Compositional Models: Iterative Structure Learning from Data

Václav Kratochvíl[1,2], Vladislav Bína[2,1], Radim Jiroušek[1,2], and Tzong-Ru Lee[3]

[1] Institute of Information Theory and Automation
Czech Academy of Sciences, Czech Republic
[2] Faculty of Management
University of Economics, Prague, Czech Republic
[3] Department of Marketing
National Chung Hsing University, Taiwan, ROC
velorex@utia.cas.cz, bina@fm.vse.cz, radim@utia.cas.cz,
trlee@dragon.nchu.edu.tw

Abstract. Multidimensional probability distributions that are too large to be stored in computer memory can be represented by a compositional model - a sequence of low-dimensional probability distributions that when composed together try to faithfully estimate the original multidimensional distribution. The decomposition to the compositional model is not satisfactorily resolved. We offer an approach based on search traversal through the decomposable model class using likelihood-test statistics. The paper is a work sketch of the current research.

Keywords: Compositional models, Structure learning, Decomposability, Likelihood-ratio, Test statistics

1 Introduction

Many real-life problems can be solved using a decomposing strategy, excellently summarized by George Pólya in his famous book [1]: *If you cannot solve a problem, then there is an easier problem you can solve: find it.* The basic idea is simple. A problem, or a complex system, can be decomposed into a subproblems/subsystems that are easier to describe/understand. Unfortunately, the art of decomposing is not always straightforward.

Thanks to the massive use of computers, we have a huge amount of data in various areas of human activity. Using these data sets we can describe complex systems that may appear as black-boxes to us. The key is to extract knowledge from the data set and use it as a support for future decision making or predictions.

A set of vital tools to work with large data-sets is accessible through a probability framework where records from the data set are considered to be realizations of random variables. In this paper, we assume problems/systems that can be described using a set of random variables. By an event, we understand a moment when we measure values of the random variables and we assume the existence of

a data set with records of such measurements in history. As a typical example can serve a patient in a hospital with a database of various diseases, symptoms, and related laboratory test results. It is very difficult to cover all the dependencies between symptoms, test results, and diseases. Still, people are trying to do exactly that. A desire for a tool that, for example, automatically alerts you to a possible threat based on the results of a common medical test is obvious. Similarly, we can imagine the area of financial markets with records of stock market movements and related tool for automatic trading, etc.

In our case, we assume random variables with a discrete finite domain. Each random variable has a probability distribution, which specifies the probability of its values. Set of random variables has a joint probability distribution.

1.1 Knowledge Representation

Suppose that knowledge can be represented using a probability distribution defined over a set of corresponding random variables. Of course, the size of such a probability distribution would be enormous. Moreover, even if we were able to store it, we would need a similarly large amount of data to estimate its parameters well. This phenomenon is called *curse of dimensionality*. Here comes the concept of conditional independence. It is well known that in case of independence among variables we can express the corresponding probabilistic distribution as a product of smaller probability distributions (i.e. distributions defined over a smaller set of variables). To save even more space, some weak (conditional) dependencies can be modeled by independencies as well.

1.2 Compositional Models

The basic idea of compositional models is simple - to describe global knowledge from an application area using pieces of local knowledge. Local knowledge can be easily obtained, easily stored in a computer, and easily understand by a user/expert. On the other hand, is some cases, the global knowledge of the problem of interest is so complicated that it is beyond human capabilities to describe it. Note that the word *compositional* stands for the fact that probability distribution representing the knowledge about the system is *composed* from a set of low-dimensional distributions and *model* because the composed probability distribution is of course just a simplification/estimate of the original multidimensional distribution. (To simplify the model, some weak conditional dependencies are modeled by independence relations.)

To handle the knowledge hidden in a compositional model (decomposed probability distribution) one can use the standard tools from probability framework like marginalization, conditioning, and inference. The methods are of course customized to handle the decomposed structure and efficiently implemented using *local computations*.

2 Notation and Essentials

Let us consider a finite system of random variables with indices from a non-empty set N. Each variable from this system $\{X_i\}_{i\in N}$ has a finite (and non-empty) set of values $\mathbf{X}_i$. All the probability distributions discussed in the paper will be denoted by Greek letters. For $K \subset N$, $\kappa(x_K)$ denotes a distribution of variables $X_K = \{X_i\}_{i\in K}$, which is defined on all subsets of a Cartesian product $\mathbf{X}_K = \prod_{i\in K} \mathbf{X}_i$. Thus x_K denotes a $|K|$-dimensional vector of variable values $\{X_i\}_{i\in K}$ and $\mathbf{X}_K$ represents the set of all such vectors. Having a probability distribution $\kappa(x_K)$ and $L \subset K$ we shall denote its *marginal distribution* by $\kappa(x_L)$. To emphasize the marginalization process, we can also use $\kappa^{\downarrow L}$.

The symbol $\Pi^{(K)}$ denotes the set of all probability distributions defined for variables X_K. For two distributions defined over the same set of variables $\kappa, \lambda \in \Pi^{(K)}$ we say that λ *dominates* κ ($\kappa \ll \lambda$) if $\forall x \in \mathbf{X}_N : (\lambda(x) = 0 \implies \kappa(x) = 0)$. The distributions $\kappa \in \Pi^{(K)}$ and $\lambda \in \Pi^{(L)}$ are said to be *consistent* if for all $x \in \mathbf{X}_{K\cap L}$ $\kappa(x) = \lambda(x)$.

Definition 1 (Operator of Composition). *For arbitrary two distributions $\kappa \in \Pi^{(K)}$ and $\lambda \in \Pi^{(L)}$ for which* [4] $\kappa^{\downarrow \mathbf{K}\cap\mathbf{L}} \ll \lambda^{\downarrow \mathbf{K}\cap\mathbf{L}}$ *their* composition *is defined by the following formula*

$$(\kappa \triangleright \lambda)(x) = \frac{\kappa(x^{\downarrow \mathbf{K}})\lambda(x^{\downarrow \mathbf{L}})}{\lambda^{\downarrow \mathbf{K}\cap\mathbf{L}}(x^{\downarrow \mathbf{K}\cap\mathbf{L}})}. \tag{1}$$

Otherwise, it remains undefined.

The operator of composition is used to construct multidimensional compositional models. Composing two distributions, we can define a distribution of a dimensionality higher than any of the original ones. The resulting distribution is defined over the union of involved random variables.

By a compositional model of a multidimensional probability distribution we understand a sequence of low-dimensional distributions that assembled together using the operator of composition represent a multidimensional distribution that would be difficult to handle otherwise. In another words, the multidimensional distribution which can be written in the following way

$$\kappa_1 \triangleright \kappa_2 \triangleright \kappa_3 \triangleright \ldots \triangleright \kappa_n = (\ldots((\kappa_1 \triangleright \kappa_2) \triangleright \kappa_3) \triangleright \ldots) \triangleright \kappa_n. \tag{2}$$

where we expect κ_i to be defined over variables with indices from K_i. The sequence $\kappa_1, \kappa_2, \ldots, \kappa_n$ is called the *generating sequence* of the model.

In this paper, we will focus on the models composed from the marginal distributions of an input distribution obtained from data. Thus there are no inconsistent distributions and the operator of composition is always defined. The sequence of sets of variables (or precisely their indices) $K_1, \ldots, K_n$ is called the

[4] $\kappa(\mathbf{M}) \ll \lambda(\mathbf{M})$ denoted that the distribution κ is absolutely continues with respect to distribution λ, which in our finite settings means that whenever κ is positive also $\lambda mustbepositive$.

structure of the model. Note that the ordering of sets is important since operator $\triangleright$ is neither commutative nor associative. Because of the nature of the paper, we can simplify the notation and highlight the structure we denote the model from (2) in the following manner:

$$(K_1 \cdot K_2 \cdot \dots \cdot K_n)_\kappa$$

Note that compositional models represent a generalization of Bayesian networks. In other words, every Bayesian network can be represented using an equivalent compositional model. Note that structure $K_1, \dots, K_n$ has a similar meaning as graphs in case of Bayesian networks. It represents the system of conditional independencies valid for the model.

For the purpose of the following text, we will introduce a degenerated model, the so-called full model:

Definition 2 (Full model). *Compositional model κ of the form $\kappa = \kappa(x_N)$ is called* full model.

In the case of the full model, the sequence of sets of variable indices is formed only by one set N. It means that no composition is performed. Thus, the original data distribution (containing all variables) is a full model.

Definition 3 (Running Intersection Property). *The sets $L_1, L_2, \dots, L_n$ fulfill the* Running Intersection Property (RIP) *if*

$$\forall i \in \{2, \dots, n\} \quad \exists k < i \quad L_i \cap \left(\bigcup_{j<i} L_j \right) \subseteq L_k.$$

Definition 4 (Decomposability). *The compositional model $(K_1 \cdot K_2 \cdot \dots \cdot K_n)_\kappa$ is said to be* decomposable *if the ordering of sets in its structure $K_1, K_2, \dots, K_n$ fulfills the RIP property.*

Definition 5 (Conditional independence (CI)). *For distribution $\kappa(x_K)$ and for mutually disjoint $A, B, C \subseteq K$ such that $A \neq \emptyset$ and $B \neq \emptyset$ we write $X_A \perp\!\!\!\perp X_B | X_C[\kappa]$ (groups of variables X_A and X_B are* conditionally independent *given X_C with respect to the distribution κ) if*

$$\kappa(x_{A \cup B \cup C})\kappa(x_C) = \kappa(x_{A \cup C})\kappa(x_{B \cup C})$$

for all $x_{A \cup B \cup C} \in \mathbf{X}_{A \cup B \cup C}$. Note that in case of $C = \emptyset$ we speak about unconditional independence and we denote it as $X_A \perp\!\!\!\perp X_B[\kappa]$.

The use of the operator of composition embeds a conditional independence relation. This fact can be easily shown from both Definitions 1 and 5. (See also, e.g., Lemma 5.2 in [2] where also other basic properties of compositional models are formulated).

3 Decomposability

By a decomposition is usually understood the result of a process that, with the goal of simplification, divides an original object into its sub-objects. Thus, for example, a problem is decomposed into two (or more) simpler sub-problems, decomposition of a positive integer into prime numbers, etc. In the latter case, an elementary decomposition is a decomposition of an integer into two factors, the product of which gives the original integer. When repeating the process of decomposition long enough we end up with elementary sub-objects that cannot be further decomposed.

It can be easily deduced from the above-presented properties that the process of a repeatedly performed decomposition of an arbitrary (finite) object into elementary sub-objects (i.e., sub-objects that cannot be further decomposed) is always finite.

In case of a finite two-dimensional probability distribution $\kappa \in \Pi^{(k,l)}$ (k, l are singletons), simpler sub-objects are just one-dimensional distributions: a distribution of variable X_k and a distribution of variable X_l. The process of decomposition corresponds to marginalization - i.e. the sub-objects are $\kappa(x_k)$ and $\kappa(x_l)$. Note that the process of marginalization is well defined. Nevertheless, except for a degenerate case when $X_k \perp\!\!\!\perp X_l[\kappa]$, we cannot unambiguously reconstruct the original two-dimensional distribution from its one-dimensional marginals. In that case, a compositional model composed from one dimensional marginal would be just a very bad estimate of the original distribution.

Having a general probability distribution, one can be interested in the way how to decompose it into a set of its marginals in a way that if composed back together (using the operator of composition), it faithfully reflects the original distribution. Or in other words, if we convert a data set into a probability distribution using e.g. frequency analysis, we would like to learn its compositional model.

The following section deals with a special type of compositional models - decomposable models. The reason why we restricted ourselves to this subclass is clarified later.

4 Hierarchy in Decomposable Models Space

The notion of decomposability has been already established in a class of probabilistic models. Following Definition 4, one can notice that decomposability is a structural property in case of compositional models. I.e., it is related to the structure of a compositional model only, not to respective properties of probability distributions from its generating sequence.

Similarly, in the case of Bayesian networks (representant of another approach to probabilistic modeling), decomposability is also a structural property. Recall that in the case of Bayesian networks, a directed acyclic graph is used to represent its structure and we say that a Bayesian network is decomposable if the graph is decomposable. Graph decomposability is equivalent to many other

strong graph properties: graph chordality, graph triangularity, the existence of a perfect elimination ordering of nodes, the existence of a junction tree of graph cliques, etc. Simply said, decomposability is a very strong structural property and, what makes it so special, it is closely related to efficient local computations.

By local computation, we understand a possibility to perform complex computations with a probability distribution represented by a compositional model (like marginalization, conditioning, and inference) without the necessity to apply the operator of composition between members of the model generating sequence. Every general compositional model is converted into an equivalent decomposable model before performing any computations with it. This is one of the reasons why we have decided to restrict the current research on structure learning algorithms on the class of decomposable models only.

Assume a compositional model $(K_1 \cdot \ldots \cdot K_n)_\kappa$. We recognize the so-called *trivial sets* of the structure. We say that set $K_i, (i \in \{1, \ldots, n\})$ is trivial in the structure if $K_i \subseteq \bigcup_{j<i} K_j$. Note that probability distribution corresponding to K_i has no impact on the compositional model. Indeed, considering the definition of the operator of composition (denote $\bigcup_{j<i} K_j$ as $K_{j<i}$ to simplify the formula) then, following (1),

$$\begin{aligned}((\kappa_1 \rhd \ldots \rhd \kappa_{i-1})) \rhd \kappa_i)(x) &= \frac{(\kappa_1 \rhd \ldots \rhd \kappa_{i-1}))(x^{\downarrow \mathbf{K_{j<i}}})\kappa_i(x^{\downarrow \mathbf{K_i}})}{\kappa_i^{\downarrow \mathbf{K_{j<i}} \cap \mathbf{K_i}}(x^{\downarrow \mathbf{K_{j<i}} \cap \mathbf{K_i}})} \\ &= \frac{(\kappa_1 \rhd \ldots \rhd \kappa_{i-1}))(x^{\downarrow \mathbf{K_{j<i}}})\kappa_i(x^{\downarrow \mathbf{K_i}})}{\kappa_i^{\downarrow \mathbf{K_i}}(x^{\downarrow \mathbf{K_i}})} \\ &= (\kappa_1 \rhd \ldots \rhd \kappa_{i-1}))(x)\end{aligned} \tag{3}$$

Nevertheless, following Definition 3 of RIP, by adding a trivial set into the structure of a decomposable model, its decomposability can be violated. Nevertheless, we can add a trivial set that is a subset of another set preceding it in the sequence. See the following auxiliary property:

Lemma 1 (Redundant marginal). *Having a set $K \subseteq L_\ell$ the model $(L_1 \cdot L_2 \cdot \cdots \cdot L_n)_\kappa$ is decomposable if and only if $(L_1 \cdot \cdots \cdot L_\ell \cdot \cdots \cdot L_m \cdot K \cdot L_{m+1} \cdot \cdots \cdot L_n)_\kappa$ is decomposable.*

Proof.
Following the same reasoning as in (3), we can end up with the simplified model where the *redundant marginal* $\kappa(x_K)$ was removed. Let us emphasize that none of the compositions on the right of the considered marginal is affected by its removal since the union of variables appearing in the model before remains the same. ∎

Using the following theorem, one can create a decomposable model from a given decomposable model by introducing a new conditional independence relation into its structure. The proof is constructive. Note that the theorem has been already published in a slightly different form in [3].

Theorem 1. *Assume a decomposable compositional model* $\hat{\kappa} = (K_1 \cdot K_2 \cdot \dots \cdot K_n)_\kappa$ *where* $\exists k \in \{1, \dots, n\}$ *such that* $|K_k| > 1$. *Then there exist a pair of variables* $\ell, m \in K_k$ *such we can introduce another decomposable model* $\hat{\kappa}'$ *with one additional conditional independence relation* $\{k\} \perp\!\!\!\perp \{\ell\} | (K_k \setminus \{\ell, m\})[\hat{\kappa}']$. *We say that* $\hat{\kappa}$ *and* $\hat{\kappa}'$ *are in a neighborhood relation.*

Proof.
Without the loss of generality, we can assume that $k = n$. Indeed, because if it is not the case then we can take just the first k elements of the generating sequence and take it as the model of our interest. Such a generating sub-sequence represents always a marginal of the original model [2] and what holds for the marginal, it holds for the original model as well.

In case of a decomposable model, its structure $K_1, \dots, K_n$ must fulfil RIP property. I.e. it holds

$$\exists i < k \quad K_k \cap \left(\bigcup_{j<k} K_j \right) \subseteq K_i. \tag{4}$$

Without loss of generality let us make two assumptions:

1. Let us assume that $K_k \not\subseteq K_i$. (If the opposite was true then K_k would be a trivial column and as such it could be omitted because it does not change the model. For more detail see Lemma 1). I.e. $\exists \ell \in K_k$ such that $\ell \notin K_i$.
2. Further, assume that $|K_k| \geq 2$ because if it is not the case then it has only one element $\ell \notin \bigcup_{j<k} K_j$ (with no intersection with any other set of indices) we can move K_k to any other place without affecting the model [4].

Under these assumptions (or rearrangements of the model) we can choose another element $m \in K_k, m \neq \ell$ and change the structure of the model by introducing new conditional independence relation

$$\{\ell\} \perp\!\!\!\perp \{m\} \,|\, K_k \setminus \{\ell, m\}$$

by replacing K_k with sets $K_k \setminus \{\ell\}$ and $K_k \setminus \{m\}$. How to read conditional independence relations from a model structure can be found in [4]. Thus, we obtain a new compositional model where the only change is the replacement of the last distribution in its generating sequence by a pair of its marginals

$$\hat{\kappa}' = (K_1 \cdot K_2 \cdot \dots \cdot K_{k-1} \cdot K_k \setminus \{\ell\} \cdot K_k \setminus \{m\})_\kappa.$$

The new structure fulfills RIP property as well, which makes $\hat{\kappa}'$ decomposable. Indeed, because the first part of the structure $K_1, K_2, \dots, K_{k-1}$ remains unchanged, it is enough to check the newly added sets. Note that the intersect of $K_k \setminus \{\ell\}$ with the union of all preceding index sets is in K_i by (4). In case of the last set $K_k \setminus \{m\}$ the intersection with all prior sets lies in the set $K_k \setminus \{\ell\}$ and namely it is equal to $K_k \setminus \{\ell, m\}$.

Note that if a trivial set appears, it can be dropped without affecting decomposability of the model. ∎

For a more detailed view of decomposable models space see [3].

5 Mutual Information and Decomposibility

As it has been mentioned in the introduction, decomposable models are essential for efficient use of compositional models due to the possibility of local computations. As an example, we can take the following computations of likelihood-ratio test statistics.

Most of the machine learning methods for probabilistic models construction are, in a way, supported by notions and theoretical results from information theory. E.g. the value of mutual information helps to find pairs of variables that are tightly connected. The value of a multi-information may be used to select the best model from a considered group of models. Note that the basic notion is the famous Shannon entropy from which all the remaining ones are derived.

To help the reader to understand the notion of mutual information, it could be beneficial to highlight that it is the measure of similarity of two distributions. In probability theory, several measures of similarity for distributions have been introduced. One of them, having its origin in information theory, is a Kullback-Leibler divergence defined for $\kappa(K)$ and $\lambda(K)$ by the formula

$$Div(\kappa \parallel \lambda) = \begin{cases} \sum_{x \in \mathbf{X}_K} \kappa(x) \log \frac{\kappa(x)}{\lambda(x)}, & \text{if } \kappa \ll \lambda \\ +\infty, & \text{otherwise.} \end{cases} \tag{5}$$

It is a known fact that Kullback-Leibler divergence is always non-negative and equals 0 if and only if $\kappa = \lambda$ (see [5, 6]). Its only disadvantage is that it is not symmetric, i.e., generally $Div(\kappa \parallel \lambda) \neq Div(\lambda \parallel \kappa)$

Therefore, for testing whether the compositional model $\hat{\kappa}$ approximates faithfully original data distribution κ (both with variables from $\mathbf{X}_K$) one can use Kullback-Leibler divergence.

In our case, we take the full model for κ and we compare it with various decomposable models. Usually, the choice of the optimal model is accomplished either by the process of hypothesis testing or by using some information criterion.

In the following, we illustrate how to take the advantage of decomposability in case of compositional models to calculate Kullback-Leibler divergence using local computations while following the notion of a neighborhood of decomposable models introduced in Theorem 1.

Assume a decomposable compositional model $\hat{\kappa} = (K_1 \cdot K_2 \cdot \cdots \cdot K_n)_\kappa$ such that $\exists i \in \{1, \ldots, n\} : |K_i| \geq 2$. Following Theorem 1 one can introduce into this model one new conditional independence relation and get a new model $\hat{\kappa}'$ where the original set K_i was replaced by a pair of sets $K_i \backslash \ell$ and $K_i \backslash m$. Note that some of these sets may be trivial in the structure of the new model and appropriate probability distributions may be removed from the model generating sequence by Lemma 1 without affection the decomposability.

Following Theorem 1, the new model $\hat{\kappa}'$ can be obtained by multiplication of the formula for model $\hat{\kappa}$ by a simple factor:

$$\hat{\kappa}' = \hat{\kappa} \cdot \frac{\kappa(x_{K_i \backslash \{\ell\}}) \kappa(x_{K_i \backslash \{m\}})}{\kappa(x_{K_i \backslash \{\ell, m\}}) \kappa(x_{K_i})}. \tag{6}$$

The Kullback-Leibler divergence for full model κ and the new model $\hat{\kappa}'$ is

$$Div(\kappa \parallel \hat{\kappa}') = \sum_{x \in \mathbf{X}_K} \kappa(x) \log \frac{\kappa(x)}{\hat{\kappa}'(x)}. \tag{7}$$

Note that the divergence is always defined because we work with marginals of κ. (7) can be rewritten using (6) into

$$Div(\kappa \parallel \hat{\kappa}') = \sum_{x \in \mathbf{X}_K} \left(\kappa(x) \cdot \log \frac{\kappa(x)\kappa(x_{K_i \setminus \{\ell,m\}})\kappa(x_{K_i})}{\hat{\kappa}(x)\kappa(x_{K_i \setminus \{\ell\}})\kappa(x_{K_i \setminus \{m\}})} \right)$$

which can be further split into the sum of two logarithms

$$Div(\kappa \parallel \hat{\kappa}') = \sum_{x \in \mathbf{X}_K} \kappa(x) \log \frac{\kappa(x)}{\hat{\kappa}(x)} + \sum_{x \in \mathbf{X}_K} \kappa(x) \log \frac{\kappa(x_{K_i \setminus \{\ell,m\}})\kappa(x_{K_i})}{\kappa(x_{K_i \setminus \{\ell\}})\kappa(x_{K_i \setminus \{m\}})}.$$

Notice that the left part is a Kullback-Leibler divergence of κ and the original model $\hat{\kappa}$. The right-hand sum can be further rewritten as

$$Div(\kappa \parallel \hat{\kappa}') = Div(\kappa \parallel \hat{\kappa}) + \sum_{x \in \mathbf{X}_{K_i}} \left(\left(\sum_{x \in \mathbf{X}_{K \setminus K_i}} \kappa(x) \right) \cdot \log \frac{\kappa(x_{K_i \setminus \{\ell,m\}})\kappa(x_{K_i})}{\kappa(x_{K_i \setminus \{\ell\}})\kappa(x_{K_i \setminus \{m\}})} \right)$$

and the inner sum is equal to a marginal $\kappa(x_{K_i})$. I.e.

$$Div(\kappa \parallel \hat{\kappa}') = Div(\kappa \parallel \hat{\kappa}) + \sum_{x \in \mathbf{X}_{K_i}} \kappa(x_{K_i}) \log \frac{\kappa(x_{K_i \setminus \{\ell,m\}})\kappa(x_{K_i})}{\kappa(x_{K_i \setminus \{\ell\}})\kappa(x_{K_i \setminus \{m\}})}.$$

Following the last formula, we can easily and efficiently compute the divergence of the new model using the already computed divergence of $\hat{\kappa}$ and local computations concerning the replaced low–dimensional marginal defined by indices K_i only.

6 Model Complexity

By decomposing the original probability distribution into its marginals we reduce the number of its parameters. That is, by the way, the main reason to do the decomposition at all. The lower number of parameters, the faster the computations are, the easier one can store the model in computer memory. Realize that current models work with dozens or hundreds of variables.

In the case of our elementary approach, we will simply use the number of parameters needed to represent the compositional model in computer memory. Because every compositional model is represented using its generating sequence – a sequence of probability distributions – we will sum the size of respective probability distributions. In this paper, we restricted ourselves to discrete finitely valued random variables. Therefore, respective probability distributions can be

represented using contingency tables, where the size of each table is connected with the number of distinct values of involved random variables.

Let r_k be the number of categories for variable k ($\forall k \in K : r_k = |\mathbf{X}_k|$). Then, in case of the full model $\kappa(x_K)$, we need a probability table with $\prod_{k \in K} r_k$ cells. Note that this number can be decreased by one – probabilities must sum up to one. The number of parameters needed to represent the full model is then given by formula

$$C_F = \prod_{k \in K} r_k - 1.$$

Assume a general compositional model

$$\hat{\kappa} = (K_1 \cdot K_2 \cdot \cdots \cdot K_n)_\kappa .$$

Despite the fact that a compositional model can be expressed in the form of product of conditional distributions where the ith conditional distribution is a distribution of variables with indices from K_i not present in previous index sets and is conditioned by variables of K_i which already appeared in the previous parts of model, we use a standard representation using unconditional probability distributions. One of the reasons is that this representation makes local computations easier.

In this case, the number of parameters needed to represent compositional model $\hat{\kappa}$ is

$$df = \sum_{i \in \{1..n\}} (\prod_{j \in K_i} r_j - 1)$$

7 Information Criteria

The goal of decomposition is to get a compositional model as simple as possible which is, of course, in direct contradiction to the faithfulness of the model. To get an optimal balance between these two measures one can become inspired by various approaches used in other probabilistic learning methods.

Another approach to estimating the optimal size of the compositional model can be found in [7]. In this paper, the authors suggest to use the famous Huffman code [8] to find in a way optimum code to encode the original data set. The procedure is rather simple and it belongs to the fundamental parts of information theory. The idea is rather simple. First, find the Huffman encoding of the data or full model. Nevertheless, we do not need the encoding, we just need the number of bits necessary to encode the data or the model. Note that in the case of Huffman code one has to keep also the coding table to reconstruct the original object. The total size of both objects gives us a hint about the space needed to encode the data and we can use it to restrict the size of the compositional model as well.

On the other hand, information criteria that began to appear in the '70s of the 20th century remind a famous Occam's razor. Based on this principle the simplest model is chosen from a class of models describing the data in the same quality.

Let us recall the famous Bayesian information criterion (BIC), Schwarz criterion [9], or Akaike (AIC) criterion [10]. The criteria are generally a difference between the distance of the model from data and the size of the space needed to store it. One variant of BIC in the notation of this paper can be

$$BIC_{\hat{\kappa}} = 2 \cdot Div(\kappa \perp\!\!\!\perp \hat{\kappa}) - \log(n) \cdot df$$

where n is the number of observations in data.

Let us highlight, that we do not have any useful information criterion so far. We hope that we will receive one based on experiments performed in the next section.

8 Algorithms

Theorem 1 suggests to perform the *breadth-first search* through a tree of decomposable models where the root of the tree is the full model. Neighbors in the tree (also in the meaning of Theorem 1) differs from each other by additional conditional independence relation introduced to the structure. Using the theorem-proof, we can immediately construct the tree. The decomposability of all models is guaranteed. Moreover, we can use the advantage of local computations of test statistics needed for information criteria - *df* and Kullback-Leibler divergence.

The following questions arise: Does the tree contain all possible decomposable models? Or in other words, can we find a path from any decomposable model to a full model in such a tree? The answer is positive. Let us note that an inverse assertion to Theorem 1 can be proven (see Theorem 12 in [11]). Both together they guarantee the existence of a path of neighbor models from a full model to an arbitrary decomposable model by repeated application of Theorem 1.

The exhaustive search among all decomposable models is computationally intractable. Indeed, the number of decomposable models is enormous – numerical results for the case of mathematically equivalent chordal graphs (the structure of an arbitrary decomposable model can be represented using a chordal graph and vice-versa) can be seen in [12] or [13]. Nevertheless, we expect that the tree traversal could be significantly speedup using various techniques like gradient descent method.

Because the optimal method does not exist, we hope that a sub-optimal method can be found. The algorithm would not go through all nodes of the tree, but it will go through a restricted sub-tree with e.g. the best values of test statistics (see [14]).

Let us start with the simplest possible algorithm - a *greedy search algorithm*. It is based on the idea to take the best optimal choice in each step to eventually reach the global optimum. The algorithm picks the best solution in each step regardless of the consequences. Using Theorem 1 we can design it as follows:

Algorithm 1 (Greedy search) *Start with the full model.*

1. ***Generate all decomposable models in its neighborhood by adding a conditional independence relation.***
2. ***Choose the best model according to an information criterion (the difference between df and Kullback-Leibler divergence from the full model.***
3. ***Repeat steps 1 and 2 until the information criterion starts to increase.***

The other idea is to use a slightly modified greedy search enhanced with a history of previous search.

Algorithm 2 (Greedy search $-$ k best) *Start with the full model.*

1. ***Generate all neighboring decomposable models with an additional CI relation between a pair of variables.***
2. ***Choose k best models according to a given information criterion.***
3. ***Repeat steps 1 and 2 until the information criterion starts to increase.***

9 Experiments

We have performed several experiments to explore the possibility of usage of greedy search approach for structure learning of compositional models. We have used two data sets. The famous ASIA data set from [15] - an artificial data set generated from a probabilistic model based on a hypothetical medical situation. The data set has 8 variables (A,B,D,E,L,S,T,X) and 5.000 records. Then, because of the computational complexity of an exhaustive search in the space of decomposable models over 8 variables, we have used also 6 variables data set REINIS from [14] with 1200 records. Using a frequency analysis, we created full models. Then, the models were iteratively decomposed using Theorem 1, likelihood-test statistics, and the greedy search algorithm. In the case of REINIS data set, we have also performed the exhaustive scan through the whole space of decomposable models over 6 variables.

To generate all decomposable models, we have used the known fact that a sequence of sets satisfying RIP property corresponds to cliques in a chordal graph (ordered using maximum cardinality search algorithm). We used a catalog of all chordal graphs over six variables from [16]. Note that in case of six variables there are 18.395 decomposable models (without those with more than 4 singletons), based on 75 chordal graphs.

The run of the greedy search algorithm in case of REINIS data set is illustrated in Table 1. The first column corresponds to a structure of respective compositional model - respective probability distributions are marginals of the full model. The second column contains KL-divergence (also called relative entropy) – a measure of how one probability distribution (represented by a compositional model with a given structure) is different from a second, reference probability distribution corresponding to the full model. *df* is the above-defined number of

structure	KL-diverg.	*df*	suggest. ind.	i
(A,B,C,D,E,F)	0.00000	63	$B \perp\!\!\!\perp D \mid K_i$	1
(A,B,C,E,F)(A,C,D,E,F)	0.00479	62	$C \perp\!\!\!\perp D \mid K_i$	2
(A,B,C,E,F)(A,D,E,F)	0.00759	46	$A \perp\!\!\!\perp F \mid K_i$	2
(A,B,C,E,F)(A,D,E)	0.01009	38	$E \perp\!\!\!\perp F \mid K_i$	1
(A,B,C,E)(A,B,C,F)(A,D,E)	0.01368	37	$A \perp\!\!\!\perp B \mid K_i$	1
(A,C,E)(B,C,E)(A,B,C,F)(A,D,E)	0.01616	36	$C \perp\!\!\!\perp E \mid K_i$	2
(A,C,E)(B,C)(A,B,C,F)(A,D,E)	0.01868	32	$C \perp\!\!\!\perp F \mid K_i$	3
(A,C,E)(B,C)(A,B,F)(A,D,E)	0.02143	24	$A \perp\!\!\!\perp F \mid K_i$	3
(A,C,E)(B,C)(B,F)(A,D,E)	0.02247	20	$B \perp\!\!\!\perp F \mid K_i$	3
(A,C,E)(B,C)(F)(A,D,E)	0.02432	18	$A \perp\!\!\!\perp D \mid K_i$	4
(A,C,E)(B,C)(F)(D,E)	0.03081	14	$D \perp\!\!\!\perp E \mid K_i$	4
(A,C,E)(B,C)(F)(D)	0.03582	12	$C \perp\!\!\!\perp E \mid K_i$	1
(A,C)(A,E)(B,C)(F)(D)	0.04432	11	$A \perp\!\!\!\perp E \mid K_i$	2
(A,C)(E)(B,C)(F)(D)	0.05113	9	$A \perp\!\!\!\perp C \mid K_i$	1
(A)(B,C)(E)(F)(D)	0.06190	7	$B \perp\!\!\!\perp C \mid K_i$	2

Table 1. Greedy search over the set of all decomposable models of REINIS data set

parameters needed to represent the model. The last but one column contains the newly introduced independence relation to the structure of the model using Theorem 1 by splitting set K_i. The index i can be found in the last column. Note that the run was not stopped by any criterion and it was performed until the structure was split into a sequence of singletons.

To see the quality of greedy search approach, compare the results from Table 1 with Table 2 which contains the results of the exhaustive search in the class of all decomposable models of REINIS data set. More precisely, Table 2 contains a set of best models based on Kullback-Leibler divergence from a full model for each possible structure complexity *df*. On can see, that it is not true that for a smaller *df* the corresponding KL divergence has to be higher. Similarly, it seems to be difficult to decide which ratio of KL divergence and *df* is reasonable. There is no significant change in KL divergence considering decreasing *df*.

Table 3 illustrates the greedy search algorithm in the case of ASIA data set. Because the size of the set of all decomposable models fro eight variables was for us computationally intractable, we have added two additional lines not corresponding to the greedy search algorithm. The last but one algorithm represents the original model used for data generation. Note that the model is not decomposable. Its decomposable version [4] is in the last row of the table. You can see that with this setting, the greedy algorithm is far from finding it. The biggest problem is located in a huge KL-divergence jump in the 9th step of the algorithm. This will require further investigation of the problem. Nevertheless, even if we compute the KL-divergence globally (not employing local computations) we end up with the same numbers.

structure	KL-diverg.	*df*	structure	KL-diverg.	*df*
(A,B,C,E,F)(A,C,D,E,F)	0.00479	62	(A,D,E)(A,C,E)(B,C,E)(F)	0.02180	22
(A,B,C,E,F)(A,D,E,F)	0.00759	46	(A,D,E)(A,C,E)(B,C,F)	0.02157	21
(A,B,C,F)(A,C,E,F)(A,D,E,F)	0.01204	45	(A,C,D,E)(B,C)(B,F)	0.02160	21
(A,B,C,E,F)(A,D,E)	0.01009	38	(A,D,E)(A,C,E)(B,C)(B,F)	0.02247	20
(A,B,C,F)(A,B,C,E)(A,D,E)	0.01368	37	(A,C,D,E)(B,C)(F)	0.02346	19
(A,B,C,D,E)(B,F)	0.01457	34	(A,D,E)(A,C,E)(B,C)(F)	0.02432	18
(A,B,C,E)(A,B,D,E)(B,F)	0.01649	33	(D,E,F)(A,C,E)(B,C)	0.02801	17
(A,B,C,D,E)(F)	0.01643	32	(A,B,C,E)(F)(D)	0.03083	17
(A,B,C,E)(A,B,D,E)(F)	0.01834	31	(B,F)(B,C)(A,C,E)(D,E)	0.02895	16
(A,D,E,F)(A,B,C,E)	0.01495	30	(A,C,E)(B,C,E)(F)(D)	0.03330	16
(A,B,C,F)(A,C,E)(A,D,E)	0.01633	29	(A,C)(B,C)(B,E)(D,E)(B,F)	0.03724	15
(A,C,E)(B,C,E)(A,D,E)(D,E,F)	0.01901	28	(A,D,E)(A,B,C)(F)	0.03047	15
(A,B,C,E)(A,D,E)(B,F)	0.01747	25	(A,C,E)(B,C)(D,E)(F)	0.03081	14
(A,C,E)(A,D,E)(B,C,E)(B,F)	0.01995	24	(A,C)(B,C)(B,E)(D,E)(F)	0.03909	13
(A,B,C,E)(A,D,E)(F)	0.01932	23	(A,C,E)(B,C)(D)(F)	0.03582	12
(A,C,D,E)(B,C,F)	0.02070	22	(A,C)(B,C)(B,E)(F)(D)	0.04411	11

Table 2. Exhaustive search over the space of all decomposable models

10 Conclusion

This paper introduces a theoretic background for iterative compositional model learning. Although the introduced test statistics is feasible to efficiently compute using local computations, it seems that a simple greedy approach is not good enough. There are still several problems to be solved:

- to find a suitable criterion to stop the decomposition process,
- to check whether $k > 1$ will lead to better results and if not, to come with another algorithm, and
- to check the circumstances under which the greedy approach can provide solutions sufficiently close to the optimal solution.

To solve the problem we have to find a way how to efficiently generate the complete class of decomposable models for eight variables - probably using the catalog of chordal graphs by employing the fact that a chordal graph is just another mathematical representation of a decomposable structure. The problem is the number of permutations of eight variables, nevertheless, using [4], we should be able to determine structures from the same equivalence class and keep only one representant of each class.

Acknowledgement

The research was financially supported by grants GAČR no. 19-04579S (first author), GAČR no. 19-06569S (second author) and AV ČR no. MOST-04-18 (third and fourth author).

References

1. George Pólya. *How to solve it.* Doubleday Anchor Books, Garden City, New York, 2nd edition, 1957.

structure	KL-diverg.	*df*	*i*	independence
(A,B,D,E,L,S,T,X)	0.00000	255	1	$A \perp\!\!\!\perp E \mid K_i$
(A,B,D,L,S,T,X) (B,D,E,L,S,T,X)	0.00000	254	2	$S \perp\!\!\!\perp E \mid K_i$
(A,B,D,L,S,T,X) (B,D,E,L,T,X)	0.00000	190	2	$L \perp\!\!\!\perp X \mid K_i$
(A,B,D,L,S,T,X) (B,D,E,L,T)	0.00000	158	2	$B \perp\!\!\!\perp E \mid K_i$
(A,B,D,L,S,T,X) (D,E,L,T)	0.00000	142	2	$E \perp\!\!\!\perp D \mid K_i$
(A,B,D,L,S,T,X) (E,L,T)	0.00000	134	1	$S \perp\!\!\!\perp X \mid K_i$
(A,B,D,L,S,T) (A,B,D,L,T,X) (E,L,T)	0.00036	133	2	$B \perp\!\!\!\perp X \mid K_i$
(A,B,D,L,S,T) (A,D,L,T,X) (E,L,T)	0.00063	101	2	$A \perp\!\!\!\perp L \mid K_i$
(A,B,D,L,S,T) (A,D,T,X) (E,L,T)	0.22641	85	2	$A \perp\!\!\!\perp T \mid K_i$
(A,B,D,L,S,T) (A,D,X) (E,L,T)	0.25171	77	1	$S \perp\!\!\!\perp D \mid K_i$
(A,B,L,S,T) (A,B,D,L,T) (A,D,X) (E,L,T)	0.25207	76	1	$A \perp\!\!\!\perp S \mid K_i$
(A,B,D,L,T) (B,L,S,T) (A,D,X) (E,L,T)	0.25260	60	3	$A \perp\!\!\!\perp X \mid K_i$
(A,B,D,L,T) (B,L,S,T) (D,X) (E,L,T)	0.25346	56	2	$T \perp\!\!\!\perp L \mid K_i$
(A,B,D,L,T) (B,S,T) (D,X) (E,L,T)	0.28748	48	2	$S \perp\!\!\!\perp T \mid K_i$
(A,B,D,L,T) (B,S) (D,X) (E,L,T)	0.28825	44	1	$A \perp\!\!\!\perp T \mid K_i$
(A,B,D,L) (B,D,L,T) (B,S) (D,X) (E,L,T)	0.28916	43	1	$A \perp\!\!\!\perp D \mid K_i$
(A,B,L) (B,D,L,T) (B,S) (D,X) (E,L,T)	0.28990	35	1	$A \perp\!\!\!\perp L \mid K_i$
(A,B) (B,D,L,T) (B,S) (D,X) (E,L,T)	0.29065	31	1	$A \perp\!\!\!\perp B \mid K_i$
(A) (B,D,L,T) (B,S) (D,X) (E,L,T)	0.29109	29	2	$T \perp\!\!\!\perp L \mid K_i$
(A) (B,D,T) (B,D,L) (B,S) (D,X) (E,L,T)	0.29274	28	2	$T \perp\!\!\!\perp B \mid K_i$
(A) (D,T) (B,D,L) (B,S) (D,X) (E,L,T)	0.29685	24	2	$T \perp\!\!\!\perp D \mid K_i$
(A) (T) (B,D,L) (B,S) (D,X) (E,L,T)	0.29971	22	3	$L \perp\!\!\!\perp B \mid K_i$
(A) (T) (D,L) (B,D) (B,S) (D,X) (E,L,T)	0.31149	21	6	$X \perp\!\!\!\perp D \mid K_i$
(A) (T) (D,L) (B,D) (B,S) (X) (E,L,T)	0.32927	19	3	$L \perp\!\!\!\perp D \mid K_i$
(A) (T) (L) (B,D) (B,S) (X) (E,L,T)	0.35411	17	7	$T \perp\!\!\!\perp L \mid K_i$
(A) (E,T) (L) (B,D) (B,S) (X)	0.70100	12	2	$T \perp\!\!\!\perp E \mid K_i$
(A) (T) (E) (L) (B,D) (B,S) (X)	0.73479	11	6	$S \perp\!\!\!\perp B \mid K_i$
(A) (T) (E) (L) (B,D) (S) (X)	0.86501	9	5	$B \perp\!\!\!\perp D \mid K_i$
(A) (T) (E) (L) (B) (D) (S) (X)	1.20740	8		
(A) (S) (B,S) (L,S) (T,A) (E,L,T) (D,B,E) (X,E)	0.00829	28		
(A,T) (E,L,T) (E,L,S) (B,E,S) (B,D,E) (E,X)	0.00789	34		

Table 3. Greedy search algorithm run in case of Asia data set

2. Radim Jiroušek. Foundations of Compositional Model Theory. *International Journal of General Systems*, 40(6):623–678, 2011.
3. Vladislav Bína. *Multidimensional probability distributions: Structure and learning.* PhD thesis, University of Economics, Prague, Faculty of Management, 2011.
4. Radim Jiroušek and Václav Kratochvíl. Foundations of compositional models: structural properties. *International Journal of General Systems*, 44(1):2–25, 2015.
5. S Kullback and R A Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22:76–86, 1951.
6. S Kullback. An information-theoretic derivation of certain limit relations for a stationary markov chain. *J. SIAM Control*, 4:454–459, 1966.
7. Radim Jiroušek and Iva Krejčová. Minimum description length principle for compositional model learning. In *International Symposium on Integrated Uncertainty in Knowledge Modelling and Decision Making*, pages 254–266. Springer, 2015.

8. David A Huffman. A method for the construction of minimum-redundancy codes. *Proceedings of the IRE*, 40(9):1098–1101, 1952.
9. Gideon Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978.
10. Hirotugu Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974.
11. Vladislav Bína. Exhaustive search among compositional models of decomposable type. In *13th Czech-Japan Seminar on Data Analysis and Decision Making in Service Sciences*, pages 103–108, Otaru University of Commerce, 2010. University Hall.
12. Jun Kawahara, Toshiki Saitoh, Hirofumi Suzuki, and Ryo Yoshinaka. Enumerating all subgraphs without forbidden induced subgraphs via multivalued decision diagrams. *CoRR*, abs/1804.03822, 2018.
13. Yasuko Matsui, Ryuhei Uehara, and Takeaki Uno. Enumeration of perfect sequences of chordal graph. In Seok-Hee Hong, Hiroshi Nagamochi, and Takuro Fukunaga, editors, *Algorithms and Computation*, pages 859–870, Berlin, Heidelberg, 2008. Springer Berlin Heidelberg.
14. Tomáš Havránek. A procedure for model search in multidimensional contingency tables. *Biometrics*, 40(1):95–100, March 1984.
15. Steffen L Lauritzen and David J Spiegelhalter. Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 157–224, 1988.
16. Brendan McKay. Various simple graphs. http://users.cecs.anu.edu.au/~bdm/data/graphs.html, 2019. [Online; accessed 27-August-2019].