Brief Introduction to Probabilistic Compositional Models

Radim Jiroušek

Abstract Any field of social sciences is based on uncertain knowledge, uncertain information and uncertain data. The economics is not an exception. This is why probability theory and probabilistic modeling play an important role in econometrics. In practical applications one has to cope with the fact that even relatively small models have to take into account rather hundreds than tens of factors. This is why the methods for multidimensional probability distribution representation, like Bayesian networks, have become so popular in this field. The goal of this paper is to promote an alternative approach, so called compositional models.

1 Introduction

There are more and more fields of human activities which are giving rise to databases of enormous size. In some of them, the research data bases are a side product of other business activities, like, for example, in banking where even small banks store hundreds or rather thousands of records describing their clients' activities every day. As another example we can consider the research in the field of customer relationship management, which is based on the analysis of records describing the customer spending. On the other hand, creating large data bases has become a business of its own, as the different media research companies attest to. These companies collect data on all possible marketing activities, like data from TV-meters, or data monitoring advertising investments, such as data on advertising in journals and on the Internet. An existence of such institutions proves the fact that data have become a business product and that their analysis and processing is an important part of business life.

Radim Jiroušek
Faculty of management, University of Economics, Prague
Jarošovská 1117/II, 377 01 Jindřichův Hradec, Czech Republic, e-mail: radim@utia.cas.cz

So, it is not surprising that there is abundant literature on application techniques such as Bayesian networks [3, 14], which is perhaps the most popular tool to describe and process multidimensional probability distributions. Here we are expected to present some examples of research papers describing typical applications of Bayesian networks, but we do not dare to do it; in a few seconds Google has found more than two million incidences of 'application of Bayesian network to ...'. In this paper we do not intend to present the *two-millionth-first* paper on the Bayesian networks. On the contrary, we want to present a survey paper (summarizing some of the results published in [5, 9, 8]) on an alternative approach to multidimensional probability distribution representation and processing, an approach based on the so-called *operator of composition*.

In contrast to Bayesian networks, an advantage of the models described in the current paper, which we call *compositional models*, is that we can make do with probability theory. Though they are as powerful as Bayesian networks (they can model the same class of distributions), they do not use graphs to represent the distribution structure. For other advantages of compositional models see Conclusions.

2 Notation and Basic Concepts

We consider variables $u \in N$, each having a finite (non-empty) set of values that will be denoted by $\mathbb{X}_u$. The set of all combinations of the considered values will be denoted $\mathbb{X}_N = \times_{u\in N}\mathbb{X}_u$. Analogously, for a subset of variables $K \subset N$, $\mathbb{X}_K = \times_{u\in K}\mathbb{X}_u$.

Distributions of the considered variables will be denoted by Greek letters $\kappa, \lambda, \ldots$ with possible indices; thus for $K \subseteq N$, we can consider a distribution $\kappa(K)$, which is a $|K|$-dimensional distribution and $\kappa(x)$ denotes the value of probability distribution κ for point $x \in \mathbb{X}_K$.

For a probability distribution $\kappa(K)$ and $J \subset K$, we will often consider a *marginal distribution* $\kappa^{\downarrow J}$ of κ, which can be computed for all $x \in \mathbb{X}_J$ by

$$\kappa^{\downarrow J}(x) = \sum_{y\in\mathbb{X}_K : y^{\downarrow J}=x} \kappa(y),$$

where $y^{\downarrow J}$ denotes the *projection* of $y \in \mathbb{X}_K$ into $\mathbb{X}_J$. Note that we do not exclude situations when $J = \emptyset$. By definition, we get $\kappa^{\downarrow\emptyset} = 1$.

Having two distributions $\pi(K)$ and $\kappa(K)$, we say that κ dominates π (in symbol $\pi \ll \kappa$) if for all $x \in \mathbb{X}_K$, for which $\kappa(x) = 0$ also $\pi(x) = 0$. As a measure of similarity of these two distributions we will consider their *Kullback-Leibler divergence* [13] (or crossentropy) defined[1]

$$Div(\pi;\kappa) = \sum_{x\in\mathbb{X}_K} \pi(x)\log\frac{\pi(x)}{\kappa(x)},$$

which is known to be zero if and only if $\pi = \kappa$.

[1] In this paper we take $\frac{0.0}{0} = 0$ by definition.

The most important notion of this paper is the operator of composition, which realizes an operation in a way inverse to marginalization. For a probability distribution $\kappa(K)$ and $J \subset K$, the respective marginal distribution $\kappa^{\downarrow J}$ is unique. For a distribution $\pi(J)$ there are (infinitely) many distributions $\nu(K)$ such that $\nu^{\downarrow J} = \pi$. All these distributions ν are *extensions of π for variables K*. But if we want to find that $\nu(K)$, which is as similar as possible to a given distribution $\mu(K)$, we can take the distribution

$$\nu = \arg \min_{\lambda(K):\lambda^{\downarrow J}=\pi} Div(\lambda;\mu),$$

which is unique if the divergence is defined. In this case we say that ν is a *projection* of μ into the set (space) of all the extensions of π for variables K.

The operator of composition is designed in the way that the projection of μ into the set of all the extension of π is got as a composition of π and μ - see Property 3 of the following Proposition.

Definition 1. For two arbitrary distributions $\kappa(K)$ and $\lambda(L)$, for which $\kappa^{\downarrow K\cap L} \ll \lambda^{\downarrow K\cap L}$, their *composition* is, for each $x \in \mathbb{X}_{L\cup K}$, given by the following formula

$$(\kappa \rhd \lambda)(x) = \frac{\kappa(x^{\downarrow K})\lambda(x^{\downarrow L})}{\lambda^{\downarrow K\cap L}(x^{\downarrow K\cap L})}.$$

In case $\kappa^{\downarrow K\cap L} \not\ll \lambda^{\downarrow K\cap L}$, the composition remains undefined.

Let us summarize the most important properties of the composition operator that were proved in [5, 9]

Proposition 1. *Suppose $\kappa(K)$ and $\lambda(L)$ are probability distributions for which $\lambda^{\downarrow K\cap L} \gg \kappa^{\downarrow K\cap L}$. Then the following statements hold:*

1. Domain: *$\kappa \rhd \lambda$ is a distribution for $K \cup L$.*
2. Composition preserves first marginal: $(\kappa \rhd \lambda)^{\downarrow K} = \kappa$.
3. Projection: $\kappa \rhd \lambda = \arg \min_{\nu(K\cup L):\nu^{\downarrow K}=\kappa} Div(\nu^{\downarrow L};\lambda)$.
4. Non-commutativity: *In general, $\kappa \rhd \lambda \neq \lambda \rhd \kappa$.*
5. Commutativity under consistency: *If $\kappa^{\downarrow K\cap L} = \lambda^{\downarrow K\cap L}$, then $\kappa \rhd \lambda = \lambda \rhd \kappa$.*
6. Non-associativity: *Suppose $\mu(M)$ is a probability distribution, then, in general, $(\kappa \rhd \lambda) \rhd \mu \neq \kappa \rhd (\lambda \rhd \mu)$.*
7. Associativity under a special condition: *Suppose $\mu(M)$ is a probability distribution, and suppose $L \supset (K \cap M)$. Then, $(\kappa \rhd \lambda) \rhd \mu = \kappa \rhd (\lambda \rhd \mu)$, if the right hand side formula is defined.*
8. Stepwise composition: *Suppose M is such that $(K \cap L) \subseteq M \subseteq L$. Then $(\kappa \rhd \lambda^{\downarrow M}) \rhd \lambda = \kappa \rhd \lambda$.*
9. Simple marginalization: *Suppose M is such that $(K \cap L) \subseteq M \subseteq K \cup L$. Then $(\kappa \rhd \lambda)^{\downarrow M} = \kappa^{\downarrow K\cap M} \rhd \lambda^{\downarrow K\cap M}$.*
10. Maximum entropy extension: *If $\kappa^{\downarrow K\cap L} = \lambda^{\downarrow K\cap L}$, then $\kappa \rhd \lambda = \arg \max_{\nu \in \Pi(\kappa,\lambda)} \mathbf{H}(\nu)$, where $\Pi(\kappa,\lambda)$ is the set of all common extensions of κ and λ, and $\mathbf{H}(\nu)$ is a Shannon entropy of ν.*

3 Compositional Models

To avoid some technical problems and the necessity of repeating some assumptions to excess, let us make three conventions.

In this and the next section we will consider a system of n distributions $\kappa_1(K_1)$, $\kappa_2(K_2),\ldots,\kappa_n(K_n)$. Therefore, whenever we speak about a distribution κ_k, if not explicitly specified otherwise, the distribution κ_k will always be assumed to be a distribution of variables K_k. Thus, for example, $\kappa_2 \triangleright \kappa_1 \triangleright \kappa_4$, if it is defined, will determine the distribution of variables $K_1 \cup K_2 \cup K_4$.

Our second convention pertains to the fact that the operator of composition is neither commutative nor associative. To avoid having to write too many parentheses in the formulas, in the rest of the paper we will apply the operators from left to right. Thus

$$\kappa_1 \triangleright \kappa_2 \triangleright \kappa_3 \triangleright \ldots \triangleright \kappa_n = (\ldots((\kappa_1 \triangleright \kappa_2) \triangleright \kappa_3) \triangleright \ldots \triangleright \kappa_n),$$

and the parentheses will be used only when we want to change this default ordering. Therefore, to construct a multidimensional distribution it is sufficient to determine a sequence – we call it a *generating sequence* – of oligodimensional distributions.

The third convention is of a rather technical nature. Since in the remaining part of the paper we are interested in a construction of multidimensional models, it is quite natural that we will always assume that all the models (compositions) we speak about are defined.

3.1 Perfect Sequences

Definition 2. A generating sequence of probability distributions $\kappa_1, \kappa_2, \ldots, \kappa_n$ is called *perfect* if all the distributions from this sequence are marginals of the distribution $(\kappa_1 \triangleright \kappa_2 \triangleright \ldots \triangleright \kappa_n)$, i.e., if for all $i = 1,2,\ldots,n$

$$(\kappa_1 \triangleright \kappa_2 \triangleright \ldots \triangleright \kappa_n)^{\downarrow K_i} = \kappa_i.$$

Notice that when defining a perfect sequence, let alone a generating sequence, we have not imposed any conditions on sets of variables for which the distributions were defined. For example, considering a generating sequence where one distribution is defined for a subset of variables of another distribution (i.e., $K_j \subset K_k$) is fully sensible and may provide some information about the resulting multidimensional distribution. If, e.g., $\kappa(u), \lambda(v), \mu(u,v,w)$ is a perfect sequence, it is quite obvious that

$$\kappa(u) \triangleright \lambda(v) \triangleright \mu(u,v,w) = \mu(u,v,w)$$

(because all the elements of a perfect sequence are marginals of the resulting distribution and therefore μ must be marginal to $\kappa \triangleright \lambda \triangleright \mu$). Nevertheless, it can happen that for some reason or another, it may be more advantageous to work with the model defined by the perfect sequence than just with the distribution μ. From this model one can immediately see that variables u and v are independent, which, not knowing the numbers defining the distribution, one cannot say about distribution μ.

Let us present two important properties on perfect sequences (Theorem 10.14 and Theorem 10.15 in [9]).

Proposition 2. *If a sequence of distributions $\kappa_1, \kappa_2, \ldots, \kappa_n$ is perfect, then*

$$\mathbf{H}(\kappa_1 \triangleright \kappa_2 \triangleright \ldots \triangleright \kappa_n) \geq \mathbf{H}(\nu)$$

for any $\nu \in \{\pi(K_1 \cup K_2 \cup \ldots \cup K_n) : \pi^{\downarrow K_i} = \kappa_i \;\; \forall i = 1, 2, \ldots, n\}$.

Proposition 3. *If a sequence of distributions $\kappa_1, \ldots, \kappa_n$ and its permutation $\kappa_{i_1}, \ldots, \kappa_{i_n}$ are both perfect, then $\kappa_1 \triangleright \kappa_2 \triangleright \ldots \triangleright \kappa_n = \kappa_{i_1} \triangleright \kappa_{i_2} \triangleright \ldots \triangleright \kappa_{i_n}$.*

From the point of view of practical applications it is important to know that each generating sequence can be transformed into a perfect sequence. The process of transformation is described in the following assertion proved in [9] (Theorem 10.9).

Proposition 4. *For any generating sequence $\kappa_1, \kappa_2, \ldots, \kappa_n$, the sequence $\pi_1, \pi_2, \ldots, \pi_n$ computed by the following process*

$$\begin{aligned}
\pi_1 &= \kappa_1, \\
\pi_2 &= \pi_1^{\downarrow K_2 \cap K_1} \triangleright \kappa_2, \\
\pi_3 &= (\pi_1 \triangleright \pi_2)^{\downarrow K_3 \cap (K_1 \cup K_2)} \triangleright \kappa_3, \\
&\vdots \\
\pi_n &= (\pi_1 \triangleright \ldots \triangleright \pi_{n-1})^{\downarrow K_n \cap (K_1 \cup \ldots K_{n-1})} \triangleright \kappa_n
\end{aligned}$$

is perfect and $\kappa_1 \triangleright \ldots \triangleright \kappa_n = \pi_1 \triangleright \ldots \triangleright \pi_n$.

From the theoretical point of view, this process is simple. Unfortunately, it need not be valid from the point of view of computational complexity. The process requires marginalization of models, which are distributions represented by generating sequences, and this may be computationally very expensive [6]. To avoid these computational problems we will use decomposable generating sequences introduced in the following paragraph.

3.2 Decomposable Sequences

We call a generating sequence $\kappa_1, \kappa_2, \ldots, \kappa_n$ *decomposable* if the corresponding sequence of variable sets $K_1, K_2, \ldots, K_n$ meets the *running intersection property* (RIP), i.e., if

$$\forall i = 2, \ldots, n \;\; \exists j (1 \leq j < i) \quad \left(K_i \cap (\bigcup_{k=1}^{i-1} K_k) \subseteq K_j \right).$$

The importance of these sequences follows, among others, from the following assertion [9].

Proposition 5. *If $\kappa_1, \kappa_2, \ldots, \kappa_n$ is a sequence of pairwise consistent probability distributions such that $K_1, \ldots, K_n$ meets RIP, then this sequence is perfect.*

The reader can notice, that if the sequence $K_1, K_2, \ldots, K_n$ in Proposition 4 meets RIP, then $K_3 \cap (K_1 \cup K_2)$ equals either $K_3 \cap K_1$ or $K_3 \cap K_2$. Similarly, $K_4 \cap (K_1 \cup K_2 \cup K_3)$ equals $K_4 \cap K_j$ for some $j \leq 3$. It means that, thanks to RIP, for all $i = 3, 4, \ldots, n$ the necessary marginal distributions

$$(\pi_1 \triangleright \ldots \triangleright \pi_{i-1})^{\downarrow K_i \cap (K_1 \cup \ldots \cup K_{i-1})}$$

can be computed from some π_j as $\pi_j^{\downarrow K_i \cap K_j}$, because $\pi_1, \ldots, \pi_{i-1}$ is a perfect sequence and therefore π_j is marginal to $\pi_1 \triangleright \ldots \triangleright \pi_{i-1}$. All this means that for this type of distributions the process of perfectization can be performed locally.

4 Conditioning

In this short section we will show that the operator of composition can also serve as a tool for computation of conditional distributions. Define a *degenerated* one-dimensional probability distribution $\pi_{|u;\alpha}$ as a distribution of variable u achieving probability 1 for value $u = \alpha$, i.e.,

$$\pi_{|u;\alpha}(x) = \begin{cases} 1 & \text{if } x = \alpha, \\ 0 & \text{otherwise.} \end{cases}$$

Now, consider a probability distribution $\kappa(K)$ for which $\{u, v\} \subset K$ and compute $(\pi_{|u;\alpha} \triangleright \kappa)^{\downarrow\{v\}}$. For any $y \in \mathbb{X}_v$

$$\begin{aligned}(\pi_{|u;\alpha} \triangleright \kappa)^{\downarrow\{v\}}(y) &= ((\pi_{|u;\alpha} \triangleright \kappa)^{\downarrow\{u,v\}})^{\downarrow\{v\}}(y) = (\pi_{|u;\alpha} \triangleright \kappa^{\downarrow\{u,v\}})^{\downarrow\{v\}}(y) \\ &= \sum_{x \in \mathbf{X}_u} \frac{\pi_{|u;\alpha}(x) \cdot \kappa^{\downarrow\{u,v\}}(x,y)}{\kappa^{\downarrow\{u\}}(x)} = \frac{\kappa^{\downarrow\{u,v\}}(\alpha, y)}{\kappa^{\downarrow\{u\}}(\alpha)} = \kappa(v = y | u = \alpha).\end{aligned}$$

Thus we have got that $\kappa(v|u = \alpha) = (\pi_{|u;\alpha} \triangleright \kappa)^{\downarrow\{v\}}$.

In the same way it can be shown for any $L \subseteq K \setminus \{u\}$ that $\left(\pi_{|u;\alpha} \triangleright \kappa\right)^{\downarrow L}$ is an $|L|$-dimensional conditional distribution κ under the condition that variable u attains value α, i.e., $\kappa(L|u = \alpha)$. Proceeding analogously even further we can get that for any $v \in K \setminus (L \cup \{u\})$ and $\beta \in \mathbb{X}_v$

$$\kappa(L|u = \alpha, v = \beta) = \left(\pi_{|v;\beta} \triangleright \left(\pi_{|u;\alpha} \triangleright \kappa\right)\right)^{\downarrow L}$$

is a conditional distribution for variables from L given that variables u and v attain values α and β, respectively.

5 Local Computations

By local computations we understand a process based on the ideas published in the famous paper by Lauritzen and Spiegelhalter [15]. Here we have especially in mind the idea that when computing the required conditional probability, one performs computations only on the system of marginal distributions defining the decomposable model. It means that during the computational process one does not need to store more data than what is necessary to store for the decomposable model.

In the preceding paragraph we showed that the conditional distribution can be expressed as a composition of a degenerated distribution with the distribution for which we want to compute the conditional distribution. So, let us assume that a distribution κ is decomposable, i.e.,

$$\kappa = \kappa^{\downarrow K_1} \triangleright \kappa^{\downarrow K_2} \triangleright \ldots \triangleright \kappa^{\downarrow K_n}$$

for a sequence $K_1, K_2, \ldots, K_n$ meeting RIP, and we want to compute, say, $\kappa(L|u = \alpha, v = \beta) = \big(\pi_{|v;\beta} \triangleright \big(\pi_{|u;\alpha} \triangleright \kappa\big)\big)^{\downarrow L}$.

For this, we will have to take advantage of the famous fact (an immediate consequence of the existence of a join tree, see [1]) that if $K_1, K_2, \ldots, K_n$ can be ordered to meet RIP, then there are many of such orderings, and for each $k \in \{1, 2, \ldots, n\}$, at least one of them starts with K_k. Therefore, thanks to Proposition 3, we can consider any of these orderings. So, consider any K_k for which $u \in K_k$, and find the ordering meeting RIP which starts with this K_k. Without loss of generality let it be $K_1, K_2, \ldots, K_n$ (so, $u \in K_1$).

Thus, our goal is to compute in the first step $\big(\pi_{|u;\alpha} \triangleright \kappa\big)$

$$\pi_{|u;\alpha} \triangleright \kappa = \pi_{|u;\alpha} \triangleright \big(\kappa^{\downarrow K_1} \triangleright \kappa^{\downarrow K_2} \triangleright \ldots \triangleright \kappa^{\downarrow K_n}\big).$$

Now applying $(n-1)$ times *Associativity under a special condition* (Property 7 of Proposition 1) we get (recall that we selected the RIP ordering, for which $u \in K_1$)

$$\begin{aligned}\pi_{|u;\alpha} \triangleright \big(\kappa^{\downarrow K_1} \triangleright \kappa^{\downarrow K_2} \triangleright \ldots \triangleright \kappa^{\downarrow K_n}\big) &= \pi_{|u;\alpha} \triangleright \big(\kappa^{\downarrow K_1} \triangleright \kappa^{\downarrow K_2} \triangleright \ldots \triangleright \kappa^{\downarrow K_{n-1}}\big) \triangleright \kappa^{\downarrow K_n} \\ &= \ldots = \pi_{|u;\alpha} \triangleright \kappa^{\downarrow K_1} \triangleright \kappa^{\downarrow K_2} \triangleright \ldots \triangleright \kappa^{\downarrow K_n},\end{aligned}$$

from which the following computationally local process (see Proposition 4 and the comment in Section 3.2)

$$\begin{aligned}\nu_1 &= \pi_{|u;\alpha} \triangleright \kappa^{\downarrow K_1}, \\ \nu_2 &= \nu_1^{\downarrow K_2 \cap K_1} \triangleright \kappa^{\downarrow K_2}, \\ \nu_3 &= (\nu_1 \triangleright \nu_2)^{\downarrow K_3 \cap (K_1 \cup K_2)} \triangleright \kappa^{\downarrow K_3}, \\ &\;\;\vdots \\ \nu_n &= (\nu_1 \triangleright \ldots \triangleright \nu_{n-1})^{\downarrow K_r \cap (K_1 \cup \ldots K_{n-1})} \triangleright \kappa^{\downarrow K_n},\end{aligned}$$

yields a perfect decomposable sequence $\nu_1, \ldots, \nu_n$, such that $\pi_{|u;\alpha} \triangleright \kappa = \nu_1 \triangleright \ldots \triangleright \nu_n$.

Now, it has remained to compute in the second step the required

$$\kappa(L|u=\alpha, v=\beta) = \left(\pi_{|v;\beta} \triangleright \left(\pi_{|u;\alpha} \triangleright \kappa\right)\right)^{\downarrow L} = \left(\pi_{|v;\beta} \triangleright (\nu_1 \triangleright \ldots \triangleright \nu_n)\right)^{\downarrow L}.$$

Thanks to decomposability of the sequence $\nu_1, \ldots, \nu_n$ the computations will proceed in the same way as in the first step. First, distributions ν_i will be reordered in the way that $\nu_{j_1}, \ldots, \nu_{j_n}$ meet RIP and variable v is among the variables for which ν_{j_1} is defined. Then we can, as in the first step, due to *Associativity under a special condition* deduce that

$$\pi_{|v;\beta} \triangleright (\nu_{j_1} \triangleright \nu_{j_2} \triangleright \ldots \triangleright \nu_{j_n}) = (\pi_{|v;\beta} \triangleright \nu_{j_1}) \triangleright \nu_{j_2} \triangleright \ldots \triangleright \nu_{j_n},$$

which can be, again, converted into a perfect sequence by the computationally local process of perfectization.

6 Heuristics for Model Construction

The reader interested in other theoretical issues concerning the operator of composition and perfect sequence models is referred to [9] and the papers cited there. Here, we want to briefly introduce a possible heuristic way to create a perfect sequence model from a data file – see Figure 1. For a more detailed description of this process, as well as for an example of its application to a small data file, the reader is referred to [8]. Notice that the described process is fully driven by an expert, and thus the following decisions must be made by a human expert:

1. Selection of oligodimensional distributions at the beginning of the whole process.
2. Decision which type of "refinement" procedure should be chosen (detailed explanation is given below).
3. Stopping rule.

As it can be seen from the diagram in Figure 1, the process is initiated with definition of a system of oligodimensional distributions. Regarding the fact that the process cyclically employs steps of *verification* and *refinement*, during which this initial system is gradually changed, the result is fairly independent of the initial selection. For example, starting with all two-dimensional distributions may be quite reasonable (for application to small data files with a limited number of variables one can consider a possibility to start with three-dimensional marginal distributions). In other situations, an expert can select the initial marginal distributions from which the model should be constructed. Generally, we propose to select distributions carrying a greater amount of information. This idea is supported by the following assertion, proved in [7] (Corollary 1.). It claims that the higher information content of a perfect sequence, the better approximation of the unknown distribution.

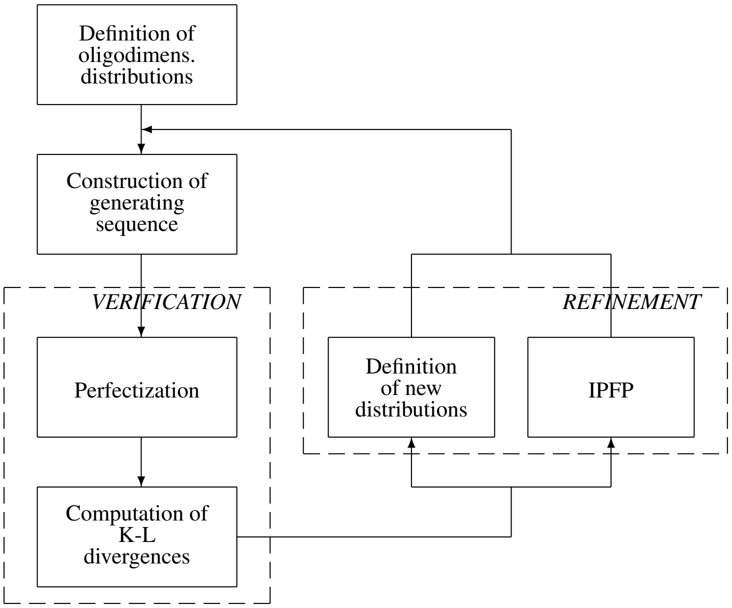
Definition of oligodimens. distributions
Construction of generating sequence
VERIFICATION
Perfectization
Computation of K-L divergences
REFINEMENT
Definition of new distributions
IPFP

Fig. 1 Process of model construction

Proposition 6. *Consider an arbitrary distribution κ, and a generating sequence consisting of its marginals $\kappa^{\downarrow K_1}, \kappa^{\downarrow K_2}, \ldots, \kappa^{\downarrow K_n}$. If this generating sequence is perfect, then*

$$Div(\kappa \| \kappa(x_{K_1}) \triangleright \ldots \triangleright \kappa(x_{K_n})) = I(\kappa) - I(\kappa(x_{K_1}) \triangleright \ldots \triangleright \kappa(x_{K_n})),$$

where the Information content *$I(\pi)$ of a distribution $\pi(J)$ is the Kullback-Leibler divergence of π and a product distribution of its one-dimensional marginal distributions:*

$$I(\pi) = Div(\pi \| \prod_{u \in J} \pi^{\downarrow\{u\}}) = \sum_{x \in \mathbb{X}_J} \pi(x) \log \frac{\pi(x)}{\prod_{u \in J} \pi^{\downarrow\{u\}}(x^{\downarrow\{u\}})}.$$

Let us stress that the information content is a generalization of a *Shannon mutual information*, which will be used in the algorithm further in this text, and which is for two disjoint (nonempty) $L, M \subset J$ defined by the formula

$$MI_\pi(K;L) = \sum_{x \in \mathbb{X}_K} \sum_{y \in \mathbb{X}_L} \pi^{\downarrow K \cup L}(x,y) \log \frac{\pi^{\downarrow K \cup L}(x,y)}{\pi^{\downarrow K}(x) \cdot \pi^{\downarrow L}(y)}.$$

If we want to construct a perfect sequence model approximating an unknown distribution κ, we have to aim at getting the model with the highest possible information content (under the assumption that the oligodimensional distributions, which the perfect sequence consists of, are marginals of the approximated distribution). In [8] we have published the following heuristic algorithm producing a sub-optimal generating sequence from a system of oligodimensional distributions.

Algorithm

```
Input: System of low-dimensional distributions $\kappa_1(K_1), \dots \kappa_n(K_n)$.
Initialization: Select a variable $u$ and a distribution $\kappa_j$ such that
    $u \in K_j$. Set $\pi_1 := \kappa_j^{\downarrow\{u\}}$, $L := \{u\}$ and $k := 1$.
Computational Cycle: While $K_1 \cup \dots \cup K_n \setminus L \neq \emptyset$ perform the following
    3 steps:
    1. for all $j = 1, \dots, n$ and all $v \in K_j \setminus L$ compute the mutual information
           $MI_{\kappa_j}(v; K_j \cap L)$.
    2. Fix $j$ and $v$ for which $MI_{\kappa_j}(v; K_j \cap L)$ achieved its maximal value.
    3. Increase $k$ by 1. Set $\pi_k := \kappa_j^{\downarrow(K_j \cap L) \cup \{v\}}$ and $L := L \cup \{m\}$.
Output: Generating sequence $\pi_1, \pi_2, \dots, \pi_k$.
```

What can be said about the resulting generating sequence $\pi_1, \pi_2, \dots, \pi_k$? Distribution $\nu = \pi_1 \triangleright \pi_2 \triangleright \dots \triangleright \pi_k$ is a probability distribution of variables $K_1 \cup K_2 \cup \dots \cup K_n$. The algorithm realizes a greedy (therefore very efficient) process, which seeks to find a sequence utilizing the information content of individual oligodimensional distributions in a maximal possible way. The result is a generating sequence which, unfortunately, need not be perfect. It means that some of the input distributions are not marginals of the resulting multidimensional model. As a rule, the expert (the model constructor) has to accept some deviations of the model marginals from the input oligodimensional distributions. To decide whether the obtained deviations are acceptable, i.e., whether the whole model construction process depicted in Figure 1 should be terminated, the expert must be provided with some additional information. To get it, the process employs the perfectization procedure described in Proposition 4. Then it is possible to compare original oligodimensional distributions with the corresponding marginals defined by the model. The comparison may be done with the help of Kullback-Leibler divergence; as already said above, its value equals 0 *iff* $\pi = \kappa$, otherwise it is always positive. Therefore, the lower this value, the closer κ to π. The goal of this step is to find all the marginal distributions which are unacceptably distorted by the model. If there is no such a marginal distribution, the process is terminated. In the opposite case, the expert proposes to perform another cycle of the whole process with a modified system of oligodimensional distributions. The described process then proceeds so that several original distributions are substituted with one *a-little-bit-more*-dimensional one in the *refinement* step.

As the reader can see from Figure 1, there are two possibilities to get these new distributions. If it is possible (i.e., the data file is large enough) the expert can decide to get them as estimations from the given data file (going along the left branch of the *refinement* box in Figure 1). However, if the data file is too small to get reliable estimations (which may happen easily if one needs to substitute several distributions with a distribution whose dimensionality is high – let us say, 6 or more), then one can take advantage of the well-known Iterative Proportional Fitting Procedure (IPFP) (see [2]; for its effective implementation, which makes it possible to compute distributions of pretty high dimensions, see [4]). In this way, when all the desired substitutions are realized, a new system of oligodimensional distributions is set up, to which the heuristic algorithm for generating sequence construction is again applied. The described cycle is repeated until the expert decides that a suitable multidimensional model representing (approximating) all the required oligodimensional distributions has been achieved.

Let us stress once more that the process shown in Figure 1 is fully controlled by the expert. The more cycles of the process are performed, the higher dimensions of the input distributions are considered. If the expert had continued ad absurdum, the process would have, in fact, finished with an application of IPFP to all of the initial oligodimensional distributions (which is, as a rule, computationally intractable in practical situations).

7 Conclusions

In this paper we summarized most of the practically oriented properties of compositional models and showed that they can be applied to multidimensional distribution representation. We also showed that conditional distributions can be computed as a composition of one or several degenerated distributions with the respective model, and that these computations can be, for decomposable models, performed locally.

Let us, now, mention another advantage of perfect compositional models that is important for another computational process that was not discussed in this paper. We have in mind the process of marginalization. Since the perfect model is composed of a system of its marginal distributions, it is not difficult to show on examples that there are number of situations when marginalization in a compositional model is simple but the same process in the corresponding Bayesian network is either computationally very expensive or even intractable. This advantageous property of compositional models is employed in algorithms described in [6].

As the last remark, let us mention that compositional models where introduced not only within the framework of probability theory, but also in possibility theory [16], the theory of belief functions [12], and recently also for the Shenoy's valuation-Based Systems [11]. Thus, most of the results presented in this paper can easily be extended into the above mentioned theoretical frameworks. For example, the content of Sections 4 and 5 have originally been published for belief functions [10], not for probability theory.

Acknowledgements This work was supported in part by the National Science Foundation of the Czech Republic by grant no. GACR 403/12/2175.

References

1. Beeri, C., Fagin, R., Maier, D., Yannakakis, M.: On the Desirability of Acyclic Database Schemes, *J. ACM*, vol. 30, no. 3, pp. 479–513 (1983)
2. Deming, W.E., Stephan, F.F.: On a least square adjustment of a sampled frequency table when the expected marginal totals are known. *Ann. Math. Stat.* 11, 427–444 (1940)
3. Jensen, F. V.: Bayesian Networks and Decision Graphs. IEEE Computer Society Press, New York (2001)
4. Jiroušek, R.: Solution of the Marginal Problem and Decomposable Distributions. *Kybernetika* 27, 5, 403-412 (1991)
5. Jiroušek, R.: Composition of probability measures on finite spaces. Uncertainty in Artificial Intelligence: Proceedings of the 13th Conference (UAI-97). D. Geiger and Prakash P. Shenoy (eds.), Morgan Kaufmann, 274–281 (1997)
6. Jiroušek, R.: Marginalization in Composed Probabilistic Models. Uncertainty in artificial intelligence. Boutilier, Craig, Goldszmidt, Moiss (eds.), San Francisco, Morgan Kaufmann Publishers, 301-308 (2000)
7. Jiroušek, R.: On approximating multidimensional probability distributions by compositional models. Proceedings of the 3rd International Symposium on Imprecise Probabilities and Their Application. Jean-Marc Bernard, Teddy Seidenfeld, Marco Zaffalon (eds.), Carleton Scientific, 305-320 (2003)
8. Jiroušek, R.: Data-Based Construction of Multidimensional Probabilistic Models with MUDIM. *Logic Journal of the IGPL* 14, 3, 501-520 (2006)
9. Jiroušek, R.: Foundations of compositional model theory. *International Journal of General Systems* 40, 6, 623–678 (2011)
10. Jiroušek, R.: Local Computations in Dempster-Shafer Theory of Evidence. To appear in: Int. J. Approx. Reason. doi: 10.1016/j.ijar.2012.06.012
11. Jiroušek, R., Shenoy, P. P.: Compositional models in valuation-based systems. *Working Paper No. 325*, School of Business, University of Kansas, Lawrence, KS (2011)
12. Jiroušek, R., Vejnarová, J., Daniel, M.: Compositional models of belief functions. Proc. of the Fifth Int. Symposium on Imprecise Probability: Theories and Applications. G. de Cooman, J. Vejnarová, M. Zaffalon (eds.), Praha, 243–252 (2007)
13. Kullback, S., Leibler, R. A.: On information and sufficiency. *Annals of Mathematical Statistics* 22, 76–86 (1951)
14. Lauritzen, S. L.: Graphical models. Oxford University Press (1996)
15. Lauritzen, S. L., Spiegelhalter, D.: Local computation with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society series B* 50, 157–224 (1988)
16. Vejnarová, J.: Composition of possibility measures on finite spaces: preliminary results. Proceedings of 7th International Conference on Information Processing and Management of Uncertainty in Knowledge-based Systems IPMU'98. B. Bouchon-Meunier, R.R. Yager (eds.), Editions E.D.K. Paris, 25–30 (1998)