Measuring Quality of Belief Function Approximations*

Radim Jiroušek[1,2][0000-0002-8982-9813] and Václav Kratochvíl[1,2][0000-0002-6013-8752]

[1] Faculty of Management, Prague University of Economics and Business, Jindřichův Hradec, Czechia
[2] Institute of Information Theory and Automation, Czech Academy of Sciences, Prague, Czechia
{radim,velorex}@utia.cas.cz

Abstract. Because of the high computational complexity of the respective procedures, the application of belief-function theory to problems of practice is possible only when the considered belief functions are approximated in an efficient way. Not all measures of similarity/dissimilarity are felicitous to measure the quality of such approximations. The paper presents results from a pilot study that tries to detect the divergences suitable for this purpose.

Keywords: Belief functions · Divergence · Approximation · Compositional models.

1 Introduction

Modeling practical problems usually requires a fair amount of random variables. Even small and simple applications require tens of variables, which complicates the application of belief-function models because the corresponding space of discernment grows super-exponentially with the number of the considered variables. As we will see, to specify a general belief function just for six binary variables, we need $2^{(2^6)} = 2^{64}$ parameters. To avoid problems arising from the high computational complexity of the respective procedures, one should restrict their attention to belief functions representable with a limited number of parameters. For this purpose, we propose models assembled from a sequence of several low-dimensional belief functions – so-called *compositional models*. In connection with this, the question arises, how to recognize whether a compositional model is an acceptable approximation of the considered multidimensional belief function.

In [7] and [6], we studied some heuristics proposed to control the model learning procedures. Inspired by the processes used in probabilistic modeling, we investigated the employment of entropy of belief functions for this purpose. Unfortunately, no belief functions entropy has the properties of probabilistic

* Financially supported by the Czech National Science Foundation under grant no. 19-06569S.

Shannon entropy that would enable us to detect the optimal approximation. Even worse, in belief function theory, there is no generally accepted measure of similarity (dissimilarity) that could help recognize which of two approximations is better. And this is the goal of the current paper. We will study which of several dissimilarity measures (divergences) are suitable for the purpose. In this paper, we consider only those divergences meeting the following two conditions:

- the values of the divergence are non-negative and equal zero only for identical belief functions (the divergence is *non-degenerative*);
- the complexity of the necessary computation is polynomial with the number of focal elements of the considered basic assignments.

Let us note at the very beginning that the achieved results depend on the fact that we consider only a specific class of approximations: the approximations of belief functions by compositional models. We admit that if considering different approximating functions, one could detect other measures of divergence as suitable.

The approximations of complex models by compositional models were first suggested for multidimensional probability distribution [15]. Similarly, the authors of some of the considered divergences also took inspiration from probability theory. And this is why we will at times turn our exposition to probability theory.

The paper is organized as follows. In the next section, we introduce basic notation and recall the idea of Perez, from whom we took the inspiration. The notation from belief function theory is briefly recollected in Section 3. Section 4 introduces the considered divergences, and Section 5 explains the class of approximations considered, i.e., the class of compositional models. The computational experiments and the achieved results are described in Section 6.

2 Basic Notation and Motivation

In this paper, we consider a finite set N of random variables, which are denoted by lower-case characters from the end of the Latin alphabet ($N = \{u, v, w, \ldots\}$). All the considered variables are assumed to be finite-valued. $\mathbb{X}_u$, $\mathbb{X}_v, \ldots$ denote the finite sets of values of variables u, $v, \ldots$. Sets of variables are denoted by upper-case characters K, L, $V, \ldots$. Thus, K may be, say, $\{u, v, w\}$. By a *state* of variables K we understand any combination of values of the respective variables, i.e., in the considered case $K = \{u, w, w\}$, a state is an element of a Cartesian product $\mathbb{X}_K = \mathbb{X}_u \times \mathbb{X}_v \times \mathbb{X}_w$. For a state $a \in \mathbb{X}_K$ and $L \subset K$, $a^{\downarrow L}$ denote a *projection* of $a \in \mathbb{X}_K$ into $\mathbb{X}_L$, i.e., $a^{\downarrow L}$ is the state from $\mathbb{X}_L$ that is got from a by dropping out all the values of variables from $K \setminus L$.

The original idea of Perez [15] was to approximate a multidimensional probability distribution $\mu(N)$ (i.e., $\mu : \mathbb{X}_N \longrightarrow [0, 1]$, for which $\sum_{a \in \mathbb{X}_N} \mu(a) = 1$) by a simpler probability distribution $\kappa(N)$. To measure the quality of such approximation he used their *relative entropy*, which is often called *Kullback-Leibler* *divergence*[3]

$$KL(\mu \parallel \kappa) = \begin{cases} \sum_{c \in \mathbb{X}_N : \kappa(c) > 0} \mu(c) \log_2 \left(\frac{\mu(c)}{\kappa(c)} \right) & \text{if } \mu \ll \kappa, \\ +\infty & \text{otherwise,} \end{cases}$$

where symbol $\mu \ll \kappa$ denotes that κ *dominates* μ, which means that for all $c \in \mathbb{X}_N$, if $\kappa(c) = 0$ then also $\mu(c) = 0$.

It is known that the Kullback-Leibler divergence is non-negative and equals 0 if and only if $\mu = \kappa$ [13]. It is also evident that it is not symmetric[4], and therefore some authors measure the non-similarity of two distributions by the arithmetic mean $\frac{1}{2}(KL(\mu \parallel \kappa) + KL(\kappa \parallel \mu))$. A more sophisticated symmetrized version of this distance is so called Jensen–Shannon divergence (JS) defined

$$JS(\mu \parallel \kappa) = \frac{1}{2} \left(KL \left(\mu \parallel \frac{\mu + \kappa}{2} \right) + KL \left(\kappa \parallel \frac{\mu + \kappa}{2} \right) \right),$$

which is, obviously, symmetric and always finite (namely, both μ and κ are dominated by $\frac{\mu+\kappa}{2}$). For more properties of this and other distances between probability measures, the reader is referred to [14], where one can learn that there is also an alternative way of expressing JS divergence using Shannon entropy

$$JS(\mu \parallel \kappa) = H \left(\frac{\mu + \kappa}{2} \right) - \frac{1}{2}(H(\mu) + H(\kappa)).$$

Recall that

$$H(\mu) = - \sum_{c \in \mathbb{X}_N} \mu(c) \log_2(\mu(c)),$$

which is known to be non-negative and less or equal to $\log_2(|\mathbb{X}_N|)$ [17].

3 Belief Functions

A basic assignment m for variables N is a function[5] $m : 2^{\mathbb{X}_N} \longrightarrow [0, 1]$, for which

- $\sum_{\mathbf{a} \subseteq \mathbb{X}_N} m(\mathbf{a}) = 1$,
- $m(\emptyset) = 0$.

We say that $\mathbf{a} \subseteq \mathbb{X}_N$ is a focal element of m if $m(\mathbf{a}) \neq 0$. We use symbols Bel_m, Pl_m, Q_m to denote belief, plausibility and commonality functions, respectively. These functions, which are known to carry the same information as the corresponding basic assignment m, are defined by the following formulas [16]

$$Bel_m(\mathbf{a}) = \sum_{\mathbf{b} \subseteq \mathbf{a}} m(\mathbf{b}); \quad Pl_m(\mathbf{a}) = \sum_{\mathbf{b} \subseteq \mathbb{X}_N : \mathbf{a} \cap \mathbf{b} \neq \emptyset} m(\mathbf{b}); \quad Q_m(\mathbf{a}) = \sum_{\mathbf{b} \subseteq \mathbb{X}_N : \mathbf{b} \supseteq \mathbf{a}} m(\mathbf{b}).$$

---

[3] We take $0 \log_2(0) = 0$.

[4] To show asymmetry of the Kullback-Leibler divergence consider $\mu = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$, and $\kappa = (\frac{1}{2}, \frac{1}{2}, 0)$.

[5] $2^{\mathbb{X}_N}$ denote the set of all subsets of $\mathbb{X}_N$.

When constructing compositional models, we need marginals of the considered basic assignments. Let m be defined for arbitrary set of variables $L \supseteq K$. Symbol $m^{\downarrow K}$ will denote the marginal of m, which is defined for variables K. Thus,

$$m^{\downarrow K}(\mathbf{b}) = \sum_{\mathbf{a} \subseteq \mathbb{X}_L :\, \mathbf{a}^{\downarrow K} = \mathbf{b}} m(\mathbf{a}).$$

for all $\mathbf{b} \subseteq \mathbb{X}_K$.

When normalizing the plausibility function on singletons, one gets a probability distribution on $\mathbb{X}_N$ called a *plausibility transform* of basic assignment m [1]. There are several other probabilistic transforms described in literature [2, 3]. In this paper we use only the above-mentioned plausibility transform λ_m and the so-called *pignistic transform* π_m strongly advocated by Philippe Smets [18], which are defined for all $a \in \mathbb{X}_N$

$$\lambda_m(a) = \frac{Pl_m(\{a\})}{\sum_{c \in \mathbb{X}_N} Pl_m(\{c\})}, \quad \text{and} \quad \pi_m(a) = \sum_{\mathbf{b} \subseteq \mathbb{X}_N : a \in \mathbf{b}} \frac{m(\mathbf{b})}{|\mathbf{b}|}.$$

Up to now, we have recalled a standard notation used in belief function theory. Rather unusual is that, to make the next exposition as simple as possible, we will sometimes view the basic assignment m also as a probability distribution on $2^{\mathbb{X}_N}$. This enables us to speak about Shannon entropy $H(m)$ of m, to say that m_1 dominates m_2, and to compute Kullback-Leibler divergence between two basic assignments.

4 Divergences

Quite a few papers suggesting different tools to measure similarity/dissimilarity of belief functions were published. The reader can find a good survey in [12]. As indicated in the Introduction, in this paper, we are interested only in those measures, the computation of which is tractable even for multidimensional belief functions if the number of focal elements of the considered basic assignments is not too high. In other words, we are interested in the formulas, the computational complexity of which depends on the number of focal elements, regardless of the number of variables, for which the respective basic assignments are defined. Given the goal of this paper, we also restrict our attention only to *non-degenerative* measures, i.e., the measures which can detect the equality of belief functions because they equal zero only for identical basic assignments. In this pilot study, we consider only the six divergences described below.

In this section, we assume that all the considered basic assignments are defined for the set of variables N.

Jousselme et al. (2001). In [11], the authors define a distance between basic assignments meeting all the metric axioms: non-negativity, non-degeneracy, symmetry, and the triangle inequality. Recall that the Kullback-Leibler divergence introduced in Section 2 meets only the first two properties; it is not symmetric, nor the triangle inequality holds for KL.

To be able use the notation of linear algebra, consider a fixed ordering of elements of $2^{|\mathbb{X}_N|}$. Then, m can be interpreted as a vector $\boldsymbol{m}$ of $2^{|\mathbb{X}_N|}$ non-negative real numbers. Jousselme et al. define their distance

$$d_{BPA}(m_1, m_2) = \sqrt{\frac{1}{2}(\boldsymbol{m_1} - \boldsymbol{m_2})^T D(\boldsymbol{m_1} - \boldsymbol{m_2})}, \tag{1}$$

where D is $2^{|\mathbb{X}_N|} \times 2^{|\mathbb{X}_N|}$ matrix defined as follows: let $\mathbf{a}_i$ be an element of $2^{\mathbb{X}_N}$ which corresponds the i-th coordinate of the vector $\boldsymbol{m}$. Then, the elements of matrix $D = (d_{ij})$ are defined

$$d_{ij} = \frac{|\mathbf{a}_i \cap \mathbf{a}_j|}{|\mathbf{a}_i \cup \mathbf{a}_j|}.$$

Note that we allow a situation of $\mathbf{a}_i = \emptyset$. In this case define $d_{ii} = 1$. Knowing the matrix D, the argument of the square root of Eq. (1) can be rewritten into the following form

$$\begin{aligned}(\boldsymbol{m_1} - \boldsymbol{m_2})^T D(\boldsymbol{m_1} - \boldsymbol{m_2}) \\ = \sum_{\mathbf{a} \subseteq \mathbb{X}_N} m_1(\mathbf{a}) \sum_{\mathbf{b} \subseteq \mathbb{X}_N} \frac{m_1(\mathbf{b})\ |\mathbf{a} \cap \mathbf{b}|}{|\mathbf{a} \cup \mathbf{b}|} + \sum_{\mathbf{a} \subseteq \mathbb{X}_N} m_2(\mathbf{a}) \sum_{\mathbf{b} \subseteq \mathbb{X}_N} \frac{m_2(\mathbf{b})\ |\mathbf{a} \cap \mathbf{b}|}{|\mathbf{a} \cup \mathbf{b}|} \\ - 2 \sum_{\mathbf{a} \subseteq \mathbb{X}_N} \sum_{\mathbf{b} \subseteq \mathbb{X}_N} \frac{m_1(\mathbf{a}) m_2(\mathbf{b})\ |\mathbf{a} \cap \mathbf{b}|}{|\mathbf{a} \cup \mathbf{b}|}.\end{aligned}$$

Xiao (2019). To define the divergence between two basic assignments m_1 and m_2, Xiao [21] makes use of the fact that a basic assignment on $\mathbb{X}_N$ is a probability measure on $2^{\mathbb{X}_N}$. Thus, she defines a belief function divergence – she calls it Belief Jensen-Shannon divergence (BJS) – which is the probabilistic Jensen-Shannon divergence of the corresponding probability measures, i.e.,

$$BJS(m_1, m_2) = \frac{1}{2}\left[KL\left(m_1 \parallel \frac{m_1 + m_2}{2}\right) + KL\left(m_2 \parallel \frac{m_1 + m_2}{2}\right)\right], \tag{2}$$

or, equivalently

$$BJS(m_1, m_2) = H\left(\frac{m_1 + m_2}{2}\right) - \frac{H(m_1) + H(m_2)}{2} \tag{3}$$

(recall, H denotes the Shannon entropy).

Song-Deng (2019a). As the authors say in [20], being inspired by Eq. (2), they replaced the arithmetic mean in Eq. (2) by the geometric mean, suggesting a new divergence BRE (perhaps from Belief Relative Entropy) defined by

$$BRE(m_1, m_2) = \sqrt{KL\left(m_1 \parallel \sqrt{m_1 \cdot m_2}\right) \cdot KL\left(m_2 \parallel \sqrt{m_1 \cdot m_2}\right)}. \tag{4}$$

In contrast to BJS, which is always finite, BRE equals $+\infty$ whenever there is at least one $\mathbf{a} \subseteq \mathbb{X}_N$, which is a focal element of only one of the two basic assignments m_1, m_2.

Song-Deng (2019b). The same pair of authors suggested also another belief function divergence related to the relative Deng entropy D_d, which is defined by the following formula

$$D_d(m_1 \parallel m_2) = \sum_{\mathbf{a} \subseteq \mathbb{X}_N : m_2(\mathbf{a})>0} \frac{1}{2^{|\mathbf{a}|}-1} m_1(\mathbf{a}) \log\left(\frac{m_1(\mathbf{a})}{m_2(\mathbf{a})}\right). \quad (5)$$

Assume that $D_d(m_1 \parallel m_2) = +\infty$ in case that there is $\mathbf{a} \subseteq \mathbb{X}_N$ for which $m_1(\mathbf{a}) > 0 = m_2(\mathbf{a})$. In [19], the authors define the divergence D_{SDM} symmetrizing the relative Deng entropy

$$D_{SDM}(m_1, m_2) = \frac{1}{2}\left(D_d(m_1 \parallel m_2) + D_d(m_2 \parallel m_1)\right). \quad (6)$$

Assuming that for $m_1 \ll m_2$, Eq. (5) defines the relative entropy, and that it equals $+\infty$ in opposite case. Then it is not difficult to show [19] that measure D_{SDM} is non-negative, non-degenerative, and symmetric.

Simple divergences. With the goal to test also some computationally cheap divergences, we, being inspired by the entropy defined in [9], consider also functions

$$Div_\lambda(m_1, m_2) = KL(\lambda_{m_1} \parallel \lambda_{m_2}) + \sum_{\mathbf{a} \subseteq \mathbb{X}_N} |m_1(\mathbf{a}) - m_2(\mathbf{a})| \cdot \log(|\mathbf{a}|), \quad (7)$$

and

$$Div_\pi(m_1, m_2) = KL(\pi_{m_1} \parallel \pi_{m_2}) + \sum_{\mathbf{a} \subseteq \mathbb{X}_N} |m_1(\mathbf{a}) - m_2(\mathbf{a})| \cdot \log(|\mathbf{a}|), \quad (8)$$

where λ and π are plausibility and pignistic transforms introduced in Section 2.

Proposition 1. *Both divergences Div_λ and Div_π are non-negative and non-degenerative.*

Proof. The non-negativity of the considered divergences follows directly from the non-negativity of Kullback-Leibler divergence.

To show their non-degenerativity, i.e., $Div_\lambda(m_1, m_2) = 0 \iff m_1 = m_2$, and $Div_\pi(m_1, m_2) = 0 \iff m_1 = m_2$, consider two basic assignments m_1 and m_2. If $m_1 = m_2$, then, trivially, $Div_\lambda(m_1, m_2) = Div_\pi(m_1, m_2) = 0$.

To show the other side of the equivalence, assume that $m_1 \neq m_2$, and

$$\sum_{\mathbf{a} \subseteq \mathbb{X}_N} |m_1(\mathbf{a}) - m_2(\mathbf{a})| \cdot \log(|\mathbf{a}|) = 0. \quad (9)$$

This equality holds if and only if $m_1(\mathbf{a}) = m_2(\mathbf{a})$ for all non-singletons $\mathbf{a} \subseteq \mathbb{X}_N$. Since,

$$\begin{aligned}\sum_{c \in \mathbb{X}_N} Pl_{m_i}(c) &= \sum_{c \in \mathbb{X}_N} m_i(c) + \sum_{c \in \mathbb{X}_N} \left(\sum_{\mathbf{a} \subseteq \mathbb{X}_N :\, c \in \mathbf{a} \,\&\, |\mathbf{a}|>1} m_i(\mathbf{a}) \right) \\ &= \left(1 - \sum_{\mathbf{a} \subseteq \mathbb{X}_N : |\mathbf{a}|>1} m_i(\mathbf{a})\right) + \sum_{c \in \mathbb{X}_N} \left(\sum_{\mathbf{a} \subseteq \mathbb{X}_N :\, c \in \mathbf{a} \,\&\, |\mathbf{a}|>1} m_i(\mathbf{a}) \right),\end{aligned}$$

we can see that $\sum_{c\in\mathbb{X}_N} Pl_{m_1}(c) = \sum_{c\in\mathbb{X}_N} Pl_{m_2}(c)$.

Since we assume that for $m_1 \neq m_2$ Eq. (9) holds, then there exists $c \in \mathbb{X}_N$, for which $m_1(c) \neq m_2(c)$, and therefore also

$$Pl_{m_1}(c) = \sum_{\mathbf{a}\subseteq\mathbb{X}_N: c\in\mathbf{a}} m_1(\mathbf{a}) \neq \sum_{\mathbf{a}\subseteq\mathbb{X}_N: c\in\mathbf{a}} m_2(\mathbf{a}) = Pl_{m_2}(c).$$

Thus,

$$\lambda_{m_1}(c) = \frac{Pl_{m_1}(c)}{\sum_{x\in\mathbb{X}_N} Pl_{m_1}(x)} \neq \frac{Pl_{m_2}(c)}{\sum_{x\in\mathbb{X}_N} Pl_{m_2}(x)} = \lambda_{m_2}(c),$$

and therefore $KL(\lambda_{m_1} \parallel \lambda_{m_2}) > 0$. This proves that Div_λ is non-degenerative because we have showed that either $KL(\pi_{m_1} \parallel \pi_{m_2})$ is positive, or Eq. (9) does not hold, whenever $m_1 \neq m_2$.

Similarly, for the considered $c \in \mathbb{X}_N$, for which $m_1(c) \neq m_2(c)$,

$$\begin{aligned} \pi_{m_1}(c) = m_1(c) &+ \sum_{\mathbf{a}\subseteq\mathbb{X}_N:\, c\in\mathbf{a}\&|\mathbf{a}|>1} \frac{m_1(\mathbf{a})}{|\mathbf{a}|} \\ \neq m_2(c) &+ \sum_{\mathbf{a}\subseteq\mathbb{X}_N:\, c\in\mathbf{a}\&|\mathbf{a}|>1} \frac{m_2(\mathbf{a})}{|\mathbf{a}|} = \pi_{m_2}(c), \end{aligned}$$

and therefore also $KL(\pi_{m_1} \parallel \pi_{m_2}) > 0$, which proves that also Div_π is non-degenerative. □

5 Compositional Models

The definition of compositional models for belief functions is analogous to that in probability theory [4]. A basic assignment of a multidimensional compositional model is assembled from a system of low-dimensional basic assignments. To do it, one needs a tool to create a more-dimensional basic assignment from two or more low-dimensional ones. In this paper, we use an *operator of composition* $\triangleright$. By this term, we understand a binary operator meeting the following four axioms (basic assignments m_1, m_2, m_3 are assumed to be defined for K, L, M, respectively):

A1 *(Domain):* $m_1 \triangleright m_2$ is a basic assignment for variables $K \cup L$.

A2 *(Composition preserves first marginal):* $(m_1 \triangleright m_2)^{\downarrow K} = m_1$.

A3 *(Commutativity under consistency):* If m_1 and m_2 are consistent, i.e., $m_1^{\downarrow K\cap L} = m_2^{\downarrow K\cap L}$, then $m_1 \triangleright m_2 = m_2 \triangleright m_1$.

A4 *(Associativity under special condition):* If $K \supset (L \cap M)$, or, $L \supset (K \cap M)$ then $(m_1 \triangleright m_2) \triangleright m_3 = m_1 \triangleright (m_2 \triangleright m_3)$.

Because of space limit we cannot discuss these axioms in details (for this we refer the reader to [5]), but roughly speaking, axioms A1, A3, A4 guarantee that the operator of composition uniquely reconstruct basic assignment $m^{\downarrow K\cup L}$ from its marginals $m^{\downarrow K}$ and $m^{\downarrow L}$, if there exists a lossless decomposition of $m^{\downarrow K\cup L}$ into $m^{\downarrow K}$ and $m^{\downarrow L}$. Surprisingly, it is axiom A4, which guarantees that no necessary information from $m^{\downarrow L}$ is lost. Axiom A2 solves the problem arising when non-consistent basic assignments are composed. Generally, there are 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 is superior to the other two possibilities from the computational point of view.

By a *compositional model*, we understand a multidimensional belief function, the basic assignment of which is assembled from a sequence of low-dimensional basic assignments with the help of the operator of composition: $m_1 \triangleright m_2 \triangleright \ldots \triangleright m_n$. Since the operator of composition is not associative, this expression is ambiguous. To avoid this ambiguity, we omit the parentheses only if the operators are to be performed from left to right, i.e.,

$$m_1 \triangleright m_2 \triangleright \ldots \triangleright m_n = (\ldots((m_1 \triangleright m_2) \triangleright m_3) \triangleright \ldots \triangleright m_{n-1}) \triangleright m_n. \qquad (10)$$

Let $m^\star = m_1 \triangleright m_2 \triangleright \ldots \triangleright m_n$, and let each m_i be defined for variables K_i. Due to axiom A2, m_1 is a marginal of $m^\star$. Similarly, $m_1 \triangleright m_2 = m^{\star \downarrow K_1 \cup K_2}$. This, however, does not mean that m_2 is also a marginal of $m^\star$. If all m_i are marginals of $m^\star$, then we say that $m^\star$ is defined by a *perfect compositional model*. The following assertion summarizes the relevant properties that were proved in [10, 8].

Proposition 2. *Let $m^\star = m_1 \triangleright m_2 \triangleright \ldots \triangleright m_n$, and let each m_i be defined for the set of variables K_i.*

- *(Compositional models can be perfectized.) There exists a perfect model $m^\star = \bar{m}_1 \triangleright \bar{m}_2 \triangleright \ldots \triangleright \bar{m}_n$ such that each $\bar{m}_i$ is defined for K_i.*
- *(Uniqueness of compositional models.) Let $m^\star = m_1 \triangleright m_2 \triangleright \ldots \triangleright m_n$ be perfect. If there is a permutation $j_1, j_2, \ldots, j_n$ such that $m_{j_1} \triangleright m_{j_2} \triangleright \ldots \triangleright m_{j_n}$ is also perfect, then $m_{j_1} \triangleright m_{j_2} \triangleright \ldots \triangleright m_{j_n} = m^\star$.*
- *(Consistent decomposable models are perfect.) If all m_i are pairwise consistent (i.e., for all $1 \leq i, j \leq n$, $m_i^{\downarrow K_i \cap K_j} = m_j^{\downarrow K_i \cap K_j}$), and the sequence $K_1, K_2, \ldots, K_n$ meets the running intersection property[6], then $m_1 \triangleright m_2 \triangleright \ldots \triangleright m_n$ is perfect.*

Now, let us express the original idea of Perez [15] in the language of compositional models: He proposed to approximate multidimensional probability distributions by compositional models and, as said above, to measure the quality of such approximations using the Kullback-Leibler divergence. He proved that if a perfect model exists, it minimizes the KL divergence (due to the *uniqueness of compositional models*, all perfect models define the identical approximation). This fact fully corresponds with our intuition. When knowing only a system of

[6] $K_1, K_2, \ldots, K_n$ meets the running intersection property if

$$\forall\ i = 2, 3, \ldots, n \ \ \exists\ j\ (1 \leq j < i) \ \ K_i \cap (K_1 \cup \ldots \cup K_{i-1}) \subseteq K_j.$$

marginals of an approximated distribution, the best approximation is a distribution having all of them for its marginals.

How to employ this idea within the framework of belief functions? Not having a generally accepted "Kulback-Leibler divergence" for belief functions at our disposal, we try to solve a problem, which is, in a sense, *inverse* to that of Perez. We accept the paradigm that the best approximation of a multidimensional basic assignment is, if it exists, a perfect compositional model assembled from the marginals of the approximated basic assignment. Based on this we test, which belief function divergences detect the optimal approximation. The corresponding computational experiments, as well as the achieved results, are described in the next section. First, however, we owe the reader a specification of the used operator of composition.

In the literature, two operators of composition meeting axioms A1 – A4 were introduced. Historically, the first was defined in [10]. Its disadvantage is that it does not comply with the Dempster-Shafer interpretation of belief function theory. The other operator, derived from Dempster's rule of combination, was designed by Shenoy in [8]. Nevertheless, because of its high computational complexity, we did not include it in the described pilot computational experiments. In the experiments described below, we used only the first operator. To present its definition, we need an additional notion.

Consider two arbitrary sets of variables K and L. By a *join* of $\mathbf{a} \subseteq \mathbb{X}_K$ and $\mathbf{b} \subseteq \mathbb{X}_L$ we understand a set

$$\mathbf{a} \bowtie \mathbf{b} = \{c \in \mathbb{X}_{K\cup L} : c^{\downarrow K} \in \mathbf{a} \ \ \& \ \ c^{\downarrow L} \in \mathbf{b}\}.$$

Realize that if K and L are disjoint, then $\mathbf{a} \bowtie \mathbf{b} = \mathbf{a} \times \mathbf{b}$, if $K = L$, then $\mathbf{a} \bowtie \mathbf{b} = \mathbf{a} \cap \mathbf{b}$, and, generally, for $\mathbf{c} \subseteq \mathbb{X}_{K\cup L}$, $\mathbf{c}$ is a subset of $\mathbf{c}^{\downarrow K} \bowtie \mathbf{c}^{\downarrow L}$, which may be proper. Notice that the sets, for which $\mathbf{c} = \mathbf{c}^{\downarrow K} \bowtie \mathbf{c}^{\downarrow L}$, were called *Z-layered rectangles* in [22, 23].

Definition 1. ***Factorizing operator of composition***
Consider two arbitrary basic assignments, m_1 and m_2 defined for sets of variables K and L, respectively. A factorizing composition $m_1 \triangleright m_2$ *is defined for each nonempty $\mathbf{c} \subseteq \mathbb{X}_{K\cup L}$ by one of the following expressions:*

(i) if $m_2^{\downarrow K\cap L}(\mathbf{c}^{\downarrow K\cap L}) > 0$ and $\mathbf{c} = \mathbf{c}^{\downarrow K} \bowtie \mathbf{c}^{\downarrow L}$, then

$$(m_1 \triangleright m_2)(\mathbf{c}) = \frac{m_1(\mathbf{c}^{\downarrow K}) \cdot m_2(\mathbf{c}^{\downarrow L})}{m_2^{\downarrow K\cap L}(\mathbf{c}^{\downarrow K\cap L})};$$

(ii) if $m_2^{\downarrow K\cap L}(\mathbf{c}^{\downarrow K\cap L}) = 0$ and $\mathbf{c} = \mathbf{c}^{\downarrow K} \times \mathbb{X}_{L\setminus K}$, then

$$(m_1 \triangleright m_2)(\mathbf{c}) = m_1(\mathbf{c}^{\downarrow K});$$

(iii) in all other cases, $(m_1 \triangleright m_2)(\mathbf{c}) = 0$.

6 Computational Experiments

As indicated in the Introduction, the goal of the described experiments is to examine which of the considered divergences can be used to (heuristically) detect the best approximations of basic assignments. To do it, we take into account only the approximations by compositional models and accept the intuitively rational and theoretically well-grounded fact that the perfect model, if it exists, is the best approximation.

In the experiments, we considered 14 binary variables ($|N| = 14$), for which we randomly generated 900 basic assignments[7] (denote them m) with 30 focal elements. For each basic assignment we randomly generated a cover of N, i.e., sets $K_1, K_2, \ldots, K_n$, and $N = K_1 \cup K_2 \cup \ldots \cup K_n$ ($5 \leq n \leq 11$, $2 \leq |K_i| \leq 4$). To assure that we can identify the best approximation, we guaranteed that this sequence met the running intersection property (RIP). Due to Proposition 2, we know that $m^{\downarrow K_1} \triangleright m^{\downarrow K_2} \triangleright \ldots \triangleright m^{\downarrow K_n}$ is perfect, and therefore it is the best approximation of m that can be composed of these marginals. To avoid misunderstanding, recall that we study the behavior of the considered divergences, and therefore, we do not mind that most of the considered approximations were much more complex (in the sense of the number of parameters defining the respective belief functions) than the approximated basic assignment.

For each perfect model, we set up also non-perfect models by randomly permuting the marginals in the sequence. Thus, for each of the 900 randomly generated 14-dimensional basic assignments, we had one RIP and several (on average about 6) non-RIP compositional models[8]. The achieved results are summarized in Table 1. From this, the reader can see that for the 900 basic assignments, we considered 6 458 approximating compositional models, 900 of which were perfect, and the remaining 5 558 were non-perfect. On the right-hand side of Table 1, the behavior of the considered distances is described. As *wrongly detected* we considered those perfect approximations $m^{\downarrow K_1} \triangleright \ldots \triangleright m^{\downarrow K_n}$, for which there was generated non-RIP model (defined by a permutation $m^{\downarrow K_{j_1}} \triangleright \ldots \triangleright m^{\downarrow K_{j_n}}$) such that

Table 1. Numbers of wrongly detected approximations.

	total	wrongly detected by			
		d_{BPA}	BJS	Div_λ	Div_π
Number of perfect approximations	900	348	216	97	9
Number of non-perfect approximations	5 558	1 613	1 007	167	15

[7] We generated basic assignments of three types: 300 of them were *nested*, 300 were *quasi-bayesian*, and the remaining 300 basic assignments had 29 fully randomly selected focal elements and the thirties one was $\mathbb{X}_N$.

[8] Precisely speaking, we know that all RIP models are perfect, but, theoretically, it may happen that also non-RIP model is perfect. However, this happens very rarely, and when assessing the results, we took that all non-RIP models were non-perfect.

$$Div(m, (m^{\downarrow K_{j_1}} \triangleright \ldots \triangleright m^{\downarrow K_{j_n}})) < Div(m, (m^{\downarrow K_1} \triangleright \ldots \triangleright m^{\downarrow K_n})), \qquad (11)$$

where Div stands for the respective divergence from Tab. 1. Analogously, *wrongly detected* non-perfect models are those non-perfect models $m^{\downarrow K_{j_1}} \triangleright \ldots \triangleright m^{\downarrow K_{j_n}}$, for which Eq. (11) holds true. It means that there is a correspondence between wrongly detected perfect and non-perfect models, however, this correspondence is not a bijection. Each wrongly detected perfect model corresponds with at least one (but often more than one) wrongly detected non-perfect model. Notice that if a perfect approximation $m^{\downarrow K_1} \triangleright \ldots \triangleright m^{\downarrow K_n}$ and its non-perfect permutation $m^{\downarrow K_{j_1}} \triangleright \ldots \triangleright m^{\downarrow K_{j_n}}$ were generated, such that the equality $Div(m, (m^{\downarrow K_{j_1}} \triangleright \ldots \triangleright m^{\downarrow K_{j_n}})) = Div(m, (m^{\downarrow K_1} \triangleright \ldots \triangleright m^{\downarrow K_n}))$ hold, none of these two approximations was recognized as wrongly detected.

Though we said in Section 4 that we would study six divergences, only four of them appear in Table 1. It is because the remaining divergences BRE and D_{SDM} (defined by Eq. (4) and Eq. (6), respectively) equal $+\infty$ whenever there is a focal element of the approximation, which is not a focal element of the originally randomly generated basic assignment. This, however, cannot be avoided for any multidimensional basic assignment and its compositional-model approximation. So, it is not surprising that all divergences computed for BRE and D_{SDM} were $+\infty$, which means that they are useless for the purpose of this study.

7 Conclusions

From Table 1 one can deduce that the simple divergences Div_λ and mainly Div_π may be recommended to identify the best approximations of multidimensional basic assignments. However, let us recall that we have achieved this conclusion when considering only approximations by f-compositional models. We have not yet, achieved any results in the case of experiments with the operator of composition derived from Dempster's rule of combination (d-composition). The main reason is the computational complexity of the operator of d-composition, the calculation of which requires conversions of low-dimensional basic assignments, from which the model is set up, from/to the respective commonality functions.

References

1. Cobb, B.R., Shenoy, P.P.: On the plausibility transformation method for translating belief function models to probability models. International Journal of Approximate Reasoning **41**(3), 314–340 (2006)
2. Cuzzolin, F.: On the relative belief transform. International Journal of Approximate Reasoning **53**(5), 786–804 (2012)
3. Daniel, M.: On transformations of belief functions to probabilities. International Journal of Intelligent Systems **21**(3), 261–282 (2006)
4. Jiroušek, R.: Foundations of compositional model theory. International Journal of General Systems **40**(6), 623–678 (2011)
5. Jiroušek, R.: A short note on decomposition and composition of knowledge. International Journal of Approximate Reasoning **120**, 24–32 (2020)

6. Jiroušek, R., Kratochvíl, V.: Approximations of belief functions using compositional models. In: European Conference on Symbolic and Quantitative Approaches with Uncertainty. pp. 354–366. Springer (2021)
7. Jiroušek, R., Kratochvíl, V., Shenoy, P.P.: Entropy-based learning of compositional models from data. In: Belief Functions: Theory and Applications (Proceedings of the 6th International Conference, BELIEF 2021), pp. 117–126. Springer (2021)
8. Jiroušek, R., Shenoy, P.P.: Compositional models in valuation-based systems. In: Belief Functions: Theory and Applications, pp. 221–228. Springer (2012)
9. Jiroušek, R., Shenoy, P.P.: A new definition of entropy of belief functions in the Dempster-Shafer theory. International Journal of Approximate Reasoning **92**(1), 49–65 (2018)
10. Jiroušek, R., Vejnarová, J., Daniel, M.: Compositional models for belief functions. In: de Cooman, G., Vejnarová, J., Zaffalon, M. (eds.) Proceedings of the Fifth International Symposium on Imprecise Probability: Theories and Applications (ISIPTA '07). pp. 243–252 (2007)
11. Jousselme, A.L., Grenier, D., Bossé, É.: A new distance between two bodies of evidence. Information fusion **2**(2), 91–101 (2001)
12. Jousselme, A.L., Maupin, P.: Distances in evidence theory: Comprehensive survey and generalizations. International Journal of Approximate Reasoning **53**(2), 118–145 (2012)
13. Kullback, S., Leibler, R.A.: On information and sufficiency. Annals of Mathematical Statistics **22**, 76–86 (1951)
14. Österreicher, F., Vajda, I.: A new class of metric divergences on probability spaces and its applicability in statistics. Annals of the Institute of Statistical Mathematics **55**(3), 639–653 (2003)
15. Perez, A.: ε-admissible simplifications of the dependence structure of a set of random variables. Kybernetika **13**(6), 439–449 (1977)
16. Shafer, G.: A Mathematical Theory of Evidence. Princeton University Press (1976)
17. Shannon, C.E.: A mathematical theory of communication. Bell System Technical Journal **27**, 379–423, 623–656 (1948)
18. Smets, P.: Constructing the pignistic probability function in a context of uncertainty. In: Henrion, M., Shachter, R., Kanal, L.N., Lemmer, J.F. (eds.) Uncertainty in Artificial Intelligence 5, pp. 29–40. Elsevier (1990)
19. Song, Y., Deng, Y.: Divergence measure of belief function and its application in data fusion. IEEE Access **7**, 107465–107472 (2019)
20. Song, Y., Deng, Y.: A new method to measure the divergence in evidential sensor data fusion. International Journal of Distributed Sensor Networks **15**(4), 1550147719841295 (2019)
21. Xiao, F.: Multi-sensor data fusion based on the belief divergence measure of evidences and the belief entropy. Information Fusion **46**, 23–32 (2019)
22. Yaghlane, B.B., Smets, P., Mellouli, K.: Belief function independence: I. the marginal case. International Journal of Approximate Reasoning **29**(1), 47–70 (2002)
23. Yaghlane, B.B., Smets, P., Mellouli, K.: Belief function independence: Ii. the conditional case. International Journal of Approximate Reasoning **31**(1-2), 31–75 (2002)