Entropy for Evaluation of Dempster-Shafer Belief Function Models

Radim Jiroušek[1,2] **& Václav Kratochvíl**[1,2] **& Prakash P. Shenoy**[3]
[1]Faculty of Management, Prague University of Economics and Business, Jindřichův Hradec
[2]Czech Academy of Sciences, Institute of Information Theory and Automation, Prague
[3]University of Kansas School of Business, Lawrence, KS 66045, USA
[radim, velorex]@utia.cas.cz; pshenoy@ku.edu

Abstract

Applications of Dempster-Shafer (D-S) belief functions to practical problems involve difficulties arising from their high computational complexity. One can use space-saving factored approximations such as graphical belief function models to solve them. Using an analogy with probability distributions, we represent these approximations in the form of compositional models. Since no theoretical apparatus similar to probabilistic information theory exists for D-S belief functions (e. g., dissimilarity measure analogous to the Kullback-Liebler divergence measure), the problems arise not only in connection with the design of algorithms seeking optimal approximations but also in connection with a criterion comparing two different approximations. In this respect, the application of the analogy with probability theory fails. Therefore, in this paper, we conduct some synthetic experiments and describe the results designed to reveal whether some belief function entropy definitions described in the literature can detect optimal approximations, i.e., that achieve their minimum for an optimal approximation.

1. Introduction

Applications of Dempster-Shafer (D-S) belief functions to practical problems involve difficulties arising from their high computational complexity. One can use space-saving factored approximations such as graphical belief function models to solve them. Using an analogy with probability distributions, we represent these approximations in the form of compositional models. However, no theoretical apparatus similar to probabilistic information theory exists for D-S belief functions. There is no dissimilarity measure possessing the properties of the Kullback-Liebler divergence measure. Thus, the problems arise in connection with the design of algorithms seeking optimal approximations because we do not have a criterion comparing two different approximations. In this respect, the application of the analogy with probability theory fails. Therefore, in this paper, we conduct some synthetic experiments and describe the results designed to reveal whether some belief function entropy definitions described in the literature can detect optimal approximations.

Like in probability theory, several entropy measures have been defined in the framework of belief functions. This paper aims to test to what extent these measures serve our needs. The tests are organized as follows. We randomly generate a set of belief functions, ensuring that one (say the first one) is superior to all others. It happens when this first belief function contains all the information from all the remaining belief functions in the set. Thus, we expect that a measure of uncertainty/conflict suitable to rate the optimality of the model should be lower for the first model compared to the remaining ones from the set. In the study, we consider their theoretical properties and computational complexity to address their applicability to real-life problems.

A theoretical foundation for the task is offered by *compositional models*. As superior belief functions, one can consider the models called *perfect decomposable models* [18]. For such models, the data-based machine learning process consists of two steps. The first problem is estimating a system of low-dimensional basic probability assignments (BPAs). Since we consider only low-dimensional BPAs, one can use one of the "statistical approaches" designed in the literature, e.g.,[7, 9]. In the second step, one must order the low-dimensional BPAs to get a joint model describing the reality. And this is the problem for which we need a suitable entropy. This step is solved using information-theoretic characteristics based on Shannon entropy when working with probabilistic compositional models. In the

case of Dempster-Shafer belief functions, we have numerous definitions of entropy similar to Shannon entropy for probability distributions [19]. In Section 8, we present results of computational experiments that should answer the question of whether one can find a belief function entropy (described in the literature) that could be used to detect an optimal solution or the extent to which a definition meets this goal. To help the reader understand what compositional models are and how they are used to construct real-life joint models, Section 2 describes the original ideas suggested in the framework of probability theory.

The remaining parts of the paper are organized as follows. Section 3 introduces the basics of D-S belief function theory. Next, Sections 4 and 5 introduce the main definitions of compositional models and their properties, respectively. A survey of entropies for belief functions is briefly presented in Section 6. In Section 7, we explain by an example how the computational experiments were conducted. Finally, Section 8 describes the experimental results.

2. Motivation

Space-complexity problems analogous to those discussed in Section 1 had to be solved in the 1970s by applying probabilistic models to support a decision. Perez suggested using a class of space-saving approximations of joint probability distributions that he called *approximations simplifying the dependence structure* [39].

Consider a large set of discrete random variables $\mathcal{W}$ with a joint probability distribution π. For an arbitrary partition $\{\mathcal{U}_1, \mathcal{U}_2, \ldots, \mathcal{U}_k\}$ of $\mathcal{W}$, one can decompose the joint distribution π using the chain rule as follow:

$$\pi(\mathcal{W}) = \pi(\mathcal{U}_1)\pi(\mathcal{U}_2|\mathcal{U}_1)\ldots\pi(\mathcal{U}_k|(\mathcal{U}_1 \cup ... \cup \mathcal{U}_{k-1})) = \prod_{i=1}^{k} \pi(\mathcal{U}_i|(\mathcal{U}_1 \cup \ldots \cup \mathcal{U}_{i-1})). \tag{1}$$

In Eq. (1), for $i = 1$, $\pi(\mathcal{U}_i|(\mathcal{U}_1 \cup \ldots \cup \mathcal{U}_{i-1}))$ is just the marginal $\pi(\mathcal{U}_1)$. For $i = 2$, $\pi(\mathcal{U}_2|\mathcal{U}_1)$ is the conditional probability table for $\mathcal{U}_2$ given $\mathcal{U}_1$, etc. In large models ($|\mathcal{W}|$ is large), it is rarely the case that the conditional marginal of $\mathcal{U}_i$ depends on all variables in $\mathcal{U}_1 \cup \ldots \cup \mathcal{U}_{i-1}$. This fact was exploited by Perez [39], who suggested using an *ε-admissible approximation by simplification of the dependence structure*[1] to overcome the computational complexity problem. His basic idea is as follows. Substitute each set $(\mathcal{U}_1 \cup \ldots \cup \mathcal{U}_{i-1})$ in Eq. (1) by its smaller subset $\mathcal{T}_i$ such that the conditional probability distribution $\pi(\mathcal{U}_i|\mathcal{T}_i)$ is almost the same as $\pi(\mathcal{U}_i|(\mathcal{U}_1 \cup \ldots \cup \mathcal{U}_{i-1}))$. The non-similarity of probability distributions π and κ defined on Ω can be measured using the Kullback-Leibler (KL) divergence [33] defined as follows[2]

$$KL(\pi\|\kappa) = \sum_{x\in\Omega:\kappa(x)>0} \pi(x) \log\left(\frac{\pi(x)}{\kappa(x)}\right). \tag{2}$$

Thus, consider a joint distribution as follows:

$$\kappa(\mathcal{W}) = \prod_{i=1}^{k} \pi(\mathcal{U}_i|\mathcal{T}_i) = (\ldots((\pi(\mathcal{U}_1) \rhd \pi(\mathcal{U}_2 \cup \mathcal{T}_2)) \rhd \pi(\mathcal{U}_3 \cup \mathcal{T}_3)) \rhd \ldots \rhd \pi(\mathcal{U}_{k-1} \cup \mathcal{T}_{k-1})) \rhd \pi(\mathcal{U}_k \cup \mathcal{T}_k), \tag{3}$$

where the binary operator, called a *composition* operator, is as follows:

$$\nu_1(\mathcal{V}_1) \rhd \nu_2(\mathcal{V}_2) = \nu_1(\mathcal{V}_1) \cdot \nu_2(\mathcal{V}_2 \backslash \mathcal{V}_1 | \mathcal{V}_1 \cap \mathcal{V}_2).$$

A probability distribution that can be expressed in the form of a multiple application of the composition operator (as that in Eq. (3)) is called a *compositional model*. If $KL(\pi\|\kappa) \leqslant \varepsilon$, then κ is an *ε-admissible approximation of π*.

Now, consider a different problem. Let $\{\mathcal{V}_1, \mathcal{V}_2, \ldots, \mathcal{V}_k\}$ be a set of subsets of $\mathcal{W}$ (generally not disjoint) such that $\bigcup_{i=1}^{k} \mathcal{V}_i = \mathcal{W}$. Given a set of low-dimensional distributions $\{\kappa_i(\mathcal{V}_i)\}_{i=1,\ldots,k}$, a question is whether there exists a joint distribution for $\mathcal{W}$ such that all κ_i's are its marginals. If such a distribution π exists, then a natural question is how

[1] The notion reflects the fact that the considered approximation extends the set of conditional independence relations holding for the probability distribution in question [49].

[2] Eq. (2) defines the KL divergence if κ dominates π, i.e., if for all $x \in \Omega$, for which $\kappa(x) = 0$, $\pi(x)$ is also 0. Otherwise, the KL divergence is defined to be $+\infty$.

to find it, or, if its computation is intractable[3], how to approximate it. Perez [39] looked for an approximation within the class of approximations simplifying the dependence structure that could be assembled from the given system of marginals $\{\kappa_i(\mathcal{V}_i)\}_{i=1,\ldots,k}$. Thus, he considered all the permutation of indices (i.e., bijections) $j : \{1,2,\ldots,k\} \to \{1,2,\ldots,k\}$, each defining a probability distribution

$$\kappa[j] = \prod_{i=1}^{k} \kappa_{j(i)}((\mathcal{V}_{j(i)} \backslash \mathcal{T}_{j(i)})|\mathcal{T}_{j(i)}),$$

where $\mathcal{T}_{j(i)} = \mathcal{V}_{j(i)} \cap (\mathcal{V}_{j(1)} \cup \ldots \cup \mathcal{V}_{j(i-1)})$. He was looking for the approximation (permutation j) minimizing $KL(\pi \| \kappa[j])$. For this, he showed [39] that[4]

$$KL(\pi \| \kappa[j]) = -H(\pi) + \sum_{i=1}^{k} \Big(H(\kappa_{j(i)}(\mathcal{V}_{j(i)})) - H(\kappa_{j(i)}(\mathcal{T}_{j(i)})) \Big); \tag{4}$$

recall that Eq. (4) holds under the assumption that all κ_i's are marginals of π. If for some permutation j, all κ_i's are also marginals of $\kappa[j]$, then Eq. (4) can be further simplified getting $KL(\pi \| \kappa[j]) = H(\kappa[j]) - H(\pi)$. Thus, regardless of whether distribution π is known or not, he proved that its best approximation (that simplifies the dependence structure), which can be set up from $\{\kappa_i(\mathcal{V}_i)\}_{i=1,\ldots,k}$, is that which minimizes $\sum_{i=1}^{k} \Big(H(\kappa_{j(i)}(\mathcal{V}_{j(i)})) - H(\kappa_{j(i)}(\mathcal{T}_{j(i)})) \Big)$. Suppose one considers only so-called perfect approximations, i.e., the approximations $\kappa[j]$ having all κ_i's for its marginals (and still assuming that all κ_i are marginals of π). In that case, the best approximation minimizes its Shannon entropy $H(\kappa[j])$. As showed in [14], such an approximation is unique in the sense that for two different permutations j and j', such that all κ_i's are marginals of both $\kappa[j]$ and $\kappa[j']$, $\kappa[j] = \kappa[j']$. The minimization of the Shannon entropy also corresponds with the intuition that the best approximation maximizes an information content, which can be expressed for the considered compositional model $\kappa[j]$

$$IC(\kappa[j]) = \sum_{X \in \mathcal{W}} H(\pi^{\downarrow X}) - H(\kappa[j]). \tag{5}$$

Example. Consider four binary variables $\mathcal{W} = \{U, X, Y, Z\}$, their subsets $\mathcal{V}_1 = \{U, X\}$, $\mathcal{V}_2 = \{X, Y\}$, $\mathcal{V}_3 = \{Y, Z\}$, and the corresponding two-dimensional probability distributions $\kappa_i(\mathcal{V}_i)$, which are pairwise consistent, i.e., $\kappa_1^{\downarrow\{X\}} = \kappa_2^{\downarrow\{X\}}$, $\kappa_2^{\downarrow\{Y\}} = \kappa_3^{\downarrow\{Y\}}$. In this simple example, all possible permutations $j : \{1,2,3\} \to \{1,2,3\}$ (we will depict them in a form of a vector $(j(1), j(2), j(3))$ in this example) define only two different approximations. If $j(3) \neq 2$, then $\kappa[j] = \kappa[(1,2,3)]$ is a perfect approximation. For the remaining two permutations $\kappa[(1,3,2)] = \kappa[(3,1,2)] = \kappa_1\kappa_3$. For this distribution, variables X and Y are independent, which means that κ_2 (see Table 1) cannot be a marginal of $\kappa[(1,3,2)]$.

Table 1: Two-dimensional probability distributions.

$\kappa_1(U,X)$		$\kappa_2(X,Y)$		$\kappa_3(Y,Z)$	
(u,x)	0.2	(x,y)	0.4	(y,z)	0.25
$(u,\bar{x})$	0.3	$(x,\bar{y})$	0.1	$(y,\bar{z})$	0.25
$(\bar{u},x)$	0.3	$(\bar{x},y)$	0.1	$(\bar{y},z)$	0.25
$(\bar{u},\bar{x})$	0.2	$(\bar{x},\bar{y})$	0.4	$(\bar{y},\bar{z})$	0.25

We have not defined the joint distribution π in this example, so we cannot compute the Kullback-Leibler divergence of the approximations from π. However, computing the value $\Gamma[j] = \sum_{i=1}^{k} \Big(H(\kappa_{j(i)}(\mathcal{V}_{j(i)})) - H(\kappa_{j(i)}(\mathcal{T}_{j(i)})) \Big)$ for the

[3] As proved by Csiszár [3], the iterative procedure described in [5] converges to the required probability distribution. Nevertheless, the computational complexity of this procedure is exponential with the number of variables, and therefore its practical application to multidimensional probability distributions may easily become intractable.

[4] In the whole paper, H without any index denotes Shannon entropy [41] of a probability distribution. Entropies and entropy-like functions for belief functions will be denoted by H with different indices.

two permutations $j = (1,2,3)$ and $j = (1,3,2)$, we get $\Gamma[(1,2,3)] = 3.69$ and $\Gamma[(1,3,2)] = 3.97$, which shows that $\kappa[(1,2,3)]$ is a better approximation of any distribution π having $\kappa_1, \kappa_2, \kappa_3$ for its marginals than the approximation $\kappa[(1,3,2)]$.

3. Belief Functions

There are several theories that use belief functions (and their equivalent representations such as basic probability assignments, plausibility functions, commonality functions, credal sets, etc.) to represent evidence. The theories differ mainly in the combination rules used to aggregate evidence. In this paper, we are interested in the Dempster-Shafer (D-S) theory of belief functions, which uses Dempster's combination rule [6].

As in Section 2, let $\mathcal{W}$ denote a set of variables with finite number of states. For $X \in \mathcal{W}$, Let Ω_X denote the set of states of variable X. Let $\Omega_\mathcal{W}$ denote $\times_{X\in W}\Omega_X$, the set of states for $\mathcal{W}$.

A *basic probability assignment* (BPA) for variables $\mathcal{U} \subseteq \mathcal{W}$ is a mapping $m_\mathcal{U} : 2^{\Omega_\mathcal{U}} \to [0,1]$, such that $\sum_{\mathbf{a}\subseteq\Omega_\mathcal{U}} m_\mathcal{U}(\mathbf{a}) = 1$ and $m_\mathcal{U}(\varnothing) = 0$.

Consider a BPA $m_\mathcal{U}$ for $\mathcal{U}$. If the set of the corresponding variables is clear from the context, we omit the subscript $\mathcal{U}$. Thus, we say that $\mathbf{a}$ is a *focal element* of m if $m(\mathbf{a}) > 0$. A BPA with only one focal element is called *deterministic*; $\iota_\mathcal{U}$ denote the deterministic BPA, for which $\iota_\mathcal{U}(\Omega_\mathcal{U}) = 1$. Since $\iota_\mathcal{U}$ represents a total ignorance, it is called *vacuous*.

A BPA m for $\Omega_\mathcal{U}$ can also be defined by the corresponding *belief function* (BEL), or by *plausibility function* (PL), or by *commonality function* (CF) [40] as follows:

$$Bel_m(\mathbf{a}) = \sum_{\mathbf{b}\subseteq\Omega_\mathcal{U}:\,\mathbf{b}\subseteq\mathbf{a}} m(\mathbf{b}),$$

$$Pl_m(\mathbf{a}) = \sum_{\mathbf{b}\subseteq\Omega_\mathcal{U}:\,\mathbf{b}\cap\mathbf{a}\neq\varnothing} m(\mathbf{b}),$$

$$Q_m(\mathbf{a}) = \sum_{\mathbf{b}\subseteq\Omega_\mathcal{U}:\,\mathbf{b}\supseteq\mathbf{a}} m(\mathbf{b}).$$

These representations are equivalent; when one of these functions is given, we can compute the others uniquely. For example:

$$Pl_m(\mathbf{a}) = 1 - Bel_m(\Omega_\mathcal{U}\backslash\mathbf{a}),$$

$$m(\mathbf{a}) = \sum_{\mathbf{b}\subseteq\mathbf{a}} (-1)^{|\mathbf{a}\backslash\mathbf{b}|} Bel_m(\mathbf{b}),$$

$$m(\mathbf{a}) = \sum_{\mathbf{b}\subseteq\Omega_\mathcal{U}:\,\mathbf{b}\supseteq\mathbf{a}} (-1)^{|\mathbf{b}\backslash\mathbf{a}|} Q_m(\mathbf{b}). \tag{6}$$

A BPA m on $\Omega_\mathcal{U}$ (or, equivalently, BEL Bel_m for $\mathcal{U}$) defines a set $\mathbb{P}_m$ of probability mass functions on $\Omega_\mathcal{U}$

$$\mathbb{P}_m = \left\{ \pi \text{ defined on } \Omega_\mathcal{U} \mid \forall \mathbf{a} \subseteq \Omega_\mathcal{U} : \sum_{c\in\mathbf{a}} \pi(c) \geqslant Bel_m(\mathbf{a}) \right\}$$

that is called a *credal* set of m. The credal set of m uniquely defines m, and vice versa.

Marginalization. For BPA $m_\mathcal{V}$, we often consider its *marginal* BPA $m_\mathcal{V}^{\downarrow\mathcal{U}}$ for $\mathcal{U} \subseteq \mathcal{V}$. A similar notation is used also for *projections*: for $a \in \Omega_\mathcal{V}$, $a^{\downarrow\mathcal{U}}$ denotes the element of $\Omega_\mathcal{U}$ that is obtained from a by omitting the values of variables in $\mathcal{V}\backslash\mathcal{U}$. Formally, if $\mathbf{a} \subseteq \Omega_\mathcal{V}$, then

$$\mathbf{a}^{\downarrow\mathcal{U}} = \{a^{\downarrow\mathcal{U}} : a \in \mathbf{a}\}.$$

The marginal $m_\mathcal{V}^{\downarrow\mathcal{U}}$ of BPA $m_\mathcal{V}$ for $\mathcal{U}$ is defined as follows:

$$m_\mathcal{V}^{\downarrow\mathcal{U}}(\mathbf{b}) = \sum_{\mathbf{a}\subseteq\Omega_\mathcal{V}:\,\mathbf{a}^{\downarrow\mathcal{U}}=\mathbf{b}} m_\mathcal{V}(\mathbf{a}).$$

for all $\mathbf{b} \subseteq \Omega_{\mathcal{U}}$.

The projection of sets enables us to define a *join* of two sets. Consider two arbitrary sets $\mathcal{U}$ and $\mathcal{V}$ of variables (they may be disjoint or overlapping, or one may be a subset of the other). Consider two sets $\mathbf{a} \subseteq \Omega_{\mathcal{U}}$ and $\mathbf{b} \subseteq \Omega_{\mathcal{V}}$. Their join is defined as:

$$\mathbf{a} \bowtie \mathbf{b} = \{c \in \Omega_{\mathcal{U}\cup\mathcal{V}} : c^{\downarrow\mathcal{U}} \in \mathbf{a} \ \ \& \ \ c^{\downarrow\mathcal{V}} \in \mathbf{b}\}.$$

Notice that if $\mathcal{U}$ and $\mathcal{V}$ are disjoint, then $\mathbf{a} \bowtie \mathbf{b} = \mathbf{a} \times \mathbf{b}$, if $\mathcal{U} = \mathcal{V}$, then $\mathbf{a} \bowtie \mathbf{b} = \mathbf{a} \cap \mathbf{b}$, and, in general, for $\mathbf{c} \subseteq \Omega_{\mathcal{U}\cup\mathcal{V}}$, $\mathbf{c}$ is a subset of $\mathbf{c}^{\downarrow\mathcal{U}} \bowtie \mathbf{c}^{\downarrow\mathcal{V}}$, which may be a proper subset.

Dempster's Combination Rule. In the D-S theory, Dempster's combination rule is used to combine distinct belief functions. Consider two distinct BPAs $m_{\mathcal{U}}$ and $m_{\mathcal{V}}$ for $\mathcal{U}$ and $\mathcal{V}$, respectively. Dempster's combination rule is defined for each $\mathbf{c} \subseteq \Omega_{\mathcal{U}\cup\mathcal{V}}$ as follows:

$$(m_{\mathcal{U}} \oplus m_{\mathcal{V}})(\mathbf{c}) = \frac{1}{K} \sum_{\mathbf{a}\subseteq\Omega_{\mathcal{U}},\mathbf{b}\subseteq\Omega_{\mathcal{V}}:\mathbf{a}\bowtie\mathbf{b}=\mathbf{c}} m_{\mathcal{U}}(\mathbf{a}) \cdot m_{\mathcal{V}}(\mathbf{b}), \tag{7}$$

where the normalization constant

$$K = \sum_{\mathbf{a}\subseteq\Omega_{\mathcal{U}},\mathbf{b}\subseteq\Omega_{\mathcal{V}}:\mathbf{a}\bowtie\mathbf{b}\neq\varnothing} m_{\mathcal{U}}(\mathbf{a}) \cdot m_{\mathcal{V}}(\mathbf{b}). \tag{8}$$

$(1-K)$ can be interpreted as the amount of conflict between $m_{\mathcal{U}}$ and $m_{\mathcal{V}}$. If $(1-K) = 1$, then we say that BPAs $m_{\mathcal{U}}$ and $m_{\mathcal{V}}$ are in total conflict and their Dempster's combination is undefined.

For this combination rule, the assumption of distinct belief functions is essential. In general $m \oplus m \neq m$. Double-counting of evidence by combining non-distinct basic assignments leads to erroneous results. In directed graphical belief function models consisting of priors and conditionals, all BPAs are distinct and we combine these using Dempster's rule. If we get some evidence represented as a BPA, and such a BPA is distinct from the BPAs in a graphical model, then we can get a posterior joint BPA by combining all BPAs using Dempster's rule.

Dempster's combination rule may also be described using the corresponding commonality functions. Consider two distinct BPAs $m_{\mathcal{U}}$, $m_{\mathcal{V}}$ and the corresponding commonality functions $Q_{m_{\mathcal{U}}}$ and $Q_{m_{\mathcal{V}}}$. Then, as shown in [40],

$$Q_{m_{\mathcal{U}}\oplus m_{\mathcal{V}}}(\mathbf{c}) = \left(\frac{1}{K}\right) Q_{m_{\mathcal{U}}}(\mathbf{c}^{\downarrow\mathcal{U}})\, Q_{m_{\mathcal{V}}}(\mathbf{c}^{\downarrow\mathcal{V}}), \tag{9}$$

where K is the same as that defined in Equation (8).

There is an important property of Dempster's combination rule and marginalization called *local computation* [44]. If $\mathcal{T} \subseteq \mathcal{U}$ and $(\mathcal{U}\backslash\mathcal{T}) \cap \mathcal{V} = \varnothing$, then

$$(m_{\mathcal{U}} \oplus m_{\mathcal{V}})^{\downarrow\mathcal{T}} = (m_{\mathcal{U}})^{\downarrow\mathcal{T}} \oplus m_{\mathcal{V}}$$

Thus, when we want to find the marginal of $m_{\mathcal{U}} \oplus m_{\mathcal{V}}$ for $\mathcal{T}$, by removing variables in $\mathcal{U}\backslash\mathcal{T}$ that does not include any variables in $\mathcal{V}$, then we can avoid combination on the state space of $\mathcal{U} \cup \mathcal{V}$ and do it instead on the smaller space of $\mathcal{T} \cup \mathcal{U}$.

Suppose m is a BPA for X, and we observe $X = x$, where $x \in \Omega_X$. Let $m_{X=x}$ denote the deterministic BPA for X such that $m_{X=x}(\{x\}) = 1$. Then, in the D-S theory, our posterior BPA for X is $m \oplus m_{X=x}$. An alternative way of dealing with the observation $X = x$ is to condition each $\pi \in \mathbb{P}_m$ using the observation $X = x$. This results in a new credal set that corresponds to m combined with $m_{X=x}$ using the so-called Fagin-Halpern combination rule [11], $m \oplus_{FH} m_{X=x}$, which, in general, is different from $m \oplus m_{X=x}$. In this sense, the credal set semantics of m is incompatible with Dempster's rule.

Removal. The inverse of Dempster's combination rule is called removal. Since Dempster's combination is defined as pointwise combination of CFs followed by normalization, removal is defined as pointwise division of CFs followed by normalization. Suppose $\mathcal{U} \subsetneq \mathcal{V}$ and $Q_{\mathcal{U}}$ is a marginal CF of $Q_{\mathcal{V}}$, i.e., $Q_{\mathcal{V}}^{\downarrow\mathcal{U}} = Q_{\mathcal{U}}$. Then, $Q_{\mathcal{V}} \ominus Q_{\mathcal{U}}$, is defined as follows:

$$(Q_{\mathcal{V}} \ominus Q_{\mathcal{U}})(\mathsf{a}) = K^{-1}\, Q_{\mathcal{V}}(\mathsf{a}) / Q_{\mathcal{U}}(\mathsf{a}^{\downarrow\mathcal{U}}), \tag{10}$$

for all $\mathsf{a} \in 2^{\Omega_\mathcal{V}}$, where K is a normalization constant given by

$$K = \sum_{\varnothing \neq \mathsf{a} \subseteq \Omega_\mathcal{V}} (-1)^{|\mathsf{a}|+1} Q_\mathcal{V}(\mathsf{a}) / Q_\mathcal{U}(\mathsf{a}^{\downarrow \mathcal{U}}). \tag{11}$$

We have defined the removal only for the case where we are removing the marginal $Q_\mathcal{U}$ from $Q_\mathcal{V}$. Thus, if $Q_\mathcal{U}(\mathsf{a}^{\downarrow \mathcal{U}}) = 0$, then $Q_\mathcal{V}(\mathsf{a}) = 0$. In this case, we define $0/0 = 1$. For more details of the properties of the removal operator $\ominus$, see [42]. In [47], the removal operator is referred to as the decombination operator.

The CF $Q_\mathcal{V} \ominus Q_\mathcal{U}$ defined in Eq. (10) and (11) may fail to be a CF because the corresponding BPA m may have negative probabilities that sum to 1. Such BPAs are called *pseudo-BPAs*. So, an important question is: Under what conditions will removal result in a CF whose corresponding BPA is not a pseudo-BPA? The following proposition answers this question.

Proposition 1. *Suppose $m_{X,Y}$ is a BPA for $\{X, Y\}$ with corresponding CF $Q_{m_{X,Y}}$. Let m_X denote the marginal of $m_{X,Y}$ for X, i.e., $m_X = (m_{X,Y})^{\downarrow X}$. Then, $Q_{m_{X,Y}} \ominus Q_{m_X}$ is a CF if and only if there exists a BPA m for $\{X, Y\}$ such that $m_{X,Y} = m_X \oplus m$, and $m^{\downarrow X}$ is the vacuous BPA for X.*

Proof. Let ι_X denote the vacuous BPA for X. If $m_{X,Y} = m_X \oplus m$, $Q_{m_{X,Y}} = Q_{m_X} \oplus Q_m$, i.e., for each $\mathsf{a} \in 2^{\Omega_{\{X,Y\}}}$, we have

$$Q_{m_{X,Y}}(\mathsf{a}) = K^{-1}\, Q_{m_X}(a^{\downarrow X})\, Q_m(\mathsf{a}) \tag{12}$$

If we marginalize both sides of Eq. (12) using the above-mentioned local computation property, we get $Q_{m_X}(\mathsf{a}^{\downarrow X}) = K^{-1}\, Q_{m_X}(\mathsf{a}^{\downarrow X})\, Q_{\iota_X}(\mathsf{a}^{\downarrow X})$. As $Q_{\iota_X}(\mathsf{a}^{\downarrow X}) = 1$, $K^{-1} = 1$.

$$\begin{aligned}
(Q_{m_{X,Y}} \ominus Q_{m_X})(\mathsf{a}) &= K_1^{-1}\, Q_{m_{X,Y}}(\mathsf{a}) / Q_{m_X}(\mathsf{a}^{\downarrow X}) \\
&= K_1^{-1}\, Q_{m_X}(\mathsf{a}^{\downarrow X})\, Q_m(\mathsf{a}) / Q_{m_X}(\mathsf{a}^{\downarrow X}) \\
&= K_1^{-1}\, Q_m(\mathsf{a}).
\end{aligned} \tag{13}$$

On the right side of Eq. (13), we have CF values for (X, Y). So, the left side of Eq. (13) must also be CF values and $K_1^{-1} = 1$.

□

The intuition behind Proposition 1 is as follows. Since pointwise multiplication of CFs represents a combination of knowledge, pointwise division of CFs represents the removal of knowledge in the denominator from the knowledge in the numerator. In the belief function literature, BPA m in Proposition 1 is called a conditional for Y given X [23]. Proposition 1 posits that if we remove knowledge m_X that is included in $m_{X,Y}$, and $m_{X,Y}$ is Dempster's combination of the marginal m_X for X and a conditional m for Y given X, then such removal never results in a pseudo-BPA.

Plausibility and Pignistic Transforms. We discuss two probability transforms of a BPA m. After normalizing the plausibility function for singleton subsets, one gets for each $\mathbf{a} \subseteq \Omega$

$$\lambda_m(\mathbf{a}) = \frac{\sum_{b \in \mathbf{a}} Pl_m(\{b\})}{\sum_{b \in \Omega} Pl_m(\{b\})} \tag{14}$$

a probability distribution on Ω. λ_m is called a *plausibility transform* of BPA m [2].

Several other probabilistic transforms of a BPA m are described in the literature (e.g., [4]). Here, we are only concerned with the *pignistic transform* [46, 48] defined as follows:

$$\pi_m(\mathbf{a}) = \sum_{a \in \mathbf{a}} \sum_{\mathbf{b} \subseteq \Omega : a \in \mathbf{b}} \frac{m(\mathbf{b})}{|\mathbf{b}|}. \tag{15}$$

The plausibility transform of $m_1 \oplus m_2$ is the same as the probabilistic combination of the plausibility transforms of m_1 and m_2. This is not true for any other probabilistic transforms. Thus, it has been argued that only the plausibility transform makes sense for the D-S theory [2].

4. Composition Operator

In Section 2, we recalled that probabilistic compositional models are joint probability distributions *composed* from a system of low-dimensional marginal probability distributions using the probabilistic composition operator. In contrast with Dempster's rule, the low-dimensional marginal distributions are not assumed to be distinct, especially if the domains of marginals are not disjoint. The composition operator aggregates low-dimensional marginals and considers common information in two non-distinct marginals.

The following axiomatic definition introduces a composition operator for (non-distinct) marginal BPA functions. We say that two BPAs $m_{\mathcal{T}}$ and $m_{\mathcal{U}}$ are consistent if $m_{\mathcal{T}}^{\downarrow \mathcal{T}\cap\mathcal{U}} = m_{\mathcal{U}}^{\downarrow \mathcal{T}\cap\mathcal{U}}$.

Definition 1. *By a* composition operator $\rhd$*, we mean a binary operator satisfying the following four axioms: Consider three (possibly non-distinct) marginal BPAs $m_{\mathcal{T}}$, $m_{\mathcal{U}}$, and $m_{\mathcal{V}}$ for $\mathcal{T}$, $\mathcal{U}$, and $\mathcal{V}$, respectively.*

A1 (Domain): *$m_{\mathcal{T}} \rhd m_{\mathcal{U}}$ is a BPA for variables $\mathcal{T} \cup \mathcal{U}$.*

A2 (Composition preserves first marginal): $(m_{\mathcal{T}} \rhd m_{\mathcal{U}})^{\downarrow \mathcal{T}} = m_{\mathcal{T}}$.

A3 (Commutativity under consistency): *If $m_{\mathcal{T}}$ and $m_{\mathcal{U}}$ are consistent, then $m_{\mathcal{T}} \rhd m_{\mathcal{U}} = m_{\mathcal{U}} \rhd m_{\mathcal{T}}$.*

A4 (Associativity under special condition): *If $\mathcal{T} \supset (\mathcal{U} \cap \mathcal{V})$, or, $\mathcal{U} \supset (\mathcal{T} \cap \mathcal{V})$, then $(m_{\mathcal{T}} \rhd m_{\mathcal{U}}) \rhd m_{\mathcal{V}} = m_{\mathcal{T}} \rhd (m_{\mathcal{U}} \rhd m_{\mathcal{V}})$.*

Notice that axioms *A1*, *A3*, *A4* guarantee that the composition operator uniquely reconstructs BPA $m_{\mathcal{T}\cup\mathcal{V}}$ from its marginals, if there exists a *lossless*[5] decomposition of $m_{\mathcal{T}\cup\mathcal{V}}$ into $m_{\mathcal{T}}$ and $m_{\mathcal{V}}$. Axiom *A3* refers to situations, when $m_{\mathcal{T}}$ and $m_{\mathcal{U}}$ are consistent, which happens when there exists their joint extension. In such a case we can construct their extension regardless we start with $m_{\mathcal{T}}$ or $m_{\mathcal{U}}$. Surprisingly, it is axiom *A4*, which guarantees that no necessary information from $m_{\mathcal{V}}$ is lost. Axiom *A2* solves the problem arising when inconsistent BPAs are composed. Generally, there could be two ways of coping with this problem. Either find a compromise (a mixture of inconsistent pieces of knowledge) or give preference to one of the sources. The solution expressed by axiom *A2* decreases the space complexity of computational algorithms necessary for handling joint models.

In the next section, we will study compositional models, i.e., a joint BPA composed from a set of low-dimensional marginal BPAs. Storing a system of low-dimensional marginal BPAs requires much less space than storing the joint BPA (assuming that the latter can be done). More importantly, one can design efficient computational procedures for making inferences from such joint BPAs without explicitly computing the joint. The properties that are a consequence of Axioms A1-A4 are summarized in the following assertion (for proofs, see [18]).

Proposition 2. *For BPAs $m_{\mathcal{T}}, m_{\mathcal{U}}, m_{\mathcal{V}}$ the following statements hold.*

1. (Reduction:) *If $\mathcal{U} \subseteq \mathcal{T}$, then $m_{\mathcal{T}} \rhd m_{\mathcal{U}} = m_{\mathcal{T}}$.*
2. (Stepwise composition): *If $(\mathcal{T} \cap \mathcal{U}) \subseteq \mathcal{V} \subseteq \mathcal{U}$, then $(m_{\mathcal{T}} \rhd m_{\mathcal{U}}^{\downarrow \mathcal{V}}) \rhd m_{\mathcal{U}} = m_{\mathcal{T}} \rhd m_{\mathcal{U}}$.*
3. (Exchangeability): *If $\mathcal{T} \supset (\mathcal{U} \cap \mathcal{V})$, then $(m_{\mathcal{T}} \rhd m_{\mathcal{U}}) \rhd m_{\mathcal{V}} = (m_{\mathcal{T}} \rhd m_{\mathcal{V}}) \rhd m_{\mathcal{U}}$.*
4. (Local computation): *If $(\mathcal{T} \cap \mathcal{U}) \subseteq \mathcal{V} \subseteq (\mathcal{T} \cup \mathcal{U})$, then $(m_{\mathcal{T}} \rhd m_{\mathcal{U}})^{\downarrow \mathcal{V}} = m_{\mathcal{T}}^{\downarrow \mathcal{T}\cap\mathcal{V}} \rhd m_{\mathcal{U}}^{\downarrow \mathcal{U}\cap\mathcal{V}}$.*

Before we discuss the composition operator, notice that Dempster's combination rule is *not* a composition operator. Dempster's rule should be applied only to distinct belief functions representing independent pieces of evidence. On the other hand, the composition operator is typically used to combine two not distinct marginals with a non-empty intersection. The composition operator is defined to avoid double counting of evidence from the two composed pieces of evidence. Thus, the two operators have different properties. Whereas Dempster's rule is always commutative and associative, the composition operator has these properties only in particular situations. On the other hand, Dempster's

[5] By this term we understand that, roughly speaking, there is a possibility to reconstruct $m_{\mathcal{T}\cup\mathcal{V}}$ from its marginals $m_{\mathcal{T}}$ and $m_{\mathcal{V}}$ without adding an additional information about $m_{\mathcal{T}\cup\mathcal{V}}$. It can be done if groups of variables $\mathcal{T}\setminus\mathcal{V}$ and $\mathcal{V}\setminus\mathcal{T}$ are *conditionally independent* given $\mathcal{T} \cap \mathcal{V}$. For a more formal introduction of this concept, see [15].

rule does not preserve the first marginal. Nevertheless, as shown below, Dempster's rule may be used to define a composition operator.

As we saw in Eq. (9), Dempster's rule can be described as pointwise multiplication of CFs followed by normalization. When composing uncertain knowledge, which may not be distinct, we must ensure that no information is double-counted. One way to ensure that knowledge is not double-counted is to remove the common knowledge using the removal operator described in Eq. (10). We have to ensure that the composition preserves the first BPA. This means we must ignore the information in the second BPA, which repeats or conflicts with the information in the first BPA. The following composition operator does this.

Definition 2. *Consider two BPAs $m_{\mathcal{U}}$, $m_{\mathcal{V}}$, and their commonality functions $Q_{m_{\mathcal{U}}}$ and $Q_{m_{\mathcal{V}}}$. Their d-composition is a BPA $m_{\mathcal{U}} \rhd_d m_{\mathcal{V}}$, the corresponding commonality function of which is given by the d-composition of their commonality functions defined for each $\mathbf{c} \subseteq \Omega_{\mathcal{U}\cup\mathcal{V}}$ by the following expression:*

$$(Q_{m_{\mathcal{U}}} \rhd_d Q_{m_{\mathcal{V}}})(\mathbf{c}) = \begin{cases} \frac{1}{L} \frac{Q_{m_{\mathcal{U}}}(\mathbf{c}^{\downarrow\mathcal{U}}) \cdot Q_{m_{\mathcal{V}}}(\mathbf{c}^{\downarrow\mathcal{V}})}{Q_{m_{\mathcal{V}}^{\downarrow\mathcal{U}\cap\mathcal{V}}}(\mathbf{c}^{\downarrow\mathcal{U}\cap\mathcal{V}})} & \text{if } Q_{m_{\mathcal{V}}^{\downarrow\mathcal{U}\cap\mathcal{V}}}(\mathbf{c}^{\downarrow\mathcal{U}\cap\mathcal{V}}) > 0, \\ 0 & \text{otherwise}, \end{cases} \tag{16}$$

where the normalization constant

$$L = \sum_{\mathbf{c}\subseteq\Omega_{\mathcal{U}\cup\mathcal{V}}: Q_{m_{\mathcal{V}}^{\downarrow\mathcal{U}\cap\mathcal{V}}}(\mathbf{c}^{\downarrow\mathcal{U}\cap\mathcal{V}})>0} (-1)^{|\mathbf{c}|+1} \frac{Q_{m_{\mathcal{U}}}(\mathbf{c}^{\downarrow\mathcal{U}}) \cdot Q_{m_{\mathcal{V}}}(\mathbf{c}^{\downarrow\mathcal{V}})}{Q_{m_{\mathcal{V}}^{\downarrow\mathcal{U}\cap\mathcal{V}}}(\mathbf{c}^{\downarrow\mathcal{U}\cap\mathcal{V}})}.$$

If $L = 0$ then $m_{\mathcal{U}}$ and $m_{\mathcal{V}}$ are in total conflict and the composition is undefined.

Remark Definition 2 is taken from [18], where the reader can find the proof that $\rhd_d$ meets all the axioms required from a composition operator, as well as the instructions on how to introduce analogous operators of composition in some other theories of uncertainty meeting the axioms of Shenoy's valuation-based systems. It can be easily shown that the computation of $m_{\mathcal{U}} \rhd_d m_{\mathcal{V}}$ may be simplified by transforming the second argument $m_{\mathcal{V}}$ into $Q_{m_{\mathcal{V}}}$, and computing the corresponding *conditional commonality function* $Q_{m_{\mathcal{V}\setminus\mathcal{U}|\mathcal{V}\cap\mathcal{U}}} = Q_{m_{\mathcal{V}}} \ominus Q_{m_{\mathcal{V}}^{\downarrow\mathcal{U}\cap\mathcal{V}}}$. This may be done using Eq. (6), transformed into the corresponding *conditional BPA* $m_{\mathcal{V}\setminus\mathcal{U}|\mathcal{V}\cap\mathcal{U}}$. Then, since

$$(Q_{m_{\mathcal{U}}} \rhd_d Q_{m_{\mathcal{V}}}) = \frac{1}{L} Q_{m_{\mathcal{U}}} \oplus Q_{m_{\mathcal{V}}} \ominus Q_{m_{\mathcal{V}}^{\downarrow\mathcal{U}\cap\mathcal{V}}}$$

we know ([40]) that

$$m_{\mathcal{U}} \rhd_d m_{\mathcal{V}} = m_{\mathcal{U}} \oplus m_{\mathcal{V}\setminus\mathcal{U}|\mathcal{V}\cap\mathcal{U}}.$$

Thus, the computations of d-composition are limited by the dimensionality of the second BPA because we do not know how to avoid transforming the second BPA into the corresponding commonality function. Notice that, as a rule, the representation of the corresponding commonality function requires the space for close to $2^{|\Omega_{\mathcal{V}}|} \geqslant 2^{(2^{|\mathcal{V}|})}$ values regardless of the number of focal elements of $m_{\mathcal{V}}$.

A disadvantage of the removal operator is that, as illustrated in the following example, the result of the d-composition of two basic assignments may be a pseudo-BPA (some focal elements may be assigned negative values), which may cause problems with the interpretation as well as when computing some entropies introduced in Section 6 [31, 35, 42].

Example of a pseudo-BPA Consider the case of composing two BPAs: m_1 defined for variable X, and m_2 defined for two variables X, Y. Consider $\Omega_X = \{x, \bar{x}\}, \Omega_Y = \{y, \bar{y}\}$, which means that $|2^{\Omega_X}| = 4$, and $|2^{\Omega_{X,Y}}| = 16$. Therefore, m_1 and m_2 may be defined by up to 3 and 15 values assigned to focal elements, respectively. In this example, each of the two considered basic assignments have only two focal elements – see Table 2. In tables, we depict only focal elements, or more precisely, if $\mathbf{a} \subseteq \Omega$ is not included in the table, then all its respective values equal 0.

The process of computing $m_1 \rhd_d m_2$ through $Q_{m_1} \rhd_d Q_{m_2}$ using Eq. (16) is shown in Table 3, and the idea described in the Remark is shown in Table 4.

In the framework of belief functions, another binary operator meeting the definition of the composition operator was defined in [21].

Table 2: m_1 and m_2

$\mathbf{a}$	$m_1(\mathbf{a})$
$\{(x)\}$	0.2
$\{(x),(\bar{x})\}$	0.8

$\mathbf{a}$	$m_2(\mathbf{a})$
$\{(x,y)\}$	0.9
$\{(x,y),(x,\bar{y}),(\bar{x},\bar{y})\}$	0.1

Table 3: Computation of $(m_1 \rhd_d m_2)(\mathbf{a})$ using Eq. (2)

$\mathbf{a}$	$m_1(\mathbf{a}^{\downarrow X})$	$Q_{m_1}(\mathbf{a}^{\downarrow X})$	$m_2(\mathbf{a})$	$Q_{m_2}(\mathbf{a})$	$m_2^{\downarrow X}(\mathbf{a}^{\downarrow X})$	$Q_{m_2}^{\downarrow X}(\mathbf{a}^{\downarrow X})$	$\frac{Q_{m_1}(\mathbf{a}^{\downarrow X})\cdot Q_{m_2}(\mathbf{a})}{Q_{m_2}^{\downarrow X}(\mathbf{a}^{\downarrow X})}$	$(m_1 \rhd_d m_2)(\mathbf{a})$
$\{(x,y)\}$	0.2	1	0.9	1	0.9	1	1	0.9
$\{(x,\bar{y})\}$	0.2	1		0.1	0.9	1	0.1	
$\{(\bar{x},\bar{y})\}$		0.8		0.1		0.1	0.8	
$\{(x,y),(x,\bar{y})\}$	0.2	1		0.1	0.9	1	0.1	-0.7
$\{(x,y),(\bar{x},\bar{y})\}$	0.8	0.8		0.1	0.1	0.1	0.8	
$\{(x,\bar{y}),(\bar{x},\bar{y})\}$	0.8	0.8		0.1	0.1	0.1	0.8	
$\Omega_{X,Y}\backslash\{(\bar{x},y)\}$	0.8	0.8	0.1	0.1	0.1	0.1	0.8	0.8

Definition 3. *Consider two BPAs $m_{\mathcal{U}}$, $m_{\mathcal{V}}$. Their f-composition is a BPA $m_{\mathcal{U}} \rhd_f m_{\mathcal{V}}$ defined for each nonempty $\mathbf{c} \subseteq \Omega_{\mathcal{U}\cup\mathcal{V}}$ by one of the following expressions:*

(i) *if $m_{\mathcal{V}}^{\downarrow\mathcal{U}\cap\mathcal{V}}(\mathbf{c}^{\downarrow\mathcal{U}\cap\mathcal{V}}) > 0$ and $\mathbf{c} = \mathbf{c}^{\downarrow\mathcal{U}} \bowtie \mathbf{c}^{\downarrow\mathcal{V}}$, then $(m_{\mathcal{U}} \rhd_f m_{\mathcal{V}})(\mathbf{c}) = \dfrac{m_{\mathcal{U}}(\mathbf{c}^{\downarrow\mathcal{U}}) \cdot m_{\mathcal{V}}(\mathbf{c}^{\downarrow\mathcal{V}})}{m_{\mathcal{V}}^{\downarrow\mathcal{U}\cap\mathcal{V}}(\mathbf{c}^{\downarrow\mathcal{U}\cap\mathcal{V}})}$;*

(ii) *if $m_{\mathcal{V}}^{\downarrow\mathcal{U}\cap\mathcal{V}}(\mathbf{c}^{\downarrow\mathcal{U}\cap\mathcal{V}}) = 0$ and $\mathbf{c} = \mathbf{c}^{\downarrow\mathcal{U}} \times \Omega_{\mathcal{V}\backslash\mathcal{U}}$, then $(m_{\mathcal{U}} \rhd_f m_{\mathcal{V}})(\mathbf{c}) = m_{\mathcal{U}}(\mathbf{c}^{\downarrow\mathcal{U}})$;*

(iii) *in all other cases, $(m_{\mathcal{U}} \rhd_f m_{\mathcal{V}})(\mathbf{c}) = 0$.*

We have specified two composition operators. Which of these should be used to construct a joint? The two composition operators differ in their theoretical properties and computational complexity. However, they yield the same results for a class of belief function models.

Since $m_{\mathcal{U}} \rhd_f m_{\mathcal{V}}$ is always a BPA (i.e., all the focal elements of this composition are assigned positive masses), we are interested only in situations when also $m_{\mathcal{U}} \rhd_d m_{\mathcal{V}}$ is a BPA. Formally, using the corresponding basic assignments, Eq. (16) can be expressed as

$$m_{\mathcal{U}} \rhd_d m_{\mathcal{V}} = m_{\mathcal{U}} \oplus m_{\mathcal{V}} \ominus m_{\mathcal{V}}^{\downarrow\mathcal{U}\cap\mathcal{V}},$$

where $\ominus$ is an inverse operator to $\oplus$. To be sure that this composition is a BPA, we deal only with situations when $(m_{\mathcal{V}} \ominus m_{\mathcal{V}}^{\downarrow\mathcal{U}\cap\mathcal{V}})$ is nonnegative. To characterize it, we prove the following simple assertion (a similar result is stated in [42] in the context of valuation-based systems).

Proposition 3. *Consider nonempty sets of variables $\mathcal{U} \subsetneq \mathcal{V}$ and BPA $m_{\mathcal{V}}$. If $(m_{\mathcal{V}} \ominus m_{\mathcal{V}}^{\downarrow\mathcal{U}\cap\mathcal{V}})$ is a BPA, then the following two properties hold:*

- $m_{\mathcal{V}} = m_{\mathcal{V}}^{\downarrow\mathcal{U}} \oplus (m_{\mathcal{V}} \ominus m_{\mathcal{V}}^{\downarrow\mathcal{U}\cap\mathcal{V}})$;
- $\left(m_{\mathcal{V}} \ominus m_{\mathcal{V}}^{\downarrow\mathcal{U}\cap\mathcal{V}}\right)^{\downarrow\mathcal{U}}$ *is vacuous.*

Proof. The first property is a direct implication of the associativity and commutativity of Dempster's rule of combination, and the latter follows immediately from the local computation property. □

To simplify the notation, and to make it a bit more lucid, let $m_{\mathcal{V}|\mathcal{U}} = m_{\mathcal{V}} \ominus m_{\mathcal{V}}^{\downarrow\mathcal{U}\cap\mathcal{V}}$. Moreover, in connection with Definition 2, we will identify situations when BPA $m_{\mathcal{V}|\mathcal{U}\cap\mathcal{V}}$ exists and is, in a way, "adapted" to $m_{\mathcal{U}}$. We will say that

Table 4: An alternative way of the computation of $(m_1 \rhd_d m_2)(\mathbf{a})$.

$\mathbf{a}$	$m_2(\mathbf{a})$	$Q_{m_2}(\mathbf{a})$	$m_2^{\downarrow X}(\mathbf{a}^{\downarrow X})$	$Q_{m_2}^{\downarrow X}(\mathbf{a}^{\downarrow X})$	$Q_{m_{2;(Y\|X)}}(\mathbf{a}) = \frac{Q_{m_2}(\mathbf{a})}{Q_{m_2}^{\downarrow X}(\mathbf{a}^{\downarrow X})}$	$m_{2;(Y\|X)}(\mathbf{a})$	$(m_1 \oplus m_2^{(Y\|X)})(\mathbf{a})$
$\{(x,y)\}$	0.9	1	0.9	1	1	0.9	0.9
$\{(x,\bar{y})\}$		0.1	0.9	1	0.1		
$\{(\bar{x},\bar{y})\}$		0.1		0.1	1		
$\{(x,y),(x,\bar{y})\}$		0.1	0.9	1	0.1	-0.9	-0.7
$\{(x,y),(\bar{x},\bar{y})\}$		0.1	0.1	0.1	1		
$\{(x,\bar{y}),(\bar{x},\bar{y})\}$		0.1	0.1	0.1	1		
$\Omega_{X,Y}\backslash\{(\bar{x},y)\}$	0.1	0.1	0.1	0.1	1	1	0.8

$m_{\mathcal{V}|\mathcal{U}\cap\mathcal{V}}$ is *tight* with respect to $m_{\mathcal{U}}$ if for all couples of focal elements $\mathbf{a}$ and $\mathbf{b}$ ($\mathbf{a}$ is a focal element of $m_{\mathcal{U}}$, and $\mathbf{b}$ is a focal element of $m_{\mathcal{V}|\mathcal{U}\cap\mathcal{V}}$) the following condition holds:

$$\text{for } \forall\, b \in \mathbf{b},\ \exists\, a \in \mathbf{a},\ \text{such that } \{a\} \bowtie \{b\} \neq \varnothing. \tag{17}$$

Proposition 4. *Let two basic assignments $m_{\mathcal{U}}, m_{\mathcal{V}}$ be such that $m_{\mathcal{V}|\mathcal{V}\cap\mathcal{U}}$ is nonnegative BPA. $m_{\mathcal{V}|\mathcal{V}\cap\mathcal{U}}$ is tight with respect to $m_{\mathcal{U}}$ if and only if*

$$m_{\mathcal{U}} \rhd_f m_{\mathcal{V}} = m_{\mathcal{U}} \rhd_d m_{\mathcal{V}}.$$

Proof. First, let us prove that both the operators coincide in situations when $m_{\mathcal{V}|\mathcal{V}\cap\mathcal{U}}$ is tight with respect to $m_{\mathcal{U}}$. For this, recall that for BPA $m_{\mathcal{V}|\mathcal{V}\cap\mathcal{U}}$, the existence of which is assumed,

$$m_{\mathcal{V}} = m_{\mathcal{V}}^{\downarrow\mathcal{V}\cap\mathcal{U}} \oplus m_{\mathcal{V}|\mathcal{V}\cap\mathcal{U}}, \tag{18}$$

and that the d-composition is defined

$$m_{\mathcal{U}} \rhd_d m_{\mathcal{V}} = m_{\mathcal{U}} \oplus m_{\mathcal{V}|\mathcal{V}\cap\mathcal{U}}.$$

What are the focal elements of $m_{\mathcal{U}} \oplus m_{\mathcal{V}|\mathcal{V}\cap\mathcal{U}}$? Let $\mathbf{a}$ and $\mathbf{b}$ be arbitrary focal elements of $m_{\mathcal{U}}$ and $m_{\mathcal{V}|\mathcal{V}\cap\mathcal{U}}$, respectively. Due to Proposition 3, $\left(m_{\mathcal{V}|\mathcal{V}\cap\mathcal{U}}\right)^{\downarrow\mathcal{V}\cap\mathcal{U}}$ is vacuous, $\mathbf{b}^{\downarrow\mathcal{V}\cap\mathcal{U}} = \Omega_{\mathcal{V}\cap\mathcal{U}}$, and $\mathbf{c} = \mathbf{a} \bowtie \mathbf{b} \neq \varnothing$ is a focal element of $m_{\mathcal{U}} \oplus m_{\mathcal{V}|\mathcal{V}\cap\mathcal{U}}$. Therefore, when computing the Dempster's rule of combination $m_{\mathcal{U}} \oplus m_{\mathcal{V}|\mathcal{V}\cap\mathcal{U}}$, the corresponding coefficient of conflict (see Eq. (8))

$$(1-K) = \sum_{\mathbf{a}\subseteq\Omega_{\mathcal{U}},\mathbf{b}\subseteq\Omega_{\mathcal{V}}:\mathbf{a}\bowtie\mathbf{b}=\varnothing} m_{\mathcal{U}}(\mathbf{a}) \cdot m_{\mathcal{V}\backslash\mathcal{U}|\mathcal{V}\cap\mathcal{U}}(\mathbf{b}) = 0, \tag{19}$$

which will also be used in the second part of the proof.

The question is whether for a focal element $\mathbf{c}$ of $m_{\mathcal{U}} \oplus m_{\mathcal{V}|\mathcal{V}\cap\mathcal{U}}$ it may happen that $\mathbf{c} = \mathbf{a} \bowtie \mathbf{b}$, and either $\mathbf{a} \neq \mathbf{c}^{\downarrow\mathcal{U}}$, or $\mathbf{b} \neq \mathbf{c}^{\downarrow\mathcal{V}}$. Since $\mathbf{b}^{\downarrow\mathcal{V}\cap\mathcal{U}} = \Omega_{\mathcal{V}\cap\mathcal{U}}$, for $\forall\, a \in \mathbf{a}$, $\exists\, b \in \mathbf{b}$, $\{a\} \bowtie \{b\}$ is a singleton from $\mathbf{c}^{\downarrow\mathcal{U}} \bowtie \mathbf{c}^{\downarrow\mathcal{V}}$ and therefore $\mathbf{a} \subseteq \mathbf{c}^{\downarrow\mathcal{U}}$. Similarly, the assumption that $m_{\mathcal{V}|\mathcal{V}\cap\mathcal{U}}$ is tight with respect to $m_{\mathcal{U}}$ guarantees that $\mathbf{b} \subseteq \mathbf{c}^{\downarrow\mathcal{V}}$. For all $c \in \mathbf{a} \bowtie \mathbf{b}$, $c^{\downarrow\mathcal{U}} \in \mathbf{a}$ from the definition of a join, and therefore $\mathbf{a} \supseteq \mathbf{c}^{\downarrow\mathcal{U}}$. Analogously, $c^{\downarrow\mathcal{V}} \in \mathbf{b}$ yields $\mathbf{b} \supseteq \mathbf{c}^{\downarrow\mathcal{V}}$. So, we have proven that each focal element $\mathbf{c}$ of $m_{\mathcal{U}} \oplus m_{\mathcal{V}|\mathcal{V}\cap\mathcal{U}}$ is created by a single pair of focal elements $\mathbf{c}^{\downarrow\mathcal{U}}$ of $m_{\mathcal{U}}$ and $\mathbf{c}^{\downarrow\mathcal{V}}$ of $m_{\mathcal{V}|\mathcal{V}\cap\mathcal{U}}$. Therefore (using definition from Eq. (7) and Eq. (19)),

$$(m_{\mathcal{U}} \oplus m_{\mathcal{V}|\mathcal{V}\cap\mathcal{U}})(\mathbf{c}) = \sum_{\mathbf{a}\subseteq\Omega_{\mathcal{U}},\mathbf{b}\subseteq\Omega_{\mathcal{V}}:\mathbf{a}\bowtie\mathbf{b}=\mathbf{c}} m_{\mathcal{U}}(\mathbf{a}) \cdot m_{\mathcal{V}|\mathcal{V}\cap\mathcal{U}}(\mathbf{b}) = m_{\mathcal{U}}(\mathbf{c}^{\downarrow\mathcal{U}}) \cdot m_{\mathcal{V}|\mathcal{V}\cap\mathcal{U}}(\mathbf{c}^{\downarrow\mathcal{V}}). \tag{20}$$

In the same way, we get from Eq. (18) also

$$m_{\mathcal{V}}(\mathbf{c}^{\downarrow\mathcal{V}}) = (m_{\mathcal{V}}^{\downarrow\mathcal{V}\cap\mathcal{U}} \oplus m_{\mathcal{V}|\mathcal{V}\cap\mathcal{U}})(\mathbf{c}^{\downarrow\mathcal{V}}) = m_{\mathcal{V}}^{\downarrow\mathcal{V}\cap\mathcal{U}}(\mathbf{c}^{\downarrow\mathcal{V}\cap\mathcal{U}}) \cdot m_{\mathcal{V}|\mathcal{V}\cap\mathcal{U}}(\mathbf{c}^{\downarrow\mathcal{V}}), \tag{21}$$

which gives that, under the given assumptions,

$$m_{\mathcal{V}|\mathcal{V}\cap\mathcal{U}}(\mathbf{c}^{\downarrow\mathcal{V}}) = \frac{m_{\mathcal{V}}(\mathbf{c}^{\downarrow\mathcal{V}})}{m_{\mathcal{V}}^{\downarrow\mathcal{V}\cap\mathcal{U}}(\mathbf{c}^{\downarrow\mathcal{V}\cap\mathcal{U}})}. \tag{22}$$

Substituting Eq. (22) into Eq. (20), we get exactly the formula from case **(i)** of Definition 3. The fact that case **(ii)** of this definition never creates a focal element of $m_{\mathcal{U}} \oplus m_{\mathcal{V}|\mathcal{V}\cap\mathcal{U}}$ follows from the fact that each couple of focal elements **a** and **b** (**a** is a focal element of $m_{\mathcal{U}}$, and **b** is a focal element of $m_{\mathcal{V}|\mathcal{U}\cap\mathcal{V}}$) gives rise of a focal element $\mathbf{a} \bowtie \mathbf{b}$ of $m_{\mathcal{U}} \oplus m_{\mathcal{V}|\mathcal{V}\cap\mathcal{U}}$. Thus, whenever case **(ii)** of Definition 3 is used (under the assumptions of this assertion), then it assigns zero.

Now, let us prove the other side of the equivalence by contradiction. Assume that $m_{\mathcal{U}} \rhd_f m_{\mathcal{V}} = m_{\mathcal{U}} \rhd_d m_{\mathcal{V}}$, and simultaneously $m_{\mathcal{V}|\mathcal{V}\cap\mathcal{U}}$ is not tight with respect to $m_{\mathcal{U}}$. It means that for some focal elements **a** and **b** (such that **a** is a focal element of $m_{\mathcal{U}}$, and **b** is a focal element of $m_{\mathcal{V}|\mathcal{U}\cap\mathcal{V}}$), and some $b \in \mathbf{b}$, $\{a\} \bowtie \{b\} = \varnothing$ for all $a \in \mathbf{a}$. It means that $b \notin (\mathbf{a} \bowtie \mathbf{b})^{\downarrow\mathcal{V}}$. Thus, we have showed that there is at least one $\mathbf{c} \subseteq \Omega_{\mathcal{U}\cup\mathcal{V}}$ (namely, $\mathbf{c} = \mathbf{a} \bowtie \mathbf{b}$), for which:

(i) $\mathbf{c}$ is a focal element of $m_{\mathcal{U}} \rhd_d m_{\mathcal{V}}$,
(ii) $\mathbf{c} = \mathbf{a} \bowtie \mathbf{b}$ for focal elements **a** and **b** such that $\mathbf{b} \supsetneq \mathbf{c}^{\downarrow\mathcal{V}}$,
(iii) $\mathbf{c} = \mathbf{a} \bowtie \mathbf{c}^{\downarrow\mathcal{V}}$.

Since we assume that $m_{\mathcal{U}} \rhd_f m_{\mathcal{V}} = m_{\mathcal{U}} \rhd_d m_{\mathcal{V}}$, $\mathbf{c}$ is also a focal element of $m_{\mathcal{U}} \rhd_f m_{\mathcal{V}}$, and $(m_{\mathcal{U}} \rhd_f m_{\mathcal{V}})(\mathbf{c}) = (m_{\mathcal{U}} \rhd_d m_{\mathcal{V}})(\mathbf{c})$. Recall, that in the first part of the proof we showed that the corresponding coefficient of conflict (Eq. (19)) equals zero, and therefore $(m_{\mathcal{U}} \rhd_d m_{\mathcal{V}})(\mathbf{c}) \geqslant m_{\mathcal{U}}(\mathbf{c}^{\downarrow\mathcal{U}}) \cdot m_{\mathcal{V}}(\mathbf{c}^{\downarrow\mathcal{V}}) + m_{\mathcal{U}}(\mathbf{c}^{\downarrow\mathcal{U}}) \cdot m_{\mathcal{V}}(\mathbf{b})$. Since $m_{\mathcal{V}}^{\downarrow\mathcal{U}\cap\mathcal{V}}$ is vacuous, $m_{\mathcal{V}}^{\downarrow\mathcal{U}\cap\mathcal{V}}(\mathbf{c}^{\downarrow\mathcal{U}\cap\mathcal{V}}) = 1$, and one can see from Definition 3 that $(m_{\mathcal{U}} \rhd_f m_{\mathcal{V}})(\mathbf{c}) > m_{\mathcal{U}}(\mathbf{c}^{\downarrow\mathcal{U}}) \cdot m_{\mathcal{V}}(\mathbf{c}^{\downarrow\mathcal{V}})$ only when $m_{\mathcal{V}}^{\downarrow\mathcal{U}\cap\mathcal{V}}(\mathbf{c}^{\downarrow\mathcal{U}\cap\mathcal{V}}) = 0$, and $\mathbf{c}^{\downarrow\mathcal{V}\setminus\mathcal{U}} = \Omega_{\mathcal{V}\setminus\mathcal{U}}$, which contradicts Property (ii) showed above. □

Corollary Let two basic assignments $m_{\mathcal{U}}, m_{\mathcal{V}}$ are such that $m_{\mathcal{V}|\mathcal{V}\cap\mathcal{U}}$ exists. If $m_{\mathcal{V}}^{\downarrow\mathcal{V}\cap\mathcal{U}}$ is vacuous, or, if $\mathcal{V} \cap \mathcal{U} = \varnothing$, then

$$m_{\mathcal{U}} \rhd_f m_{\mathcal{V}} = m_{\mathcal{U}} \rhd_d m_{\mathcal{V}}.$$

5. Compositional Models

By a *belief function compositional model* we understand a BPA $m_1 \rhd \cdots \rhd m_n$ obtained by a multiple application of the composition operator. Thus, we can speak about d-compositional or f-compositional models in correspondence, which of the two introduced operators of compositions are used[6]. Let us emphasize that if not specified otherwise by parentheses, the operators are always performed from left to right, i.e.,

$$m_1 \rhd m_2 \rhd m_3 \rhd \ldots \rhd m_n = (\ldots((m_1 \rhd m_2) \rhd m_3) \rhd \ldots \rhd m_{n-1}) \rhd m_n.$$

Consider a (finite) system $\mathbb{W}$ of small subsets of the considered variables $\mathcal{W}$. The vague assumption that $\mathcal{U} \in \mathbb{W}$ is small is accepted to avoid the computational problems connected with computations with the corresponding basic assignments. Thus, we assume that for each $\mathcal{U} \in \mathbb{W}$ we have (or we can easily get) a BPA $m_{\mathcal{U}}$. Moreover, we assume that these basic assignments, as well as the corresponding commonality functions $Q_{m_{\mathcal{U}}}$, can effectively be represented in computer memory. Thus, in context with the operator of d-composition, the cardinality of $\mathcal{U} \in \mathbb{W}$ is always less than 5.

Having a system of low-dimensional basic assignments $\{m_{\mathcal{U}}\}_{\mathcal{U}\in\mathbb{W}}$ we follow the idea of Perez described in Section 2. We assume that there exists a BPA m having all these $m_{\mathcal{U}}$ for its marginals, and we want to find the best approximation of m assembled from the given marginals. In other words, we are looking for a sequence of sets $(\mathcal{U}_i)_{i=1,\ldots,n}$ from $\mathbb{W}$ such that the compositional model $m_{\mathcal{U}_1} \rhd m_{\mathcal{U}_2} \rhd \cdots \rhd m_{\mathcal{U}_n}$ approximates the unknown BPA m best. To simplify notation, we denote $m_i = m_{\mathcal{U}_i}$. Therefore we will speak about a model $m_1 \rhd m_2 \rhd \ldots \rhd m_n$, in which BPA m_i is defined for variables $\mathcal{U}_i$, and the corresponding commonality function is Q_i.

The considered compositional model is a $|\mathcal{U}_1 \cup \ldots \cup \mathcal{U}_n|$-dimensional BPA. It is said to be *perfect* if all m_i are marginals of $m_1 \rhd m_2 \rhd \ldots \rhd m_n$. Thus, perfect models reflect all the information represented by the low-dimensional basic assignments from which they are composed. So, it is not surprising that the optimal approximation will be, as a rule, a perfect model.

If a model is not perfect, it can always be *perfectized* using the following assertion (proved in [18]).

[6] We never consider a possibility that both operators are simultaneously used in one model.

Proposition 5 (perfectization procedure). *For any compositional model $m_1 \rhd m_2 \rhd \ldots \rhd m_n$, the model $\bar{m}_1 \rhd \bar{m}_2 \rhd \ldots \rhd \bar{m}_n$ defined*

$$\begin{aligned}
\bar{m}_1 &= m_1, \\
\bar{m}_2 &= \bar{m}_1^{\downarrow \mathcal{U}_2 \cap \mathcal{U}_1} \rhd m_2, \\
&\vdots \\
\bar{m}_n &= \bar{m}_n^{\downarrow \mathcal{U}_n \cap (\mathcal{U}_1 \cup \ldots \cup \mathcal{U}_{n-1})} \rhd m_n,
\end{aligned}$$

is perfect, and $m_1 \rhd m_2 \rhd \ldots \rhd m_n = \bar{m}_1 \rhd \bar{m}_2 \rhd \ldots \rhd \bar{m}_n$.

Let us highlight that the assertion holds for both the operators introduced in the previous section. The procedure applies to any compositional model; nevertheless, its computational efficiency depends not only on the considered composition operator but also on a *structure* of the model, which is determined by the sequence of sets $\mathcal{U}_1, \mathcal{U}_2, \ldots, \mathcal{U}_n$. In this context, the most advantageous models are *decomposable models*, for which the sequence $\mathcal{U}_1, \mathcal{U}_2, \ldots, \mathcal{U}_n$ meets the so-called *running intersection property* (RIP): $\forall i = 2, \ldots, n \;\; \exists j \; (1 \leqslant j < i) : \mathcal{U}_i \cap (\mathcal{U}_1 \cup \ldots \cup \mathcal{U}_{i-1}) \subseteq \mathcal{U}_j$.

When computing with perfect models, we can take advantage that several permutations of low-dimensional basic assignments can equivalently represent a perfect model. In [18], the following assertions are proved.

Proposition 6 (testing perfectness of models). *Model $m_1 \rhd \ldots \rhd m_n$ is perfect if and only if for all $i = 2, 3, \ldots, n$, basic assignments $(m_1 \rhd \ldots \rhd m_{i-1})$ and m_i are consistent, i.e., $(m_1 \rhd \ldots \rhd m_{i-1})^{\downarrow \mathcal{U}_i \cap (\mathcal{U}_1 \cup \ldots \cup \mathcal{U}_{i-1})} = m_i^{\downarrow \mathcal{U}_i \cap (\mathcal{U}_1 \cup \ldots \cup \mathcal{U}_{i-1})}$.*

Proposition 7 (on consistent decomposable models). *Consider a decomposable model $m_1 \rhd m_2 \rhd \ldots \rhd m_n$. The model is perfect if and only if basic assignments $m_1, m_2, \ldots, m_n$ are pairwise consistent, i.e., $\forall \{i, j\} \subset \{1, 2, \ldots, n\}$, $m_i^{\downarrow \mathcal{U}_i \cap \mathcal{U}_j} = m_j^{\downarrow \mathcal{U}_i \cap \mathcal{U}_j}$.*

The verification of the perfectness of a compositional model using Proposition 6 may be computationally expensive, and therefore Proposition 7 is often used. Checking the pairwise consistency of low-dimensional basic assignments from a given sequence is computationally simple, and therefore we will rely on Proposition 7 in Section 8.

6. Entropy of Belief Functions

As explained in Section 2, to save the space necessary to represent multidimensional probability distributions, one can approximate them by probabilistic compositional models. The same idea also holds for belief functions. The economic representation of a joint BPA also often reduces the extremely high computational complexity of the necessary procedures. However, the problem arises with determining the quality of such approximations. No generally accepted similarity measure analogous to the probabilistic Kullback-Leibler divergence exists. We do not know how to recognize which of the two approximations is better than the other. We do not know how to identify an optimal approximation if the perfect one does not exist. Therefore, we set a goal to test whether some of the entropies designed in the literature can be used to detect the optimal approximations. If such entropy is found, it will be subject to a detailed future analysis.

Though we are aware of other recently introduced entropies, for the described pilot study, we considered the battery of entropy functions listed in Table 5 (primarily those described in [19]). Realize that not all of them were called *entropy* by their authors. Some authors suggest their functions to measure other characteristics of belief functions, such as a *measure of conflict* [28, 27]. Therefore, we should not be surprised that some entropy functions failed. However, having developed a software system for this purpose, we were not limited by the number of the considered entropies. The only criterion for including an entropy-like function in the experiments was its computational complexity.

In the experiments, we computed entropies for 20-dimensional models. For this, we had to keep the number of focal elements of the considered models limited (see Figure 1). Nevertheless, theoretically, the maximum number of focal elements is super-exponential with the number of variables. It is why we are not able to calculate entropies based on the conversion of BPA to Belief function (Höhle), based on the conversion to Commonality function (Smets), or the conversion to Plausibility function (Yager, Lamata & Moral). All these alternative representations are positive for

Table 5: A list of entropies of belief functions

Höhle [13]	$H_o(m_\mathcal{V}) = \sum_{\mathbf{a}\subseteq\Omega_\mathcal{V}} m_\mathcal{V}(\mathbf{a}) \log(\frac{1}{Bel_{m_\mathcal{V}}(\mathbf{a})})$
Smets [45]	$H_t(m_\mathcal{V}) = \sum_{\mathbf{a}\subseteq\Omega_\mathcal{V}} \log(\frac{1}{Q_{m_\mathcal{V}}(\mathbf{a})})$
Yager [50]	$H_y(m_\mathcal{V}) = \sum_{\mathbf{a}\subseteq\Omega_\mathcal{V}} m_\mathcal{V}(\mathbf{a}) \log(\frac{1}{Pl_{m_\mathcal{V}}(\mathbf{a})})$
Nguyen [37]	$H_n(m_\mathcal{V}) = \sum_{\mathbf{a}\subseteq\Omega_\mathcal{V}} m_\mathcal{V}(\mathbf{a}) \log(\frac{1}{m_\mathcal{V}(\mathbf{a})})$
Dubois & Prade [10]	$H_d(m_\mathcal{V}) = \sum_{\mathbf{a}\subseteq\Omega_\mathcal{V}} m_\mathcal{V}(\mathbf{a}) \log(\lvert\mathbf{a}\rvert)$
Lamata & Moral [34]	$H_l(m_\mathcal{V}) = H_y(m_\mathcal{V}) + H_d(m_\mathcal{V})$
Klir & Ramer [30]	$H_k(m_\mathcal{V}) = H_d(m_\mathcal{V}) - \sum_{\mathbf{a}\subseteq\Omega_\mathcal{V}} m_\mathcal{V}(\mathbf{a}) \log\left(1 - \sum_{\mathbf{b}\subseteq\Omega_\mathcal{V}} m_\mathcal{V}(\mathbf{b}) \frac{\lvert\mathbf{b}\setminus\mathbf{a}\rvert}{\lvert\mathbf{b}\rvert}\right)$
Klir & Parviz [29]	$H_k(m_\mathcal{V}) = H_d(m_\mathcal{V}) - \sum_{\mathbf{a}\subseteq\Omega_\mathcal{V}} m_\mathcal{V}(\mathbf{a}) \log\left(1 - \sum_{\mathbf{b}\subseteq\Omega_\mathcal{V}} m_\mathcal{V}(\mathbf{b}) \frac{\lvert\mathbf{a}\setminus\mathbf{b}\rvert}{\lvert\mathbf{a}\rvert}\right)$
Pal et al. [38]	$H_b(m_\mathcal{V}) = \sum_{\mathbf{a}\subseteq\Omega_\mathcal{V}} m_\mathcal{V}(\mathbf{a}) \log(\frac{\lvert\mathbf{a}\rvert}{m_\mathcal{V}(\mathbf{a})})$
Harmanec & Klir [12]	$H_h(m_\mathcal{V}) = \max\{H(\pi) \mid \pi \in \mathbb{P}(m_\mathcal{V})\}$
Maeda & Ichihashi [36]	$H_i(m_\mathcal{V}) = H_h(m_\mathcal{V}) + H_d(m_\mathcal{V})$
Abellán & Moral [1]	$H_a(m_\mathcal{V}) = H_i(m_\mathcal{V}) + KL(\pi\|\kappa)$ for specific $\pi, \kappa \in \mathbb{P}(m_\mathcal{V})$
Jousselme et al. [26]	$H_j(m_\mathcal{V}) = H(\pi_{m_\mathcal{V}})$
Deng [8]	$H_g(m_\mathcal{V}) = H_n(m_\mathcal{V}) + \sum_{\mathbf{a}\subseteq\Omega_\mathcal{V}} m_\mathcal{V}(\mathbf{a}) \log(2^{\lvert\mathbf{a}\rvert} - 1)$
Jiroušek & Shenoy [19]	$H_\lambda(m_\mathcal{V}) = H(\lambda_{m_\mathcal{V}}) + H_d(m_\mathcal{V})$
Jiroušek et al. [17]	$H_\pi(m_\mathcal{V}) = H(\pi_{m_\mathcal{V}}) + H_d(m_\mathcal{V})$
Decomposable [20]	$H_S(m_\mathcal{V}) = \sum_{\mathbf{a}\subseteq\Omega_\mathcal{V}} (-1)^{\lvert\mathbf{a}\rvert} Q_{m_\mathcal{V}}(\mathbf{a}) \log(Q_{m_\mathcal{V}}(\mathbf{a}))$

a substantial part of all possible states of variables, and their space complexity is, therefore, super-exponential to the number of variables. Similarly, we had to exclude also entropies requiring maximization of Shannon entropy over the credal set, i.e., Harmanec & Klir, Maeda & Ishibashi and Abellán & Moral definitions.

Except for special situations, the number of focal elements of compositional models exponentially increases with the length of models. It holds even when the number of focal elements of BPAs, from which the models are composed is limited. It is visible from Figure 1 depicting the situation of 10 representative compositional models (5 f-compositional, and 5 d-compositional). Each of these models is composed of 13 – 16 low-dimensional basic assignments, and each graph describes how the number of the focal elements grows with the length i of the considered prefix model $m_1 \rhd m_2 \rhd \ldots \rhd m_i$ (for the definition of various model types, see Section 8). Therefore, in the experiments described in Section 8, we considered only the entropies, the computational complexity of which is a linear function of the number of focal elements. This condition is also violated by Klir & Ramer, and Klir & Parviz's definitions. Therefore, we eventually computed only $H_n, H_d, H_b, H_j, H_g, H_\lambda, H_\pi$, and H_S.

The last decomposable entropy H_S is some kind of an exception. It is defined using the commonality function:

$$H_S(m_\mathcal{V}) = \sum_{\mathbf{a}\subseteq\Omega_\mathcal{V}} (-1)^{\lvert\mathbf{a}\rvert} Q_{m_\mathcal{V}}(\mathbf{a}) \log(Q_{m_\mathcal{V}}(\mathbf{a})). \tag{23}$$

Even though $H_S(m_\mathcal{V})$ is not always nonnegative, its merit is that it is the only definition of belief function entropy that satisfies an additivity property in the sense that $H_S(m_X \oplus m_{Y|X}) = H_S(m_X) + H_S(m_{Y|X})$ (here, m_X is a BPA for X, and $m_{Y|X}$ is a conditional BPA for Y given X such that its marginal for X is vacuous). This additivity, which is one of the fundamental properties of probabilistic information theory, makes the computation of the entropy for high-dimensional perfect d-compositional models possible. It is the only entropy that can be computed for general d-compositional models for more than ten variables.

As mentioned above, we restricted the number of focal elements for our experiments so that the joint BPA has a "reasonable" number of focal elements. It makes the computation of other entropies possible directly for the joint

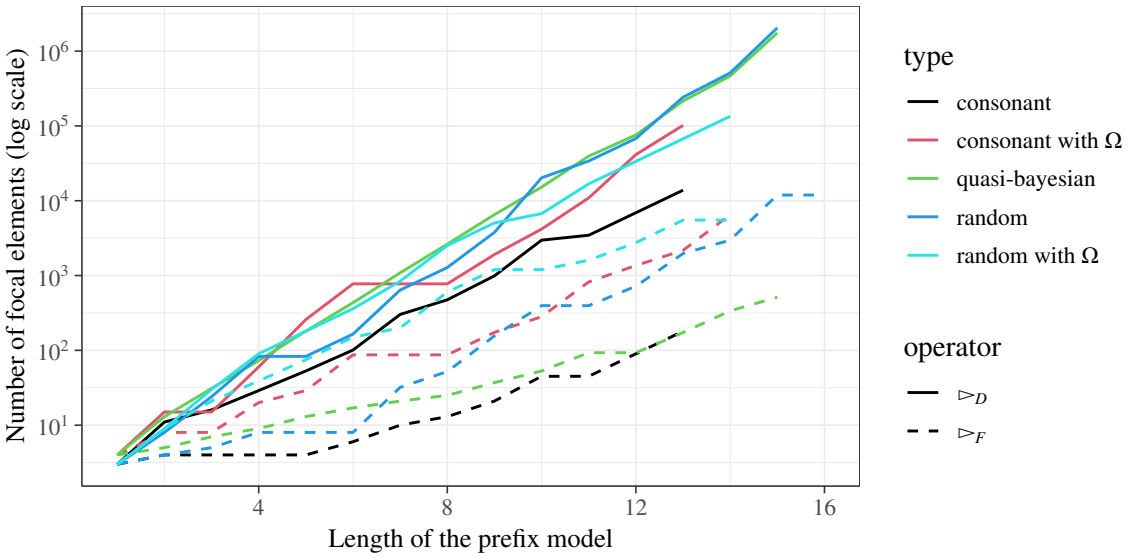

Figure 1: Number of focal elements for compositional models and their prefix sub-models

BPAs. To make the computation of the decomposable entropy possible, another restriction is given for the size of low-dimensional BPAs from which the models are assembled. They should not be defined for more than four variables because, when computing the entropy for the d-composition of two consistent BPAs $m_{\mathcal{U}}$ and $m_{\mathcal{V}}$,

$$H_S(m_{\mathcal{U}} \rhd_d m_{\mathcal{V}}) = H_S(m_{\mathcal{U}}) + H_S(m_{\mathcal{V}\setminus\mathcal{U}|\mathcal{V}\cap\mathcal{U}}), \tag{24}$$

one has to compute the conditional entropy according to the following formula

$$H_S(m_{\mathcal{V}\setminus\mathcal{U}|\mathcal{V}\cap\mathcal{U}}) = \sum_{\mathbf{a}\subseteq\Omega_{\mathcal{V}}} (-1)^{|\mathbf{a}|} Q_{m_{\mathcal{V}}}(\mathbf{a}) \log(Q_{m_{\mathcal{V}\setminus\mathcal{U}|\mathcal{V}\cap\mathcal{U}}}(\mathbf{a})), \tag{25}$$

which requires the transformation of $m_{\mathcal{V}}$ into the commonality function. Having this commonality function, one can compute $Q_{m_{\mathcal{V}\setminus\mathcal{U}|\mathcal{V}\cap\mathcal{U}}}(\mathbf{a}) = Q_{m_{\mathcal{V}}}(\mathbf{a}) / Q_{m_{\mathcal{V}}^{\downarrow\mathcal{V}\cap\mathcal{U}}}(\mathbf{a}^{\downarrow\mathcal{V}\cap\mathcal{U}})$ for all $\mathbf{a} \subseteq \Omega_{\mathcal{V}}$. Note that for $\mathcal{V} \cap \mathcal{U} = \varnothing$, $H_S(m_{\mathcal{V}\setminus\mathcal{U}|\mathcal{V}\cap\mathcal{U}}) = H_S(m_{\mathcal{V}})$.

7. An Example

Consider 20 binary variables $A, B, C, \ldots, T$, and a (randomly generated) perfect decomposable model assembled from 15 low-dimensional BPAs:

$$\begin{aligned}\mathfrak{M}_1 : \quad & m_{\{AGK\}} \rhd m_{\{AKS\}} \rhd m_{\{BDK\}} \rhd m_{\{AQT\}} \rhd m_{\{AP\}} \rhd m_{\{OP\}} \rhd m_{\{GH\}} \rhd m_{\{NT\}} \rhd m_{\{IS\}} \\ & \rhd m_{\{ALQ\}} \rhd m_{\{IFR\}} \rhd m_{\{AGM\}} \rhd m_{\{AEM\}} \rhd m_{\{CQT\}} \rhd m_{\{BJ\}}.\end{aligned}$$

It is an easy task to verify that the model is decomposable. The random generator was set to generate pairwise consistent BPAs. Therefore, due to Proposition 7, the perfectness of $\mathfrak{M}_1$ is guaranteed.

Using other permutations of the generated BPAs, we can set up a lot of other models, mostly non-decomposable. Consider just (randomly selected) two of them:

$$\begin{aligned}\mathfrak{M}_2 : \quad & m_{\{CQT\}} \rhd m_{\{BJ\}} \rhd m_{\{ALQ\}} \rhd m_{\{BDK\}} \rhd m_{\{AEM\}} \rhd m_{\{AQT\}} \rhd m_{\{IS\}} \rhd m_{\{AGK\}} \rhd m_{\{GH\}} \\ & \rhd m_{\{IFR\}} \rhd m_{\{OP\}} \rhd m_{\{AKS\}} \rhd m_{\{AGM\}} \rhd m_{\{NT\}} \rhd m_{\{AP\}}, \\ \mathfrak{M}_3 : \quad & m_{\{AQT\}} \rhd m_{\{AGM\}} \rhd m_{\{GH\}} \rhd m_{\{BDK\}} \rhd m_{\{AKS\}} \rhd m_{\{AEM\}} \rhd m_{\{IFR\}} \rhd m_{\{NT\}} \rhd m_{\{AGK\}} \\ & \rhd m_{\{BJ\}} \rhd m_{\{ALQ\}} \rhd m_{\{OP\}} \rhd m_{\{AP\}} \rhd m_{\{CQT\}} \rhd m_{\{IS\}}.\end{aligned}$$

Though assembled from the same system of low-dimensional BPA, these three models differ from each other. For example, due to Proposition 7, we know that all the considered low-dimensional basic assignments are marginals of model $\mathfrak{M}_1$, which does not hold for $\mathfrak{M}_2$ and $\mathfrak{M}_3$.

Table 6: Entropies of models $\mathfrak{M}_1$, $\mathfrak{M}_2$, and $\mathfrak{M}_3$.

	f-compositional models			d-compositional models		
Entropy	$\mathfrak{M}_1$	$\mathfrak{M}_2$	$\mathfrak{M}_3$	$\mathfrak{M}_1$	$\mathfrak{M}_2$	$\mathfrak{M}_3$
H_s	-	-	-	3.442	3.835	3.642
H_n	12.351	12.405	11.771	-	-	-
H_d	15.297	16.105	15.792	15.612	15.954	14.894
H_b	27.648	28.511	27.563	-	-	-
H_j	18.542	19.017	18.840	18.223	18.247	19.323
H_g	69475	101676	90577	-	-	-
H_λ	34.613	35.620	35.287	34.466	34.789	34.790
H_π	33.840	35.123	34.633	33.835	34.201	34.217

Since the condition of Proposition 4 is rarely satisfied for randomly generated BPAs, these models also differ depending on whether we consider f- or d-compositional models. Thus, we consider six compositional models. In the left-hand side of Table 6, we see all the values of entropies that can be (using our software system) computed for the considered three f-compositional models. As explained above, it is impossible to calculate the decomposable entropy H_S for f-compositional models; therefore, the left side of the respective row is empty. On the other hand, the computation of H_n, H_b, and H_g for d-compositional models failed in this particular example. The reason is that there is at least one focal element with negative mass in d-compositional models $\mathfrak{M}_1$, $\mathfrak{M}_2$, and $\mathfrak{M}_3$ and the respective logarithm is thus undefined. As illustrated by results in Section 8, similar failures may appear for H_j, H_λ, and H_π because even the respective probabilistic transforms may have some negative masses. It does not happen for the three considered models $\mathfrak{M}_1$, $\mathfrak{M}_2$, $\mathfrak{M}_3$. Both the pignistic and plausibility transforms for these models are nonnegative BPAs, and we could compute H_j, H_λ, H_π. Realize that the Dubois-Prade entropy is calculable regardless of whether the considered BPA is nonnegative.

Let us illustrate how the results recorded in Table 6 are interpreted in the process of the entropies assessment described in the next section. Consider, say, f-compositional models. Based on the values of entropies from the left-hand side of Table 6, we see that the Nguyen's entropy

$$H_n(\mathfrak{M}_3) < H_n(\mathfrak{M}_1) < H_n(\mathfrak{M}_2),$$

which is not what we expect from the entropy used to detect the optimal approximations. On the other side, the values of Dubois-Prade entropy minimizes the optimal compositional model

$$H_d(\mathfrak{M}_1) < H_d(\mathfrak{M}_3) < H_d(\mathfrak{M}_2).$$

Thus, f-compositional model $\mathfrak{M}_1$ is in the next section counted among those models that are successfully detected by H_d (and also by H_j, H_g, H_λ, H_π and H_S) . Similarly, one can see from Table 6 that d-compositional model $\mathfrak{M}_1$ is counted among those models that are successfully detected by H_j, H_λ, and H_π.

Even though it does not happen for the models considered in this example, it is not rare that an entropy achieves the same value for the optimal and some non-optimal models. Then, if no non-optimal model has lower entropy than the optimal one, we count the model among those weakly detected by the considered entropy.

8. Experimental Results

Using software developed in R-studio, we randomly generated 357 perfect decomposable compositional models for 20 variables. To generate a decomposable model, we generated a sequence of sets of variables satisfying the running intersection property. Then, we generated random BPAs for these sets of variables and ran the perfectization

procedure described in Proposition 5. Inspired by [32], we designed our random generator of BPAs to be able to produce assignments with a specific structure and a predefined number of focal elements. Specifically, we generated five types of models where all BPAs were as follows:

1. consonant (81 models) – all focal elements are nested;
2. consonant with Ω (68 models) – all focal elements are nested, and the whole space of discernment Ω is among them;
3. quasi-Bayesian (90 models) – Ω is one focal element, all others are singletons;
4. random (68 models) – no restriction was imposed on the structure randomly generated focal elements;
5. random with Ω (50 models) – the only restriction of the structure is that Ω is among the focal elements.

Notice that the given type determines the low-dimensional basic assignments from which the joint is defined. It does not mean that the joint is of the given type. After composition, the type is generally not preserved. It is easy to see that the f-composition of two quasi-Bayesian BPAs is not quasi-Bayesian, whenever case (ii) of Definition 3 finds its use. On the other hand, it is not difficult to show that the d-composition of two quasi-Bayesian BPAs is quasi-Bayesian.

In addition to the type of generated BPAs, the random generator was controlled by two parameters. The first is an upper bound for the number of variables for which BPAs are defined (we used 3, 4, and 5), and the second parameter is the maximum number of focal elements (we used 3,4,5 and 6).

These parameters were mainly used to control the computational complexity of d-compositional models. Note that we are able to calculate $(m_{\mathcal{U}} \rhd_d m_{\mathcal{V}})$ if $|\mathcal{V}| \leqslant 4$ (in case of $|\mathcal{V}| = 5$ the corresponding conditional CF is determined by 2^{32} parameters and its conversion to BPA requires 2^{64} comparisons.)

For each perfect decomposable model, we created several non-decomposable models assembled from the same system of low-dimensional BPAs. We destroyed the running intersection property by randomly permuting the sequence in which the generated BPAs were composed. In this way, similar to the example described in the previous section, we obtained for each randomly generated decomposable model several non-decomposable models (on average, about six). Altogether, we had 2077 different models. Then each model was converted into a joint BPA by applying the composition operator. Note that this was possible because we significantly limited the number of focal elements of generated low- dimensional BPAs. As a result, the number of focal elements of the joint BPA remained reasonable, and it was possible to calculate them. If the number of focal elements of the joint BPA began to grow (which happened a few times), it was impossible to calculate the joint BPA within a reasonable time. Therefore we did not include these models in the results.

To test the behavior of each entropy function, we first considered f-compositional models. For each model we calculated 7 different entropies H_λ, H_d, H_π, H_n, H_b, H_j, and H_g of the joint. Then we calculated the relative difference of the model entropy from the entropy of the corresponding decomposable model.

Recall that an ideal entropy function should always have a smaller value for a decomposable model than for a non-decomposable one assembled from the same marginals. Therefore, the entropy of each non-decomposable model, minus the entropy of the corresponding decomposable model, should be nonnegative, ideally positive. We created histograms of these differences in Figure 2 for each entropy definition. In the case of an ideal entropy, the region corresponding to the negative difference should be empty. Unfortunately, this does not happen for any considered entropy. Nevertheless, if accepting a reasonable error, one can use entropies H_d, H_π, and H_j for detecting the decomposable models as the best ones.

In another way, these results for f-compositional models are also presented in Table 7. The first row repeats the numbers of decomposable models generated for the respective type of low-dimensional BPA. The remaining numbers express the ability of the individual entropy to detect the optimal approximation. The first number (of each couple) reads the percentage of decomposable models (strongly) detected by the respective entropy (entropy value for RIP model is lower than values for all non-RIP models). The second number (in the parentheses) reads the percentage of decomposable models weakly detected by the respective entropy (entropy value for the RIP model is lower or equal than those for all non-RIP models).

From Table 7, one can identify some differences in the behavior of different entropies when applied to different types of models (though, as mentioned above, it is not a type of a model but the type of low-dimensional BPAs from which the model is composed). For example, while H_λ is not so bad in the case of consonant models, it fails

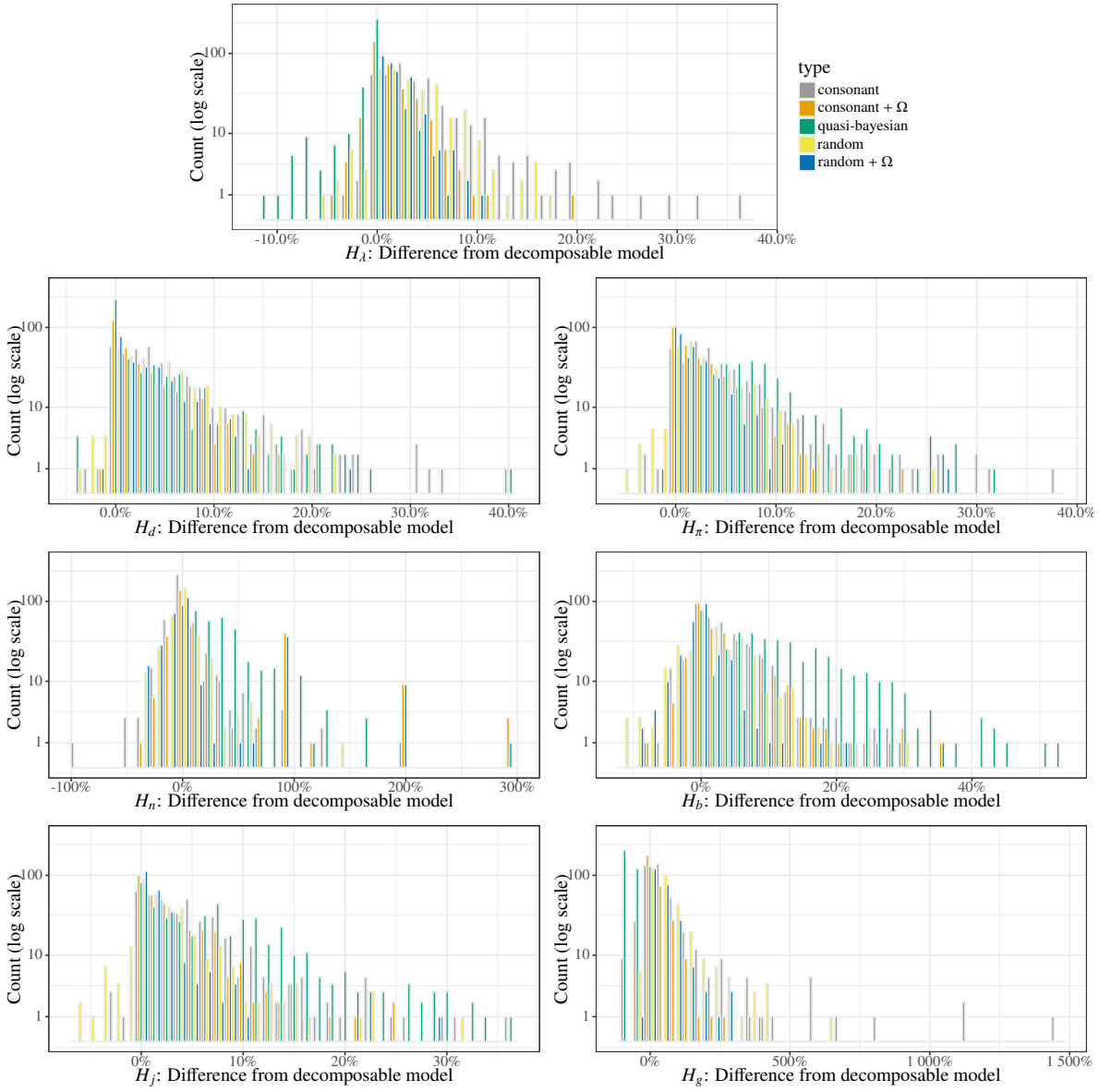

Figure 2: Frequencies of differences of entropies of non-decomposable models and the corresponding decomposable models. F-compositional models.

for quasi-Bayesian models. Also, the significant differences between the numbers of strongly and weakly detected decomposable models deserve further analysis – see, for example, H_d in the case of quasi-Bayesian models.

We do not have enough data to perform an earnest statistical analysis. Nevertheless, the best entropies may seem to be for this purpose H_j and H_π. Notice that both are based on the Shannon entropy of the pignistic transform. Despite its simplicity, H_d detects the best model relatively successfully in all cases except quasi-Bayesian models. Note that H_λ, which is grounded in Shannon entropy of the plausibility transform, was not so bad generally; however, it fails for quasi-Bayesian models. On the other hand, the Nguyen entropy, which is the Shannon entropy of the BPA values, does not seem to be the proper tool for this purpose.

Table 7: Detection rate (%) for f-compositional models. (In parentheses, percentage of weakly detected decomposable models.)

	consonant	consonant with Ω	quasi Bayesian	random	random with Ω	all types
# of models	81	68	90	68	50	357
H_λ	68 (94)	49 (74)	17 (31)	54 (82)	50 (98)	46 (73)
H_d	69 (98)	50 (99)	24 (96)	60 (91)	50 (100)	50 (96)
H_π	69 (95)	56 (99)	63 (99)	56 (84)	48 (96)	60 (95)
H_n	17 (49)	22 (59)	63 (100)	13 (32)	4 (16)	27 (56)
H_b	51 (72)	34 (72)	63 (99)	31 (51)	6 (18)	41 (67)
H_j	68 (94)	54 (94)	63 (99)	49 (76)	44 (86)	57 (91)
H_g	65 (93)	40 (56)	0 (4)	57 (84)	50 (94)	40 (62)

The results achieved for d-compositional models are summarized in Table 8, which is organized similarly as Table 7. Nevertheless, the issue is that the computations with d-compositional models are much more time- and space-demanding than with f-compositional models. To get the selected entropies, one must find all focal elements and compute assigned probability masses. Usually, there are too many of them (much more than those for f-compositional models). Our software could do it only in (about) one-fifth of all generated decomposable models.

Moreover, since some focal elements of d-compositional models are assigned negative values, not all the entropies could be computed. From the first row of the table, one can see the number of models for which we succeeded in computing all three entropies H_λ, H_d, H_S. These entropies could be computed even when some focal elements are assigned negative values. Since the numbers of analyzed models are too low, we do not dare to make any conclusions from the contents of Table 8. We take it mainly as a clue for future research.

Table 8: Detection rate (%) for d-compositional models. (In parentheses, percentage of weakly detected decomposable models.)

	consonant	consonant with Ω	quasi Bayesian	random	random with Ω	all types
# of models	20	12	16	12	9	69
H_λ	35 (75)	25 (42)	44 (44)	42 (50)	56 (100)	39 (61)
H_d	35 (75)	25 (75)	6 (56)	42 (58)	56 (100)	30 (71)
H_S	5 (75)	8 (83)	19 (88)	8 (75)	11 (78)	10(80)

However, the high number of d-compositional models for which we could not compute their entropy deserves our attention. It refutes our original conception that d-compositional models, based on the idea of Dempster's rule of combination, may be exploited regardless of the negative masses assigned to some focal elements. It convinced us that d-compositional models should only be used when defining a regular (i.e., nonnegative) BPA. Due to Proposition 1, it is guaranteed, for example, when the resulting models are equivalent to graphical belief function models. Then, as shown in [24], it happens surprisingly often that the d- and f-compositional models coincide.

The experiments were designed to detect possible differences among the considered entropies. As the reader can see from the respective tables and graphs, none of the tested entropy definitions appeared to be a universal tool recommendable for model learning. On the contrary, the experiments suggest that a further study of narrower classes of models (e.g., those where Proposition 4 applies) will be necessary. For them, finding a suitable definition of entropy may still be possible (regardless of their computational complexity).

9. Summary & Conclusions

The paper contributes to an effort to find a criterion to assess the quality of belief function models. It is based on the following idea: If a criterion is sound for machine learning procedures, it should detect a model containing all given information from models containing only a part of this information. This idea, translated into the terminology of compositional models, means that a sound criterion should distinguish a perfect compositional model from non-perfect models set up from the same system of building blocks, regardless of whether it is a measure of information or inner conflict.

Shannon entropy, being the first entropy introduced in probability theory, has many properties that predetermine its role in machine-learning procedures. In the theory of belief functions, the situation is much more complicated. In this theoretical framework, most authors distinguish characteristics indistinguishable in probability theory. In this paper, we consider mainly a measure of entropy and a measure of inner conflict. This fact manifests in connection with the design of machine learning procedures.

To realize computational experiments verifying which of the entropies described in literature meets the paradigm mentioned above best, one has to cope with the problem of which of two composition operators is to be used for the purpose. The d-composition operator is designed for the Dempster-Shafer theory of belief functions, the semantics of which are clearly understood. Although the f-composition operator has some nice mathematical properties, we do not know which theory of belief function it corresponds to and the semantics of such a theory. The solution to this problem is a topic for future work.

In this paper, oriented towards computational experiments, we had to consider the computational complexity of the necessary procedures. It was the primary criterion for selecting different definitions of entropy of belief functions. Thus, the results are partially devalued because we did not care that some of these definitions are designed specifically for specific theories of belief functions. For example, the decomposable entropy [25, 20] is explicitly designed for the Dempster-Shafer theory, where belief functions constitute a graphical model. Some definitions, such as Maeda & Ichihashi, Jousselme et al., Jirousek et al, make sense for the Fagin-Halpern theory of belief functions and its corresponding semantics of credal sets. The d-composition operator is not appropriate for these definitions of entropy. A composition operator corresponding to the Fagin-Halpern combination rule is yet to be defined. It is a topic for future work.

It is well known that the Dempster-Shafer theory can be used for large graphical models with many variables [43, 44]. It is suggested in [20, 22] that the decomposable entropy can be tractably computed for large belief function graphical models even though the decomposable entropy is defined in terms of the commonality functions. How to use it to assess the models' quality is also a topic for future work.

Acknowledgement

The Czech Science Foundation financially supported this study under Grant No. 19-06569S, and the Ronald G. Harper Professorship at the University of Kansas supported the third author. A part of this paper previously appeared as [16].

References

[1] J. Abellán and S. Moral. Completing a total uncertainty measure in Dempster-Shafer theory. *International Journal of General Systems*, 28(4–5):299–314, 1999.

[2] B. R. Cobb and P. P. Shenoy. On the plausibility transformation method for translating belief function models to probability models. *International Journal of Approximate Reasoning*, 41(3):314–340, 2006.

[3] I. Csiszár. I-divergence geometry of probability distributions and minimization problems. *The Annals of Probability*, pages 146–158, 1975.

[4] F. Cuzzolin. On the relative belief transform. *International Journal of Approximate Reasoning*, 53(5):786–804, 2012.

[5] W. E. Deming and F. F. Stephan. On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *The Annals of Mathematical Statistics*, 11(4):427–444, 1940.

[6] A. P. Dempster. A generalization of Bayesian inference. *Journal of the Royal Statistical Society, Series B*, 30(2):205–247, 1968.

[7] A. P. Dempster. The Dempster-Shafer calculus for statisticians. *International Journal of approximate reasoning*, 48(2):365–377, 2008.

[8] Y. Deng. Deng entropy. *Chaos, Solitons & Fractals*, 91:549–553, 2016.

[9] T. Denoeux. Likelihood-based belief function: justification and some extensions to low-quality data. *International Journal of Approximate Reasoning*, 55(7):1535–1547, 2014.

[10] D. Dubois and H. Prade. Properties of measures of information in evidence and possibility theories. *Fuzzy Sets and Systems*, 24(2):161–182, 1987.

[11] R. Fagin and J. Y. Halpern. A new approach to updating beliefs. In P. Bonissone, M. Henrion, L. Kanal, and J. Lemmer, editors, *Uncertainty in Artificial Intelligence 6*, pages 347–374. North-Holland, 1991.

[12] D. Harmanec and G. J. Klir. Measuring total uncertainty in Dempster-Shafer theory: A novel approach. *International Journal of General Systems*, 22(4):405–419, 1994.

[13] U. Höhle. Entropy with respect to plausibility measures. In *Proceedings of the 12th IEEE Symposium on Multiple-Valued Logic*, pages 167–169, 1982.

[14] R. Jiroušek. Foundations of compositional model theory. *International Journal of General Systems*, 40(6):623–678, 2011.

[15] R. Jiroušek. A short note on decomposition and composition of knowledge. *International Journal of Approximate Reasoning*, 120:24–32, 2020.

[16] R. Jiroušek and V. Kratochvíl. Approximations of belief functions using compositional models. In *European Conference on Symbolic and Quantitative Approaches with Uncertainty*, pages 354–366. Springer, 2021.

[17] R. Jiroušek, V. Kratochvíl, and P. P. Shenoy. Entropy-based learning of compositional models from data. In T. Denœux, E. Lefèvre, Z. Liu, and F. Pichon, editors, *Belief Functions: Theory and Applications, Proceedings of the 6th International Conference, BELIEF 2021,*, volume 12915 of *Lecture Notes in Artificial Intelligence*, pages 117–126. Springer Nature, Switzerland, 2021.

[18] R. Jiroušek and P. P. Shenoy. Compositional models in valuation-based systems. *International Journal of Approximate Reasoning*, 55(1):277–293, 2014.

[19] R. Jiroušek and P. P. Shenoy. A new definition of entropy of belief functions in the Dempster-Shafer theory. *International Journal of Approximate Reasoning*, 92(1):49–65, 2018.

[20] R. Jiroušek and P. P. Shenoy. On properties of a new decomposable entropy of Dempster-Shafer belief functions. *International Journal of Approximate Reasoning*, 119(4):260–279, 2020.

[21] R. Jiroušek, J. Vejnarová, and M. Daniel. Compositional models for belief functions. In G. de Cooman, J. Vejnarová, and M. Zaffalon, editors, *Proceedings of the Fifth International Symposium on Imprecise Probability: Theories and Applications (ISIPTA '07)*, pages 243–252, 2007.

[22] R. Jiroušek, V. Kratochvíl, and P. P. Shenoy. Computing the decomposable entropy of graphical belief function models. In M. Studený, N. Ay, G. Coletti, G. D. Kleiter, and P. P. Shenoy, editors, *Proceedings of the 12th Workshop on Uncertainty Processing*, pages 111–122. MatfyzPress, 2022.

[23] R. Jiroušek, V. Kratochvíl, and P. P. Shenoy. On conditional belief functions in the Dempster-Shafer theory. In S. L. Hégarat-Mascle, I. Bloch, and E. Aldea, editors, *Belief Functions: Theory and Applications, 7th International Conference, BELIEF 2022*, Lecture Notes in Artificial Intelligence, page in press, Switzerland, 2022. Springer Nature.

[24] R. Jiroušek, V. Kratochvíl, and P. P. Shenoy. Two composition operators for belief functions revisited. In M. Studený, N. Ay, G. Coletti, G. D. Kleiter, and P. P. Shenoy, editors, *Proceedings of the 12th Workshop on Uncertainty Processing*, pages 123–134. MatfyzPress, 2022.

[25] R. Jiroušek and P. P. Shenoy. A decomposable entropy for belief functions in the Dempster-Shafer theory. In S. Destercke, T. Denoeux, F. Cuzzolin, and A. Martin, editors, *Belief Functions: Theory and Applications*, volume 11069 of *Lecture Notes in Artificial Intelligence*, pages 146–154. Springer Nature, Switzerland, 2018.

[26] A.-L. Jousselme, C. Liu, D. Grenier, and E. Bossé. Measuring ambiguity in the evidence theory. *IEEE Transactions on Systems, Man, and Cybernetics, Part A: Systems and Humans*, 36(5):890–903, 2006.

[27] A.-L. Jousselme, F. Pichon, N. Ben Abdallah, and S. Destercke. A note about entropy and inconsistency in evidence theory. In *International Conference on Belief Functions*, pages 215–223. Springer, 2021.

[28] G. J. Klir. Uncertainty and information: foundations of generalized information theory. *Kybernetes*, 2006.

[29] G. J. Klir and B. Parviz. A note on the measure of discord. In D. Dubois, M. P. Wellman, B. D'Ambrosio, and P. Smets, editors, *Uncertainty in Artificial Intelligence: Proceedings of the Eighth Conference*, pages 138–141. Morgan Kaufmann, 1992.

[30] G. J. Klir and A. Ramer. Uncertainty in the Dempster-Shafer theory: A critical re-examination. *International Journal of General Systems*, 18(2):155–166, 1990.

[31] C. T. A. Kong. *Multivariate belief functions and graphical models*. PhD thesis, Harvard University, Department of Statistics, Cambridge, Massachusetts, 1986.

[32] Z. Kuang and M. Arnaud. *ibelief: Belief Function Implementation*, 2021. R package version 1.3.1.

[33] S. Kullback and R. A. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22:76–86, 1951.

[34] M. T. Lamata and S. Moral. Measures of entropy in the theory of evidence. *International Journal of General Systems*, 14(4):297–305, 1988.

[35] S. L. Lauritzen and F. V. Jensen. Local computation with valuations from a commutative semigroup. *Annals of Mathematics and Artificial Intelligence*, 21(1):51–69, 1997.

[36] Y. Maeda and H. Ichihashi. An uncertainty measure under the random set inclusion. *International Journal of General Systems*, 21(4):379–392, 1993.

[37] H. T. Nguyen. On entropy of random sets and possibility distributions. In J. C. Bezdek, editor, *The Analysis of Fuzzy Information*, pages 145–156. CRC Press, 1985.

[38] N. R. Pal, J. C. Bezdek, and R. Hemasinha. Uncertainty measures for evidential reasoning II: A new measure of total uncertainty. *International Journal of Approximate Reasoning*, 8(1):1–16, 1993.

[39] A. Perez. ε-admissible simplifications of the dependence structure of a set of random variables. *Kybernetika*, 13(6):439–449, 1977.

[40] G. Shafer. *A Mathematical Theory of Evidence*. Princeton University Press, 1976.

[41] C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423, 623–656, 1948.

[42] P. P. Shenoy. Conditional independence in valuation-based systems. *International Journal of Approximate Reasoning*, 10(3):203–234, 1994.

[43] P. P. Shenoy and G. Shafer. Propagating belief functions with local computations. *IEEE Expert*, 1(3):43–52, 1986.

[44] P. P. Shenoy and G. Shafer. Axioms for probability and belief-function propagation. In R. D. Shachter, T. Levitt, J. F. Lemmer, and L. N. Kanal, editors, *Uncertainty in Artificial Intelligence 4*, pages 169–198. North-Holland, 1990.

[45] P. Smets. Information content of an evidence. *International Journal of Man-Machine Studies*, 19:33–43, 1983.

[46] P. Smets. Constructing the pignistic probability function in a context of uncertainty. In M. Henrion, R. Shachter, L. N. Kanal, and J. F. Lemmer, editors, *Uncertainty in Artificial Intelligence 5*, pages 29–40. Elsevier, 1990.

[47] P. Smets. The canonical decomposition of a weighted belief. In *Proceedings of the 1995 IJCAI Conference*, volume 95, pages 1896–1901, 1995.

[48] P. Smets and R. Kennes. The transferable belief model. *Artificial intelligence*, 66(2):191–234, 1994.

[49] M. Studený. Formal properties of conditional independence in different calculi of ai. In *European Conference on Symbolic and Quantitative Approaches to Reasoning and Uncertainty*, pages 341–348. Springer, 1993.

[50] R. Yager. Entropy and specificity in a mathematical theory of evidence. *International Journal of General Systems*, 9(4):249–260, 1983.