} = 0.$
Hence, owing to $\sigma(0) = \sigma(1) = 0$ and $n \geq 2$ we obtain

$$\sum_{r=1}^{n-1} \beta(r)\cdot\Delta(r) = \sum_{j=2}^{n} l(j)\cdot\sigma(j) = \sigma(n) = M(N)\,.$$

□

If one considers a subset $A \subset N$ in the role of N in the preceding statement, then one obtains

$$M(A\,\|\,P) = \sum_{r=1}^{\mathrm{card}\,A-1} \beta(r,\mathrm{card}\,A)\cdot\Delta(r,A\,\|\,P) \tag{8}$$

for every $A \subset N$, $\mathrm{card}\,A \geq 2$. One can interpret it in the following way. Whenever $[\xi_i]_{i\in A}$ is a random subvector of $[\xi_i]_{i\in N}$, then $M(A\,\|\,P)$ is a measure of global dependence among factors in A, and the value $\beta(r,\mathrm{card}\,A)\cdot\Delta(r,A\,\|\,P)$ expresses the contribution of dependences of level r among factors in A. In this sense, the coefficient $\beta(r,\mathrm{card}\,A)$ then reflects the relationship between the level r and the number of factors. Thus, the 'weights' of different levels (and their mutual ratios, too) depend on the number of factors in consideration.

The formula (8) leads to the following proposal. We propose to measure the strength of stochastic dependence among factors $A \subset N$ ($\mathrm{card}\, A \geq 2$) of level r ($1 \leq r \leq \mathrm{card}\, A - 1$) by means of the number:

$$\lambda(r, A \,\|\, P) \;=\; \beta(r, \mathrm{card}\, A) \cdot \Delta(r, A \,\|\, P)\,.$$

The symbol of P is omitted whenever it is suitable. By Proposition 4.1 $\lambda(r, A)$ is nonnegative and vanishes just in case of absence of interactions of degree r within A. The formula (8) says that $M(A)$ is just the sum of $\lambda(r, A)$s. To have a direct formula one can rewrite the definition of $\lambda(r, A)$ using Lemma 4.1 as follows:

$$\begin{aligned}\lambda(r, A) \;=\; & (a-r) \cdot \binom{a}{r+1}^{-1} \cdot \sigma(r+1, A) \\ & - 2 \cdot (a-r) \cdot \binom{a}{r}^{-1} \cdot \sigma(r, A) + (a-r) \cdot \binom{a}{r-1}^{-1} \cdot \sigma(r-1, A)\,,\end{aligned}$$

where $a = \mathrm{card}\, A$, $1 \leq r \leq a - 1$.

Let us clarify the relation to Han's measure [8] $\Delta^2 \mathbf{e}_r^{(n)}$ of level r among $n = \mathrm{card}\, N$ variables. It holds:

$$\lambda(r, N) \;=\; (n-r) \cdot \Delta^2 \mathbf{e}_r^{(n)} \quad \text{for every } 1 \leq r \leq n-1,\; n \geq 2\,.$$

We did not study the computational complexity of calculating particular characteristics introduced in this section — this can be a subject of future, more applied research.

5. Axiomatic characterization

The aim of this section is to demonstrate that the multiinformation function can be used to derive theoretical results concerning formal properties of conditional independence. For this purpose we recall the proof of the result from [20]. Moreover, we enrich the proof by introducing several concepts which (as we hope) clarify all the proof and indicate which steps are substantial. The reader may surmise that our proof is based on Consequence 2.1 and the formula from Lemma 2.3. However, these facts by themselves are not sufficient, one needs something more.

Let us describe the structure of this long section. Since the mentioned result says that probabilistic independency models cannot be characterized by means of a finite number of formal properties of (= axioms for) independency models one has to clarify thoroughly what is meant by such a formal property. This is done in subsection 5.1: first (in 5.1.1) syntactic records of those properties are introduced and illustrated by examples, and then (in

5.1.2) their meaning is explained. The aim to get rid of superfluous formal properties motivates the rest of the subsection 5.1: the situation when a formal property of independency models is a consequence of other such formal properties is analyzed in 5.1.3, 'pure' formal properties having in every situation a nontrivial meaning are treated in 5.1.4.

The subsection 5.2 is devoted to specific formal properties of probabilistic independency models. We show by an example that their validity (= probabilistic soundness) can be sometimes derived by means of the multiinformation function. The analysis in 5.2.1 leads to the proposal to limit attention to certain 'perfect' formal properties of probabilistic independency models in 5.2.2.

Finally, the subsection 5.3 contains the proof of the described nonaxiomatizability result. The method of the proof is described in 5.3.1: one has to find an infinite collection of perfect probabilistically sound formal properties of independency models. Their probabilistic soundness is verified in 5.3.2, their perfectness in 5.3.3.

5.1. FORMAL PROPERTIES OF INDEPENDENCY MODELS

We have already introduced the concept of an independency model over N as a subset of the class $\mathcal{T}(N)$ (see subsection 2.2.). This is too general a concept to be of much use. One needs to restrict oneself to special independency models which satisfy certain reasonable properties. Many authors dealing with probabilistic independency models formulated certain reasonable properties in the form of formal schemata which they named *axioms*. Since we want to prove that probabilistic independency models cannot be characterized by means of a finite number of such axioms we have to specify meticulously what is the exact meaning of such formal schemata. Thus, we both describe the syntax of those schemata and explain their semantics.

Let us start with an example. *Semigraphoid* [14] is an independency model which satisfies four formal properties expressed by the following schemata having the form of inference rules.

$\langle A,B\|C\rangle \ \rightarrow \ \langle B,A\|C\rangle$	symmetry
$\langle A,BC\|D\rangle \ \rightarrow \ \langle A,C\|D\rangle$	decomposition
$\langle A,BC\|D\rangle \ \rightarrow \ \langle A,B\|CD\rangle$	weak union
$[\langle A,B\|CD\rangle \ \wedge \ \langle A,C\|D\rangle] \ \rightarrow \ \langle A,BC\|D\rangle$	contraction.

Roughly said, the schemata should be understood as follows: if an independency model contains the triplets before the arrow, then it contains the triplet after the arrow. Thus, we are interested in formal properties of independency models of such a type.

5.1.1. *Syntax of an inference rule*
Let us start with a few technical definitions. Supposing $\mathcal{S}$ is a given fixed nonempty finite set of symbols, the formulas $\langle \mathcal{K}_1, \mathcal{K}_2 | \mathcal{K}_3 \rangle$, where $\mathcal{K}_1, \mathcal{K}_2, \mathcal{K}_3$ are disjoint subsets of $\mathcal{S}$ represented by juxtapositions of their elements, will be called *terms* over $\mathcal{S}$.

We write $\mathcal{K} \approx \mathcal{L}$ to denote that $\mathcal{K}$ and $\mathcal{L}$ are juxtapositions of all elements of the same subset of $\mathcal{S}$ (they can differ in their order). We say that a term $\langle \mathcal{K}_1, \mathcal{K}_2 | \mathcal{K}_3 \rangle$ over $\mathcal{S}$ is an *equivalent version* of the term $\langle \mathcal{L}_1, \mathcal{L}_2 | \mathcal{L}_3 \rangle$ over $\mathcal{S}$ if $\mathcal{K}_i \approx \mathcal{L}_i$ for every $i = 1, 2, 3$. We say that $\langle \mathcal{K}_1, \mathcal{K}_2 | \mathcal{K}_3 \rangle$ is a *symmetric version* of $\langle \mathcal{L}_1, \mathcal{L}_2 | \mathcal{L}_3 \rangle$ if $\mathcal{K}_1 \approx \mathcal{L}_2$, $\mathcal{K}_2 \approx \mathcal{L}_1$, $\mathcal{K}_3 \approx \mathcal{L}_3$. For example, the term $\langle AE, BC | D \rangle$ over $\mathcal{S} = \{A, B, C, D, E, F\}$ is an equivalent version of the term $\langle AE, CB | D \rangle$ and a symmetric version of the term $\langle BC, EA | D \rangle$.

Regular inference rule with r antecedents and s consequents is specified by

(a) positive integers r, s,

(b) a finite set of symbols $\mathcal{S}$, possibly including a special symbol $\emptyset$,

(c) a sequence of ordered triplets $[\mathcal{S}_1^k, \mathcal{S}_2^k, \mathcal{S}_3^k]$, $k = 1, \ldots, r+s$ of nonempty subsets of $\mathcal{S}$ such that for every k the sets $\mathcal{S}_1^k$, $\mathcal{S}_2^k$, $\mathcal{S}_3^k$ are pairwise disjoint.

Moreover, we have several technical requirements:

- $\mathcal{S}$ has at least three symbols,
- if $\mathcal{S}_i^k$ contains the symbol $\emptyset$, then no other symbol from $\mathcal{S}$ is involved in $\mathcal{S}_i^k$ (for every $k = 1, \ldots, r+s$ and every $i = 1, 2, 3$),
- if $k, l \in \{1, \ldots, r+s\}$, $k \neq l$, then $\mathcal{S}_i^k \neq \mathcal{S}_i^l$ for some $i \in \{1, 2, 3\}$,
- every $\sigma \in \mathcal{S}$ belongs to some $\mathcal{S}_i^k$,
- there is no couple of different symbols $\sigma, \tau \in \mathcal{S}$ such that $\forall k = 1, \ldots, r+s \ \ \forall i = 1, 2, 3 \quad [\sigma \in \mathcal{S}_i^k \Rightarrow \tau \in \mathcal{S}_i^k]$.

A *syntactic record* of the corresponding inference rule is then

$$[\langle \mathcal{S}_1^1, \mathcal{S}_2^1 | \mathcal{S}_3^1 \rangle \wedge \ldots \wedge \langle \mathcal{S}_1^r, \mathcal{S}_2^r | \mathcal{S}_3^r \rangle] \rightarrow [\langle \mathcal{S}_1^{r+1}, \mathcal{S}_2^{r+1} | \mathcal{S}_3^{r+1} \rangle \vee \ldots \vee \langle \mathcal{S}_1^{r+s}, \mathcal{S}_2^{r+s} | \mathcal{S}_3^{r+s} \rangle]$$

where each $\mathcal{S}_i^k$ is represented by a juxtaposition of involved symbols. Here the terms $\langle \mathcal{S}_1^k, \mathcal{S}_2^k | \mathcal{S}_3^k \rangle$ for $k = 1, \ldots, r$ are the *antecedent terms*, while $\langle \mathcal{S}_1^k, \mathcal{S}_2^k | \mathcal{S}_3^k \rangle$ for $k = r+1, \ldots, r+s$ are the *consequent terms.*

Example 5.1 Take $r = 2$, $s = 1$, and $\mathcal{S} = \{A, B, C, D\}$. Moreover, let us put $[\mathcal{S}_1^1, \mathcal{S}_2^1, \mathcal{S}_3^1] = [\{A\}, \{B\}, \{C, D\}]$, $[\mathcal{S}_1^2, \mathcal{S}_2^2, \mathcal{S}_3^2] = [\{A\}, \{C\}, \{D\}]$, $[\mathcal{S}_1^3, \mathcal{S}_2^3, \mathcal{S}_3^3] = [\{A\}, \{B, C\}, \{D\}]$. All our technical requirements are satisfied. One possible corresponding syntactic record was already mentioned

under the label 'contraction' in the definition of semigraphoid. Thus, contraction is a regular inference rule with two antecedents and one consequent. Note that another possible syntactic record can be obtained for example by replacing the first antecedent term by its equivalent version:
$[\langle A,B|DC\rangle \;\wedge\; \langle A,C|D\rangle] \rightarrow \langle A,BC|D\rangle$. $\diamond$

Of course, the remaining semigraphoid schemata are also regular inference rules in the sense of our definition.

Remark Our technical requirements in the above definition anticipate the semantics of the symbols. The symbols from $\mathcal{S}$ are interpreted as (disjoint) subsets of a factor set N and the special symbol $\emptyset$ is reserved for the empty set. Terms are interpreted as elements of $\mathcal{T}(N)$. The third requirement ensures that no term in a syntactic record of an inference rule is an equivalent version of another (different) term. Further requirements avoid redundancy of symbols in $\mathcal{S}$: the fourth one means that no symbol is unused, while the fifth one prevents their doubling, as for example in the 'rule': $[\langle A,BE|CD\rangle \;\wedge\; \langle A,C|D\rangle] \;\rightarrow\; \langle A,EBC|D\rangle$
where the symbol B is doubled by the symbol E.

5.1.2. *Semantics of an inference rule*
Let us consider a regular inference rule $\boldsymbol{\alpha}$ with r antecedents and s consequents. What is its meaning for a fixed nonempty factor set N? A *substitution mapping* (for N) is a mapping m which assigns a set $m(\sigma)\subset N$ to every symbol $\sigma\in\mathcal{S}$ in such a way that:

- $m(\emptyset)$ is the empty set,
- $\{\, m(\sigma)\,;\, \sigma\in\mathcal{S}\,\}$ is a disjoint collection of subsets of N,
- $\bigcup_{\sigma\in\mathcal{S}_1^k}\; m(\sigma)\neq\emptyset$ for every $k=1,\ldots,r+s$,
- $\bigcup_{\sigma\in\mathcal{S}_2^k}\; m(\sigma)\neq\emptyset$ for every $k=1,\ldots,r+s$.

Of course, it may happen that no such substitution mapping exists for a factor set N; for example in case of contraction for N with $\mathrm{card}\,N=2$. However, in case such a mapping m exists an *inference instance* of the considered inference rule (induced by m) is $(r+s)$-tuple $[t_1,\ldots,t_{r+s}]$ of elements of $\mathcal{T}(N)$ defined as follows:

$$t_k=\langle \bigcup_{\sigma\in\mathcal{S}_1^k} m(\sigma)\,,\; \bigcup_{\sigma\in\mathcal{S}_2^k} m(\sigma)\;|\; \bigcup_{\sigma\in\mathcal{S}_3^k} m(\sigma)\,\rangle \quad \text{for } k=1,\ldots,r+s\,.$$

The $(r+s)$-tuple $[t_1,\ldots,t_r\,|\,t_{r+1},\ldots,t_{r+s}]$ is formally divided into the r-tuple made of the triplets $t_1,\ldots,t_r$ which are called *antecedents*, and the s-tuple made of the triplets $t_{r+1},\ldots,t_{r+s}$ which are called *consequents*.

Example 5.2 Let us continue with Example 5.1 and consider contraction and $N = \{1, 2, 3\}$. Put $m(A) = \{1\}$, $m(B) = \{2\}$, $m(C) = \{3\}$, $m(D) = \emptyset$. It is a substitution mapping for N. The corresponding inference instance (induced by m) is then $[t_1, t_2 \,|\, t_3]$ where
$t_1 = \langle\{1\}, \{2\}|\{3\}\rangle$, $t_2 = \langle\{1\}, \{3\}|\emptyset\rangle$, $t_3 = \langle\{1\}, \{2, 3\}|\emptyset\rangle$.
Here t_1, t_2 are the antecedents and t_3 is the consequent. However, there are other inference instances, induced by other possible substitution mappings for N. In this case one finds 5 other ones:
$\tilde{t}_1 = \langle\{1\}, \{3\}|\{2\}\rangle$, $\tilde{t}_2 = \langle\{1\}, \{2\}|\emptyset\rangle$, $\tilde{t}_3 = \langle\{1\}, \{2, 3\}|\emptyset\rangle$,
$\hat{t}_1 = \langle\{2\}, \{1\}|\{3\}\rangle$, $\hat{t}_2 = \langle\{2\}, \{3\}|\emptyset\rangle$, $\hat{t}_3 = \langle\{2\}, \{1, 3\}|\emptyset\rangle$,
$\check{t}_1 = \langle\{2\}, \{3\}|\{1\}\rangle$, $\check{t}_2 = \langle\{2\}, \{1\}|\emptyset\rangle$, $\check{t}_3 = \langle\{2\}, \{1, 3\}|\emptyset\rangle$,
$\bar{t}_1 = \langle\{3\}, \{1\}|\{2\}\rangle$, $\bar{t}_2 = \langle\{3\}, \{2\}|\emptyset\rangle$, $\bar{t}_3 = \langle\{3\}, \{1, 2\}|\emptyset\rangle$,
$\ddot{t}_1 = \langle\{3\}, \{2\}|\{1\}\rangle$, $\ddot{t}_2 = \langle\{3\}, \{1\}|\emptyset\rangle$, $\ddot{t}_3 = \langle\{3\}, \{1, 2\}|\emptyset\rangle$. ◇

Of course, the number of possible substitution mappings is finite for a fixed regular inference rule and a fixed factor set. Therefore, the number of all inference instances of a regular inference rule for a given factor set is always finite and the following definition is sensible.

Having a fixed factor set N we say that an independency model $\mathcal{I} \subset \mathcal{T}(N)$ is *closed under* a regular inference rule $\boldsymbol{\alpha}$ with r antecedents and s consequents iff for every inference instance $[t_1, \ldots, t_{r+s}] \in \mathcal{T}(N)^{r+s}$ (of $\boldsymbol{\alpha}$ for N) $\{\, t_1, \ldots, t_r \,\} \subset \mathcal{I}$ implies $\{t_{r+1}, ..., t_{r+s}\} \cap \mathcal{I} \neq \emptyset$.

Example 5.3 Let us continue with Example 5.2. The independency model $\mathcal{I}$ over $N = \{1, 2, 3\}$ consisting of the triplet $\langle\{1\}, \{2\}|\emptyset\rangle$ only is closed under contraction since no inference instance for N has both antecedents in $\mathcal{I}$. On the other hand, the model $\mathcal{M} = \{\, \langle\{1\}, \{2\}|\emptyset\rangle \;,\; \langle\{1\}, \{3\}|\{2\}\rangle \,\}$ is not closed under contraction. Indeed, one has $\tilde{t}_1, \tilde{t}_2 \in \mathcal{M}$ but $\tilde{t}_3 \notin \mathcal{M}$ for the inference instance $[\, \tilde{t}_1, \tilde{t}_2 \,|\, \tilde{t}_3 \,]$ ◇

5.1.3. *Logical implication of inference rules*

The aim of regular inference rules is to sketch formal properties of independency models, especially probabilistic independency models. In fact, one can have in mind another reasonable class of independency models instead of the class of probabilistic independency models. For example the class of graph-isomorphic independency models [14] or the class of EMVD-models [20, 9] or various classes of possibilistic independency models [1, 6]. Well, such an approach hides a deeper wish or hope to characterize the respective class of independency models as the class of those closed under a collection of regular inference rules. We can speak about the *axiomatic characterization* of the respective class of independency models.

For example, in the case of probabilistic independency models such a characterization would make it possible to recognize them without laborious construction of an inducing probability distribution. Indeed, the process of verification of whether a given independency model is closed under a finite number of known inference rules is completely automatic and can be done by a computer. Of course such a desired collection of inference rules should be minimal (a finite collection would be an ideal solution). One needs a criterion for removing superfluous regular inference rules from such a desired collection. Therefore, we are interested in the following relation among inference rules.

We say that a collection of regular inference rules Υ *logically implies* a regular inference $\boldsymbol{\omega}$ and write $\Upsilon \models \boldsymbol{\omega}$ if for every (nonempty finite) factor set N and for every independency model $\mathcal{M}$ over N the following holds: whenever $\mathcal{M}$ is closed under every inference rule $\boldsymbol{\upsilon} \in \Upsilon$, then $\mathcal{M}$ is closed under $\boldsymbol{\omega}$.

Usually, an easy sufficient condition for logical implication is (syntactic) derivability. We give an illustrative example to explain what we have in mind. We hope that it gives a better insight than a pedantic definition, which would be too complicated.

Example 5.4 Let us consider the following regular inference rule $\boldsymbol{\omega}$ with three antecedents and one consequent:
$[\, \langle A,B\,|\,E\rangle \;\wedge\; \langle A,C\,|\,BE\rangle \;\wedge\; \langle A,D\,|\,CE\rangle \,] \;\rightarrow\; \langle A,D\,|\,E\rangle$.
This inference rule is logically implied by the semigraphoid inference rules. To show it we construct a special *derivation sequence* of terms over the corresponding set of symbols $\mathcal{S} = \{A,B,C,D,E\}$. Here is the derivation sequence:

1. $\langle A,B\,|\,E\rangle$,
2. $\langle A,C\,|\,BE\rangle$,
3. $\langle A,D\,|\,CE\rangle$,
4. $\langle A,BC\,|\,E\rangle$ is directly derived from 2. and 1. by contraction,
5. $\langle A,C\,|\,E\rangle$ is directly derived from 4. by decomposition,
6. $\langle A,CD\,|\,E\rangle$ is directly derived from 3. and 5. by contraction,
7. $\langle A,D\,|\,E\rangle$ is directly derived from 6. be decomposition.

The last term is the consequent term of $\boldsymbol{\omega}$. Every term in the derivation sequence is either an antecedent term of $\boldsymbol{\omega}$, or it is 'directly derived' from preceding terms (in the derivation sequence) by virtue of a semigraphoid inference rule.

Now, let us consider a fixed factor set N and a semigraphoid $\mathcal{M} \subset \mathcal{T}(N)$ (i.e. an independency model over N closed under all semigraphoid inference rules). To show that $\mathcal{M}$ is closed under $\boldsymbol{\omega}$ let us consider an inference instance $[\, t_1, t_2, t_3 \,|\, t_4 \,]$ of $\boldsymbol{\omega}$ for N induced by a substitution mapping m. So, we can construct a sequence $u_1, \ldots, u_7$ of elements of $\mathcal{T}(N)$ which 'copies' the derivation sequence:

$$u_1 = \langle\, m(A)\,,\, m(B) \,|\, m(E) \rangle \equiv t_1\,,$$

$$u_2 = \langle\, m(A)\,,\, m(C) \,|\, m(B) \cup m(E) \rangle \equiv t_2\,,$$

$$u_3 = \langle\, m(A)\,,\, m(D) \,|\, m(C) \cup m(E) \rangle \equiv t_3\,,$$

$$u_4 = \langle\, m(A)\,,\, m(B) \cup m(C) \,|\, m(E) \rangle\,,$$

$$u_5 = \langle\, m(A)\,,\, m(C) \,|\, m(E) \rangle\,,$$

$$u_6 = \langle\, m(A)\,,\, m(C) \cup m(D) \,|\, m(E) \rangle\,,$$

$$u_7 = \langle\, m(A)\,,\, m(D) \,|\, m(E) \rangle \equiv t_4.$$

Owing to the fact that $\mathcal{M}$ is closed under every semigraphoid inference rule one can derive from the assumption $\{t_1, t_2, t_3\} \subset \mathcal{M}$ by induction on $j = 1, \ldots, 7$ that $\{u_1, \ldots, u_j\} \subset \mathcal{M}$. Especially, $t_4 \in \mathcal{M}$, which was the desired conclusion. Thus, $\mathcal{M}$ is closed under $\boldsymbol{\omega}$. $\diamond$

5.1.4. *Pure inference rules*

It may happen that an inference instance of a regular inference rule is trivial in the sense that it has as a consequent one of its antecedents (for example in the case of decomposition for a substitution mapping m with $m(B) = \emptyset$). Thus, we wish to concentrate on a class of 'pure' inference rules which have only 'informative' inference instances. For technical reasons (which will become clear later - see 5.2.2) we would also like to avoid those inference rules which possibly may have an inference instance whose consequent is the symmetric image of an antecedent, as demonstrated by the following example.

Example 5.5 Let us consider the following regular inference rule:
$[\,\langle A, BC|D\rangle \;\wedge\; \langle B, D|AC\rangle\,] \;\rightarrow\; \langle B, A|D\rangle$.
Take $N = \{1, 2\}$ and put $m(A) = \{1\}$, $m(B) = \{2\}$, $m(C) = \emptyset$, $m(D) = \{3\}$. It induces the inference instance $[t_1, t_2 \,|\, t_3]$ with $t_1 = \langle\, \{1\}\,,\, \{2\} \,|\, \{3\}\,\rangle$, $t_2 = \langle\, \{2\}\,,\, \{3\} \,|\, \{1\}\,\rangle$, $t_3 = \langle\, \{2\}\,,\, \{1\} \,|\, \{3\}\,\rangle$. Here the consequent t_3 is the symmetric image of the antecedent t_1. $\diamond$

Thus, we say that a regular inference rule $\boldsymbol{\omega}$ is *pure* if there is no inference instance of $\boldsymbol{\omega}$ (for arbitrary factor set N) in which a consequent either coincides with an antecedent or with the symmetric image of an antecedent.

Such a definition is not suitable for verification. We need a sufficient condition formulated by means of syntactic concepts from 5.1.1. To formulate it we give two definitions. Suppose that $\boldsymbol{\omega}$ is a regular inference rule with a syntactic record having $\mathcal{S}$ as the set of symbols. We say that the symbol sets $\mathcal{K}, \mathcal{L} \subset \mathcal{S}$ are *distinguished* in $\boldsymbol{\omega}$ if
$\exists\, k \in \{1, \ldots, r+s\}\ \exists\, j \in \{1,2\}\ \ \mathcal{S}_j^k \subset (\mathcal{K} \setminus \mathcal{L}) \cup (\mathcal{L} \setminus \mathcal{K})$. A term $\langle \mathcal{K}_1, \mathcal{K}_2 | \mathcal{K}_3 \rangle$ over $\mathcal{S}$ is *distinguished* in $\boldsymbol{\omega}$ from a term $\langle \mathcal{L}_1, \mathcal{L}_2 | \mathcal{L}_3 \rangle$ over $\mathcal{S}$ if $\mathcal{K}_i$ and $\mathcal{L}_i$ are distinguished in $\boldsymbol{\omega}$ for some $i = 1, 2, 3$.

Lemma 5.1 *A regular inference rule $\boldsymbol{\omega}$ is pure if every consequent term of $\boldsymbol{\omega}$ is distinguished in $\boldsymbol{\omega}$ both from all antecedent terms of $\boldsymbol{\omega}$ and from their symmetric versions.*

Proof: At first realize this: whenever symbol sets $\mathcal{K}$ and $\mathcal{L}$ are distinguished in $\boldsymbol{\omega}$, then for every substitution mapping m one has
$\emptyset \neq m(\mathcal{S}_j^k) \subset m(\mathcal{K} \setminus \mathcal{L}) \cup m(\mathcal{L} \setminus \mathcal{K}) \subset (m(\mathcal{K}) \setminus m(\mathcal{L})) \cup (m(\mathcal{L}) \setminus m(\mathcal{K}))$,
which implies $m(\mathcal{K}) \neq m(\mathcal{L})$. Hence, terms distinguished in $\boldsymbol{\omega}$ are transformed to distinct elements of $\mathcal{T}(N)$ by any substitution mapping. Therefore, under the mentioned assumption, no consequent of a respective inference instance can coincide either with an antecedent or with its symmetric image. $\square$

We leave it to the reader to verify by means of Lemma 5.1 that contraction is a pure inference rule. On the other hand one can easily see that decomposition and weak union are not pure rules.

5.2. PROBABILISTICALLY SOUND INFERENCE RULES

We say that a regular inference rule $\boldsymbol{\omega}$ is *probabilistically sound* if every probabilistic independency model is closed under $\boldsymbol{\omega}$.

That means, every probabilistically sound inference rule expresses a formal property which is shared by all probabilistic independency models. Is it difficult to verify probabilistic soundness of a given regular inference rule? The multiinformation function is a good tool for this purpose, although maybe not universal. In the effort to characterize all probabilistic independency models over four factors [10, 11] a lot of probabilistically sound inference rules was found whose soundness was not verified with help of the multiinformation function. However, it has appeared lately that at least some of them can be regarded as a consequence of deeper properties of the multiinformation function, namely of a certain 'conditional' inequalities for the multiinformation (or entropic) function [27, 12]. Thus, the question whether every probabilistically sound inference rule can be derived by means of the multiinformation function remains open. However, to support our arguments about its usefulness we give an illustrative example. We believe that an example is more didactic than a technical description of the method.

Example 5.6 To show the probabilistic soundness of weak union one has to verify for arbitrary factor set N, for any probability distribution P over N, and for any collection of disjoint sets $A, B, C, D \subset N$ which are nonempty with possible exceptions of C and D, that

$$A \perp\!\!\!\perp BC|D\,(P) \;\Rightarrow\; A \perp\!\!\!\perp B|CD\,(P)\,.$$

The assumption $A \perp\!\!\!\perp BC|D\,(P)$ can be rewritten by Consequence 2.1(b) and Lemma 2.3 in terms of the multiinformation function M induced by the distribution P:

$$0 = M(ABCD) + M(D) - M(AD) - M(BCD)\,.$$

Then one can 'artificially' add and subtract the terms $M(CD) - M(ACD)$ and by Lemma 2.3 derive:

$$\begin{aligned} 0 &= \{M(ABCD) + M(CD) - M(ACD) - M(BCD)\} \\ &\quad + \{M(ACD) + M(D) - M(AD) - M(CD)\} \\ &= I(A;B|CD) + I(A;C|D)\,. \end{aligned}$$

By Consequence 2.1(a) both $I(A;B|CD)$ and $I(A;C|D)$ are nonnegative, and therefore they vanish! But that implies by Consequence 2.1(b) that $A \perp\!\!\!\perp B|CD\,(P)$. $\diamond$

Note that one can easily see using the method shown in the preceding example that every semigraphoid inference rule is probabilistically sound.

5.2.1. *Redundant rules*

However, some probabilistically sound inference rules are superfluous for the purposes of providing an axiomatic characterization of probabilistic independency models. The following consequence follows directly from given definitions.

Consequence 5.1 *If $\boldsymbol{\omega}$ is a regular inference rule which is logically implied by a collection of probabilistically sound inference rules, then $\boldsymbol{\omega}$ is probabilistically sound.*

A clear example of a superfluous rule is an inference rule with redundant antecedent terms.

Example 5.7 The inference rule
$[\langle A, BC \,|\, D\rangle \;\wedge\; \langle C, B \,|\, A\rangle\,] \;\rightarrow\; \langle A, B \,|\, CD\rangle$
is a probabilistically sound regular inference rule. But it can be ignored since it is evidently logically implied by weak union. $\diamond$

Therefore we should limit ourselves to 'minimal' probabilistically sound inference rules, i.e. to such probabilistically sound inference rules that no antecedent term can be removed without violating the probabilistic soundness of the resulting reduced inference rule. However, even such a rule can be logically implied by probabilistically sound rules with fewer antecedents. We need the following auxiliary construction of a probability distribution to give an easy example.

Construction B Supposing $A \subset N$, $\mathrm{card}\, A \geq 2$, there exists a probability distribution P over N such that

$$M(B \,\|\, P) = \max\{0, \mathrm{card}\,(A \cap B) - 1\} \cdot \ln 2 \quad \text{for } B \subset N\,.$$

Proof: Let us put $\mathbf{X}_i = \{0,1\}$ for $i \in A$, $\mathbf{X}_i = \{0\}$ for $i \in N \setminus A$. Define P on $\mathbf{X}_N$ as follows:

$$\begin{aligned} P([x_i]_{i \in N}) &= \tfrac{1}{2} && \text{whenever } [\forall i, j \in A \;\; x_i = x_j]\,, \\ P([x_i]_{i \in N}) &= 0 && \text{otherwise.} \end{aligned}$$

$\square$

Example 5.8 We have already verified earlier that the inference rule $\boldsymbol{\omega}$ from Example 5.4 is logically implied by the semigraphoid inference rules. Hence, $\boldsymbol{\omega}$ is probabilistically sound by Consequence 5.1.

Let us consider a 'reduced' inference rule made by a removal of an antecedent term:
$[\langle A, B|E\rangle \;\wedge\; \langle A, C|BE\rangle\,] \;\rightarrow\; \langle A, D|E\rangle$.
It is a regular inference rule with 2 antecedents and one consequent. To disprove its probabilistic soundness one has to find a probabilistic independency model over a factor set N which is not closed under this rule. Use Construction B with $N = \{1,2,3,4\}$ and $A = \{1,4\}$. By Consequence 2.1 one verifies that $\{1\} \perp\!\!\!\perp \{2\}|\emptyset\,(P)$, $\{1\} \perp\!\!\!\perp \{3\}|\{2\}\,(P)$, but $\neg[\,\{1\} \perp\!\!\!\perp \{4\}|\emptyset\,(P)\,]$ for the constructed distribution P. As concerns an alternative 'reduced' inference rule
$[\langle A, B \,|\, E\rangle \;\wedge\; \langle A, D \,|\, CE\rangle\,] \;\rightarrow\; \langle A, D \,|\, E\rangle$
use Construction B with $A = \{1,3,4\}$ and a distribution P over N such that $\{1\} \perp\!\!\!\perp \{2\}|\emptyset\,(P)$, $\{1\} \perp\!\!\!\perp \{4\}|\{3\}\,(P)$, but $\neg[\,\{1\} \perp\!\!\!\perp \{4\}|\emptyset\,(P)\,]$. As concerns the third possible 'reduced' inference rule

$[\langle A,C\,|\,BE\rangle \;\wedge\; \langle A,D\,|\,CE\rangle\,] \;\to\; \langle A,D\,|\,E\rangle$
use again Construction B with $A=\{1,2,3,4\}$. Thus, one has a distribution P with $\{1\} \perp\!\!\!\perp \{3\}|\{2\}\,(P)$, $\{1\} \perp\!\!\!\perp \{4\}|\{3\}\,(P)$, but $\neg[\,\{1\} \perp\!\!\!\perp \{4\}|\emptyset\,(P)\,]$. $\diamondsuit$

5.2.2. *Perfect rules*

Thus, one should search for conditions which ensure that an inference rule is not logically implied by probabilistically sound inference rules with fewer antecedents. We propose the following condition.

We say that a probabilistically sound regular inference rule with r antecedents (and s consequents) is *perfect* if there exists a factor set N and an inference instance $[t_1,\ldots,t_r\,|\,t_{r+1},\ldots,t_{r+s}] \in \mathcal{T}(N)^{r+s}$ such that the symmetric closure of every proper subset of $\{t_1,\ldots,t_r\}$ is a probabilistic independency model over N.

Lemma 5.2 *Let $\boldsymbol{\omega}$ be a perfect, probabilistically sound, pure inference rule with r antecedents, $r \geq 1$. Then there exists a factor set N and an independency model $\mathcal{M}$ over N such that*

- *$\mathcal{M}$ is closed under every probabilistically sound regular inference rule with at most $r-1$ antecedents,*
- *$\mathcal{M}$ is not closed under $\boldsymbol{\omega}$.*

Proof: Let $[t_1,\ldots,t_{r+s}] \in \mathcal{T}(N)$ be the inference instance of $\boldsymbol{\omega}$ mentioned in the definition of perfectness. Define $\mathcal{M} \subset \mathcal{T}(N)$ as the symmetric closure of the set of antecedents $\{t_1,\ldots,t_r\}$. Let us show that $\mathcal{M}$ is closed under all probabilistically sound inference rules with at most $r-1$ antecedents.

Suppose for a contradiction that $[\tilde{t}_1,\ldots,\tilde{t}_{\tilde{r}+\tilde{s}}] \in \mathcal{T}(N)^{\tilde{r}+\tilde{s}}$ is an inference instance of such an inference rule $\boldsymbol{\upsilon}$ (with $\tilde{r} \leq r-1$ antecedents and $\tilde{s}$ consequents) for N with $\{\tilde{t}_1,\ldots,\tilde{t}_{\tilde{r}}\} \subset \mathcal{M}$ and $\{\tilde{t}_{\tilde{r}+1},\ldots,\tilde{t}_{\tilde{r}+\tilde{s}}\} \cap \mathcal{M} = \emptyset$. However, owing to the fact that $\tilde{r} < r$ and the assumption (of perfectness) the symmetric closure $\mathcal{I}$ of $\{\tilde{t}_1,\ldots,\tilde{t}_{\tilde{r}}\}$ is a probabilistic independency model. So, (by the definition of probabilistic soundness) $\mathcal{I}$ is closed under $\boldsymbol{\upsilon}$, and therefore $\{\tilde{t}_{\tilde{r}+1},\ldots,\tilde{t}_{\tilde{r}+\tilde{s}}\} \cap \mathcal{I} \neq \emptyset$ which contradicts the fact that $\mathcal{I} \subset \mathcal{M}$. Therefore $\mathcal{M}$ has to be closed under any such inference rule $\boldsymbol{\upsilon}$.

Owing to the assumption that the inference rule $\boldsymbol{\omega}$ is pure by definition one has $\{t_{r+1},\ldots,t_{r+s}\} \cap \mathcal{M} = \emptyset$. Since $\mathcal{M}$ was defined to contain $\{t_1,\ldots,t_r\}$, it is not closed under $\boldsymbol{\omega}$. $\square$

The preceding lemma implies the following consequence with help of the definition of logical implication.

Consequence 5.2 *No perfect probabilistically sound pure inference rule is logically implied by a collection of probabilistically sound inference rules with fewer antecedents.*

Contraction is an example of a perfect pure regular inference rule.

5.3. NO FINITE AXIOMATIC CHARACTERIZATION

5.3.1. *Method of the proof*

It is clear in the light of Consequence 5.2 how to disprove the existence of a finite system of regular inference rules characterizing probabilistic independency models.

Lemma 5.3 *Let us suppose that we have found for every $r \geq 3$ a perfect, probabilistically sound, pure inference rule with at least r antecedents. Then every system Υ of regular inference rules characterizing probabilistic independency models as independency models closed under rules in Υ is infinite.*

Proof: Let us suppose for a contradiction that there exists a finite system Υ of regular inference rules such that for every factor set N an independency model $\mathcal{M} \subset \mathcal{T}(N)$ is a probabilistic independency model (over N) iff it is closed under all rules in Υ. Hence, every rule in Υ must be probabilistically sound. We choose $\tilde{r} \geq 3$ which exceeds the maximal number of antecedents of rules in Υ. According to the assumption there exists a perfect, probabilistically sound, pure inference rule $\boldsymbol{\omega}$ with r antecedents, where $r \geq \tilde{r}$.

By Lemma 5.2 we find a factor set N and an independency model $\mathcal{M}$ over N which is closed under every probabilistically sound inference rule with at most $r-1$ antecedents but not under $\boldsymbol{\omega}$. Since every inference rule from Υ has at most $r-1$ antecedents, $\mathcal{M}$ is closed under every inference rule from Υ. Therefore $\mathcal{M}$ is a probabilistic independency model over N. However, $\mathcal{M}$ is not closed under $\boldsymbol{\omega}$ which contradicts the fact that $\boldsymbol{\omega}$ is probabilistically sound. $\square$

Thus, we need to verify the assumptions of the preceding lemma. Let us consider for each $n \geq 3$ the following inference rule $\boldsymbol{\gamma}(n)$ with n antecedents and one consequent:

$$[\,\langle A, B_1|B_2\rangle \;\wedge \ldots \wedge\; \langle A, B_{n-1}|B_n\rangle \;\wedge\; \langle A, B_n|B_1\rangle\,] \;\rightarrow\; \langle A, B_2|B_1\rangle\,. \quad \boldsymbol{\gamma}(n)$$

It is no problem to verify that each $\boldsymbol{\gamma}(n)$ is indeed a regular inference rule. Moreover, one can verify easily using Lemma 5.1 that each $\boldsymbol{\gamma}(n)$ is a pure rule.

5.3.2. *Soundness*

To show their probabilistic soundness we use the properties of the multiinformation function.

Lemma 5.4 *Each above mentioned rule $\gamma(n)$ is probabilistically sound.*

Proof: Let us fix $n \geq 3$. We have to show for arbitrary factor set N, any distribution P over N, and any collection of nonempty disjoint subsets $A, B_1, \ldots, B_n \subset N$ that (under convention $B_{n+1} \equiv B_1$) the assumption

$$[\forall j = 1, \ldots, n \quad A \perp\!\!\!\perp B_j | B_{j+1} \, (P) \,]$$

implies that $A \perp\!\!\!\perp B_2 | B_1 \, (P)$. By Consequence 2.1(b) with Lemma 2.3 one has for every $j = 1, \ldots, n$ (M is the corresponding multiinformation function):

$$M(AB_jB_{j+1}) + M(B_{j+1}) - M(AB_{j+1}) - M(B_jB_{j+1}) = 0 \, .$$

Hence we get by summing, the above mentioned convention and Lemma 2.3:

$$\begin{aligned}
0 = & \sum_{j=1}^{n} \{ \, M(AB_jB_{j+1}) + M(B_{j+1}) - M(AB_{j+1}) - M(B_jB_{j+1}) \, \} \\
= & \sum_{j=1}^{n} M(AB_jB_{j+1}) + \sum_{j=1}^{n} M(B_{j+1}) - \sum_{j=1}^{n} M(AB_{j+1}) - \sum_{j=1}^{n} M(B_jB_{j+1}) \\
= & \sum_{j=1}^{n} M(AB_jB_{j+1}) + \sum_{j=1}^{n} M(B_j) - \sum_{j=1}^{n} M(AB_j) - \sum_{j=1}^{n} M(B_jB_{j+1}) \\
= & \sum_{j=1}^{n} \{ \, M(AB_jB_{j+1}) + M(B_j) - M(AB_j) - M(B_jB_{j+1}) \, \} \\
= & \sum_{j=1}^{n} I(A; B_{j+1} | B_j) \; .
\end{aligned}$$

Owing to Consequence 2.1(a) necessarily $I(A; B_{j+1} | B_j \parallel P) = 0$ for every $j = 1, \ldots, n$. Hence by Consequence 2.1(b) $A \perp\!\!\!\perp B_2 | B_1 \, (P)$. □

5.3.3. *Perfectness*

To verify perfectness of a rule one needs some method for showing that an independency model is a probabilistic independency model. We again Constructions A and B.

Lemma 5.5 *Suppose that $N = \{0, 1, \ldots, n\}$, $n \geq 3$ and $\mathcal{M} \subset \mathcal{T}(N)$ be the symmetric closure of the set $\{ \langle \{0\}, \{i\} \,|\, \{i+1\} \rangle \, ; \; i = 1, \ldots, n-1 \}$. Then $\mathcal{M}$ is a probabilistic independency model over N.*

Proof: It suffices to find a probabilistic independency model $\mathcal{M}_t$ with $\mathcal{M} \subset \mathcal{M}_t$ and $t \notin \mathcal{M}_t$ for every $t \in \mathcal{T}(N) \setminus \mathcal{M}$. Indeed, then $\mathcal{M} \equiv \bigcap_{t \in \mathcal{T}(N) \setminus \mathcal{M}} \mathcal{M}_t$, and by Lemma 2.1 $\mathcal{M}$ is a probabilistic independency model.

Moreover, one can limit oneself to the triplets of the form $\langle a,b|C\rangle \in \mathcal{T}(N)\setminus\mathcal{M}$ where a,b are singletons. Indeed, for a given general $\langle A,B|C\rangle \in \mathcal{T}(N)\setminus\mathcal{M}$ choose $a\in A$, $b\in B$ and find the respective probabilistic independency model $\mathcal{M}_t$ for $t=\langle a,b|C\rangle$. Since $\mathcal{M}_t$ is a semigraphoid, $t\notin\mathcal{M}_t$ implies $\langle A,B|C\rangle\notin\mathcal{M}_t$.

In the sequel we distinguish 5 cases for a given fixed $\langle a,b|C\rangle\in\mathcal{T}(N)\setminus\mathcal{M}$. Each case requires a different construction of the respective probabilistic independency model $\mathcal{M}_t$, that is a different construction of a probability distribution P over N such that $\{0\}\perp\!\!\!\perp\{i\}\,|\,\{i+1\}\,(P)$ for $i=1,\ldots,n-1$, but $\neg[\,\{a\}\perp\!\!\!\perp\{b\}\,|\,C\,(P)\,]$. One can verify these statements about P through the multiinformation function induced by P. If the multiinformation function is known (as it is in the case of our constructions) one can use Consequence 2.1(b) and Lemma 2.3 for this purpose. We leave this to the reader. Here is the list of cases.

I. $\forall i=1,\ldots,n-1\quad \{a,b\}\neq\{0,i\}$ (C arbitrary).
In this case use Construction A where $A=\{a,b\}$.

II. $[\exists j\in\{1,\ldots,n-1\}\quad \{a,b\}=\{0,j\}]$ and $C\setminus\{j-1,j+1\}\neq\emptyset$.
In this case choose $r\in C\setminus\{j-1,j+1\}$ and use Construction A where $A=\{0,j,r\}$.

III. $[\exists j\in\{2,\ldots,n-1\}\quad \{a,b\}=\{0,j\}]$ and $C=\{j-1,j+1\}$.
In this case use Construction A where $A=\{0,j-1,j,j+1\}$.

IV. $[\exists j\in\{2,\ldots,n-1\}\quad \{a,b\}=\{0,j\}]$ and $C=\{j-1\}$.
Use Construction B where $A=\{0,j,j+1,\ldots,n\}$.

V. $[\exists j\in\{1,\ldots,n-1\}\quad \{a,b\}=\{0,j\}]$ and $C=\emptyset$.
Use Construction B where $A=N$.

□

Consequence 5.3 *Each above mentioned rule $\boldsymbol{\gamma}(n)$ is perfect.*

Proof: Let us fix $n\geq 3$, put $N=\{0,1,\ldots,n\}$ and $t_j=\langle\{0\},\{j\}|\{j+1\}\rangle$ for $j=1,\ldots,n$ (convention $n+1\equiv 1$), $t_{n+1}=\langle\{0\},\{2\}|\{1\}\rangle$. Evidently, $[\,t_1,\ldots,t_n\,|\,t_{n+1}\,]$ is an inference instance of $\boldsymbol{\gamma}(n)$. To show that the symmetric closure of every proper subset of $\{t_1,\ldots,t_n\}$ is a probabilistic independency model it suffices to verify it only for every subset of cardinality $n-1$ (use Lemma 2.1). However, owing to possible cyclic re-indexing of N it suffices to prove (only) that the symmetric closure $\mathcal{M}$ of $\{t_1,\ldots,t_{n-1}\}$ is a probabilistic independency model. This follows from Lemma 5.5. □

Proposition 5.1 *There is no finite system Υ of regular inference rules characterizing probabilistic independency models as independency models closed under rules in Υ .*

Proof: An easy consequence of Lemmas 5.3, 5.4 and Consequence 5.3. □

Conclusions

Let us summarize the paper. Several results support our claim that conditional mutual information $I(A;B|C)$ is a good measure of stochastic conditional dependence between random vectors ξ_A and ξ_B given ξ_C. The value of $I(A;B|C)$ is always nonnegative and vanishes iff ξ_A is conditionally independent of ξ_B given ξ_C. On the other hand, the upper bound for $I(A;B|C)$ is $\min\{H(A|C), H(B|C)\}$, and the value $H(A|C)$ is achieved just in case ξ_A is a function of ξ_{BC}. A transformation of ξ_{ABC} which saves ξ_{AC} and ξ_{BC} increases the value of $I(A;B|C)$. On the other hand, if ξ_A is transformed while ξ_{BC} is saved, then $I(A;B|C)$ decreases. Note that the paper [29] deals with a more practical use of conditional mutual information: it is applied to the problem of finding relevant factors in medical decision-making.

Special level-specific measures of dependence were introduced. While the value $M(A)$ of the multiinformation function is viewed as a measure of global stochastic dependence within $[\xi_i]_{i\in A}$, the value of $\lambda(r,A)$ (for $1 \leq r \leq \mathrm{card}\, A-1$) is interpreted as a measure of the strength of dependence of level r among variables $[\xi_i]_{i\in A}$. The value of $\lambda(r,A)$ is always nonnegative and vanishes iff ξ_i is conditionally independent of ξ_j given ξ_K for arbitrary distinct $i,j \in A$, $K \subset A$, $\mathrm{card}\, K = r-1$. And of course, the sum of $\lambda(r,A)$s is just $M(A)$. Note that measures $\lambda(r,A)$ are certain multiples of Han's [8] measures of multivariate symmetric correlation.

Finally, we have used the multiinformation function as a tool to show that conditional independence models have no finite axiomatic characterization. A didactic proof of this result, originally shown in [20], is given. We analyze thoroughly syntax and semantics of inference rule schemata (= axioms) which characterize formal properties of conditional independence models. The result of the analysis is that two principal features of such schemata are pointed out: the inference rules should be (probabilistically) *sound* and *perfect*. To derive the nonaxiomatizability result one has to find an infinite collection of sound and perfect inference rules. In the verification of both soundness and perfectness the multiinformation function was proved to be an effective tool.

Let us add a remark concerning the concept of perfect rule. We have used this concept just in the proof of the nonaxiomatizability result. However, our aim is a little bit deeper, in fact. We (vaguely) guess that probabilistic independency models have certain uniquely determined 'minimal' axiomatic characterization, which is of course infinite. In particular, we conjecture that the semigraphoid inference rules and perfect probabilistically sound pure inference rules form together the desired axiomatic characterization of probabilistic independency models.

Acknowledgments

We would like to express our gratitude to our colleague František Matúš who directed our attention to the paper [8]. We also thank to both reviewers for their valuable comments and correction of grammatical errors. This work was partially supported by the grant VŠ 96008 of the Ministry of Education of the Czech Republic and by the grant 201/98/0478 "Conditional independence structures: information theoretical approach" of the Grant Agency of Czech Republic.

References

1. de Campos, L.M. (1995) Independence relationships in possibility theory and their application to learning in belief networks, in G. Della Riccia, R. Kruse and R. Viertl (eds.), *Mathematical and Statistical Methods in Artificial Intelligence*, Springer-Verlag, 119–130.
2. Csiszár, I. (1975) I-divergence geometry of probability distributions and minimazition problems, *Ann. Probab.*, **3**, 146–158.
3. Cover, T.M., and Thomas, J.A. (1991) *Elements of Information Theory,* John Wiley, New York.
4. Darroch, J.N., Lauritzen, S.L., and Speed, T.P. (1980) Markov fields and log-linear interaction models for contingency tables, *Ann. Statist.*, **8**, 522–539.
5. Dawid, A.P. (1979) Conditional independence in statistical theory, *J. Roy. Stat. Soc. B*, **41**, 1–31.
6. Fonck P. (1994) Conditional independence in possibility theory, in R.L. de Mantaras and D. Poole (eds.), *Uncertainty in Artificial Intelligence: proceedings of the 10th conference,* Morgan Kaufman, San Francisco, 221–226.
7. Gallager, R.G. (1968) *Information Theory and Reliable Communication*, John Wiley, New York.
8. Han T.S. (1978) Nonnegative entropy of multivariate symmetric correlations, *Information and Control*, **36**, 113–156.
9. Malvestuto, F.M. (1983) Theory of random observables in relational data bases, *Inform. Systems*, **8**, 281–289.
10. Matúš, F., and Studený, M. (1995) Conditional independencies among four random variables I., *Combinatorics, Probability and Computing*, **4**, 269–278.
11. Matúš, F. (1995) Conditional independencies among four random variables II., *Combinatorics, Probability and Computing*, **4**, 407–417.
12. Matúš, F. (1998) Conditional independencies among four random variables III., submitted to *Combinatorics, Probability and Computing.*
13. Pearl, J., and Paz, A. (1987) Graphoids: graph-based logic for reasoning about relevance relations, in B. Du Boulay, D. Hogg and L. Steels (eds.), *Advances in Artificial Intelligence - II*, North Holland, Amsterdam, pp. 357–363.
14. Pearl, J. (1988) *Probabilistic Reasoning in Intelligent Systems: networks of plausible inference*, Morgan Kaufmann, San Mateo.
15. Perez, A. (1977) ε-admissible simplifications of the dependence structure of a set of random variables, *Kybernetika*, **13**, 439–449.
16. Rényi, A. (1959) On measures of dependence, *Acta Math. Acad. Sci. Hung.*, **10**, 441–451.

17. Spohn, W. (1980) Stochastic independence, causal independence and shieldability, *J. Philos. Logic*, **9**, 73–99.
18. Studený, M. (1987) Asymptotic behaviour of empirical multiinformation, *Kybernetika*, **23**, 124–135.
19. Studený, M. (1989) Multiinformation and the problem of characterization of conditional independence relations, *Problems of Control and Information Theory*, **18**, 3–16.
20. Studený, M. (1992) Conditional independence relations have no finite complete characterization, in S. Kubík and J.Á. Víšek (eds.), *Information Theory, Statistical Decision Functions and Random Processes: proceedings of the 11th Prague conference - B*, Kluwer, Dordrecht (also Academia, Prague), pp. 377–396.
21. Studený, M. (1987) The concept of multiinformation in probabilistic decision-making (in Czech), PhD. thesis, Institute of Information Theory and Automation, Czechoslovak Academy of Sciences, Prague.
22. Vejnarová, J. (1994) A few remarks on measures of uncertainty in Dempster-Shafer theory, *Int. J. General Systems*, **22**, pp. 233–243.
23. Vejnarová J. (1997) Measures of uncertainty and independence concept in different calculi, accepted to *EPIA'97*.
24. Watanabe, S. (1960) Information theoretical analysis of multivariate correlation, *IBM Journal of research and development*, **4**, pp. 66–81.
25. Watanabe, S. (1969) *Knowing and Guessing: a qualitative study of inference and information*, John Wiley, New York.
26. Xiang, Y., Wong, S.K.M., and Cercone, N. (1996) Critical remarks on single link search in learning belief networks, in E. Horvitz and F. Jensen (eds.), *Uncertainty in Artificial Intelligence: proceedings of 12th conference,* Morgan Kaufman, San Francisco, 564–571.
27. Zhang, Z., and Yeung, R. (1997) A non-Shannon type conditional information inequality, to appear in *IEEE Transactions on Information Theory.*
28. Zvárová, J. (1974) On measures of statistical dependence, *Časopis pro pěstování matematiky*, **99**, 15–29.
29. Zvárová, J., and Studený, M. (1997) Information-theoretical approach to constitution and reduction of medical data, *Int. J. Medical Informatics*, **45**, 65–74.