KYBERNETIKA — VOLUME *43* (2007), NUMBER 5, PAGES 591 – 618

COMPARISON OF TWO METHODS FOR APPROXIMATION OF PROBABILITY DISTRIBUTIONS WITH PRESCRIBED MARGINALS

Albert Perez and Milan Studený

Let P be a discrete multidimensional probability distribution over a finite set of variables N which is only partially specified by the requirement that it has prescribed given marginals $\{P_A;\ A \in \mathcal{S}\}$, where $\mathcal{S}$ is a class of subsets of N with $\bigcup \mathcal{S} = N$. The paper deals with the problem of approximating P on the basis of those given marginals. The divergence of an approximation $\hat{P}$ from P is measured by the relative entropy $H(P|\hat{P})$. Two methods for approximating P are compared. One of them uses formerly introduced concept of *dependence structure simplification* (see Perez [4]). The other one is based on an *explicit expression*, which has to be normalized. We give examples showing that neither of these two methods is universally better than the other. If one of the considered approximations $\hat{P}$ really has the prescribed marginals then it appears to be the distribution P with minimal possible multiinformation. A simple condition on the class $\mathcal{S}$ implying the existence of an approximation $\hat{P}$ with prescribed marginals is recalled. If the condition holds then both methods for approximating P give the same result.

Keywords: marginal problem, relative entropy, dependence structure simplification, explicit expression approximation, multiinformation, decomposable model, asteroid

AMS Subject Classification: 68T37, 62C25

PREFACE: MEMORIES OF THE SECOND AUTHOR

This paper was written particularly for this Special Issue of Kybernetika in honour of Albert Perez. I had the opportunity to be the last doctoral student of Dr. Perez. I joined the Institute of Information Theory and Automation in 1983 to start my studies for a CSc degree[1] under his supervision. I am indebted to him for directing me towards the interesting topic of probabilistic decision making. What I learned from him during my doctoral studies was the base of my later research on probabilistic conditional independence. For example, the basic idea of using information-theoretical tools in this field was inspired by his paper [4]. After defending my CSc thesis in 1987 I became a regular member of the department led by Albert Perez.

[1] This is the official name of the scientific degree conferred in Czechoslovakia in the 1980s. Nowadays, doctoral students get PhD degree.

He tried to stimulate the activity of his colleagues in the department by organizing a weekly seminar (I also attended). Moreover, he himself continued in research activity until his retirement in 1990.

We renewed our contacts in November 2001 when I invited him to a small celebration in a restaurant. During the celebration, we agreed to have another meeting, this time in the Institute, together with two other my colleagues and former co-workers, Radim Jiroušek and Otakar Kříž. Otakar, Radim and I expected an informal meeting over some refreshment but when Albert Perez came he wanted us to discuss with him on scientific theme. We learned that he returned to the research in the area of probabilistic decision making. Thus, in the period 2001 − 2003, we had a chance to discuss with Albert Perez in more details his latest ideas. I personally visited him a few times in his flat. We mainly discussed former preliminary versions of the manuscript [7], he planned to publish after all relevant changes.

When I phoned him in December 2003 to arrange giving him my comments on the last version of [7] he did not answer the phone. My colleagues and I learned later that it was because he was already dead. After the funeral, Radim Jiroušek came with an idea to prepare in future a Special Issue of Kybernetika in honour of Albert Perez. I promised to write a paper based on [7] and submit it to the volume. Of course, the present paper differs from the original manuscript quite a lot: I changed the structure of the paper and omitted some points. Nevertheless, since the paper is very substantially based on the results and ideas of Albert Perez, he is the first author.

1. INTRODUCTION

The paper deals with the following problem. Let N be a finite non-empty set of variables, $\mathcal{S}$ be a class of subsets of N whose union is N and $\mathcal{M} = \{P_A; A \in \mathcal{S}\}$ a given system of marginals of a discrete probability distribution P over N.[2] In general, P is not uniquely determined by $\mathcal{M}$. Thus, we only know that P belongs to the class $\mathcal{K}_{\mathcal{M}}$ of discrete probability distributions over N that have the prescribed system of marginals $\mathcal{M}$. We are interested in the problem of approximating P on the basis of $\mathcal{M}$. More specifically, we consider special approximations $\hat{P}$ of P. These are probability distributions over N "constructed" from $\mathcal{M}$ by means of "multiplication" in a special way. Actually, we deal with and compare two special methods for constructing approximations of this kind. The first approach leads to *dependence structure simplifications*, introduced already in [4]. In the present paper, we introduce an alternative method based on a certain *explicit expression*, which has to be normalized. To compare the quality of approximations we use the relative entropy $H(P|\hat{P})$ as the measure of divergence of an approximation $\hat{P}$ from P. The point is that the quality of an approximation $\hat{P}$ of the considered kind actually does not depend on the choice of $P \in \mathcal{K}_{\mathcal{M}}$. This is because, for any $P \in \mathcal{K}_{\mathcal{M}}$ and any approximation $\hat{P}$ of this kind, the following formula holds:

$$H(P|\hat{P}) = I(P) - I_{\mathcal{M}}(\hat{P}), \tag{1}$$

[2] Of course, P_A is a distribution over A where $A \subseteq N$.

where $I(P)$ is the multiinformation of P and $I_{\mathcal{M}}(\hat{P})$ an expression, called the *information content* of $\hat{P}$, that does not depend on particular $P \in \mathcal{K}_{\mathcal{M}}$. The motivation for this problem comes from probabilistic decision making. More specifically, the considered approximations can be utilized in multi-symptom diagnosis making.

The structure of the paper is as follows. Section 2 is an overview of basic concepts and facts. We recall some information-theoretical concepts and describe the considered situation in detail and in mathematical terms. In Section 3 we introduce the concept of an $\mathcal{M}$-construct, which is the above mentioned approximation of $P \in \mathcal{K}_{\mathcal{M}}$ constructed from $\mathcal{M}$ by "multiplication". We also derive the formula (1) there and explain the idea of application of $\mathcal{M}$-constructs in multi-symptom diagnosis making. The concept of a dependence structure simplification (DSS) is dealt with in Section 4. We recall the definition from [8] and the respective formula for the information content. We also discuss the problem of finding an optimal DSS and a possible modification of the definition of a DSS. Section 5 is devoted to approximating P by means of a special explicit expression. We explain the role of a normalizing constant and give a formula for the respective information content $I_{\mathcal{M}}(\hat{P})$. Section 6 is devoted to the case of fitting marginals. This is the fortunate case when $\hat{P}$ falls within $\mathcal{K}_{\mathcal{M}}$. We show that then $\hat{P}$ is the probability distribution from $\mathcal{K}_{\mathcal{M}}$ which has minimal multiinformation.[3] Section 7 gives a simple sufficient condition on $\mathcal{S}$ which ensures that the approximation $\hat{P}$ falls in $\mathcal{K}_{\mathcal{M}}$. The condition, named the *running intersection property*, is strongly related to well-known decomposable graphical models [3]. In Section 8 we discuss the barycenter principle for the choice of a representative of $\mathcal{K}_{\mathcal{M}}$ introduced in [5] and show that the choice of an optimal DSS is in concordance with this principle. Open problems are formulated in Conclusions. The Appendix contains several examples including the crucial ones showing that none of two described methods for approximating P is better than the other in the sense of the information content.

2. BASIC CONCEPTS

Throughout the paper we will assume the situation described in the following subsection.

2.1. The considered situation

Let N be a non-empty finite set of variables. Every $i \in N$ has assigned the respective *individual sample space* $\boldsymbol{X}_i$, which is a non-empty finite set of its possible values. Given a set $A \subseteq N$, by a *configuration* of values for A we mean any list $[x_i]_{i\in A}$ such that $x_i \in \boldsymbol{X}_i$ for any $i \in A$. Of course, if $A \neq \emptyset$ then a configuration for A is nothing but an element of the Cartesian product $\prod_{i\in A} \boldsymbol{X}_i$. However, the above definition also formally introduces a configuration for the empty set; it is simply the empty list. We will denote the set of configurations for $A \subseteq N$ by $\boldsymbol{X}_A$ and call it the sample space for A. The *joint sample space* is then $\boldsymbol{X}_N$.

[3] This is equivalent to the requirement that it has maximal entropy within $\mathcal{K}_{\mathcal{M}}$.

Two basic operations with configurations are as follows. Given $A \subseteq B \subseteq N$ and $x = [x_i]_{i \in B} \in \boldsymbol{X}_B$, the *marginal configuration* (of x) for A, denoted by x_A, is the restriction of the list x to the items that correspond the variables in A: $x_A \equiv [x_i]_{i \in A}$. Given $A, C \subseteq N$, $A \cap C = \emptyset$, by *concatenation* of $x = [x_i]_{i \in A} \in \boldsymbol{X}_A$ and $y = [y_i]_{i \in C} \in \boldsymbol{X}_C$ we will understand the configuration $z = [z_i]_{i \in A \cup C}$ for $A \cup C$ obtained by merging the lists x and y: that is, $z_i = x_i$ for $i \in A$ and $z_i = y_i$ for $i \in C$. It will be denoted by $[x, y]$.

Further assumption is that a class $\mathcal{S}$ of subsets of N is given whose union is N. The symbol $\mathcal{S}^{\downarrow}$ will denote the class $\{B\,;\; B \subseteq A \text{ for } A \in \mathcal{S}\}$ of subsets of sets in $\mathcal{S}$. If $\mathcal{A} \subseteq \mathcal{S}$ is a non-empty subclass of $\mathcal{S}$ then the symbol $\bigcup \mathcal{A}$, respectively $\bigcap \mathcal{A}$, will be used to denote the union, respectively the intersection, of sets in $\mathcal{A}$.

A basic concept is the concept of a probability measure on $\boldsymbol{X}_N$. A probability measure of this kind is given by its *density*, which is a function $p : \boldsymbol{X}_N \to [0,1]$ such that $\sum \{p(x)\,;\; x \in \boldsymbol{X}_N\} = 1$. The respective probability measure is then a set function on subsets of $\boldsymbol{X}_N$ which ascribes $P(\boldsymbol{T}) = \sum \{p(x)\,;\; x \in \boldsymbol{T}\}$ to every $\boldsymbol{T} \subseteq \boldsymbol{X}_N$.[4] By a discrete *probability distribution over* N we will understand a probability measure on any joint sample sample space $\boldsymbol{X}_N$ of the above-mentioned kind.

Given a probability measure P on $\boldsymbol{X}_N$ and $A \subseteq N$, the symbol P^A will denote the *marginal* of P for A, that is, the probability measure on $\boldsymbol{X}_A$ given by:

$$P^A(\boldsymbol{Y}) \;=\; P(\{x \in \boldsymbol{X}_N\,;\; x_A \in \boldsymbol{Y}\}) \quad \text{for } \boldsymbol{Y} \subseteq \boldsymbol{X}_A\,.$$

It is easy to see that P^A is determined by the *marginal density* p^A for A, given by

$$p^A(y) \;=\; \sum \{\, p([x,y])\,;\; x \in \boldsymbol{X}_{N \setminus A}\} \quad \text{for } y \in \boldsymbol{X}_A\,.$$

In particular, $p^N = p$ and $p^{\emptyset} \equiv 1$. Observe that marginal densities comply with the following *vanishing principle*:

$$\text{if } A \subseteq B \subseteq N \text{ and } z \in \boldsymbol{X}_B \;\text{ then }\; p^A(z_A) = 0 \text{ implies } p^B(z) = 0\,. \tag{2}$$

The last assumption is that a collection of marginals of a probability measure on $\boldsymbol{X}_N$ is given. More specifically, we assume that a collection of probability measures $\mathcal{M} = \{P_A\,;\; A \in \mathcal{S}\}$ is given, where P_A is a probability measure on $\boldsymbol{X}_A$ for $A \in \mathcal{S}$ and there exists at least one probability measure P on $\boldsymbol{X}_N$ such that

$$\forall\, A \in \mathcal{S} \quad P_A = P^A\,. \tag{3}$$

The last assumption on $\mathcal{M}$ is the requirement of its *strong consistency*.[5] We will use the symbol $\mathcal{K}_{\mathcal{M}}$ to denote the class of all probability measures P on $\boldsymbol{X}_N$ such that (3) holds. The assumption of strong consistency of $\mathcal{M}$ means that $\mathcal{K}_{\mathcal{M}}$ is non-empty. Of course, $\mathcal{K}_{\mathcal{M}}$ may contain more than one probability measure in general.

[4] Of course, then $P(\emptyset) = 0$ by a convention.

[5] As $\mathcal{M}$ is supposed to be a class of marginals of a probability distribution over N it is denoted by the letter $\mathcal{M}$.

Remark 1. One can assume without loss of generality that $\mathcal{S}$ consists of incomparable sets, that is, $A \setminus B \neq \emptyset \neq B \setminus A$ for any pair of distinct sets $A, B \in \mathcal{S}$. This is because otherwise $\mathcal{S}$ can be reduced to

$$\mathcal{S}^{\max} = \{\, A \in \mathcal{S}\,;\ \neg(\exists\, B \in \mathcal{S} \ \text{ with } A \subset B)\,\},$$[6]

and $\mathcal{M}$ to $\mathcal{M}^{\max} = \{P_A\,;\, A \in \mathcal{S}^{\max}\}$. Owing to strong consistency assumption the collection $\mathcal{M}$ can be reconstructed from $\mathcal{M}^{\max}$ (and $\mathcal{S}$) and one has $\mathcal{K}_{\mathcal{M}} = \mathcal{K}_{\mathcal{M}^{\max}}$.

2.1.1. The question of checking consistency

An important question is how to verify the assumption of strong consistency of $\mathcal{M}$. In general, it is not an easy task. The only general method for its verification is to find $P \in \mathcal{K}_{\mathcal{M}}$ directly, but no universal instructions how to do it are available. To show that (3) is not fulfilled the following concept is suitable. We say that $\mathcal{M}$ is *weakly consistent* if

$$\forall\, A, B \in \mathcal{S} \quad (P_A)^{A\cap B} = (P_B)^{A\cap B}\,. \tag{4}$$

Evidently, strong consistency of $\mathcal{M}$ implies its weak consistency. As weak consistency is easy to verify the condition (4) can be used to disprove strong consistency. On the other hand, the weak consistency does not imply the strong one as the following example shows.

Example 1. Put $N = \{a, b, c\}$ and $\boldsymbol{X}_i = \{0, 1\}$ for every $i \in N$. Consider the class of two-element subsets of N, that is, $\mathcal{S} = \{A \subseteq N\,;\, |A| = 2\}$. The density p_A of P_A for any $A \in \mathcal{S}$ is given as follows:

$$p_A(0,0) = p_A(1,1) = \frac{1}{10}\,, \quad p_A(0,1) = p_A(1,0) = \frac{2}{5}\,.$$

As $(p_A)^{\{i\}}(0) = (p_A)^{\{i\}}(1) = 1/2$ for both $i \in A$, the collection $\mathcal{M} = \{P_A\,;\ A \in \mathcal{S}\}$ is weakly consistent. However, (3) is not valid for any P on $\boldsymbol{X}_N$. To see this assume for a contradiction that $P \in \mathcal{K}_{\mathcal{M}}$ with density p exists and put $x \equiv p(1,1,1) \geq 0$. The fact $p_{\{b,c\}}(1,1) = 1/10$ and (3) implies $p(0,1,1) = (1/10) - x$. Hence, by $p_{\{a,b\}}(0,1) = 2/5$ observe $p(0,1,0) = 2/5 - [(1/10) - x] = (3/10) + x$. Finally, by $p_{\{a,c\}}(0,0) = 1/10$ get $p(0,0,0) = 1/10 - [(3/10) + x] = -(2/10) - x$. The fact $p(0,0,0) \geq 0$ gives $x \leq -2/10$, which contradicts the assumption $x \geq 0$.

Fortunately, the condition (4) implies strong consistency under an additional assumption on the class $\mathcal{S}$, namely that $\mathcal{S}$ satisfies so-called *running intersection property* – for detail see Section 7. Moreover, even if that additional condition is not fulfilled strong consistency can sometimes be verified as follows. Provided that (4) holds, an approximation $\hat{P}$ is constructed on the basis of $\mathcal{M}$. Then one can try to check whether $\hat{P}$ has $\mathcal{M}$ as the collection of marginals. This may happen even if $\mathcal{S}$ does not satisfy the running intersection property – see Example 4 in Section 6, where we use the approximations $\hat{P}$ described later in this paper.

[6] Here, $\subset$ denotes strict inclusion of sets.

2.2. Some related concepts and notation

In this subsection we introduce some concepts used systematically in the rest of the paper.

2.2.1. The greatest support

Given a probability measure P on $\boldsymbol{X}_N$ with density p, by the *support* of P will be meant the set $N_P \equiv \{x \in \boldsymbol{X}_N \,;\, p(x) > 0\}$. It is the least subset $\boldsymbol{T} \subseteq \boldsymbol{X}_N$ such that P is concentrated on $\boldsymbol{T}$, that is, $P(\boldsymbol{X}_N \setminus \boldsymbol{T}) = 0$. As $\mathcal{K}_\mathcal{M}$ is a convex set[7] and $\boldsymbol{X}_N$ has finitely many subsets there exists a probability measure $R \in \mathcal{K}_\mathcal{M}$ which has the greatest support in $\mathcal{K}_\mathcal{M}$.[8] It will be denoted by the symbol $N_\mathcal{M}$.

2.2.2. Relative entropy

Given two probability measures P, Q on $\boldsymbol{X}_N$ we say that P is *absolutely continuous* with respect to Q and write $P \ll Q$ if $Q(\boldsymbol{T}) = 0$ implies $P(\boldsymbol{T}) = 0$ for each $\boldsymbol{T} \subseteq \boldsymbol{X}_N$.[9] We also say that Q *dominates* P.

A well-known result is Radon–Nikodym theorem which says that $P \ll Q$ iff there exists a function $\frac{\mathrm{d}P}{\mathrm{d}Q} : \boldsymbol{X}_N \to [0, \infty)$, called the *Radon–Nikodym derivative* of P with respect to Q, such that

$$P(\boldsymbol{T}) = \sum_{x \in \boldsymbol{T}} \frac{\mathrm{d}P}{\mathrm{d}Q}(x) \cdot q(x) \quad \text{for any } \boldsymbol{T} \subseteq \boldsymbol{X}_N,$$

where q is the density of Q. Of course, $\mathrm{d}P/\mathrm{d}Q$ is uniquely determined on N_Q, in particular, on N_P.

The *relative entropy* of P with respect to Q is defined by the formula

$$H(P|Q) \equiv \sum_{x \in \boldsymbol{X}_N,\, p(x)>0} p(x) \cdot \ln \frac{\mathrm{d}P}{\mathrm{d}Q}(x) = \sum_{x \in \boldsymbol{X}_N} q(x) \cdot \frac{\mathrm{d}P}{\mathrm{d}Q}(x) \cdot \ln \frac{\mathrm{d}P}{\mathrm{d}Q}(x)\,,$$

provided that $P \ll Q$ and $H(P|Q) = \infty$ otherwise. A well-known fact is that $H(P|Q) \geq 0$ and $H(P|Q) = 0$ iff $P = Q$ – see § A.6.3 in [9]. Thus, $H(P|Q)$ can be understood as a measure of distinction between P and Q.[10] Observe that, in the considered discrete case, one has $H(P|Q) < \infty$ iff $P \ll Q$. In particular, it follows from the previous observation from Section 2.2.1:

Proposition 1. There exists $R \in \mathcal{K}_\mathcal{M}$ such that $\forall\, P \in \mathcal{K}_\mathcal{M} \quad H(P|R) < \infty$.

[7] This means that it is closed under convex combinations: if $P, Q \in \mathcal{K}_\mathcal{M}$, $\alpha \in [0, 1]$ then $\alpha \cdot P + (1 - \alpha) \cdot Q \in \mathcal{K}_\mathcal{M}$.

[8] Realize that whenever $R = \alpha \cdot P + (1 - \alpha) \cdot Q$ with $\alpha \in (0, 1)$ then $N_R = N_P \cup N_Q$.

[9] Note that in the considered case of a finite joint sample space $\boldsymbol{X}_N$ this is equivalent to the inclusion $N_P \subseteq N_Q$.

[10] However, because it may happen $H(P|Q) \neq H(Q|P)$ even if $P \ll Q \ll P$, it is not a distance.

2.2.3. Dominating product measure

The first step is to realize that a given collection of marginals $\mathcal{M}$ can uniquely be extended to a system of marginals $\mathcal{M}^{\downarrow} = \{P_B\,;\ B \in \mathcal{S}^{\downarrow}\}$. Indeed, given $B \in \mathcal{S}^{\downarrow}$ there exists $A \in \mathcal{S}$ with $B \subseteq A$ and we put $P_B = (P_A)^B$. The weak consistency condition (4) implies that the definition does not depend on the choice of $A \in \mathcal{S}$, it only depends on $\mathcal{M}$. Actually, the fact that every $P \in \mathcal{K}_{\mathcal{M}}$ satisfies (3) implies $P_B = P^B$ for every $P \in \mathcal{K}_{\mathcal{M}}$ and $B \in \mathcal{S}^{\downarrow}$. Given $\mathcal{M}^{\downarrow}$ and $B \in \mathcal{S}^{\downarrow}$ the symbol p_B will denote the density of P_B.

Given $i \in N$, the assumption $\bigcup \mathcal{S} = N$ implies that $\{i\} \in \mathcal{S}^{\downarrow}$ for every $i \in N$. Let us put $P_i = P_{\{i\}}$ then. The product of these probability measures $\prod_{i\in N} P_i$ will be called the *dominating product measure* and denoted by L. It is a probability measure on $\boldsymbol{X}_N$ with density l is given by

$$l(x) = \prod_{i\in N} p_{\{i\}}(x_i) \qquad \text{for every } x = [x_i]_{i\in N} \in \boldsymbol{X}_N\,.$$

The terminology is justified because one can easily observe that $P \ll L$ for every $P \in \mathcal{K}_{\mathcal{M}}$.[11] This allows one to derive $P_B \ll L^B$ for every $B \in \mathcal{S}^{\downarrow}$.[12] In particular, the Radon–Nikodym derivative $\mathrm{d}P_B/\mathrm{d}L^B$ exists for every $B \in \mathcal{S}^{\downarrow}$ and is uniquely determined on the support of L^B – it will be denoted by the symbol f_B in the sequel. Of course,

$$f_B(x_B) \;=\; p_B(x_B)\cdot \prod_{j\in B} p_{\{j\}}(x_j)^{-1} \qquad \text{for any } x \in \boldsymbol{X}_N \text{ with } l(x) > 0 \text{ and } B \subseteq N\,.$$

Remark 2. Note that we can assume without loss of generality $l(x) > 0$ for every $x \in \boldsymbol{X}_N$. Indeed, otherwise replace every $\boldsymbol{X}_i$, by $\boldsymbol{X}'_i = \{y \in \boldsymbol{X}_i\,;\ p_{\{i\}}(y) > 0\,\}$ for any $i \in N$. Then every $P \in \mathcal{K}_{\mathcal{M}}$ is concentrated on $\boldsymbol{X}'_N = \prod_{i\in N} \boldsymbol{X}'_i$.

2.2.4. Multiinformation and entropy

Given a probability measure P on $\boldsymbol{X}_N$, the relative entropy $H(P|\prod_{i\in N} P^{\{i\}})$ will be called its *multiinformation* and denoted by $I(P)$. In the considered discrete case one always has $P \ll \prod_{i\in N} P^{\{i\}}$, which implies that $I(P) < \infty$. Of course, if $P \in \mathcal{K}_{\mathcal{M}}$ then $I(P) = H(P|L)$.

The *entropy* of a probability measure P on $\boldsymbol{X}_N$, denoted by $H(P)$, is given by the following formula:

$$H(P) = \sum_{x\in \boldsymbol{X}_N,\, p(x)>0} p(x)\cdot \ln \frac{1}{p(x)}\,.$$ [13]

Note that entropy is a non-negative (finite) real number. The following lemma recalls basic facts on multiinformation and entropy in the considered situation.

[11] Observe that $p^{\{i\}}(x_{\{i\}}) = 0$ implies $p(x) = 0$ for $x \in \boldsymbol{X}_N$, $i \in N$ by vanishing principle (2).

[12] Realize that $P_B = P^B$ and $P \ll L$ gives $P^B \ll L^B$.

[13] Of course, the given definition only makes sense in the discrete case.

Lemma 2. There exists uniquely determined $P_* \in \mathcal{K}_{\mathcal{M}}$ such that

$$H(P_*) = \max \{ H(P) \, ; \, P \in \mathcal{K}_{\mathcal{M}} \} .$$

It coincides with unique $P_* \in \mathcal{K}_{\mathcal{M}}$ such that $I(P_*) = \min \{I(P); \, P \in \mathcal{K}_{\mathcal{M}}\}$. Moreover, there exists (at least one) $P_\dagger \in \mathcal{K}_{\mathcal{M}}$ with $I(P_\dagger) = \max \{I(P); \, P \in \mathcal{K}_{\mathcal{M}} \} < \infty$.

Proof. Let us introduce an auxiliary (continuous) real function $h : \mathbb{R} \to \mathbb{R}$ as follows:

$$h(y) = \begin{cases} y \cdot \ln y & \text{if } y > 0, \\ 0 & \text{otherwise.} \end{cases}$$

Observe that $-H(P) = \sum_{x \in \boldsymbol{X}_N} h(p(x))$ for every probability measure P on $\boldsymbol{X}_N$. As h is strictly convex on $[0, \infty)$ the function $P \mapsto -H(P)$ is a strictly convex continuous function on $\mathcal{K}_{\mathcal{M}}$. Moreover, $\mathcal{K}_{\mathcal{M}}$ is a convex compact subset of $\mathbb{R}^{\boldsymbol{X}_N}$. Thus, the function achieves both the maximum and the minimum on $\mathcal{K}_{\mathcal{M}}$ and the $P_* \in \mathcal{K}_{\mathcal{M}}$ in which the minimum is achieved is uniquely determined. The second basic fact is that

$$I(P) = -H(P) + \sum_{i \in N} H(P^{\{i\}}) \quad \text{for every } P \in \mathcal{K}_{\mathcal{M}} \, . \tag{5}$$

Since one-dimensional marginals are shared within $\mathcal{K}_{\mathcal{M}}$, the second sum in (5) is constant. This observation implies the remaining statements in the lemma. □

3. $\mathcal{M}$–CONSTRUCT

The following definition is a modification of the concept introduced in [7].

Definition 1. Let $\mathcal{M} = \{P_A ; \, A \in \mathcal{S}\}$ be a strongly consistent collection of probability measures. By an $\mathcal{M}$-*construct* we will understand any probability measure Q on $\boldsymbol{X}_N$ which is absolutely continuous with respect to the dominating product measure L and whose Radon–Nikodym derivative $\mathrm{d}Q/\mathrm{d}L$ satisfies the condition

$$\forall \, x \in N_{\mathcal{M}} \quad \frac{\mathrm{d}Q}{\mathrm{d}L}(x) \; = \; k \cdot \prod_{B \in \mathcal{S}^{\downarrow}} f_B(x_B)^{\nu(B)} \, , \tag{6}$$

where $k \in (0, \infty)$ and $\nu(B) \in \mathbb{Z}$, $B \in \mathcal{S}^{\downarrow}$ are the respective parameters of Q.[14]

The *multiinformation content* of the $\mathcal{M}$-construct Q given by (6) is the following number, denoted by $I_{\mathcal{M}}(Q)$,

$$I_{\mathcal{M}}(Q) \; = \; \ln k + \sum_{B \in \mathcal{S}^{\downarrow}} \nu(B) \cdot I(P_B) \, . \tag{7}$$

[14] Recall that the functions f_B, $B \in \mathcal{S}^{\downarrow}$, which are introduced in Section 2.2.3, are uniquely determined by $\mathcal{M}$.

Note that the multiinformation content depends solely on the $\mathcal{M}$-construct Q and not on its particular parameters from (6) – this follows from later formula (9), where p is the density of arbitrary $P \in \mathcal{K}_{\mathcal{M}}$.

An example of an $\mathcal{M}$-construct is the dominating product measure L – it suffices to put $k = 1$, $\nu(\{i\}) = 1$ for $i \in N$ and $\nu(B) = 0$ for remaining $B \in \mathcal{S}^{\downarrow}$.[15] However, there are other examples of $\mathcal{M}$-constructs, namely the approximations of $P \in \mathcal{K}_{\mathcal{M}}$ mentioned in Sections 4 and 5. The following lemma says that every $\mathcal{M}$-construct gives a lower estimate of the minimal multiinformation in $\mathcal{K}_{\mathcal{M}}$.

Lemma 3. Let $\mathcal{M} = \{P_A;\ A \in \mathcal{S}\}$ be a strongly consistent collection of probability measures and Q be an $\mathcal{M}$-construct. Then $P \ll Q \ll L$ for every $P \in \mathcal{K}_{\mathcal{M}}$. Moreover,

$$\min\{I(P)\,;\ P \in \mathcal{K}_{\mathcal{M}}\} \geq I_{\mathcal{M}}(Q)\,, \tag{8}$$

and the equality in (8) occurs iff $Q \in \mathcal{K}_{\mathcal{M}}$, in which case $I_{\mathcal{M}}(Q) = I(Q)$. Actually, one has

$$H(P|Q) = I(P) - I_{\mathcal{M}}(Q) \quad \text{for any } P \in \mathcal{K}_{\mathcal{M}} \text{ and an } \mathcal{M}\text{-construct } Q.$$

Proof. The fact $Q \ll L$ follows directly from Definition 1. To show $P \ll Q$ it suffices to verify $p(x) > 0 \;\Rightarrow\; q(x) > 0$ for $x \in \boldsymbol{X}_N$. If $p(x) > 0$ then $l(x) > 0$ and to get $q(x) > 0$ one needs to show that $(\mathrm{d}Q/\mathrm{d}L)(x) > 0$.[16] However, then $x \in N_P \subseteq N_{\mathcal{M}}$ and the formula (6) for $\mathrm{d}Q/\mathrm{d}L(x)$ can be used. The vanishing principle for marginal densities (2) implies $p_B(x_B) > 0$ for every $B \subseteq N$ and this gives $f_B(x_B) > 0$ for any $B \in \mathcal{S}^{\downarrow}$.[17] In particular, (6) gives $(\mathrm{d}Q/\mathrm{d}L)(x) > 0$, which was needed.

The next step is to observe that

$$\sum_{x \in \boldsymbol{X}_N,\, p(x)>0} p(x) \cdot \ln \frac{\mathrm{d}Q}{\mathrm{d}L}(x) \;=\; I_{\mathcal{M}}(Q)\,. \tag{9}$$

Indeed, whenever $x \in \boldsymbol{X}_N$, $p(x) > 0$ then $x \in N_P \subseteq N_{\mathcal{M}}$ and (6) can be used, which gives:

$$\sum_{x \in \boldsymbol{X}_N,\, p(x)>0} p(x) \cdot \ln \frac{\mathrm{d}Q}{\mathrm{d}L}(x) = \sum_{p(x)>0} p(x) \cdot \ln k + \sum_{B \in \mathcal{S}^{\downarrow}} \nu(S) \cdot \sum_{p(x)>0} p(x) \cdot \ln f_B(x_B)\,.$$

To get the expression in (7) write the last internal sum as follows:

$$\begin{aligned}
\sum_{x \in \boldsymbol{X}_N,\, p(x)>0} p(x) \cdot \ln f_B(x_B) \;&=\; \sum_{y \in \boldsymbol{X}_B,\, p_B(y)>0} \;\; \sum_{z \in \boldsymbol{X}_{N\setminus B},\, p([y,z])>0} p([y,z]) \cdot \ln f_B(y) \\
&= \sum_{y \in \boldsymbol{X}_B,\, p_B(y)>0} \ln f_B(y) \cdot \sum_{z \in \boldsymbol{X}_{N\setminus B},\, p([y,z])>0} p([y,z]) \\
&= \sum_{y \in \boldsymbol{X}_B,\, p_B(y)>0} \ln f_B(y) \cdot p_B(y)\,,
\end{aligned}$$

[15] Note that $f_B \equiv 1$ whenever $|B| = 1$.

[16] Realize that $q(x) = \mathrm{d}Q/\mathrm{d}L(x) \cdot l(x)$.

[17] Recall that $p_B(x_B) = (\mathrm{d}P_B/\mathrm{d}L^B)(x_B) \cdot l^B(x_B) = f_B(x_B) \cdot l^B(x_B)$.

and realize that $f_B = \mathrm{d}P_B/\mathrm{d}L^B$.

Now, (9) can be used to derive (8). Consider $P \in \mathcal{K}_{\mathcal{M}}$. The fact $P \ll Q \ll L$ implies that $\frac{\mathrm{d}P}{\mathrm{d}Q}(x) = \frac{\mathrm{d}P}{\mathrm{d}L}(x)/\frac{\mathrm{d}Q}{\mathrm{d}L}(x)$ for every $x \in N_Q$.[18] This allow one to write using (9):

$$\begin{aligned} 0 \quad \leq \quad & H(P|Q) = \sum_{x \in \boldsymbol{X}_N,\, p(x)>0} p(x) \cdot \ln \frac{\mathrm{d}P}{\mathrm{d}Q}(x) \\ = \quad & \sum_{p(x)>0} p(x) \cdot \ln \frac{\mathrm{d}P}{\mathrm{d}L}(x) - \sum_{p(x)>0} p(x) \cdot \ln \frac{\mathrm{d}Q}{\mathrm{d}L}(x) = I(P) - I_{\mathcal{M}}(Q)\,. \end{aligned}$$

This gives $I(P) \geq I_{\mathcal{M}}(Q)$ and (8). Moreover, the equality $I(P) = I_{\mathcal{M}}(Q)$ means that $H(P|Q) = 0$ and this occurs iff $P = Q$. However, $P = Q$ implies $Q \in \mathcal{K}_{\mathcal{M}}$. Conversely, if $Q \in \mathcal{K}_{\mathcal{M}}$ then we put $P' = Q \in \mathcal{K}_{\mathcal{M}}$ and repeat the above consideration to get $0 = H(P'|Q) = I(P') - I_{\mathcal{M}}(Q)$. The formula (8) allows us to write

$$I(P') \geq \min\{I(P);\, P \in \mathcal{K}_{\mathcal{M}}\} \geq I_{\mathcal{M}}(Q) = I(P')\,,$$

which implies that the equality in (8) occurs and $I(Q) = I(P') = I_{\mathcal{M}}(Q)$. The last equality mentioned in Lemma 3 was verified above. □

3.1. The idea of application to diagnosis making

In this subsection we describe how $\mathcal{M}$-constructs can possibly be utilized in multisymptom diagnosis making. Let us consider the following special situation. Let $d \in N$ be a distinguished *diagnostic variable*, that is, a variable whose value we would like to "determine" on the basis of remaining variables. The variables in $S \equiv N \setminus \{d\}$ are, therefore, called *symptom variables.*

Our decision should be based on an "observed" configuration of values $x_S \equiv [x_i]_{i \in S}$, where $x_i \in \boldsymbol{X}_i$ for $i \in S$. On the basis of the configuration x_S, we would like to determine the most probable value of the diagnostic variable. That means, we would like to find $y \in \boldsymbol{X}_d$ with maximal conditional probability $P_{d|S}(y|x_S)$.[19] The complication is that we do not know the "actual" distribution P which describes the probabilistic relationships among variables in N. Therefore, we try to replace P by its approximation $\hat{P}$ based on a given system of marginals $\mathcal{M} = \{P_A;\, A \in \mathcal{S}\}$ with $d \in A$ for every $A \in \mathcal{S}$.[20]

There are two methodological procedures that can be applied in this situation. The first approach is based on *direct approximation* of P: we use an approximation

[18] Observe that $\frac{\mathrm{d}Q}{\mathrm{d}L}(x) > 0$ for every $x \in N_Q$ and use the definition of the Radon–Nikodym derivative.

[19] Of course, this problem is equivalent to the problem of finding $y \in \boldsymbol{X}_d$ which maximizes $P([y, x_S])$. This alternative formulation formally avoids assuming that the marginal probability $P^S(x_S)$ of the observed configuration is strictly positive, which assumption is needed to define the conditional probability $P_{d|S}(\star|x_S)$.

[20] This is an additional assumption we made in the considered special situation of diagnosis making.

$\hat{P}$ instead of P which leads to the following estimator of the value y of the diagnostic variable:

$$\psi_1(x_S) \;=\; \operatorname{argmax}\,\{\hat{P}([y,x_S])\,;\; y\in \boldsymbol{X}_d\,\}\,.$$[21]

The second approach is a Bayesian one. It is based on the idea that a prior distribution Q_d is given on $\boldsymbol{X}_d$. In this case, we use $Q_d\cdot\hat{P}_{S|d}$ instead of P, where $\hat{P}_{S|d}$ is an estimate of the respective conditional probability. For fixed $y\in\boldsymbol{X}_d$, we consider the system of probability distributions over subsets of $S\equiv N\setminus\{d\}$, namely $\mathcal{M}[y]=\{P_{A\setminus\{d\}|d}(\star|y)\,;\,A\in\mathcal{S}\}$, which should be the system of marginals of the conditional probability $P_{S|d}(\star|y)$.[22] Now, on the basis of $\mathcal{M}[y]$, we can analogously construct an approximation $\hat{P}_{[y]}$ of $P_{S|d}(\star|y)$.[23] This leads to the following estimator:

$$\psi_2(x_S) \;=\; \operatorname{argmax}\,\{\,Q_d(y)\cdot\hat{P}_{[y]}(x_S)\,;\; y\in\boldsymbol{X}_d\}\,.$$

4. DEPENDENCE STRUCTURE SIMPLIFICATIONS

This is one of the ways to approximate measures from $\mathcal{K}_{\mathcal{M}}$, already proposed in the 1970s by the first author in [4]. Dependence structure simplifications were also dealt with in the CSc thesis of the second author [8]. The following is a minor modification of the definition from [8].

Definition 2. Let $\mathcal{M}=\{P_A;\,A\in\mathcal{S}\}$ be a strongly consistent collection of probability measures. Let us choose a total ordering $\tau:S_1,\dots,S_n$, $n\geq 1$ of elements of $\mathcal{S}$ and put $F_j\equiv S_j\cap\bigcup\{S_k;k<j\}$ and $G_j\equiv S_j\setminus F_j$ for $1\leq j\leq n$.[24] By a *choice* for $\mathcal{M}$ and τ we will understand a mapping ϑ which assigns a conditional density $p_{G_j|F_j}$ on $\boldsymbol{X}_{G_j}$ given $\boldsymbol{X}_{F_j}$ consonant with p_{S_j} to every $1\leq j\leq n$.[25]

By a *dependence structure simplification* (DSS) for $\mathcal{M}$ determined by ordering τ and the choice ϑ will be understood a probability measure on $\boldsymbol{X}_N$ whose density $\overline{p}_{\tau,\vartheta}$ is given by

$$\overline{p}_{\tau,\vartheta}(x) \;=\; \prod_{j=1}^{n} p_{G_j|F_j}(x_{G_j}|x_{F_j}) \qquad \text{for every } x\in\boldsymbol{X}_N\,. \tag{10}$$[26]

The class of all DSSs for $\mathcal{M}$ (determined by any possible τ and ϑ) will be denoted by $\mathcal{D}_{\mathcal{M}}$.

[21] The symbol $\operatorname{argmax}\,\{f(y)\,;\;y\in\boldsymbol{Y}\,\}$ denotes any $z\in\boldsymbol{Y}$ such that $f(z)=\max\,\{f(y)\,;\;y\in\boldsymbol{Y}\,\}$.

[22] We implicitly assume that $P_d(y)>0$ for every $y\in\boldsymbol{X}_d$ for otherwise $\boldsymbol{X}_d$ can be reduced to $y\in\boldsymbol{x}_d\,;\;P_d(y)>0\}$.

[23] Indeed, the situation is completely analogous to the problem of approximating P on the basis of $\mathcal{M}$ – the only difference is that N is replaced by S and $\mathcal{M}$ by $\mathcal{M}[y]$.

[24] In particular, $F_1=\emptyset$ and $G_1=S_1$.

[25] By a *conditional density* on $\boldsymbol{X}_A$ given $\boldsymbol{X}_C$ is meant a function of two variables $[y,z]\mapsto p_{A|C}(y|z)$, $y\in\boldsymbol{X}_A$, $z\in\boldsymbol{X}_C$ such that $\forall\,z\in\boldsymbol{X}_C$ its restriction $y\mapsto p_{A|C}(y|z)$, $y\in\boldsymbol{X}_A$ is a density of a probability measure on $\boldsymbol{X}_A$. It is called *consonant* with a density q on $\boldsymbol{X}_{AC}$ if $p_{A|C}(y|z)=q([y,z])/q^C(z)$ whenever $q^C(z)>0$.

[26] It can be shown by induction on n that (10) indeed defines a density of a probability measure on $\boldsymbol{X}_N$.

Remark 3. The concept of a “choice for $\mathcal{M}$ and τ” is a technical concept which is needed to overcome some troubles one can come across if densities of given distributions from $\mathcal{M}$ vanish for certain marginal configurations.

Of course, if $p_{F_j} > 0$ on X_{F_j} for some $j \in \{1,\ldots,n\}$ then the conditional density $p_{G_j|F_j}$ consistent with p_{S_j} is uniquely determined as the ratio p_{S_j}/p_{F_j}.[27] Therefore, if the ordering τ is such that $p_{F_j} > 0$ on $\boldsymbol{X}_{F_j}$ for any $j = 1,\ldots,n$,[28] then all terms in (10) are uniquely determined and the formula takes the form

$$\overline{p}_{\tau}(x) = \prod_{j=1}^{n} \frac{p_{S_j}(x_{S_j})}{p_{F_j}(x_{F_j})} \qquad \text{for any } x \in \boldsymbol{X}_N\,. \tag{11}$$

In that special case the concept of choice (for $\mathcal{M}$ and τ) is superfluous and can be omitted.

However, on the other hand, if $p_{F_j}(x_{F_j}) = 0$ for at least one $j \in \{1,\ldots,n\}$ and $x \in \boldsymbol{X}_N$ then the respective term $p_{S_j}(x_{S_j})/p_{F_j}(x_{F_j})$ in (11) is an undefined ratio $0/0$! It may even happen that no other term $p_{S_k}(x_{S_k})/p_{F_k}(x_{F_k})$ for $k \neq j$ vanishes for that particular configuration $x \in \boldsymbol{X}_N$, which means that $\overline{p}_{\tau}(x)$ is not defined then – see Example 6 from Section A1. Therefore, some additional “conventions” are needed to ensure that the formula (11) defines a density on $\boldsymbol{X}_N$. One of the methods to settle the matter is to choose and fix versions of conditional densities. Surprisingly, this choice appears not to influence the quality of the resulting approximation from the point of view we consider – see Lemma 4. Another possible approach to deal with the above problem is mentioned in Remark 4.

Another interesting observation is that whenever $S_j \subseteq \bigcup\{S_k; k < j\}$ for some $j \in \{1,\ldots,n\}$ then p_{S_j} does not influence the value of $\overline{p}_{\tau,\vartheta}$.[29] The following is a basic observation concerning DSSs.

Lemma 4. Assume that $l(x) > 0$ for every $x \in \boldsymbol{X}_N$.[30] Then every $Q \in \mathcal{D}_{\mathcal{M}}$ is an $\mathcal{M}$-construct and, provided that its density $\overline{p}_{\tau,\vartheta}$ is given by (10), its multiinformation content is

$$I_{\mathcal{M}}(Q) = \sum_{A\in\mathcal{S}} I(P_A) - \sum_{j=2}^{n} I(P_{F_j}) = \prod_{B\in\mathcal{S}^{\downarrow}} \nu(B)\cdot I(P_B)\,, \tag{12}$$

where

$$\nu(B) = |\{j; S_j = B\}| - |\{j; F_j = B\}| \qquad \text{for any } B \in \mathcal{S}^{\downarrow}\,. \tag{13}$$

In particular, the multiinformation content of Q does not depend on the choice ϑ for $\mathcal{M}$ and τ.

[27] Observe that p_{F_j} belongs to the extended system $\mathcal{M}^{\downarrow}$ mentioned in Section 2.2.3 and that if $F_j = \emptyset$ the $p_{F_j} > 0$ on $\boldsymbol{X}_{F_j} = \boldsymbol{X}_{\emptyset}$ owing to our convention from Section 2.1.

[28] This happens whenever $p_S > 0$ on $\boldsymbol{X}_S$ for every $S \in \mathcal{S}$, by vanishing principle.

[29] This is because then $G_j = \emptyset$ and $p_{G_j|F_j}(x_{G_j}|x_{F_j}) = p_{\emptyset|S_j}(x_{\emptyset}|x_{S_j}) = 1$ for any $x \in \boldsymbol{X}_N$.

[30] This unrestrictive assumption – see Remark 2 – is needed to ensure $Q \ll L$ for every $Q \in \mathcal{D}_{\mathcal{M}}$. Alternatively, we can modify Definition 2 and restrict our choices to conditional densities $p_{G_j|F_j}$ on $\boldsymbol{X}'_{G_j}$ given $\boldsymbol{X}'_{F_j}$.

Proof. As $l(x) > 0$ for every $x \in \boldsymbol{X}_N$, the claim $Q \ll L$ is evident. We can express the Radon–Nikodym derivative $\mathrm{d}Q/\mathrm{d}L$ as the ratio of respective densities $\overline{p}_{\tau,\vartheta}$ and l. To verify (6) let us choose $P \in \mathcal{K}_{\mathcal{M}}$ such that $N_P = N_{\mathcal{M}}$. Thus, given $x \in N_{\mathcal{M}}$ one has $p(x) > 0$ and this implies by the vanishing principle $p_{F_j}(x_{F_j}) > 0$ for every $j = 1, \ldots, n$. Another point is that the density l of the dominating product measure L can formally be written as follows:

$$l(x) \;=\; \prod_{i \in N} l_i(x_i) = \prod_{j=1}^{n} l_{G_j}(x_{G_j}) = \prod_{j=1}^{n} \frac{l_{S_j}(x_{S_j})}{l_{F_j}(x_{F_j})} \quad \text{for } x \in \boldsymbol{X}_N\,.$$

Therefore, we can write for $x \in N_{\mathcal{M}}$ by (10) and the above formula:

$$\frac{\mathrm{d}Q}{\mathrm{d}L}(x) = \frac{\overline{p}_{\tau,\vartheta}(x)}{l(x)} \;=\; \prod_{j=1}^{n} \frac{p_{S_j}(x_{S_j}) \,\cdot\, l_{F_j}(x_{F_j})}{l_{S_j}(x_{S_j}) \,\cdot\, p_{F_j}(x_{F_j})} = \prod_{j=1}^{n} \frac{f_{S_j}(x_{S_j})}{f_{F_j}(x_{F_j})} \;=\; \prod_{B \in \mathcal{S}^{\downarrow}} f_B(x_B)^{\nu(B)}\,,$$

where $\nu(B)$ is given by (13). Thus, (6) holds with $k = 1$. By substituting $\nu(B)$, $B \in \mathcal{S}^{\downarrow}$ to (7) and realizing that $I(P_{F_1}) = I(P_{\emptyset}) = 0$ we get (12). □

Note that the multiinformation content $I_{\mathcal{M}}(Q)$ of a DSS Q may differ from its multiinformation $I(Q)$ – see Example 5 in Section A1. Lemmas 4 and 3 allow one to derive the following corollary, already given in [8].

Corollary 1. Provided $l(x) > 0$ for every $x \in \boldsymbol{X}_N$, $Q \in \mathcal{D}_{\mathcal{M}}$ corresponding to $\tau : S_1, \ldots, S_n$, $n \geq 1$ and $P \in \mathcal{K}_{\mathcal{M}}$ one has

$$H(P|Q) = I(P) - I_{\mathcal{M}}(Q) = I(P) - \sum_{A \in \mathcal{S}} I(P_A) + \sum_{j=2}^{n} I(P_{F_j})\,.$$

This corollary substantially simplifies the task of finding an optimal DSS.

Definition 3. Let $\mathcal{M} = \{P_A;\, A \in \mathcal{S}\}$ be a strongly consistent collection of probability measures. A DSS $Q \in \mathcal{D}_{\mathcal{M}}$ will be called *optimal* relative to $P \in \mathcal{K}_{\mathcal{M}}$ if

$$H(P|Q) \;=\; \min\{\, H(P|Q');\; Q' \in \mathcal{D}_{\mathcal{M}}\}\,.$$

It follows from the formula in Corollary 1 that $Q = \overline{P}_{\tau,\vartheta} \in \mathcal{D}_{\mathcal{M}}$ is optimal iff it maximizes the multiinformation content $I_{\mathcal{M}}(Q)$ given by (12). Of course, this occurs it τ minimizes the value of the function $\tau \mapsto \iota(\tau) \equiv \sum_{j=2}^{n} I(P_{F_j})$. In particular, the fact that $Q \in \mathcal{D}_{\mathcal{M}}$ is optimal relative to a particular $P \in \mathcal{K}_{\mathcal{M}}$ actually does not depend on P! Note that the problem of finding an ordering yielding an optimal DSS was dealt with in more detail in [8]. The following example illustrates the procedure. In this example, an optimal DSS is uniquely determined.[31]

[31] On the other hand, all three different possible DSSs in Example 5 from Section A1 are optimal.

Example 2. Put $N = \{a, b, c\}$, $\boldsymbol{X}_i = \{0, 1\}$ for any $i \in N$, $\mathcal{S} = \{A \subseteq N\,;\ |A| = 2\}$. These are the densities of probability measures from $\mathcal{M} = \{P_A;\ A \in \mathcal{S}\}$:

$$p_{\{a,b\}}(0,0) = p_{\{a,b\}}(0,1) = \frac{1}{4}\,, \quad p_{\{a,b\}}(1,0) = \frac{1}{8} \quad p_{\{a,b\}}(1,1) = \frac{3}{8}\,,$$

$p_{\{a,c\}}(x) = 1/4$ for every $x \in \boldsymbol{X}_{\{a,c\}}$, and

$$p_{\{b,c\}}(0,0) = p_{\{b,c\}}(1,0) = \frac{1}{4}\,, \quad p_{\{b,c\}}(0,1) = \frac{1}{8} \quad p_{\{b,c\}}(1,1) = \frac{3}{8}\,.$$

To show that $\mathcal{M}$ is strongly consistent consider a density p on $\boldsymbol{X}_{\{a,b,c\}}$ given as follows: $p(0,0,0) = p(0,1,1) = p(1,1,0) = 1/4$ and $p(1,0,1) = p(1,1,1) = 1/8$.

For example, the ordering $\tau_1 : S_1 = \{a, b\}, S_2 = \{a, c\}, S_3 = \{b, c\}$ gives $F_2 = \{a\}$ and $F_3 = \{b, c\}$ and this leads to the value $\iota(\tau_1) = I(P_a) + I(P_{bc}) = I(P_{bc})$. Clearly, the value of $\iota(\tau)$ is the multiinformation of the last marginal in the ordering τ. As $I(P_{ac}) = 0$ and $I(P_{ab}) = I(P_{bc}) = \frac{3}{2} \cdot \ln 2 - \frac{5}{8} \cdot \ln 5 > 0$ there are two "optimal" orderings, namely $\{a, b\}$, $\{b, c\}$, $\{a, c\}$ and $\{b, c\}$, $\{a, b\}$, $\{a, c\}$. They both lead to the same DSS, given by this density $q \equiv p_{\{a,b\}} \cdot p_{\{b,c\}}/p_{\{b\}}$:

$$\begin{aligned} &q(0,0,0) = \frac{1}{6}, \quad q(0,0,1) = \frac{1}{12}, \quad q(0,1,0) = \frac{1}{10}, \quad q(0,1,1) = \frac{3}{20}, \\ &q(1,0,0) = \frac{1}{12}, \quad q(1,0,1) = \frac{1}{24}, \quad q(1,1,0) = \frac{3}{20}, \quad q(1,1,1) = \frac{9}{40}. \end{aligned}$$

Remark 4. An alternative formal definition of a DSS, mentioned implicitly in the manuscript [7], is as follows. The convention $(0/0) \equiv 0$ is accepted. Then (11) defines "density" of a non-negative measure on $\boldsymbol{X}_N$. However, in general, $0 < d \equiv \sum_{x \in \boldsymbol{X}_N} \overline{p}_\tau(x) \leq 1$.[32] One can introduce a density q by the formula $q(x) = d^{-1} \cdot \overline{p}_\tau(x)$ for $x \in \boldsymbol{X}_N$. The point is that this alternative definition of a DSS[33] leads to a different formula for the multiinformation content, namely $\ln d^{-1} + \sum_{A \in \mathcal{S}} I(P_A) - \sum_{j=2}^{n} I(P_{F_j})$; see (7). Paradoxically, this can give better approximation of $P \in \mathcal{K}_\mathcal{M}$ than the DSS introduced in Definition 2 – because the multiinformation content is enlarged by the factor $\ln d^{-1}$. Nevertheless, this only can happen in "non-standard" situations. For example, as mentioned in Remark 3, if $p_S > 0$ for any $S \in \mathcal{S}$ then all terms in (11) are defined and there is no difference between those two formal definitions of a DSS.

5. EXPLICIT EXPRESSION APPROXIMATION

This is a method for approximating measures from $\mathcal{K}_\mathcal{M}$ proposed newly in [7]. The motivation for this proposal was to utilize maximally the information given by $\mathcal{M}$ and, moreover, impose the minimal possible amount of dependencies between variables. The idea was elicited by the first author when he tried to solve the approximation problem described in Section 1 by the method of Lagrange multipliers.

[32] The fact $d > 0$ can be derived from strong consistency of $\mathcal{M}$. Indeed, consider the density p of $P \in \mathcal{K}_\mathcal{M}$ and $x \in \boldsymbol{X}_N$ with $p(x) > 0$. Then, by (2), all nominators and denominators in (11) are positive and $\overline{p}_\tau(x) > 0$.

[33] It is also an $\mathcal{M}$-construct – one can modify the arguments from the proof of Lemma 4.

Definition 4. Given $n \in \mathbb{Z}^+$, the symbol $\mathit{odd}\,(n)$ will be used as a shorthand for $(-1)^{n+1}$; it is a kind of "oddness" indicator: $\mathit{odd}\,(n) = +1$ for odd n and $\mathit{odd}\,(n) = -1$ for even n. Let $\mathcal{M} = \{P_A;\ A \in \mathcal{S}\}$ be a strongly consistent collection of probability measures. Let us put

$$\mathsf{Exe}\,(x) \;=\; \prod_{\emptyset \neq \mathcal{A} \subseteq \mathcal{S}} p_{\bigcap \mathcal{A}}(x_{\bigcap \mathcal{A}})^{\;\mathit{odd}\,(|\mathcal{A}|)} \qquad \text{for every } x \in \boldsymbol{X}_N,^{[34]} \tag{14}$$

where we accept the convention that $0^{-1} \equiv 0$. Then we put $c = \sum_{x \in \boldsymbol{X}_N} \mathsf{Exe}\,(x)$,[35] and define

$$\overline{\mathsf{Exe}}(x) \;=\; c^{-1} \cdot \mathsf{Exe}\,(x) \equiv c^{-1} \cdot \prod_{\emptyset \neq \mathcal{A} \subseteq \mathcal{S}} p_{\bigcap \mathcal{A}}(x_{\bigcap \mathcal{A}})^{\;\mathit{odd}\,(|\mathcal{A}|)} \quad \text{for every } x \in \boldsymbol{X}_N. \tag{15}$$

Of course, $\overline{\mathsf{Exe}}$ is a density of a probability measure on $\boldsymbol{X}_N$, which will be denoted below by P_{exe}. The number c will be called the *norm* (of the explicit expression Exe) and denoted by $|\mathsf{Exe}\,|$.

Note that some factors in the formula (14) can cancel out. We decided to introduce Exe by formally redundant but elegant formula to make subsequent proofs easy to follow. The norm $|\mathsf{Exe}\,|$ could be both higher and lower than 1 – Example 7 in Section A2 shows that it may happen $|\mathsf{Exe}\,| > 1$ while Example 8 shows that it may happen $|\mathsf{Exe}\,| < 1$. Nevertheless, even if $|\mathsf{Exe}\,| = 1$ then the respective explicit expression approximation P_{exe} need not belong to $\mathcal{K}_{\mathcal{M}}$ as the following example shows.

Example 3. Consider the system of marginals $\mathcal{M}$ from Example 2. Then $p_{\{a\}}(0) = p_{\{a\}}(1) = 1/2 = p_{\{c\}}(0) = p_{\{c\}}(1)$ and $p_{\{b\}}(0) = 3/8$, $p_{\{b\}}(1) = 5/8$; this allows one to write by (14):

$$\begin{aligned}\mathsf{Exe}\,(0,0,0) \;&=\; \frac{p_{\{a,b\}}(0,0)\cdot p_{\{a,c\}}(0,0)\cdot p_{\{b,c\}}(0,0)\,\cdot \qquad\qquad\qquad\qquad \cdot p_{\emptyset}(-)}{\qquad\qquad\qquad p_{\{a\}}(0)\cdot p_{\{b\}}(0)\cdot p_{\{c\}}(0)} \\ &=\; \frac{\frac{1}{4}\cdot\frac{1}{4}\cdot\frac{1}{4}\,\cdot \qquad \cdot 1}{\qquad \frac{1}{2}\cdot\frac{3}{8}\cdot\frac{1}{2}} \;=\; \frac{2\cdot 8\cdot 2}{4\cdot 4\cdot 4\cdot 3} \;=\; \frac{1}{6}\,.\end{aligned}$$

Actually, the result of detailed calculation of Exe is the density q of the optimal DSS mentioned in Example 2. In particular, $|\mathsf{Exe}\,| = 1$ and P_{exe} has density q. However, $q^{\{a,c\}}(0,0) = (1/6) + (1/10) = 8/30 \neq 1/4 = p_{\{a,c\}}(0,0)$, which means $P_{exe} \notin \mathcal{K}_{\mathcal{M}}$. On the other hand, the example also shows that P_{exe} can coincide with an optimal DSS.

[34] Observe that Exe defines a "density" of a non-negative non-zero measure EXE on $\boldsymbol{X}_N$ such that $P \ll \mathsf{EXE}$ for every $P \in \mathcal{K}_{\mathcal{M}}$. Indeed, (2) implies that whenever $p(x) > 0$ for $x \in \boldsymbol{X}_N$ then $p_{\bigcap \mathcal{A}}(x_{\bigcap \mathcal{A}}) > 0$ for every $\emptyset \neq \mathcal{A} \subseteq \mathcal{S}$.

[35] The assumption of strong consistency of $\mathcal{M}$ implies that $c > 0$ – use what it says in the preceding footnote.

Lemma 5. Let $\mathcal{M} = \{P_A;\ A \in \mathcal{S}\}$ be a strongly consistent collection of probability measures. Then the probability measure P_{exe} is an $\mathcal{M}$-construct. Its multiinformation content is

$$I_{\mathcal{M}}(P_{exe}) \ =\ -\ln|\mathsf{Exe}\,| + \sum_{B\in\mathcal{S}^{\downarrow}} \nu(B)\cdot I(P_B)\,, \tag{16}$$

where

$$\nu(B) \ =\ \sum\{\,odd\,(|\mathcal{A}|)\,;\ \emptyset\neq\mathcal{A}\subseteq\mathcal{S},\ \bigcap\mathcal{A}=B\} \quad \text{for any } B\in\mathcal{S}^{\downarrow}\,. \tag{17}$$

Proof. The first observation is that

$$\forall\, i\in N \quad \sum\{\,odd\,(|\mathcal{A}|)\,;\ \emptyset\neq\mathcal{A}\subseteq\mathcal{S},\ i\in\bigcap\mathcal{A}\} = +1\,. \tag{18}$$

Indeed, consider a fixed $i\in N$, denote by $\mathcal{H}$ the class of $A\in\mathcal{S}$ with $i\in A$ and write using the definition of $odd\,(n)$ and binominal formula:

$$\begin{aligned}\sum_{\emptyset\neq\mathcal{A}\subseteq\mathcal{H}} odd\,(|\mathcal{A}|) \ &=\ \sum_{\emptyset\neq\mathcal{A}\subseteq\mathcal{H}} (-1)^{|\mathcal{A}|+1} = \sum_{\ell=1}^{|\mathcal{H}|}\ \sum_{\mathcal{A}\subseteq\mathcal{H},|\mathcal{A}|=\ell} (-1)^{\ell+1}\\ &=\ \sum_{\ell=1}^{|\mathcal{H}|}(-1)^{\ell+1}\cdot|\{\mathcal{A}\subseteq\mathcal{H};\ |\mathcal{A}|=\ell\}| = \sum_{\ell=1}^{|\mathcal{H}|}(-1)^{\ell+1}\cdot\binom{|\mathcal{H}|}{\ell}\\ &=\ +1-\sum_{\ell=0}^{|\mathcal{H}|}(-1)^{\ell}\cdot 1^{|\mathcal{H}|-\ell}\cdot\binom{|\mathcal{H}|}{\ell} \ =\ +1-(-1+1)^{|\mathcal{H}|} \ =\ +1\,.\end{aligned}$$

The main step is to introduce a non-negative measure Q on $\boldsymbol{X}_N$ such that $Q \ll L$ and its Radon–Nikodym derivative $\mathrm{d}Q/\mathrm{d}L$ has the following form:

$$\frac{\mathrm{d}Q}{\mathrm{d}L}(x) \ =\ \prod_{\emptyset\neq\mathcal{A}\subseteq\mathcal{S}} f_{\bigcap\mathcal{A}}(x_{\bigcap\mathcal{A}})^{\ odd\,(|\mathcal{A}|)} \quad \text{for any } x\in\boldsymbol{X}_N\,. \tag{19}$$

To show that P_{exe} is an $\mathcal{M}$-construct it suffices to show that the "density" q of Q coincides with Exe.[36] This is easy to see for $x\in\boldsymbol{X}_N$ with $l(x)=0$. Then $p_{\{i\}}(x_i)=0$ for some $i\in N$ and the assumption $\bigcup\mathcal{S}=N$ forces the existence of $A\in\mathcal{S}$ with $i\in A$. Therefore, the vanishing principle (2) implies that at least one factor in (14) vanishes and $\mathsf{Exe}\,(x)=0$.

To verify $q(x)=\mathsf{Exe}\,(x)$ for $x\in\boldsymbol{X}_N$ with $l(x)>0$ we first observe that

$$\prod_{\emptyset\neq\mathcal{A}\subseteq\mathcal{S}}\ \prod_{j\in\bigcap\mathcal{A}} p_{\{j\}}(x_j)^{-odd\,(|\mathcal{A}|)} \ =\ \prod_{i\in N} p_{\{i\}}(x_i)^{-1}\,. \tag{20}$$

Indeed, one can write it with the help of (18) as follows:

$$\begin{aligned}\prod_{\emptyset\neq\mathcal{A}\subseteq\mathcal{S}}\ \prod_{j\in\bigcap\mathcal{A}} p_{\{j\}}(x_j)^{-odd\,(|\mathcal{A}|)} \ &=\ \prod_{i\in N}\ \prod_{\emptyset\neq\mathcal{A}\subseteq\mathcal{S},\ i\in\bigcap\mathcal{A}} p_{\{i\}}(x_i)^{-odd\,(|\mathcal{A}|)}\\ &=\ \prod_{i\in N} p_{\{i\}}(x_i)^{-\sum\{\,odd\,(|\mathcal{A}|)\,;\ \emptyset\neq\mathcal{A}\subseteq\mathcal{S},\, i\in\bigcap\mathcal{A}\}} \ =\ \prod_{i\in N} p_{\{i\}}(x_i)^{-1}\,.\end{aligned}$$

[36] Recall that, since Q need not be a probability measure, one can have $\sum\{q(x), x\in\boldsymbol{X}_N\}\neq 1$.

The formulas (19), $f_B = p_B \cdot \prod_{j\in B} p_{\{j\}}^{-1}$ for $B \subseteq N$ (see Section 2.2.3) and (20) now allow one to write $q(x)$ as follows:

$$\begin{aligned}
q(x) &= \frac{\mathrm{d}Q}{\mathrm{d}L}(x)\cdot l(x) \;=\; \prod_{\emptyset\neq\mathcal{A}\subseteq\mathcal{S}} f_{\bigcap\mathcal{A}}(x_{\bigcap\mathcal{A}})^{\;odd\,(|\mathcal{A}|)} \cdot \prod_{i\in N} p_{\{i\}}(x_i)\\
&= \prod_{\emptyset\neq\mathcal{A}\subseteq\mathcal{S}} \{\, p_{\bigcap\mathcal{A}}(x_{\bigcap\mathcal{A}})^{\;odd\,(|\mathcal{A}|)} \cdot \prod_{j\in\bigcap\mathcal{A}} p_{\{j\}}(x_j)^{-odd\,(|\mathcal{A}|)} \,\} \cdot \prod_{i\in N} p_{\{i\}}(x_i)\\
&= \prod_{\emptyset\neq\mathcal{A}\subseteq\mathcal{S}} p_{\bigcap\mathcal{A}}(x_{\bigcap\mathcal{A}})^{\;odd\,(|\mathcal{A}|)} \cdot \prod_{\emptyset\neq\mathcal{A}\subseteq\mathcal{S}}\;\prod_{j\in\bigcap\mathcal{A}} p_{\{j\}}(x_j)^{-odd\,(|\mathcal{A}|)} \cdot \prod_{i\in N} p_{\{i\}}(x_i)\\
&= \prod_{\emptyset\neq\mathcal{A}\subseteq\mathcal{S}} p_{\bigcap\mathcal{A}}(x_{\bigcap\mathcal{A}})^{\;odd\,(|\mathcal{A}|)} \cdot \prod_{i\in N} p_{\{i\}}(x_i)^{-1} \cdot \prod_{i\in N} p_{\{i\}}(x_i)\\
&= \prod_{\emptyset\neq\mathcal{A}\subseteq\mathcal{S}} p_{\bigcap\mathcal{A}}(x_{\bigcap\mathcal{A}})^{\;odd\,(|\mathcal{A}|)} \cdot 1 \;=\; \mathsf{Exe}\,(x)\,.
\end{aligned}$$

The observation $q = \mathsf{Exe}$ means that P_{exe} is c^{-1}-multiple of Q where $c = |\mathsf{Exe}\,|$. In particular, by (19), $P_{exe} \ll L$ and (17) write:

$$\begin{aligned}
c\cdot\frac{\mathrm{d}P_{exe}}{\mathrm{d}L}(x) &= \prod_{\emptyset\neq\mathcal{A}\subseteq\mathcal{S}} f_{\bigcap\mathcal{A}}(x_{\bigcap\mathcal{A}})^{\;odd\,(|\mathcal{A}|)}\\
&= \prod_{B\in\mathcal{S}^{\downarrow}} f_B(x_B)^{\sum\{\,odd\,(|\mathcal{A}|)\,;\;\emptyset\neq\mathcal{A}\subseteq\mathcal{S},\,\bigcap\mathcal{A}=B\}} \;=\; \prod_{B\in\mathcal{S}^{\downarrow}} f_B(x_B)^{\,\nu(B)}\,.
\end{aligned}$$

Thus, by Definition 1, P_{exe} is an $\mathcal{M}$-construct with $k = c^{-1}$ and $\nu(B)$, $B \subseteq N$ given by (17). The formula (16) follows from (7). □

Corollary 2. Given $P \in \mathcal{K}_{\mathcal{M}}$ one has

$$H(P|P_{exe}) \;=\; I(P) - I_{\mathcal{M}}(P_{exe}) \;=\; I(P) + \ln|\mathsf{Exe}\,| - \sum_{B\in\mathcal{S}^{\downarrow}} \nu(B)\cdot I(P_B)\,,$$

where $\nu(B)$, $B \in \mathcal{S}^{\downarrow}$ is given by (17). In particular, $\min_{P\in\mathcal{K}_{\mathcal{M}}} I(P) \geq I_{\mathcal{M}}(P_{exe})$ and the equality occurs iff $P_{exe} \in \mathcal{K}_{\mathcal{M}}$, in which case $I_{\mathcal{M}}(P_{exe}) = I(P_{exe})$.

Proof. This follows from Lemma 3: put $Q = P_{exe}$ and use the formula (16). □

Remark 5. An useful observation concerning explicit expression approximation was made in [7]. If we consider the multi-symptom diagnostic problem mentioned in Section 3.1 and base our estimator on direct approximation of P by means of the explicit expression $\hat{P} = P_{exe}$, then it is *not necessary* to compute the norm $|\mathsf{Exe}\,|$. This is because $\overline{\mathsf{Exe}}$ and Exe only differ in a multiplicative positive factor and always achieve their maxima in same configurations. Thus, in this particular case, one has

$$\psi_1(x_S) \;=\; \operatorname{argmax}\{\mathsf{Exe}\,([y, x_S])\,;\; y \in \boldsymbol{X}_d\,\}\,.$$

5.1. Comparison of DSSs and explicit expression approximations

In general, it is not possible to claim that one of the above-mentioned methods for approximation of a distribution P with prescribed marginals is better than the other, if one takes the relative entropy $H(P|\hat{P})$ as the measure of divergence of an approximation $\hat{P}$ from P. The respective Examples 7 and 8 are given in the Appendix, Section A2.

6. THE CASE OF FITTING MARGINALS

It may happen that an approximation $\hat{P}$ of measures from $\mathcal{K}_{\mathcal{M}}$ fits the prescribed marginals, that is, $\hat{P}$ really has the measures from $\mathcal{M}$ as marginals and, therefore, it belongs to $\mathcal{K}_{\mathcal{M}}$. The following example shows that both methods for approximation mentioned in this paper may result in a distribution from $\mathcal{K}_{\mathcal{M}}$.

Example 4. Let us put $N = \{a,b,c\}$, $\boldsymbol{X}_a = \boldsymbol{X}_c = \{0,1\}$, $\boldsymbol{X}_b = \{0,1,2\}$ and $\mathcal{S} = \{A \subseteq N;\ |A| = 2\}$. The densities of measures from $\mathcal{M} = \{P_A\,;\ A \in \mathcal{S}\}$ are given as follows:

$$p_{\{a,b\}}(0,0) = \frac{2}{9}\,, \quad p_{\{a,b\}}(0,1) = \frac{1}{9}, \quad p_{\{a,b\}}(1,1) = p_{\{a,b\}}(1,2) = \frac{1}{3}\,,$$

$$p_{\{a,c\}}(0,0) = p_{\{a,c\}}(1,1) = \frac{2}{9}\,, \quad p_{\{a,c\}}(0,1) = \frac{1}{9}, \quad p_{\{a,c\}}(1,0) = \frac{4}{9}\,,$$

and, finally

$$p_{\{b,c\}}(0,0) = p_{\{b,c\}}(0,1) = p_{\{b,c\}}(2,0) = \frac{1}{9}\,, \quad p_{\{b,c\}}(1,0) = \frac{4}{9}, \quad p_{\{b,c\}}(2,1) = \frac{2}{9}\,.$$

Detailed calculation of Exe gives this

$$\mathsf{Exe}\,(0,0,0) = \mathsf{Exe}\,(0,0,1) = \mathsf{Exe}\,(0,1,0) = \mathsf{Exe}\,(1,2,0) = \frac{1}{9},$$

$$\mathsf{Exe}\,(1,1,0) = \frac{1}{3}, \ \ \mathsf{Exe}\,(1,2,1) = \frac{2}{9}, \ \ \text{and } \mathsf{Exe}\,(x) = 0 \text{ for remaining } x \in \boldsymbol{X}_N.$$

In particular, $|\mathsf{Exe}\,| = 1$ and the density p of P_{exe} coincides with Exe. It is easy to see that $p^A = p_A$ for $A \in \mathcal{S}$. Moreover, the calculation of DSS for $\tau : S_1 = \{a,b\}$, $S_2 = \{b,c\}$, $S_3 = \{a,c\}$ gives the same result.

Note that if a DSS has the prescribed marginals then it is optimal.

Corollary 3. Assume $l(x) > 0$ for every $x \in \boldsymbol{X}_N$. If $Q^* \in \mathcal{D}_{\mathcal{M}} \cap \mathcal{K}_{\mathcal{M}}$ then Q^* is an optimal DSS (relative to any $P \in \mathcal{K}_{\mathcal{M}}$).

Proof. By Lemma 4, Q^* is an $\mathcal{M}$-construct and Lemma 3 says that $Q^* \in \mathcal{K}_{\mathcal{M}}$ implies $\min\{I(P);\ P \in \mathcal{K}_{\mathcal{M}}\} = I_{\mathcal{M}}(Q^*)$. Given arbitrary $Q \in \mathcal{D}_{\mathcal{M}}$, again by Lemmas 4 and 3, observe that

$$I_{\mathcal{M}}(Q^*) \ = \ \min\{I(P);\ P \in \mathcal{K}_{\mathcal{M}}\} \ \geq \ I_{\mathcal{M}}(Q)\,.$$

Therefore, $I_{\mathcal{M}}(Q^*) = \max\{I_{\mathcal{M}}(Q);\ Q \in \mathcal{D}_{\mathcal{M}}\}$. However, this means Q^* is optimal – see the explanation after Definition 3. □

The approximations should be reasonable in the sense that if an estimate $\hat{P}$ incidentally has the prescribed marginals from $\mathcal{M}$ then it is a distinguished representative of the class $\mathcal{K}_{\mathcal{M}}$. There are more principles for the choice of a representative of a class of distributions suitable from the point of view of probabilistic decision-making. One of them is the *maximum entropy principle.*[37] The idea is to choose $P \in \mathcal{K}_{\mathcal{M}}$ which maximizes the entropy $H(P)$ in $\mathcal{K}_{\mathcal{M}}$. By Lemma 2, this distribution is uniquely determined. The results from Sections 4 and 5 imply that both approximation methods dealt with in this paper are in concordance with this principle.

Corollary 4. Let $\mathcal{M} = \{P_A;\ A \in \mathcal{S}\}$ be a strongly consistent collection of probability measures. If $P_{exe} \in \mathcal{K}_{\mathcal{M}}$ then $\hat{P} = P_{exe}$ is the measure maximizing entropy in $\mathcal{K}_{\mathcal{M}}$. Assuming $l(x) > 0$ for all $x \in \boldsymbol{X}_N$ and $Q \in \mathcal{D}_{\mathcal{M}} \cap \mathcal{K}_{\mathcal{M}}$ the distribution $\hat{P} = Q$ maximizes entropy in $\mathcal{K}_{\mathcal{M}}$.

Proof. Lemmas 5 and 4 imply that the considered approximation $\hat{P}$ is an $\mathcal{M}$-construct. Then, Lemma 3 says that $\hat{P} \in \mathcal{K}_{\mathcal{M}}$ implies the equality in (8); that is, $\min\{I(P);\ P \in \mathcal{K}_{\mathcal{M}}\} = I_{\mathcal{M}}(\hat{P})$ and, moreover, $I_{\mathcal{M}}(\hat{P}) = I(\hat{P})$. Thus, $\hat{P}$ minimizes the multiinformation in $\mathcal{K}_{\mathcal{M}}$ and, by Lemma 2, it maximizes the entropy. □

7. SIMPLE SUFFICIENT CONDITION FOR STRONG CONSISTENCY

Of course, as mentioned in Section 6, the ideal case is when the approximation has prescribed marginals from $\mathcal{M}$. The problem is often to ensure this situation. There exists simple strong sufficient condition for this in terms of the class $\mathcal{S}$. The condition has close connection to graphical models [3], more precisely, to so-called *decomposable graphical models.* Even more special and simpler case is the case of so-called *asteroid*, which is the concept introduced in the manuscript [7] by the first author.

Definition 5. Let $\mathcal{S}$ be a class of subsets of N such that $\bigcup \mathcal{S} = N$. We say that it is *decomposable* if there exists an ordering $\tau : S_1, \ldots, S_n$, $n \geq 1$ of sets in $\mathcal{S}$ that satisfies the *running intersection property*:

$$\forall\, j > 1 \quad \exists\, \ell < j \quad F_j \equiv S_j \cap (S_1 \cup \ldots S_{j-1}) \subseteq S_\ell\,. \tag{21}$$

Given a partitioning $\{E_0, \ldots, E_r\}$, $r \geq 2$ of the set N, an *asteroid* with core $C = E_0$ (generated by that partitioning) is the class of sets

$$\mathcal{S} \;=\; \{E_0 \cup E_i;\ i = 1, \ldots, r\}\,.$$

[37] An alternative *barycenter principle* is mentioned in Section 8.

It is evident that every asteroid is a decomposable class; actually, any ordering of sets of an asteroid satisfies the running intersection property.[38] The point is that the decomposability condition is a necessary and sufficient condition for the equivalence of weak and strong consistency of any system $\mathcal{M}$ of probability measures which has $\mathcal{S}$ as the class of "indexing" sets – see [8] and [2]. However, in the context of this paper, the following observations are crucial.

Proposition 6. Let $\mathcal{S}$ be a decomposable class of subsets of N with $\bigcup \mathcal{S} = N$ and $\mathcal{M} = \{P_A;\ A \in \mathcal{S}\}$ be a (strongly) consistent collection of probability measures. Then any total ordering $\tau : S_1, \ldots, S_n,\ n \geq 1$ of sets in $\mathcal{S}$ satisfying the running intersection property (21) yields an optimal DSS. The respective optimal DSS coincides with P_{exe} and has fitting prescribed marginals from $\mathcal{M}$. Thus, it coincides with the distribution chosen from $\mathcal{K}_{\mathcal{M}}$ by the maximum entropy principle.

Proof. To show the first claim it suffices to verify that the respective DSS has prescribed marginals from $\mathcal{M}$ and apply Corollary 3. The statement that if τ satisfies (21) then the density $\overline{p}_{\tau,\vartheta}$ given by (10) has $p_{S_1}, \ldots, p_{S_n}$ as marginal densities can be proved by induction on n.[39] It is evident for $n = 1$. If $n > 1$ then we denote $R = S_1 \cup \ldots \cup S_{n-1}$, consider a shortened ordering $\tau' : S_1, \ldots, S_{n-1}$, a restricted choice ϑ' and derive from (10):

$$\overline{p}_{\tau,\vartheta}(x) = \prod_{j=1}^{n} p_{G_j|F_j}(x_{G_j}|x_{F_j}) = \overline{p}_{\tau',\vartheta'}(x_R) \cdot p_{G_n|F_n}(x_{G_n}|x_{F_n}) \quad \text{for } x \in \boldsymbol{X}_N\,. \tag{22}$$

Hence, $(\overline{p}_{\tau,\vartheta})^R = \overline{p}_{\tau',\vartheta'}$,[40] which allows one to observe by the induction assumption that $\overline{p}_{\tau,\vartheta}$ has $p_{S_1}, \ldots, p_{S_{n-1}}$ as marginal densities:

$$\forall\, j < n \quad (\overline{p}_{\tau,\vartheta})^{S_j} = ((\overline{p}_{\tau,\vartheta})^R)^{S_j} = (\overline{p}_{\tau',\vartheta'})^{S_j} = p_{S_j}\,.$$

To show that it has p_{S_n} as marginal density find $\ell < n$ with $F_n \subseteq S_\ell$. Now, the induction assumption says $(\overline{p}_{\tau',\vartheta'})^{S_\ell} = p_{S_\ell}$ which allows one to observe that $\overline{p}_{\tau',\vartheta'}$ has p_{F_n} as marginal density:

$$(\overline{p}_{\tau',\vartheta'})^{F_n} = ((\overline{p}_{\tau',\vartheta'})^{S_\ell})^{F_n} = (p_{S_\ell})^{F_n} = p_{F_n}\,.$$

Therefore, by (22), the marginal density of $\overline{p}_{\tau,\vartheta}$ for S_n can be written as follows:

$$(\overline{p}_{\tau,\vartheta})^{S_n} = (\overline{p}_{\tau',\vartheta'})^{F_n} \cdot p_{G_n|F_n} = p_{F_n} \cdot p_{G_n|F_n} = p_{S_n}\,,$$

because the conditional density $p_{G_n|F_n}$ is consonant with p_{S_n}. This completes the induction step.

To show that the respective optimal DSS coincides with P_{exe} we first observe that if $\tau : S_1, \ldots, S_n,\ n \geq 1$ satisfies (21) then the concept of choice for $\mathcal{M}$ and τ is not

[38] This is because then the core $C \equiv E_0$ coincides with the set $S_j \cap (S_1 \cup \ldots S_{j-1})$ for any $j > 1$ no matter what ordering $S_1, \ldots, S_r$ of $\mathcal{S} = \{E_0 \cup E_i;\ i = 1, \ldots, r\}$ is chosen.

[39] This holds irrespective of what choice ϑ for $\mathcal{M}$ and τ is considered.

[40] Use the definition of conditional density.

needed because the density $\overline{p}_{\tau,\vartheta}$ given by (10) does not depend on ϑ. Actually, the density of the respective DSS is then given by (11) where we accept the convention $0^{-1} \equiv 0$.[41] Thus, (11) implies that the density $\overline{p}_\tau$ has the form:

$$\overline{p}_\tau(x) \;=\; \prod_{B\in\mathcal{S}^\downarrow} p_B(x_B)^{\nu(B)} \qquad \text{for } x \in \boldsymbol{X}_N\,,$$

where $\nu(B)$, $B \in \mathcal{S}^\downarrow$ is given by (13) and the convention $0^{-1} = 0$ is accepted. Now, the formula (14) implies

$$\mathsf{Exe}\,(x) \;=\; \prod_{B\in\mathcal{S}^\downarrow} p_B(x_B)^{\nu(B)} \qquad \text{for } x \in \boldsymbol{X}_N\,,$$

where $\nu(B)$, $B \in \mathcal{S}^\downarrow$ is given by (17) and the same convention holds. The point is that if τ satisfies the running intersection property (21) then the formulas (13) and (17) give the same result – this is what is proved in Lemma 7.2 in [9].[42] In particular, $\overline{p}_\tau = \mathsf{Exe}$. As $\overline{p}_\tau$ is a density of a probability measure $|\mathsf{Exe}\,| = 1$ and one has $\overline{p}_\tau = \overline{\mathsf{Exe}}$. Thus, the respective DSS Q coincides with P_{exe}. We have already shown that Q has prescribed marginals.

The last statement in Proposition 6 follows from Corollary 4. □

Remark 6. Of course, if one considers the family of all classes of sets $\mathcal{S}$ with $\bigcup\mathcal{S} = N$ then not many of them are decomposable. However, the point is that, in the context of probabilistic decision making, the final goal is the respective decision procedure, that is, the estimator – see Section 3.1. Thus, one has some freedom in the choice of the system $\mathcal{S}$ and can, therefore, intentionally choose a decomposable class.

8. BARYCENTER PRINCIPLE

Another principle for the choice of a representative of a class of probability distributions, different from the maximum entropy principle, is the *barycenter principle*. It was proposed by the first author in the 1980s (see [5, 6]). It is also closely related to information projections as studied in [1]. The following restricted definition is suitable for the purpose of this paper.

Definition 6. Let $\mathcal{K}$ and $\mathcal{T}$ are two classes of probability measures on the same sample space, say, on $\boldsymbol{X}_N$. A *barycenter* of $\mathcal{K}$ (taken) in $\mathcal{T}$ is any probability measure $R_* \in \mathcal{T}$ which minimizes the function

$$R \mapsto \mu(R) \equiv \max_{P\in\mathcal{K}} H(P|R), \quad R \in \mathcal{T}, \tag{23}$$

[41] Given $x \in \boldsymbol{X}_N$ consider the first (possible) $j \geq 2$ with $p_{F_j}(x_{F_j}) = 0$ and, by (21), find $1 \leq \ell < j$ with $F_j \subseteq S_\ell$. As $\mathcal{M}$ is strongly consistent, by (2), $p_{S_\ell}(x_{S_\ell}) = 0$. However, as $p_{F_\ell}(x_{F_\ell}) > 0$ one certainly has $p_{G_\ell|F_\ell}(x_{G_\ell}|x_{F_\ell}) = 0$ and $\overline{p}_{\tau,\vartheta}(x) = 0$, no matter what choice ϑ was considered.

[42] It can be verified by the induction on n.

that is, in other words, it is obtained by the following “mini-max” procedure:

$$\max_{P\in\mathcal{K}} H(P|R_*) = \min_{R\in\mathcal{T}} \max_{P\in\mathcal{K}} H(P|R)\,.$$

An implicit technical requirement is that the clases $\mathcal{K}$ and $\mathcal{T}$ are such that the maxima in (23) exist and the function μ is finite for at least one $R\in\mathcal{T}$.

The interpretation is that $\mathcal{T}$ is the class of approximations of distributions from $\mathcal{K}$. Thus, we typically have in mind the set $\mathcal{K}_{\mathcal{M}}$ in place of $\mathcal{K}$. If we put $\mathcal{T}=\mathcal{D}_{\mathcal{M}}$ the concept of barycenter reduces to the concept of an optimal DSS.

Proposition 7. Let $\mathcal{M}$ be a strongly consistent collection of probability measures. Assume $l(x)>0$ for every $x\in \boldsymbol{X}_N$. Then a probability measure Q on $\boldsymbol{X}_N$ is an optimal DSS (for $\mathcal{M}$) iff it is a barycenter of $\mathcal{K}_{\mathcal{M}}$ in $\mathcal{D}_{\mathcal{M}}$.

Proof. It follows from Lemma 2 that $\max_{P\in\mathcal{K}_{\mathcal{M}}} I(P)<\infty$ and that at least one $P_\dagger$ in $\mathcal{K}_{\mathcal{M}}$ exists with $I(P_\dagger)=\max_{P\in\mathcal{K}_{\mathcal{M}}} I(P)$. Moreover, it follows from Lemmas 4 and 3 that $H(P|Q)=I(P)-I_{\mathcal{M}}(Q)$ for any $P\in\mathcal{K}_{\mathcal{M}}$ and $Q\in\mathcal{D}_{\mathcal{M}}$. In particular, given $Q\in\mathcal{D}_{\mathcal{M}}$, one has

$$\max_{P\in\mathcal{K}_{\mathcal{M}}} H(P|Q) = \max_{P\in\mathcal{K}_{\mathcal{M}}} I(P)-I_{\mathcal{M}}(Q) = I(P_\dagger)-I_{\mathcal{M}}(Q),$$

and the task to minimize $Q\mapsto \max_{P\in\mathcal{K}_{\mathcal{M}}} H(P|Q)$, $Q\in\mathcal{D}_{\mathcal{M}}$ is equivalent to the task to maximize $I_{\mathcal{M}}(Q)$ on $\mathcal{D}_{\mathcal{M}}$. However, as explained after Definition 3, Q is an optimal DSS iff it maximizes the multiinformation content $I_{\mathcal{M}}(Q)$ on $\mathcal{D}_{\mathcal{M}}$. □

The above definition of barycenter is general enough: one can even put $\mathcal{T}\equiv\mathcal{K}$, which means that one is looking for a *barycenter of a class of distributions $\mathcal{K}$ in itself.* Actually, this is an alternative to the maximum entropy principle, proposed already in [6]. It was shown there that in several common situations, the maximum entropy principle and (this special) barycenter principle yield the same result. However, this is not always the case. Example 9 in Section A3 shows that, if we consider the case of $\mathcal{K}=\mathcal{K}_{\mathcal{M}}$, then the barycenter principle and the maximum entropy principle may result in different approximations.

9. CONCLUSIONS AND OPEN PROBLEMS

Let us summarize the results of the paper. We have compared two methods for approximation of probability distributions with prescribed marginals: the optimal DSS approximation and the explicit expression approximation. Both these methods can be applied to multi-symptom diagnosis making as explained in Section 3.1. The conclusion is that none of these two methods is universally better than the other – we gave the respective examples in Section A2. As mentioned in [7], the formal advantage of the explicit expression approximation is that if we use this approach then we automatically avoid the optimization procedure needed in the case of DSS approximations.

Moreover, in the case of fitting marginals, both methods result in the distribution chosen by the maximum entropy principle – see Section 6. A simple sufficient condition for this in terms of $\mathcal{S}$ was recalled in Section 7. Finally, in Section 8, we compared the barycenter principle and the maximum entropy principle and showed that they differ in the considered special case; actually, this disproves one of the conjectures from [7].

Of course, some questions remain open. One of them is as follows. Is it true that if $|\mathsf{Exe}\,| = 1$ then P_{exe} coincides with an optimal DSS approximation? This was also mentioned in [7] as a conjecture. The second author tried to verify or disprove that conjecture but he has not succeeded so far. The conjecture was verified in the case $|\boldsymbol{X}_i| = 2$ for $i \in N$ and $|\mathcal{S}| \leq 3$ – this was done with the essential help of a computer program Mathematica. Another open question is mentioned in the end of Section A2: is it true that if $P_{exe} \in \mathcal{K}_{\mathcal{M}}$ then $\mathcal{K}_{\mathcal{M}} \cap \mathcal{D}_{\mathcal{M}} \neq \emptyset$?[43]

APPENDIX: EXAMPLES

A1. Examples related to dependence structure simplifications

The following example shows that the multiinformation content of a DSS Q need not equal to its multiinformation.

Example 5. Put $N = \{a,b,c,d\}$, $\boldsymbol{X}_i = \{0,1\}$ for every $i \in N$ and consider the class of sets $\mathcal{S} = \{S_1, S_2, S_3\}$, where $S_1 = \{a,b\}$, $S_2 = \{a,c\}$ and $S_3 = \{b,c,d\}$. The collection of probability measures $\mathcal{M} = \{P_A;\ A \in \mathcal{S}\}$ is introduced by means of densities:

$$p_A(0,0) = p_A(1,1) = \frac{1}{5}\,, \quad p_A(0,1) = p_A(1,0) = \frac{3}{10} \qquad \text{for } A = S_1 \text{ and } A = S_2\,,$$

while for $B = S_3 = \{b,c,d\}$

$$p_B(0,0,0) = p_B(1,1,0) = \frac{1}{5}, \quad p_B(0,1,1) = p_B(1,0,1) = \frac{3}{10}\,.$$

To see that $\mathcal{M}$ is strongly consistent consider a density $p : \boldsymbol{X}_N \to [0,1]$, where $p(0,0,0,0) = p(1,1,1,0) = 1/20$ and $p(x) = 3/20$ for any of the following six configurations: $(0,0,1,1)$, $(0,1,0,1)$, $(0,1,1,0)$, $(1,0,0,0)$, $(1,0,1,1)$ and $(1,1,0,1)$.

Take the ordering $\tau : S_1, S_2, S_3$ and observe that $p_{F_j} > 0$ for $j = 2,3$. Therefore, the density $q = \overline{p}_\tau$ of the respective DSS Q is unambiguously defined. It has the same support as the above mentioned joint density p. More specifically, $q(0,0,0,0) = q(1,1,1,0) = 2/25$, $q(0,1,1,0) = q(1,0,0,0) = 9/50$ and $q(x) = 3/25$ for the following four configurations: $(0,0,1,1)$, $(0,1,0,1)$, $(1,0,1,1)$ and $(1,1,0,1)$. Hence, one has for $B = \{b,c,d\}$:

$$q^B(0,0,0) = q^B(1,1,0) = \frac{13}{50}, \quad q^B(0,1,1) = q^B(1,0,1) = \frac{6}{25}\,.$$

[43] Note that if $P_{exe} \in \mathcal{K}_{\mathcal{M}}$ then $\mathcal{K}_{\mathcal{M}} \cap \mathcal{D}_{\mathcal{M}} \neq \emptyset$ is equivalent to $P_{exe} \in \mathcal{D}_{\mathcal{M}}$ – use Corollary 4.

To express the difference $I(Q) - I_{\mathcal{M}}(Q)$ we first write the multiinformation of Q as follows:

$$I(Q) \;=\; I(Q^{ab}) + I(Q^{ac}) + I(Q^{bcd}) - I(Q^{a}) - I(Q^{bc})\,.$$[44]

Now, by (12), $I_{\mathcal{M}}(Q)$ has the same form, but Q^{A} is replaced by P_{A} for respective sets $A \subseteq N$ there. As $Q^{ab} = P_{ab}$ and $Q^{ac} = P_{ac}$ one has

$$I(Q) - I_{\mathcal{M}}(Q) \;=\; [I(Q^{bcd}) - I(Q^{bc})] - [I(P_{bcd}) - I(P_{bc})]\,,$$

and the reader can obtain by direct computation[45] $I(Q^{bcd}) - I(Q^{bc}) = \frac{13}{25} \cdot \ln \frac{25}{13} + \frac{12}{25} \cdot \ln \frac{25}{12}$ and $I(P_{bcd}) - I(P_{bc}) = \frac{2}{5} \cdot \ln \frac{5}{2} + \frac{3}{5} \cdot \ln \frac{5}{3}$. Hence, $I(Q) - I_{\mathcal{M}}(Q) = -\frac{14}{25} \cdot \ln 2 + \frac{3}{25} \cdot \ln 3 + \ln 5 - \frac{13}{25} \cdot \ln 13 \neq 0$.

The next example illustrates what was mentioned in Remark 3, namely that an undefined expression can occur in the formula (11) defining a DSS.

Example 6. Put $N = \{a,b,c,d\}$ and $\boldsymbol{X}_i = \{0,1\}$ for every $i \in N$. Consider a class of sets $\mathcal{S} = \{S_1, S_2, S_3\}$, where $S_1 = \{a,b\}$, $S_2 = \{a,c\}$ and $S_3 = \{b,c,d\}$. The densities of probability measures from $\mathcal{M} = \{P_A;\, A \in \mathcal{S}\}$ are given as follows: $p_{\{a,b\}}(x) = 1/4$ for any $x \in \boldsymbol{X}_{\{a,b\}}$, $p_{\{a,c\}}(x) = 1/4$ for any $x \in \boldsymbol{X}_{\{a,c\}}$ and $p_{\{b,c,d\}}$ has the value $1/4$ for any of the following four configurations: $(0,0,0)$, $(0,0,1)$, $(1,1,0)$ and $(1,1,1)$. To see that $\mathcal{M}$ is strongly consistent consider a density p on $\boldsymbol{X}_{\{a,b,c,d\}}$ such that $p(x) = 1/8$ for any configuration x of the following eight ones: $(0,0,0,0)$, $(0,0,0,1)$, $(0,1,1,0)$, $(0,1,1,1)$, $(1,0,0,0)$, $(1,0,0,1)$, $(1,1,1,0)$ and $(1,1,1,1)$.

If we consider the ordering $\tau : S_1, S_2, S_3$ then $F_2 = \{a\}$ and $F_3 = \{b,c\}$. The point is that $p_{\{b,c\}}(0,1) = p_{\{b,c\}}(1,0) = 0$. Therefore, one has:

$$\overline{p}_{\tau}(0,0,1,0) = \frac{p_{\{a,b\}}(0,0) \cdot p_{\{a,c\}}(0,1) \cdot p_{\{b,c,d\}}(0,1,0)}{p_{\{a\}}(0) \cdot p_{\{b,c\}}(0,1)} = \frac{\frac{1}{4} \cdot \frac{1}{4} \cdot 0}{\frac{1}{2} \cdot 0}\,,$$

which is an undefined expression. Actually, the sum of the defined terms in (11), that is, $\overline{p}_{\tau}(x)$ with $x_{\{b,c\}} = (0,0)$ or $x_{\{b,c\}} = (1,1)$, is $1/2$. This indicates that the idea to put $\overline{p}_{\tau}(x) = 0$ whenever the expression is not defined does not solve the problem.

A2. Examples related to the comparison of approximations

The following example shows that the optimal DSS approximation could be better than the explicit expression approximation. Actually, it this particular example, the optimal DSS approximation has fitting marginals. The example also shows that it can be the case that $|\mathsf{Exe}| > 1$.

[44] To see this one can utilize the concept of conditional independence and the formula (2.17) in [9]. Indeed, by construction one has $d \perp\!\!\!\perp a \,|\, bc\; [Q]$ and $b \perp\!\!\!\perp c \,|\, a\; [Q]$.

[45] Actually, $I(Q^{bcd}) - I(Q^{bc}) = H(Q^{bcd} | Q^{bc} \times Q^{d})$ and $I(P_{bcd}) - I(P_{bc}) = H(P_{bcd} | P_{bc} \times P_{d})$ and one use the above formulas for q^{B} and p_{B} with $B = \{b,c,d\}$.

Example 7. Put $N = \{a, b, c\}$, $\boldsymbol{X}_i = \{0, 1\}$ for any $i \in N$, $\mathcal{S} = \{A \subseteq N\,;\, |A| = 2\}$. Densities of measures from $\mathcal{M} = \{P_A;\; A \in \mathcal{S}\}$ are given as follows:

$$p_A(0,0) = p_A(0,1) = p_A(1,1) = \frac{1}{3} \quad \text{for } A = \{a,c\} \text{ and } A = \{b,c\}\,,$$

while $p_{\{a,b\}}(0,0) = 2/3$, $p_{\{a,b\}}(1,1) = 1/3$. Clearly, $\mathcal{M}$ is strongly consistent; consider the density p which ascribes $1/3$ to any of the following three configurations of $x_{\{a,b,c\}}$: $(0,0,0)$, $(0,0,1)$ and $(1,1,1)$. Actually, if one takes the ordering $\tau_* : S_1 = \{a,b\}, S_2 = \{b,c\}, S_3 = \{a,c\}$ then the respective DSS has just the density p. In particular, $\mathcal{D}_{\mathcal{M}} \cap \mathcal{K}_{\mathcal{M}} \neq \emptyset$ and p defines an optimal DSS by Corollary 3.

Direct calculation of Exe gives this result:

$$\mathsf{Exe}\,(0,0,0) = \frac{1}{2}, \quad \mathsf{Exe}\,(0,0,1) = \frac{1}{4}, \quad \mathsf{Exe}\,(1,1,1) = \frac{1}{2},$$

and $\mathsf{Exe}\,(x) = 0$ for remaining configurations $x \in \boldsymbol{X}_N$. Therefore, $|\mathsf{Exe}\,| = 5/4 > 1$ and the respective explicit expression approximation has the form

$$\overline{\mathsf{Exe}}(0,0,0) = \frac{2}{5}, \quad \overline{\mathsf{Exe}}(0,0,1) = \frac{1}{5}, \quad \overline{\mathsf{Exe}}(1,1,1) = \frac{2}{5},$$

and $\overline{\mathsf{Exe}}(x) = 0$ for other configurations $x \in \boldsymbol{X}_N$. Hence, $\overline{\mathsf{Exe}}_{\{a,b\}}(0,0) = \frac{3}{5} \neq \frac{2}{3} = p_{\{a,b\}}(0,0)$ implies that $P_{exe} \notin \mathcal{K}_{\mathcal{M}}$. The formulas (12) and (16) allow us to compare the multiinformation contents of Q and the explicit expression P_{exe} directly:

$$\begin{aligned} I_{\mathcal{M}}(Q) - I_{\mathcal{M}}(P_{exe}) &= -I(\{a,c\}) + \ln|\mathsf{Exe}\,| \;=\; -(\ln 3 - \frac{4}{3}\cdot\ln 2) + \ln\frac{5}{4} \\ &= \ln 5 - \ln 3 - \frac{2}{3}\cdot\ln 2 > 0\,. \end{aligned}$$

On the other hand, the next example shows that the explicit expression approximation could be better than the optimal DSS approximation. Moreover, it also shows that it may happen $|\mathsf{Exe}\,| < 1$.

Example 8. Put $N = \{a, b, c\}$, $\boldsymbol{X}_i = \{0, 1\}$ for $i \in N$, $\mathcal{S} = \{A \subseteq N\,;\, |A| = 2\}$. The density p_A of P_A for any $A \in \mathcal{S}$ is given as follows:

$$p_A(0,0) = \frac{2}{3}, \quad p_A(0,1) = p_A(1,0) = \frac{1}{6}, \quad p_A(1,1) = 0\,.$$

To see that $\mathcal{M} = \{P_A; A \in \mathcal{S}\}$ is strongly consistent consider the density p given as follows:

$$p(0,0,0) = \frac{1}{2}, \quad p(0,0,1) = p(0,1,0) = p(1,0,0) = \frac{1}{6}\,,$$

and $p(x) = 0$ for remaining $x \in \boldsymbol{X}_N$. Since $I(P_A) = \frac{7}{3}\cdot\ln 2 + \ln 3 - \frac{5}{3}\cdot\ln 5 \equiv k > 0$ for any $A \in \mathcal{S}$, every ordering τ gives an optimal DSS. For example, the ordering

$S_1 = \{a,b\}$, $S_2 = \{b,c\}$, $S_3 = \{a,c\}$ leads to the following density q of an optimal DSS:

$$q(0,0,0) = \frac{8}{15}, \quad q(0,0,1) = q(1,0,0) = \frac{2}{15}, \quad q(0,1,0) = \frac{1}{6}, \quad q(1,0,1) = \frac{1}{30},$$

and $q(x) = 0$ for remaining $x \in \boldsymbol{X}_N$. Direct computation of Exe gives this result:

$$\mathsf{Exe}\,(0,0,0) = \frac{64}{125}\,, \quad \mathsf{Exe}\,(0,0,1) = \mathsf{Exe}\,(0,1,0) = \mathsf{Exe}\,(1,0,0) = \frac{20}{125}\,,$$

and $\mathsf{Exe}\,(x) = 0$ for remaining configurations $x \in \boldsymbol{X}_N$. In particular, $|\mathsf{Exe}\,| = 124/125 < 1$. Therefore,

$$\overline{\mathsf{Exe}}(0,0,0) = \frac{16}{31}\,, \quad \overline{\mathsf{Exe}}(0,0,1) = \overline{\mathsf{Exe}}(0,1,0) = \overline{\mathsf{Exe}}(1,0,0) = \frac{5}{31}\,,$$

and $\overline{\mathsf{Exe}}(x) = 0$ for other $x \in \boldsymbol{X}_N$. Of course, $P_{exe} \notin \mathcal{K}_{\mathcal{M}}$ as $\overline{\mathsf{Exe}}_{\{a,b\}}(0,0) = \frac{21}{31} \neq \frac{2}{3} = p_{\{a,b\}}(0,0)$. Formulas (16) and (12) allow one to compare multiinformation contents of both (types) of approximation:

$$I_{\mathcal{M}}(P_{exe}) - I_{\mathcal{M}}(Q) \;=\; (-\ln|\mathsf{Exe}\,| + 3k) \,-\, (3k-k) \;=\; k - \ln|\mathsf{Exe}\,| = k + \ln\frac{125}{124} > 0\,,$$

which means that P_{exe} is better.

Note that so far no example was found that $P_{exe} \in \mathcal{K}_{\mathcal{M}}$ and $\mathcal{K}_{\mathcal{M}} \cap \mathcal{D}_{\mathcal{M}} = \emptyset$.

A3. Example related to the barycenter principle

The following example shows that the barycenter of $\mathcal{K}_{\mathcal{M}}$ in itself may differ from the distribution maximizing entropy in $\mathcal{K}_{\mathcal{M}}$.

Example 9. Put $N = \{a,b\}$, $\boldsymbol{X}_a = \boldsymbol{X}_b = \{0,1\}$ and $\mathcal{S} = \{A \subseteq N\,;\, |A| = 1\}$. The collection $\mathcal{M} = \{P_A;\, A \in \mathcal{S}\}$ is given by respective marginal densities:

$$p_{\{a\}}(0) = \frac{1}{3},\; p_{\{a\}}(1) = \frac{2}{3}, \quad p_{\{b\}}(0) = \frac{1}{4},\; p_{\{b\}}(1) = \frac{3}{4}\,.$$

We omit the proof of the fact that $\mathcal{K}_{\mathcal{M}}$ consists of convex combinations of two probability measures, namely the measure R^1 given by the density

$$r^1(0,0) = 0,\;\; r^1(0,1) = \frac{1}{3},\;\; r^1(1,0) = \frac{1}{4},\;\; r^1(1,1) = \frac{5}{12}\,,$$

and the measure R^2 given by the density

$$r^2(0,0) = \frac{1}{4},\;\; r^2(0,1) = \frac{1}{12},\;\; r^2(1,0) = 0,\;\; r^2(1,1) = \frac{2}{3}\,.$$

In particular, the product measure $Q = P_{\{a\}} \times P_{\{b\}}$ with density

$$q(0,0) = \frac{1}{12},\;\; q(0,1) = \frac{1}{4},\;\; q(1,0) = \frac{1}{6},\;\; q(1,1) = \frac{1}{2}$$

has the form $Q = \frac{2}{3} \cdot R^1 + \frac{1}{3} \cdot R^2$. Note that this measure minimizes the multiinformation in $\mathcal{K}_{\mathcal{M}}$ and, therefore, it maximizes the entropy – see Lemma 2. To show that Q differs from the measure chosen by the barycenter principle it suffices to find at least one $R \in \mathcal{K}_{\mathcal{M}}$ such that

$$\mu(Q) \equiv \max_{P \in \mathcal{K}_{\mathcal{M}}} H(P|Q) > \max_{P \in \mathcal{K}_{\mathcal{M}}} H(P|R) \equiv \mu(R)\,.$$

A basic observation is that, given $Q' \in \mathcal{K}_{\mathcal{M}}$ with strictly positive density, the function $P \mapsto H(P|Q')$, $P \in \mathcal{K}_{\mathcal{M}}$ is convex on $\mathcal{K}_{\mathcal{M}}$ and achieves its minimum 0 at $P = Q'$. Moreover, in the considered case, $\mathcal{K}_{\mathcal{M}}$ is an "interval" between R^1 and R^2, for which reason the maximum of the function $P \mapsto H(P|Q')$ is achieved in one of the "extreme" measures R^1 and R^2. In particular,

$$\max_{P \in \mathcal{K}_{\mathcal{M}}} H(P|Q') = \max\{\, H(R^1|Q'), H(R^2|Q')\,\}.$$

Now, direct computation gives

$$H(R^2|Q) = \frac{4}{3} \cdot \ln 2 - \frac{1}{2} \cdot \ln 3 > \frac{-1}{2} \cdot \ln 3 + \frac{5}{12} \cdot \ln 5 = H(R^1|Q)\,,$$

which means that $\mu(Q) = H(R^2|Q) = \frac{4}{3} \cdot \ln 2 - \frac{1}{2} \cdot \ln 3$. We put $R = \frac{1}{3} \cdot R^1 + \frac{2}{3} \cdot R^2$ and observe it has the following density:

$$r(0,0) = r(0,1) = \frac{1}{6},\;\; r(1,0) = \frac{1}{12},\;\; r(1,1) = \frac{7}{12}\,.$$

Thus, we can analogously get

$$H(R^1|R) = \frac{1}{3} \cdot \ln 2 + \frac{1}{4} \cdot \ln 3 + \frac{5}{12} \cdot \ln 5 - \frac{5}{12} \cdot \ln 7 > \frac{5}{3} \cdot \ln 2 + \frac{1}{4} \cdot \ln 3 - \frac{2}{3} \cdot \ln 7 = H(R^2|R)\,,$$

which means that $\mu(R) = H(R^1|R) = \frac{1}{3} \cdot \ln 2 + \frac{1}{4} \cdot \ln 3 + \frac{5}{12} \cdot \ln 5 - \frac{5}{12} \cdot \ln 7$. It is straightforward to observe by detailed computation that $\mu(Q) > \mu(R)$.

ACKNOWLEDGEMENT

The second author is indebted to Imre Csiszár and Fero Matúš, whose expertise helped him to find the counterexample from Section A3, and to the reviewers, whose comments helped him to improve the structure of the paper. The work of the second author had been supported by the Grant Agency of the Academy of Sciences of the Czech Republic under Grant IAA 100750603 and by the Ministry of Education, Youth and Sports of the Czech Republic under project 1M0572.

(Received March 3, 2006.)

REFERENCES

[1] I. Csiszár and F. Matúš: Information projections revisited. IEEE Trans. Inform. Theory *49* (2003), 1474–1490.

[2] H. G. Kellerer: Verteilungsfunktionen mit gegebenem Marginalverteilungen (in German, translation: Distribution functions with given marginal distributions). Z. Wahrsch. verw. Gerbiete *3* (1964), 247–270.

[3] S. L. Lauritzen: Graphical Models. Clarendon Press, Oxford 1996.

[4] A. Perez: ε-admissible simplifications of the dependence structure of random variables. Kybernetika *13* (1979), 439–449.

[5] A. Perez: The barycenter concept of a set of probability measures as a tool in statistical decision. In: The book of abstracts of the 4th Internat. Vilnius Conference on Probability Theory and Mathematical Statistics 1985, pp. 226–228.

[6] A. Perez: Princip maxima entropie a princip barycentra při integraci dílčích znalostí v expertních systémech (in Czech, translation: The maximum entropy principle and the barycenter principle in partial knowledge integration in expert systems). In: Metody umělé inteligence a expertní systémy III (V. Mařík and Z. Zdráhal, eds.), ČSVT – FEL ČVUT, Prague 1987, pp. 62–74.

[7] A. Perez: Explicit expression Exe – containing the same multiinformation as that in the given marginal set – for approximating probability distributions. A manuscript in Word, 2003.

[8] M. Studený: Pojem multiinformace v pravděpodobnostním rozhodování (in Czech, translation: The notion of multiinformation in probabilistic decision-making). CSc Thesis, Czechoslovak Academy of Sciences, Institute of Information Theory and Automation, Prague 1987.

[9] M. Studený: Probabilistic Conditional Independence Structures. Springer–Verlag, London 2005.

Albert Perez (deceased) and Milan Studený, Institute of Information Theory and Automation – Academy of Sciences of the Czech Republic, Pod Vodárenskou věží 4, 182 08 Praha 8. Czech Republic.
e-mail: studeny@utia.cas.cz