On Approximating Multidimensional Probability Distributions by Compositional Models*

R. JIROUŠEK
Academy of Science, Czech Republic

Abstract

Because of computational problems, multidimensional probability distributions must be approximated by distributions which can be defined by a reasonable number of parameters. As a rule, distributions with a special dependence structure (i.e., complying with a system of conditional independence relations) are considered; graphical Markov models and especially Bayesian networks are often used. This paper proposes application of compositional models for this purpose. In addition to a theoretical background, a heuristic algorithm solving one part of a model learning process is presented. Its basic idea, construction of an approximation exploiting informational content of given low-dimensional distributions in a maximal possible way, was proposed by Albert Perez as early as in 1977.

Keywords

multidimensional distributions, approximations, conditional independence, operator of composition

1 Introduction

Data-driven methods for probability model construction usually suffer from a lack of data. This is why one must always keep in mind that any probability estimate is imprecise and the more probabilities, the less precise their estimates. Moreover, it would be absurd to try to get estimates of (let us say) 2^{50} probabilities defining a 50-dimensional distribution (of binary variables) from a file whose size is only several Mbytes. Such an effort would also be in contradiction with the *Minimum Description Length* principle often employed in the field of AI. Therefore, application of probabilistic models to problems of practice, when the dimensionality

*This work has been supported in part by GA AV ČR, under grant A2075302.

of considered multidimensional probability distributions is expressed in hundreds rather than tens, quite naturally leads to the necessity of approximations.

The present paper proposes to look for an approximation of a probability distribution in a class of so-called *compositional models* (CM), which is an alternative apparatus to that usually called *Graphical Markov Modeling* (GMM). GMM is used as a general term describing any of the approaches representing multidimensional probability distributions by means of graphs and systems of quantitative parameters like Bayesian networks (BN), decomposable and graphical models, influence diagrams and chain graph models.

The main idea of CM is the same as that of GMM: not to strive for estimating multidimensional distribution but only its oligo-dimensional marginals, from which the multidimensional model is subsequently composed. In a way this model resembles a jigsaw puzzle that has a great number of parts, each bearing a local piece of a picture, and the goal is to find how to assemble them in a way that the global picture makes sense, reflecting all the individual small parts. Naturally, the whole task can be split into two subproblems: how to find which oligo-dimensional distributions are to be estimated and how to compose them into a multidimensional model. Though the present paper concentrates exclusively on the latter one, let us mention that, to be consistent with the apparatus employed in this paper, the problem of selection of oligodimensional distributions should be solved with the help of information theoretic quantities; distributions with the highest informational content (see section 5) should be selected.

Before introducing the apparatus of CM let us mention that both GMM and CM are based on the very idea published by Albert Perez as early as 1977 in his unfortunately neglected paper [10]. In this paper Perez calls these probability distributions *dependence structure simplification approximations* and studies increase of risk connected with statistical decision problem when, instead of Bayes optimal solution, ε-Bayes optimal solution (ie., Bayes optimal with respect to ε-approximation) is accepted.

2 Notation

In this text, we will deal with a finite system of finite-valued random variables. Let N be an arbitrary finite index set, $N \neq \emptyset$. Each variable from $\{X_i\}_{i\in N}$ is assumed to have a finite (non-empty) set of values $\mathbf{X}_i$. Distributions of these variables will be denoted by Greek letters (π, κ); thus for $K \subseteq N$, we can consider a distribution $\pi((X_i)_{i\in K})$. To make the formulae more lucid, the following simplified notation will be used. Symbol $\pi(x_K)$ will denote both a $|K|$-dimensional distribution and a value of a probability distribution π (when several distributions are considered, we shall distinguish between them by indices), which is defined for variables $(X_i)_{i\in K}$ at a combination of values x_K; x_K thus represents a $|K|$-dimensional vector of values of variables $\{X_i\}_{i\in K}$. Analogously, we shall also denote the set of all these

vectors $\mathbf{X}_K$:

$$\mathbf{X}_K = \times_{i\in K}\mathbf{X}_i.$$

For a probabilistic distribution $\pi(x_K)$ and $J \subset K$ we will often consider a *marginal distribution* $\pi(x_J)$ of distribution $\pi(x_K)$, which can be computed by

$$\pi(x_J) = \sum_{x_{K\setminus J}\in\mathbf{X}_{K\setminus J}} \pi(x_K) = \sum_{x_{K\setminus J}\in\mathbf{X}_{K\setminus J}} \pi(x_{K\setminus J}, x_J).$$

In this simple formula we have introduced a notation used throughout this article: a vector x_K is composed of two subvectors $x_{K\setminus J}$ and x_J, where x_J is a *projection* of x_K into $\mathbf{X}_J$, and, analogously $x_{K\setminus J}$ is a projection of x_K into $\mathbf{X}_{K\setminus J}$. For computation of marginal distributions we need not exclude situations when $J = \emptyset$. In accordance with the above-introduced formula we get $\pi(x_\emptyset) = 1$.

In some situations we will want to stress that we are dealing with a marginal distribution of a distribution π; we will use symbol $\pi^{(J)}$ to denote the marginal distribution of π for variables $(X_i)_{i\in J}$. That is, for $J \subseteq K$ and a distribution $\pi(x_K)$,

$$\pi^{(J)} = \pi(x_J).$$

For a distribution $\pi(x_K)$ and two disjoint subsets $J, L \subseteq K$ we will also speak about a *conditional distribution* $\pi(x_J|x_L)$, which is, for each fixed $x_L \in \mathbf{X}_L$, a $|J|$-dimensional probability distribution, for which $\pi(x_J|x_L)\pi(x_L) = \pi(x_{J\cup L})$. (Notice that this definition is ambiguous if $\pi(x_L) = 0$ for some combination(s) of values $x_L \in \mathbf{X}_L$.) The reader can immediately see that if $J = \emptyset$ then $\pi(x_J|x_L) = 1$, and if $L = \emptyset$ then $\pi(x_J|x_L) = \pi(x_J)$.

Consider $K \subseteq L \subseteq N$ and a probability distribution $\pi(x_K)$. With $\Pi^{(L)}$ we shall denote the set of all probability distributions defined for variables X_L. Similarly, $\Pi^{(L)}(\pi)$ will denote the system of all *extensions* of the distribution π to L-dimensional distributions:

$$\Pi^{(L)}(\pi) = \left\{\kappa \in \Pi^{(L)} : \kappa(x_K) = \pi(x_K)\right\},$$

(where $\kappa(x_K)$ naturally denotes the marginal distribution of κ for variables X_K). Having a system

$$\Xi = \{\pi_1(x_{K_1}), \pi_2(x_{K_2}), \ldots, \pi_n(x_{K_n})\},$$

of oligo-dimensional distributions ($K_1 \cup \ldots \cup K_n \subseteq L$), the symbol $\Pi^{(L)}(\Xi)$ denotes the system of distributions that are extensions of all the distributions from Ξ:

$$\Pi^{(L)}(\Xi) = \left\{\kappa \in \Pi^{(L)} : \kappa^{(K_i)} = \pi_i \ \ \forall i = 1, \ldots, n\right\} = \bigcap_{i=1}^{n} \Pi^{(L)}(\pi_i).$$

3 Operator of composition

To be able to compose low-dimensional distributions to get a distribution of a higher dimension we will introduce an *operator of composition*.

To make this construction clear from the very beginning, let us stress that it is just a generalization of the idea of computing the three-dimensional distribution from two two-dimensional ones introducing the conditional independence:

$$\pi(x_1,x_2)\triangleright\kappa(x_2,x_3)=\frac{\pi(x_1,x_2)\kappa(x_2,x_3)}{\kappa(x_2)}=\pi(x_1,x_2)\kappa(x_3|x_2).$$

Consider two probability distributions $\pi(x_K)$ and $\kappa(x_L)$, such that $\kappa(x_{L\cap K})$ dominates[1] $\pi(x_{L\cap K})$; in symbol: $\pi(x_{L\cap K})\ll\kappa(x_{L\cap K})$. The *composition* of these two distributions is defined by the formula

$$\pi(x_K)\triangleright\kappa(x_L)=\frac{\pi(x_K)\kappa(x_L)}{\kappa^{(L\cap K)}}.$$

Since we assume $\pi^{(L\cap K)}\ll\kappa^{(L\cap K)}$, if for any $x\in\mathbf{X}_{(L\cap K)}$ $\kappa^{(L\cap K)}(x)=0$ then there is a product of two zeros in the nominator and we take $0.0/0=0$. If $L\cap K=\emptyset$ then $\kappa^{(L\cap K)}=1$ and the formula degenerates to a simple product of π and κ.

Let us stress that in the case $\pi^{(L\cap K)}\not\ll\kappa^{(L\cap K)}$, the expression $\pi\triangleright\kappa$ remains undefined.

Thus, the formal definition of the operator $\triangleright$ is as follows.

Definition 1 *For two arbitrary distributions $\pi\in\Pi^{(K)}$ and $\kappa\in\Pi^{(L)}$ their* composition *is given by the following formula*

$$\pi(x_K)\triangleright\kappa(x_L)=\begin{cases}\frac{\pi(x_K)\kappa(x_L)}{\kappa(x_{K\cap L})} & \text{if } \pi(x_{K\cap L})\ll\kappa(x_{K\cap L}),\\ \text{undefined} & \text{otherwise}.\end{cases}$$

The following simple assertion proven in [5] answers the question: what is the result of the composition of two distributions?

Theorem 1 *If $\pi(x_{L\cap K})\ll\kappa(x_{L\cap K})$ (i.e., if $\pi(x_K)\triangleright\kappa(x_L)$ is defined) then $\pi(x_K)\triangleright\kappa(x_L)$ is a probability distribution from $\Pi^{(L\cup K)}(\pi)$, i.e., it is a probability distribution of $X_{K\cup L}$ and its marginal distribution for variables X_K equals π: $(\pi\triangleright\kappa)(x_K)=\pi(x_K)$.*

An importance of this operator arises from the fact that, when applied iteratively, it defines a multidimensional distribution from a system of low-dimensional ones.

[1] The concept of dominance (or absolute continuity) $\pi\ll\kappa$ in finite case simplifies to

$$\forall x\in\mathbf{X}\ \ (\kappa(x)=0\Longrightarrow\pi(x)=0)\,.$$

4 Generating sequences

Let us now consider a system of n low-dimensional distributions $\pi_1(x_{K_1}), \pi_2(x_{K_2}), \ldots, \pi_n(x_{K_n})$, and start studying a distribution $\pi_1 \triangleright \pi_2 \triangleright \ldots \triangleright \pi_n$, which (if defined) is a distribution of variables $X_{K_1 \cup K_2 \cup \ldots \cup K_n}$. Regarding the fact that the operator is neither commutative nor associative, let us stress that we always apply the operators from left to right:

$$\pi_1 \triangleright \pi_2 \triangleright \ldots \triangleright \pi_n = (\ldots((\pi_1 \triangleright \pi_2) \triangleright \pi_3) \triangleright \ldots \triangleright \pi_n).$$

Therefore, in order to construct a multidimensional distribution it is sufficient to determine a sequence – we call it a *generating sequence* – of low-dimensional distributions.

Example 1 *In agreement with what has just been said, the generating sequence*

$$\pi_1(x_1,x_3), \pi_2(x_3,x_5), \pi_3(x_1,x_4,x_5,x_6), \pi_4(x_2,x_5,x_6)$$

defines distribution

$$\begin{aligned}(\pi_1 \triangleright \pi_2 \triangleright \pi_3 \triangleright \pi_4)&(x_1,x_2,x_3,x_4,x_5,x_6)\\ &= \big((\pi_1(x_1,x_3) \triangleright \pi_2(x_3,x_5)) \triangleright \pi_3(x_1,x_4,x_5,x_6)\big) \triangleright \pi_4(x_2,x_5,x_6)\\ &= \pi_1(x_1,x_3)\pi_2(x_5|x_3)\pi_3(x_4,x_6|x_1,x_5)\pi_4(x_2|x_5,x_6).\end{aligned}$$ ◇

Not all generating sequences are equally efficient in their representations of multidimensional distributions. Among them, the so-called perfect sequences hold an important position.

Definition 2 *A generating sequence of probability distributions $\pi_1, \pi_2, \ldots, \pi_n$ is called* perfect *if for all $k = 2, \ldots, n$ distributions $\pi_1 \triangleright \ldots \triangleright \pi_k$ are defined and*

$$\pi_1 \triangleright \ldots \triangleright \pi_k = \pi_k \triangleright (\pi_1 \triangleright \ldots \triangleright \pi_{k-1}).$$

This definition enables us to check whether a generating sequence is perfect[2] but one can hardly see from it the importance of perfect sequences. This importance becomes clearer from the following characterization theorem (Theorem 2 in [7]).

Theorem 2 *A sequence of distributions $\pi_1, \pi_2, \ldots, \pi_n$ is perfect* iff *all the distributions from this sequence are marginals of the distribution $(\pi_1 \triangleright \pi_2 \triangleright \ldots \triangleright \pi_n)$.*

What is the main message conveyed by this characterization theorem? Considering that low-dimensional distributions π_k are carriers of local information, the constructed multidimensional distribution represents global information, faithfully reflecting all of the local input.

Let us briefly summarize the main properties of distributions represented by perfect sequences and their relation to the well-known concepts of GMM.

[2] A sequence is perfect *iff* for all $k = 2, \ldots, n$, $(\pi_1 \triangleright \ldots \triangleright \pi_{k-1})^{(K_n \cap (K_1 \cup \ldots \cup K_{i-1}))} = \pi_k^{(K_n \cap (K_1 \cup \ldots \cup K_{i-1}))}$.

(i) It was shown that perfect sequences are equivalent to BNs in the sense that any distribution representable by a perfect sequence can be represented by BN (and vice versa) and both of these strucures are defined with the same number of parameters – probabilities (for details see [7]) .

(ii) In analogy to BN, for each distribution represented by a perfect sequence a list of conditional independence relations holds true. For a BN, one can read all these relations from its graph by the famous d-separation criterion. How to determine them for CM was shown in [8].

(iii) Let us stress that whether a generating sequence is perfect does not depend only on structural properties (those corresponding to sets $K_1,\ldots,K_n$ and their ordering), but also on probabilities. To make this remark clearer notice the two extreme sufficient conditions, guaranteeing perfecness of a generating sequence:

- **(a)** if distributions $\pi_1(x_{K_1}),\ldots,\pi_n(x_{K_n})$ are pairwise consistent ($\pi_i^{(K_i\cap K_j)} = \pi_j^{(K_1\cap K_j)}$) and the sequence $K_1,\ldots,K_n$ meets the *running intersection property*[3] then $\pi_1(x_{K_1}),\ldots,\pi_n(x_{K_n})$ is perfect;
- **(b)** if all the distributions $\pi_k(x_{K_k})$ are uniform then $\pi_1(x_{K_1}),\ldots,\pi_n(x_{K_n})$ is always perfect.

(iv) Distributions represented by perfect sequences are unique in the following sense: if two permutations $\pi_{i_1},\ldots,\pi_{i_n}$ and $\pi_{j_1},\ldots,\pi_{j_n}$ of a system of oligodimensional distributions are perfect then $\pi_{i_1}\triangleright\ldots\triangleright\pi_{i_n} = \pi_{j_1}\triangleright\ldots\triangleright\pi_{j_n}$. This property, somehow resembling decomposable distributions, is especially important for designing computational procedures.

(v) Notice that we have not imposed any conditions on sets K_k. For example, considering a generating sequence where one distribution is defined for a subset of variables of another distribution (ie., $K_j \subset K_k$) is fully sensible and may enrich a system of considered multidimensional distributions (cf. Algorithm in Section 6.3).

5 Information-theoretic notions

In Section 6 several notions characterizing probability distributions and their relationship will be used. The first is the well-known *Shannon entropy* defined (for $\pi\in\Pi^{(N)}$)

$$H(\pi) = -\sum_{x\in\mathbf{X}_N}\pi(x)\log\pi(x).$$

[3] $\forall k=2,\ldots,n\quad \exists j(1\le j<k)\quad K_k\cap(K_1\cup\ldots\cup K_{k-1})\subset K_j$.

Recall that for two disjoint index sets $K, L \subset N$ one can also define a *conditional entropy* $H(\pi(x_K|x_L)$ using the expression:

$$H(\pi(x_K|x_L)) \quad = \quad -\sum_{x\in\mathbf{X}_{K\cup L}} \pi(x)\log\pi(x_K|x_L).$$

To compare two distributions defined for the same system of variables (i.e. $\pi, \kappa \in \Pi^N$) we will use *Kullback-Leibler divergence* (in literature sometimes called I-divergence, or cross-entropy). It is in fact a relative entropy of the first distribution with respect to the other:

$$Div(\pi\|\kappa) = \begin{cases} \sum\limits_{x\in\mathbf{X}_N} \pi(x)\log\frac{\pi(x)}{\kappa(x)} & \text{if } \pi \ll \kappa, \\ +\infty & \text{otherwise.} \end{cases}$$

The reader can immediately see that if $\pi = \kappa$ then $Div(\pi\|\kappa) = 0$. It is a well-known property of Kullback-Leibler divergence (and not too difficult to be proven) that its value is always non-negative and equals 0 if and only if $\pi = \kappa$. (Recall also that this divergence is not symmetric, i.e., generally $Div(\pi\|\kappa) \neq Div(\kappa\|\pi)$.)

One of the fundamental notions of information theory is a *mutual information*. Having a distribution $\pi(x_N)$ and two disjoint subsets $K, L \subset N$, it expresses how much one group of variables X_K influences the other one – X_L. It is defined

$$MI_\pi(X_K; X_L) = \sum_{x_{K\cup L}\in\mathbf{X}_{K\cup L}} \pi(x_{K\cup L})\log\frac{\pi(x_{K\cup L})}{\pi(x_K)\pi(x_L)},$$

and equals 0 if and only if the groups of variables X_K and X_L are independent under the distribution π. Otherwise, it is always positive.

The last notion, which will be of great importance, but which is not as famous as Shannon entropy or mutual information, is an *informational content* of a distribution defined by the formula:

$$I(\pi) = \sum_{x\in\mathbf{X}_N} \pi(x)\log\frac{\pi(x)}{\prod\limits_{j\in N}\pi(x_j)}.$$

Notice that this formula is nothing but a Kullback-Leibler divergence of two distributions: $\pi(x_N)$ and $\prod_{j\in N}\pi(x_j)$. Therefore, it is always non-negative and equals 0 if and only if $\pi(x_N) = \prod_{j\in N}\pi(x_j)$. In fact, this value expresses how much individual variables are dependent under the distribution π. Therefore the higher this value, the more dependent the variables, and consequently, the greater amount of information carried by the distribution.

One can also immediately see that for a 2-dimensional distribution $\pi(x_1, x_2)$

$$I(\pi) = MI_\pi(X_1; X_2).$$

6 Approximations

Let us consider an arbitrary multidimensional distribution $\kappa \in \Pi^{(N)}$ and assume that for one reason or another we are looking for its approximation in the form of a compositional model. Such situations appear quite often in practical problems; κ can be, for example, a sample distribution of a large database, or it can be an unknown theoretical distribution, from which some data file has been generated. In any case, we need its approximation.

6.1 Criterion function

For a candidate compositional distribution $\pi = \pi_1 \triangleright \pi_2 \triangleright \ldots \triangleright \pi_n \in \Pi^{(N)}$, the Kullback-Leibler divergence $Div(\kappa\|\pi)$ will be used as a criterion function. Naturally, the smaller the value of the Kullback-Leibler divergence, the better approximation π.

For compositional models this divergence can be expressed in a special form, which enables us to analyze individual factors of the divergence. To make the formulae more transparent we will use the following notation: for each $i = 1, \ldots, n$ set K_i is split into two disjoint parts

$$R_i = K_i \setminus (K_1 \cup \ldots \cup K_{i-1}), \quad S_i = K_i \cap (K_1 \cup \ldots \cup K_{i-1}).$$

(Naturally, $R_1 = K_1$ and $S_1 = \emptyset$.) In the following computations we shall use a standard trick, according to which

$$\begin{aligned} \sum_{x \in \mathbf{X}_N} \kappa(x) \log \kappa(x_K) &= \sum_{x_K \in \mathbf{X}_K} \kappa(x_K) \log \kappa(x_K) \sum_{x_{N\setminus K} \in \mathbf{X}_{N\setminus K}} \kappa(x_{N\setminus K}|x_K) \\ &= \sum_{x_K \in \mathbf{X}_K} \kappa(x_K) \log \kappa(x_K) \end{aligned}$$

because $\sum_{x_{N\setminus K} \in \mathbf{X}_{N\setminus K}} \kappa(x_{N\setminus K}|x_K) = 1$. Thus, assuming $Div(\kappa\|\pi)$ is finite, we can compute

$$\begin{aligned} Div(\kappa\|\pi) &= \sum_{x \in \mathbf{X}_N} \kappa(x) \log \frac{\kappa(x)}{\pi_1(x_{K_1}) \triangleright \ldots \triangleright \pi_n(x_{K_n})} \\ &= \sum_{x \in \mathbf{X}_N} \kappa(x) \log \kappa(x) - \sum_{x \in \mathbf{X}_N} \kappa(x) \log \prod_{i=1}^{n} \pi_i(x_{R_i}|x_{S_i}) \\ &= -H(\kappa) - \sum_{i=1}^{n} \sum_{x \in \mathbf{X}_N} \kappa(x) \log \pi_i(x_{R_i}|x_{S_i}) \\ &= -H(\kappa) - \sum_{i=1}^{n} \sum_{x_{K_i} \in \mathbf{X}_{K_i}} \kappa(x_{K_i}) \log \pi_i(x_{R_i}|x_{S_i}) \end{aligned}$$

$$= -H(\kappa) + \sum_{i=1}^{n} \sum_{x_{K_i} \in \mathbf{X}_{K_i}} \kappa(x_{K_i}) \log \frac{\kappa(x_{R_i}|x_{S_i})}{\pi_i(x_{R_i}|x_{S_i})} - \sum_{i=1}^{n} \sum_{x_{K_i} \in \mathbf{X}_{K_i}} \kappa(x_{K_i}) \log \kappa(x_{R_i}|x_{S_i})$$

$$= -H(\kappa) + \sum_{i=1}^{n} Div(\kappa(x_{R_i}|x_{S_i}) \| \pi_i(x_{R_i}|x_{S_i})) + \sum_{i=1}^{n} H(\kappa(x_{R_i}|x_{S_i})).$$

Now, let us have a look at the meaning of the expression

$$\sum_{i=1}^{n} H(\kappa_i(x_{R_i}|x_{S_i})) - H(\kappa).$$

First, for each $i = 1, \ldots, n$ we get

$$\begin{aligned} H(\kappa_i(x_{R_i}|x_{S_i})) &= -\sum_{x_{K_i} \in \mathbf{X}_{K_i}} \kappa(x_{K_i}) \log \kappa(x_{R_i}|x_{S_i}) \\ &= -\sum_{x_{K_i} \in \mathbf{X}_{K_i}} \kappa(x_{K_i}) \log \frac{\kappa(x_{K_i})}{\kappa(x_{S_i})} \frac{\prod_{j \in K_i} \kappa(x_j)}{\prod_{j \in K_i} \kappa(x_j)} \\ &= -I(\kappa(x_{K_i})) + I(\kappa(x_{S_i})) + \sum_{j \in R_i} H(\kappa(x_j)). \end{aligned}$$

Since all sets R_i are mutually disjoint and their union is the whole set N we are getting

$$\begin{aligned} \sum_{i=1}^{n} H(\kappa_i(x_{R_i}|x_{S_i})) - H(\kappa) &= \sum_{i=1}^{n} (I(\kappa(x_{S_i})) - I(\kappa(x_{K_i}))) + \sum_{j \in N} H(\kappa(x_j)) - H(\kappa) \\ &= \sum_{i=1}^{n} (I(\kappa(x_{S_i})) - I(\kappa(x_{K_i}))) + I(\kappa). \end{aligned}$$

In this way we have deduced that

$$\begin{aligned} &Div(\kappa \| \pi) \\ &\quad = \sum_{i=1}^{n} Div(\kappa(x_{R_i}|x_{S_i}) \| \pi_i(x_{R_i}|x_{S_i})) + \sum_{i=1}^{n} (I(\kappa(x_{S_i})) - I(\kappa(x_{K_i}))) + I(\kappa), \end{aligned} \quad (1)$$

which is a result that is worth being formulated as a theorem.

Theorem 3 *Let a distribution $\kappa \in \Pi^{(N)}$ and a sequence of distributions $\pi_1(x_{K_1})$, $\pi_2(x_{K_2}), \ldots, \pi_n(x_{K_n})$, for which $\bigcup_{i=1}^{n} K_i = N$, be such that $Div(\kappa \| \pi_1 \triangleright \ldots \triangleright \pi_n)$ is finite. Then, denoting $\pi = \pi_1 \triangleright \pi_2 \triangleright \ldots \triangleright \pi_n$, for the Kullback-Leibler divergence $Div(\kappa \| \pi)$ the equation (1) holds true.*

So, the divergence of distributions κ and π consists of two parts. The first one

$$\sum_{i=1}^{n} Div(\kappa(x_{R_i}|x_{S_i})\|\pi_i(x_{R_i}|x_{S_i}))$$

describes the "local" difference between κ and π (more precisely it renders the difference between conditional distributions $\kappa(x_{R_i}|x_{S_i})$ and $\pi_i(x_{R_i}|x_{S_i})$), and the second part

$$I(\kappa) - \sum_{i=1}^{n} (I(\kappa(x_{K_i})) - I(\kappa(x_{S_i})))$$

describes the difference resulting from the application of a compositional model. As it will be shown below, in the case that $\kappa(x_{K_1}), \kappa(x_{K_2}), \ldots, \kappa(x_{K_n})$ is a perfect sequence, it is exactly a difference between the informational content of the distributions κ and $\kappa(x_{K_1}) \triangleright \ldots \triangleright \kappa(x_{K_n})$.

Corollary 1 *If for a distribution κ a generating sequence of its marginals $\kappa(x_{K_1})$, $\kappa(x_{K_2}), \ldots, \kappa(x_{K_n})$ is perfect then*

$$I(\kappa(x_{K_1}) \triangleright \kappa(x_{K_2}) \triangleright \ldots \triangleright \kappa(x_{K_n})) = \sum_{i=1}^{n} (I(\kappa(x_{K_i})) - I(\kappa(x_{S_i}))),$$

and therefore also

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

Proof. The first equation can immediately be obtained by substituting $\kappa(x_{K_1}) \triangleright \kappa(x_{K_2}) \triangleright \ldots \triangleright \kappa(x_{K_n})$ for both κ and π in equation (1), because then the Kullback-Leibler divergence must equal 0. The second one is a direct consequence of the first equality following from (1). □

6.2 Perfect sequence approximations

Problem of model learning in context of CM means that one wants to find a properly ordered system of oligodimensional distributions. It is evident from the expression (1) that the best approximations are defined by generating sequences consisting of distributions which are marginals[4] of the approximated distribution κ. In this case, namely, for all $i = 1, \ldots n$, $Div(\kappa(x_{R_i}|x_{S_i})\|\pi_i(x_{R_i}|x_{S_i}))$ equal 0 and the formula (1) simplifies to

$$Div(\kappa\|\pi) = I(\kappa) - \sum_{i=1}^{n} (I(\kappa(x_{K_i})) - I(\kappa(x_{S_i}))), \tag{2}$$

[4] In fact, it is enough when all $\pi_i(x_{R_i}|x_{S_i}) = \kappa(x_{R_i}|x_{S_i})$.

which does not depend on values of distributions π_i (quite naturally, because they are marginals of κ) but only on the system, or more precisely sequence, $K_1, K_2, \ldots, K_n$. In the following example we shall show that different orderings of the distributions in generating sequences can result in different values of the Kullback-Leibler divergence.

Example 2 *Consider a* 4*-dimensional distribution* $\kappa(x_1,x_2,x_3,x_4)$ *and its three marginal distributions denoted* π_1,π_2,π_3*:*

$$\pi_1(x_1,x_2) = \kappa(x_1,x_2), \quad \pi_2(x_2,x_3) = \kappa(x_2,x_3), \quad \pi_3(x_3,x_4) = \kappa(x_3,x_4).$$

Compute $Div(\kappa\|\pi)$ *and* $Div(\kappa\|\hat{\pi})$ *for* $\pi = \pi_1 \triangleright \pi_2 \triangleright \pi_3$ *and* $\hat{\pi} = \pi_1 \triangleright \pi_3 \triangleright \pi_2$*. For the first distribution it is*

$$\begin{aligned} Div(\kappa\|\pi) &= I(\kappa) - \big(I(\kappa(x_{\{1,2\}})) + I(\kappa(x_{\{2,3\}})) + I(\kappa(x_{\{3,4\}}))\big) \\ &\qquad + \big(I(\kappa(x_\emptyset)) + I(\kappa(x_{\{2\}})) + I(\kappa(x_{\{3\}}))\big) \\ &= I(\kappa) - I(\kappa(x_{\{1,2\}})) - I(\kappa(x_{\{2,3\}})) - I(\kappa(x_{\{3,4\}})), \end{aligned}$$

whereas for $\hat{\pi}$ *we get*

$$\begin{aligned} Div(\kappa\|\hat{\pi}) &= I(\kappa) - \big(I(\kappa(x_{\{1,2\}})) + I(\kappa(x_{\{3,4\}})) + I(\kappa(x_{\{2,3\}}))\big) \\ &\qquad + \big(I(\kappa(x_\emptyset)) + I(\kappa(x_\emptyset)) + I(\kappa(x_{\{2,3\}}))\big) \\ &= I(\kappa) - I(\kappa(x_{\{1,2\}})) - I(\kappa(x_{\{3,4\}})) - I(\kappa(x_{\{2,3\}})) + I(\kappa(x_{\{2,3\}})) \\ &= Div(\kappa\|\pi) + I(\kappa(x_{\{2,3\}})). \end{aligned}$$

The reader probably noticed that, for the sake of simplicity, we introduced a situation corresponding to a decomposable model. It is perhaps worth mentioning that even in this case it may happen that both of the sequences defining distributions π *and* $\hat{\pi}$ *are perfect. In correspondence with the assertion mentioned in Section 4 (item* (iv)*), it happens only when* $\pi = \hat{\pi}$ *and therefore also* $Div(\kappa\|\pi) = Div(\kappa\|\hat{\pi})$*, from which we get that* $I(\kappa(x_{\{2,3\}})) = 0$*. This means that variables* X_2 *and* X_3 *are independent.* ◇

In the example we have shown that a quality of a compositional approximation depends not only on the selected system of low-dimensional distributions (possibly marginals of the approximated distribution) but also on their ordering. We could see that κ was better approximated by perfect sequence π_1,π_2,π_3 than by π_1,π_3,π_2, in case that the latter one was not perfect. From the following assertion we will see that perfect sequences are always, in a sense, the best approximations.

Theorem 4 *If* $\pi_1,\pi_2,\ldots,\pi_n$ *is a perfect sequence of marginal distributions of* κ *(*$\kappa \in \Pi^{(K_1\cup\ldots\cup K_n)}$*) then*

$$Div(\kappa\|\pi_1 \triangleright \pi_2 \triangleright \ldots \triangleright \pi_n) \leq Div(\kappa\|\pi_{i_1} \triangleright \pi_{i_2} \triangleright \ldots \triangleright \pi_{i_n})$$

for any permutation $i_1, i_2, \ldots, i_n$ *of indices* $1, 2, \ldots, n$*.*

Proof. Since $\pi_1, \pi_2, \ldots, \pi_n$ is a perfect sequence of marginals of κ, we get from Corollary 1

$$Div(\kappa \| \pi_1 \triangleright \pi_2 \triangleright \ldots \triangleright \pi_n)) = I(\kappa) - I(\pi_1 \triangleright \pi_2 \triangleright \ldots \triangleright \pi_n),$$

and, because the Kullback-Leibler divergence is always nonnegative,

$$I(\kappa) \geq I(\pi_1 \triangleright \pi_2 \triangleright \ldots \triangleright \pi_n).$$

We assume that $\pi_1, \pi_2, \ldots, \pi_n$ are marginals of κ, and since they form a perfect sequence (due to Theorem 2) they are also marginals of $\pi_1 \triangleright \pi_2 \triangleright \ldots \triangleright \pi_n$. Therefore, equation (2) can be applied to both $Div(\kappa \| \pi_{i_1} \triangleright \ldots \triangleright \pi_{i_n}))$ and $Div(\pi_1 \triangleright \ldots \triangleright \pi_n \| \pi_{i_1} \triangleright \ldots \triangleright \pi_{i_n}))$:

$$Div(\kappa \| \pi_{i_1} \triangleright \ldots \triangleright \pi_{i_n})) = I(\kappa) - \sum_{\ell=1}^{n} \left(I(\kappa(x_{K_{i_\ell}})) - I(\kappa(x_{S_{i_\ell}})) \right), \tag{3}$$

$$Div(\pi_1 \triangleright \ldots \triangleright \pi_n \| \pi_{i_1} \triangleright \ldots \triangleright \pi_{i_n})) = I(\pi_1 \triangleright \ldots \triangleright \pi_n) - \sum_{\ell=1}^{n} \left(I(\kappa(x_{K_{i_\ell}})) - I(\kappa(x_{S_{i_\ell}})) \right).$$

The latter equality gives (respecting again the fact that the Kullback-Leibler divergence value must be nonnegative)

$$I(\pi_1 \triangleright \ldots \triangleright \pi_n) \geq \sum_{\ell=1}^{n} \left(I(\kappa(x_{K_{i_\ell}})) - I(\kappa(x_{S_{i_\ell}})) \right).$$

Combining this with equality (3) we get

$$Div(\kappa \| \pi_{i_1} \triangleright \ldots \triangleright \pi_{i_n})) \geq I(\kappa) - I(\pi_1 \triangleright \ldots \triangleright \pi_n),$$

where the right-hand side part of the inequality equals, as mentioned at the very beginning of the proof, $Div(\kappa \| \pi_1 \triangleright \ldots \triangleright \pi_n)$. □

6.3 Heuristic algorithm

Regarding the above-mentioned fact that perfect sequence models are equivalent to Bayesian networks, it is obvious that all the methods for Bayesian network learning can be adapted to CM construction (see eg. [1]). Another very simple and effective possibility, though far from being optimal, is the process discussed in the rest of the paper.

We split the model construction process into two steps. The first one, which is not discussed in this paper, is selection of oligodimensional distributions, from which the model will be constructed. In some situations one can be quite naturally relieved of necessity to perform this step. For example, when the data file is too small and only 2-dimensional distributions can be estimated, then all these

2-dimensional distributions can be considered. In other situations, an expert can select the distributions from which the model should be constructed. Otherwise, informational content of low-dimensional distributions should be taken as a criterion for selection of a system of oligodimensional distributions.

The second step of the model construction process is to find a proper ordering of the selected oligodimensional distributions. The properties presented in the above sections theoretically support a heuristic algorithm, which arranges low-dimensional distributions into a generating sequence in a manner that utilizes its informational content as much as possible. In this section its simplest version is presented that enables the reader to understand the basic principle of exploiting the informational content of individual input low-dimensional distributions.

The reader will see that the procedure considers not only the given system of distributions but also their marginals; this can, in some situations, improve exploitation of the informational content of distributions, since it considers a greater variety of conditional independence structures.

Algorithm

Input: System of low-dimensional distributions $\pi_1(x_{K_1}), \ldots \pi_n(x_{K_n})$.

Initialization: Select a variable X_m and a distribution π_j such that $m \in K_j$. Set $\kappa_1 := \pi_j(x_m)$, $L := \{m\}$ and $k := 1$.

Computational Cycle: While $K_1 \cup \ldots \cup K_n \setminus L \neq \emptyset$ perform the following 3 steps:

1. for all $j = 1, \ldots, n$ and all $m \in K_j \setminus L$ compute the mutual information
$$MI_{\pi_j}(X_m; X_{K_j \cap L}).$$
2. Fix j and m for which $MI_{\pi_j}(X_m; X_{K_j \cap L})$ achieved its maximal value.
3. Increase k by 1. Set $\kappa_k := \pi_j(X_{(K_j \cap L) \cup \{m\}})$ and $L := L \cup \{m\}$.

Output: Generating sequence $\kappa_1, \kappa_2, \ldots, \kappa_k$.

What can be said about the resulting generating sequence $\kappa_1, \kappa_2, \ldots, \kappa_k$? Distribution $\kappa^* = \kappa_1 \triangleright \kappa_2 \triangleright \ldots \triangleright \kappa_k$ is a probability distribution of $X_{K_1 \cup K_2 \cup \ldots \cup K_n}$. The goal of the algorithm is to get a distribution with the highest possible informational content $I(\kappa^*)$ (we know that the higher informational content, the lower the criterion function – Kullback-Leibler divergence). Important questions concern the facts whether the resulting sequence $\kappa_1, \kappa_2, \ldots, \kappa_k$ is perfect and contains all the distributions from $\pi_1, \pi_2, \ldots, \pi_n$. Unfortunately, answers to both these questions are negative.

Though the heuristics employed in the algorithm do not guarantee that a perfect sequence will always be found when it does exist, the advantage is in its efficiency and in the fact that it always suggests a subset of distributions that may form a perfect sequence, exploiting the available information in a suboptimal way[5]. One should realize, however, that a distribution from such a perfect sequence, though defined for groups of variables for which some input distribution π_j is defined, can differ from this input distribution π_j. In such a case, we employ a *process of verification and refinement*.

The detailed description of this process is beyond the scope (and extent) of this paper. Briefly said, verification consists of computation of Kullback-Leibler divergence of model distributions and the respective input distributions (or their marginals). If we find that some of the distributions defining the perfect model are too far from the required marginals (assuming that input distributions are marginals of the approximated distribution), then refinement is applied. This is realized by substituting a group of input marginal distributions by one distribution defined for all of the variables which are arguments of the deleted distributions. Naturally, this must be applied carefully, to avoid too much increase in the dimension of input distributions. New, more-dimensional input distributions are either estimated from data, or often computed from the original input distributions by the well-known Iterative Proportional Fitting Procedure ([3]). Then, having a new group of input distributions, the process starts from the very beginning by application of Algorithm.

The same process of verification and refinement is also applied when some of the input distributions are not included in the model.

Let us illustrate this process by a simple example.

Example 3 *Let us consider the following* 10 3*-dimensional distributions (their values were estimated from a data file):*

$$\begin{array}{lll} \pi_1(x_1,x_2,x_4), & \pi_2(x_1,x_2,x_6), & \pi_3(x_1,x_4,x_6), \\ \pi_4(x_3,x_6,x_{11}), & \pi_5(x_3,x_{10},x_{11}), & \pi_6(x_4,x_6,x_{11}), \\ \pi_7(x_5,x_6,x_8), & \pi_8(x_6,x_8,x_{11}), & \pi_9(x_7,x_{10},x_{11}), \\ & \pi_{10}(x_9,x_{10},x_{11}). & \end{array}$$

The algorithm (starting with variable X_1 and distribution π_1) produced the sequence

$$\pi_1(x_1),\pi_1(x_1,x_4),\pi_3(x_1,x_4,x_6),\pi_6(x_4,x_6,x_{11}),\pi_8(x_6,x_8,x_{11}),\pi_4(x_3,x_6,x_{11}),$$
$$\pi_7(x_5,x_6,x_8),\pi_5(x_3,x_{10},x_{11}),\pi_{10}(x_9,x_{10},x_{11}),\pi_9(x_7,x_{10},x_{11}),\pi_1(x_1,x_2,x_4).$$

[5] Any generating sequence can be converted into a perfect sequence according to the following assertion ([5, 6]).

Theorem 5 *Let $\pi_1 \triangleright \ldots \triangleright \pi_n$ be defined and let the sequence $\kappa_1,\ldots,\kappa_n$ be: $\kappa_1 = \pi_1$, $\kappa_2 = \kappa_1^{(K_2 \cap K_1)} \triangleright \pi_2$, and generally $\kappa_j = (\kappa_1 \triangleright \ldots \triangleright \kappa_{j-1})^{(K_j \cap (K_1 \cup \ldots \cup K_{j-1}))} \triangleright \pi_j$. Then $\kappa_1,\ldots,\kappa_n$ is perfect and $\pi_1 \triangleright \ldots \triangleright \pi_n = \kappa_1 \triangleright \ldots \triangleright \kappa_n$.*

There are two points that can be made about this sequence. First, since all the distributions were estimated from one data file (with no missing values), all the distributions were pairwise consistent, and thus both $\pi_1(x_1) = \pi_3(x_1)$ *and* $\pi_1(x_1,x_4) = \pi_3(x_1,x_4)$*, and therefore also*

$$\pi_1(x_1) \triangleright \pi_1(x_1,x_4) \triangleright \pi_3(x_1,x_4,x_6) = \pi_3(x_1,x_4,x_6).$$

Therefore, the result of the algotihm was, in fact, a generating sequence

$$\pi_3, \pi_6, \pi_8, \pi_4, \pi_7, \pi_5, \pi_{10}, \pi_9, \pi_1,$$

which was perfect (see assertion (iiia) *in Section 4).*

The negative property of this result was the fact that the algorithm finished before exploiting distribution $\pi_2(x_1,x_2,x_6)$*. Since we are looking for an approximation of a distribution from which the data file was generated, (following the verification and refinement process) we have to assess how much omitting* π_2 *influences the quality of the achieved result. This is done by considering the Kullback-Leibler divergence* $Div(\pi_2(x_1,x_2,x_6) \| \pi_{appr}(x_1,x_2,x_6))$*, for*

$$\pi_{appr} = \pi_3 \triangleright \pi_6 \triangleright \pi_8 \triangleright \pi_4 \triangleright \pi_7 \triangleright \pi_5 \triangleright \pi_{10} \triangleright \pi_9 \triangleright \pi_1$$

(let us mention that in this case $\pi_{appr}(x_1,x_2,x_6) = (\pi_3 \triangleright \pi_1)(x_1,x_2,x_6)$*). If we are not satisfied, refinement results in getting a distribution* $\pi_{11}(x_1,x_2,x_4,x_6)$ *and substituting it for* π_1, π_2 *and* π_3*. Subsequent application of the algorithm to the set of distributions* $\pi_4, \pi_5, \pi_6, \pi_7, \pi_8, \pi_9, \pi_{10}, \pi_{11}$ *resulted in obtaining the perfect sequence*

$$\pi_{11}, \pi_6, \pi_8, \pi_4, \pi_7, \pi_5, \pi_{10}, \pi_9.$$ ◇

7 Conclusions

We have presented theoretical results showing that if an approximation of a probability distribution is looked for in a family of compositional distributions then the Kullback-Leibler divergence representing a quality of the approximation can be expressed as a sum of two contributions. The first one, which can easily be suppressed by considering only marginals of the approximated distribution, describes "local" differences, while the other one corresponds to the loss of information resulting from the compositional model (from introducing the respective conditional independence relations). This knowledge was exploited for designing a heuristic algorithm based on an effort to maximize informational content of the constructed approximation.

Let us conclude the paper by a brief comment advocating CMs. Based on de Cooman approach to conditionning [2], J. Vejnarová introduced the operator of composition also in possibility theory [11], which made it possible to extend the whole approach beyond probabilistic framework.

References

[1] R.R. Boukaert, *Bayesian belief networks – from construction to inference*, PhD. thesis, University of Utrecht (Netherlands), 1995.

[2] G. de Cooman, Possibility theory I – III *International Journal of General Systems* **25** (1997), pp. 291–371.

[3] W.E. Deming and F.F. Stephan, On a least square adjustment of a sampled frequency table when the expected marginal totals are known, *Ann. Math. Stat.* **11** (1940), pp. 427-444.

[4] F. V. Jensen, *Introduction to Bayesian Network*, UCL Press 1996.

[5] R. Jiroušek, Composition of probability measures on finite spaces. In: *Proc. of the 13th Conf. Uncertainty in Artificial Intelligence UAI'97* (D. Geiger and P. P. Shenoy, eds.), Morgan Kaufmann Publ., San Francisco, California, 1997, pp. 274-281.

[6] Jiroušek, R. Graph Modelling without Graphs. In *Proc. of the 17th Int. Conf. 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, 1998, pp. 809-816.

[7] R. Jiroušek, Marginalization in composed probabilistic models. In: *Proc. of the 16th Conf. Uncertainty in Artificial Intelligence UAI'00* (C. Boutilier and M. Goldszmidt eds.), Morgan Kaufmann Publ., San Francisco, California, 2000, pp. 301-308.

[8] Jiroušek, R. Detection of independence relations from persegrams. In: *Proceedings of the 9th Int. Conf. on Information Processing and Management of Uncertainty in Knowledge Based Systems* (Bernadette Bouchon-Meunier, Ronald R. Yager (ed)), ESIA, Annecy, France, 2002, pp. 1261-1267.

[9] S.L. Lauritzen, *Graphical Models*. Clarendon Press, Oxford, 1996.

[10] A. Perez, ε-admissible simplification of the dependence structure of a set of random variables, *Kybernetika* **13** (1977), pp. 439–450.

[11] J. Vejnarová, Possibilistic independence and operators of composition of possibility measures. In *Prague Stochastics'98* (M. Hušková, J. Á. Víšek and P. Lachout eds.), JČMF 1998, pp. 575–580.

Radim Jiroušek is with the Institute of Information Theory and Automation of the Academy of Sciences of the Czech Republic. He is also a visiting Professor at Faculty of Management, Jindřichův Hradec, Czech Republic. E-mail: radim@utia.cas.cz