Compositional Models for Data Mining: an Example

Radim JIROUŠEK[1], Václav KRATOCHVÍL[1], and Tzong-Ru LEE[2]

[1] *Czech Academy of Sciences, Inst. Information Theory and Automation*
Pod Vodárenskou věží 4, CZ-182 08 Praha 8,
Czech Republic
radim@utia.cas.cz, velorex@utia.cas.cz

[2] *Department of Marketing, National Chung Hsing University*
145 Xingda Rd., South Dist., Taichung City 402,
Taiwan (R.O.C.)
trlee@dragon.nchu.edu.tw

Abstract

Like Bayesian networks, compositional models may also be used for data mining. Nevertheless, one can find several reasons why to prefer compositional models for this purpose. Perhaps the most important is the fact that compositional models are assembled from low-dimensional (unconditional) distributions so that computationally advantageous formulas are known for information theoretic characteristics computation. The other reason is that a decomposition is a natural way of complex tasks simplification. Therefore, the inverse process of composition is easily understandable for specialists from many fields of applications regardless of their level of mathematical education.
Keywords: Data mining, Mutual information, Compositional model, Conditional independence, Probability theory.

1 Introduction

The basic idea of compositional models is very simple: it is beyond human capabilities to describe global knowledge from an application area - one always works only with pieces of local knowledge. Such local knowledge can be, within probability theory, easily represented by low-dimensional distributions and a multidimensional distribution is (in a special way) composed from a number of such local pieces of knowledge – low-dimensional distributions. This analogy also explains why the compositional models are (relatively) easily understandable to specialists from the area of application - non-mathematicians. And it is also the reason why this technique can be, like Bayesian networks [1, 2], included among the methods of data-mining.

The goal of a data mining process [3] is not a model itself but its interpretation in the form of a distilled knowledge. Nevertheless, as we will see below, a greater part of knowledge is gained already during the process of model construction. The supervised approach to model construction enables the researchers to influence the resulting models in the way that these models are easily comprehensible and interpretable. Further, the user can have some knowledge about data, based on which the model is constructed. They may know that the data are not well stratified and some properties should be suppressed some others highlighted. Quite often, they want to adapt the constructed model to the purpose for which the model is constructed. Therefore, it is natural that we cannot give general instructions on how to proceed when constructing a model. It is the reason why we present in this paper just a simple example.

In this paper, we consider only finite-valued variables, which are denoted by upper-case Latin characters. Groups of variables are denoted by bold-face characters: i.e., for example, $\mathbf{M} = \{X, Y, Z, V, W\}$. The set of values of variable X is denoted by $\mathbb{X}_X$. Similarly, we use $\mathbb{X}_Y, \mathbb{X}_Z$, and so on.

By a *state* of a group of variables we understand any combination of values of the respective variables, i.e., for a group of three variables X, Z, V, a state is an element of a Cartesian product $\mathbb{X}_X \times \mathbb{X}_Z \times \mathbb{X}_V$. For the sake of simplicity, this Cartesian product is often denoted $\mathbb{X}_{\{X,Z,V\}}$. For a state $y \in \mathbb{X}_{\{X,Z,V\}}$ and a subset of the respective variables, say, for variables X, V, by $y^{\downarrow\{X,V\}}$

we denote a *projection* of y into $\mathbb{X}_{\{X,V\}}$, i.e., $y^{\downarrow\{X,V\}}$ is the state from $\mathbb{X}_{\{X,V\}}$ that is got from y by dropping out the value of variable Z.

Probability distributions are denoted by characters of Greek alphabet (κ, ν, π) (with possible indices). Recall that it means that $\kappa(Y,V) : \mathbb{X}_{\{Y,V\}} \longrightarrow [0,1]$, for which[1] $\sum_{x \in \mathbb{X}_{\{Y,V\}}} \kappa(x) = 1$.

Having a probability distribution $\kappa(X,Z,V)$, and a subset of variables $\mathbf{L} \subset \{X,Z,V\}$, $\kappa^{\downarrow\mathbf{L}}$ denote a *marginal distribution* of κ defined for each $x \in \mathbb{X}_{\mathbf{L}}$ by the formula

$$\kappa^{\downarrow\mathbf{L}}(x) = \sum_{y \in \mathbb{X}_{\mathbf{K}} : y^{\downarrow\mathbf{L}} = x} \kappa(y).$$

Note that we do not exclude situations when $\mathbf{L} = \emptyset$, for which we get $\kappa^{\downarrow\emptyset} = 1$.

Everybody knows that two variables X and Z are *independent* with respect to probability distribution $\pi(X,Z)$ if $\pi(X,Z) = \pi(X) \cdot \pi(Z)$. This is because, in this case[2],

$$\pi(X|Z) = \frac{\pi(X,Z)}{\pi(Z)} = \frac{\pi(X) \cdot \pi(Z)}{\pi(Z)} = \pi(X),$$

which expresses the fact that the knowledge of the value of variable Z does not bear any new information about the value of variable X, i.e., the conditional probability of X given Z equals the (un)conditional probability of X. For the purpose of data mining, a generalization of the notion of independence is very important. Therefore we introduce it in its general form.

Definition 1 *Consider a probability distribution $\pi(\mathbf{N})$, and three disjoint subsets of variables $\mathbf{K},\mathbf{L},\mathbf{M}$ ($\mathbf{K} \cup \mathbf{L} \cup \mathbf{M} \subseteq \mathbf{N}$). Let $\mathbf{K}$ and $\mathbf{L}$ be nonempty. We say that groups of variables $\mathbf{K}$ and $\mathbf{L}$ are* conditionally independent given $\mathbf{M}$ for distribution π *if*[3]

$$\pi^{\downarrow\mathbf{K}\cup\mathbf{L}\cup\mathbf{M}} \cdot \pi^{\downarrow\mathbf{M}} = \pi^{\downarrow\mathbf{K}\cup\mathbf{M}} \cdot \pi^{\downarrow\mathbf{L}\cup\mathbf{M}}.$$

In symbol, this property is expressed by $\mathbf{K} \perp\!\!\!\perp \mathbf{L}|\mathbf{M}\ [\pi]$.

Notice that in case of $\mathbf{M} = \emptyset$ we use only $\mathbf{K} \perp\!\!\!\perp \mathbf{L}\ [\pi]$ and speak about an unconditional independence (some authors call it marginal independence).

2 Decomposability

By a decomposition, we usually understand 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. General properties of such decomposition can be viewed on an example familiar to everybody: decomposition of a positive integer into prime numbers. In this case, an elementary decomposition is a decomposition of an integer into two factors, the product of which gives the original integer. For this example, we see that

- the result of the decomposition are two objects of the same type as the decomposed object – an integer is decomposed into two integers;
- both these sub-objects are simpler (smaller) than the original object – both factors are smaller than the original integer, we do not consider $1 \times n$ to be a decomposition of n;
- not all objects can be decomposed – prime numbers cannot be decomposed;
- there exists an inverse operation (we will call it a composition) yielding the original object from its decomposed parts – the composition of two integers is their product.

[1] Notice that symbol $\kappa(Y,V)$ is used to express the fact that probability distribution κ is defined for variables Y and V. $\kappa(x)$ for $x \in \mathbb{X}_{\{Y,V\}}$ is a probability of state $x \in \mathbb{X}_{\{Y,V\}}$.

[2] Naturally, this computation is valid only for positive distributions.

[3] This expression means that for all $x \in \mathbb{X}_{\mathbf{K}\cup\mathbf{L}\cup\mathbf{M}}$

$$\pi^{\downarrow\mathbf{K}\cup\mathbf{L}\cup\mathbf{M}}(x) \cdot \pi^{\downarrow\mathbf{M}}(x^{\downarrow\mathbf{M}}) = \pi^{\downarrow\mathbf{K}\cup\mathbf{M}}(x^{\downarrow\mathbf{K}\cup\mathbf{M}}) \cdot \pi^{\downarrow\mathbf{L}\cup\mathbf{M}}(x^{\downarrow\mathbf{L}\cup\mathbf{M}}).$$

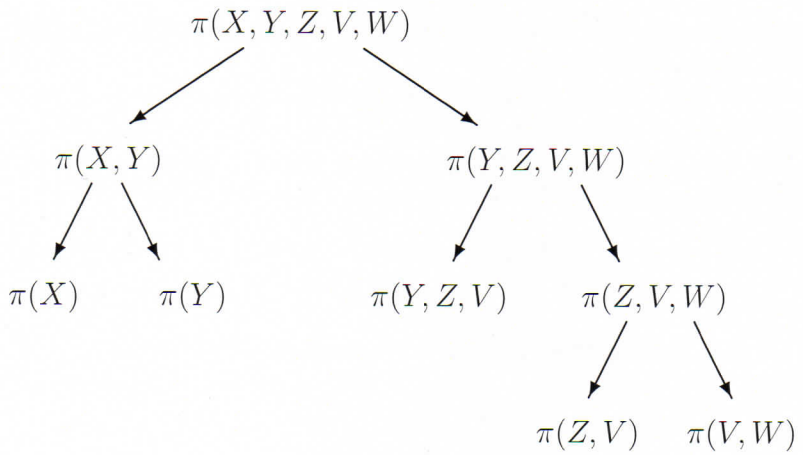


Figure 1: Hierarchical decomposition of $\pi(X,Y,Z,V,W)$.

What is a decomposition of a finite probability distribution? Consider a two-dimensional distribution $\pi(X,Y)$. Simpler sub-objects are just one-dimensional distributions: a distribution of a variable X and a distribution of a variable Y. To have a chance to reconstruct the original two-dimensional distribution π from these one-dimensional distributions, we have to consider marginals of π: $\pi^{\downarrow X}$ and $\pi^{\downarrow Y}$. Generally, the process of marginalization is unique, but, with the exception of a degenerate distribution, we cannot unambiguously reconstruct the original two-dimensional distribution from its one-dimensional marginals. To bypass this fact, we restrict the decomposition of two-dimensional distributions $\pi(X,Y)$ into their one-dimensional marginals only for the case of independence: $X \perp\!\!\!\perp Y\ [\pi]$. In this case, $\pi(X,Y)$ can easily be reconstructed from its marginals $\pi^{\downarrow X}$ and $\pi^{\downarrow Y}$: $\pi(X,Y) = \pi^{\downarrow X} \cdot \pi^{\downarrow Y}$, where "$\cdot$" denotes pointwise multiplication, i.e., $\pi(x,y) = \pi^{\downarrow X}(x) \cdot \pi^{\downarrow Y}(y)$ for all states $(x,y) \in \mathbb{X}_{\{X,Y\}}$.

Analogously, three-dimensional distribution $\pi(X,Y,Z)$ can be decomposed into two simpler probability distributions (marginals of $\pi(X,Y,Z)$) only if either a couple of variables (say X,Y) is independent of the remaining third variable (in this case Z), or, if two variables (say X and Z) are conditionally independent given the remaining third variable (in this case Y):

- $\{X,Y\} \perp\!\!\!\perp Z\ [\pi]$, then $\pi(X,Y,Z)$ can be reconstructed from $\pi^{\downarrow\{X,Y\}}$ and $\pi^{\downarrow Z}$,
- $X \perp\!\!\!\perp Z|Y\ [\pi]$, then $\pi(X,Y,Z)$ can be reconstructed from $\pi^{\downarrow\{X,Y\}}$ and $\pi^{\downarrow\{Y,Z\}}$.

This leads us to the following general definition.

Definition 2 *We say that a probability distribution* $\pi(\mathbf{M})$ *is* decomposed *into its marginals* $\pi^{\downarrow\mathbf{K}}$ *and* $\pi^{\downarrow\mathbf{L}}$ *if*

1. $\mathbf{K} \cup \mathbf{L} = \mathbf{M}$;
2. $\mathbf{K} \neq \mathbf{M}$, $\mathbf{L} \neq \mathbf{M}$;
3. $\pi(\mathbf{M}) \cdot \pi^{\downarrow\mathbf{K}\cap\mathbf{L}} = \pi^{\downarrow\mathbf{K}} \cdot \pi^{\downarrow\mathbf{L}}$.

Notice that the third condition is nothing else than $\mathbf{K} \setminus \mathbf{L} \perp\!\!\!\perp \mathbf{L} \setminus \mathbf{K} | \mathbf{K} \cap \mathbf{L}\ [\pi]$, and that the original distribution $\pi(\mathbf{M})$ can be uniquely reconstructed from the marginals $\pi^{\downarrow\mathbf{K}}$ and $\pi^{\downarrow\mathbf{L}}$.

Analogously to the decomposition of integers to prime numbers, even probability distributions can be hierarchically decomposed into a system of distributions that cannot be further decomposed. An example of such a hierarchical process represented by a corresponding tree structure can be seen in Figure 1, where distribution $\pi(X,Y,Z,V,W)$ is decomposed into a system of its marginal distributions: $\pi^{\downarrow X}$, $\pi^{\downarrow Y}$, $\pi^{\downarrow\{Y,Z,V\}}$, $\pi^{\downarrow\{Z,V\}}$, $\pi^{\downarrow\{V,W\}}$. Each decomposition was made possible by the fact that the respective conditional independence relation holds for distribution π. The decomposition process from Figure 1 was made possible by the assumption that the following system of conditional relations holds for distribution π (or, in other words, that the independence structure [4] of distribution π is the following):

- $X \perp\!\!\!\perp \{Z, V, W\} | Y \ [\pi]$;
- $X \perp\!\!\!\perp Y \ [\pi]$;
- $Y \perp\!\!\!\perp W | \{Z, V\} \ [\pi]$;
- $Z \perp\!\!\!\perp W | V \ [\pi]$.

Definition 3 *A probability distribution $\pi(\mathbf{N})$ is said to be* decomposable *if it can be decomposed into a system of its marginals $\pi^{\downarrow \mathbf{M}_1}, \pi^{\downarrow \mathbf{M}_2}, \ldots, \pi^{\downarrow \mathbf{M}_m}$, such that the variable sets $\mathbf{M}_1, \mathbf{M}_2, \ldots, \mathbf{M}_m$ can be reordered so that they meet the* Running Intersection Property *(RIP):*

$$\forall j = 2, 3, \ldots, m \quad \exists k (1 \leq k < j) \text{ for which } \mathbf{M}_j \cap (\mathbf{M}_1 \cup \ldots \cup \mathbf{M}_{j-1}) \subseteq \mathbf{M}_k.$$

3 Compositional models

This section introduces an operator of composition, originally introduced in [5], which realizes a process inverse to the process of decomposition discussed in previous section. For this, we need a notion of a *dominance*: $\pi(N) \ll \nu(N)$ if

$$\forall y \in \mathbb{X}_{\mathbf{N}} \quad \pi(y) > 0 \implies \nu(y) > 0.$$

Definition 4 *For arbitrary two distributions $\kappa(\mathbf{K})$ and $\lambda(\mathbf{L})$, for which $\kappa^{\downarrow \mathbf{K} \cap \mathbf{L}} \ll \lambda^{\downarrow \mathbf{K} \cap \mathbf{L}}$, their* composition *is for each $x \in \mathbb{X}_{\mathbf{K} \cup \mathbf{L}}$ given by the following formula*[4]

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

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

The reader certainly noticed that the presented definition is slightly more general than just an inverse operation to decomposition discussed above. We do not require that both $\mathbf{K}$ and $\mathbf{L}$ are proper subsets of $\mathbf{K} \cup \mathbf{L}$. The main reason is that this generalization makes the formulation of some theoretical properties simpler. Moreover, abandoning this requirement appears advantageous when constructing compositional models and when reading a knowledge from the resulting models. As we will see in Section 5, it enables the user to specify the required relations of (conditional) independence, which would not be otherwise representable in a model.

This operator of composition enables us to set up *multidimensional compositional models*, i.e. multidimensional probability distributions assembled from sequences of low-dimensional distributions with the help of the operators of composition [6, 7, 8]. Considering a systems of low-dimensional distributions $\kappa_1(\mathbf{K}_1), \kappa_2(\mathbf{K}_2), \ldots, \kappa_n(\mathbf{K}_n)$, the formula $\kappa_1 \triangleright \kappa_2 \triangleright \ldots \triangleright \kappa_n$, if defined, specifies a distribution of variables $\mathbf{K}_1 \cup \mathbf{K}_2 \cup \ldots \cup \mathbf{K}_n$. However, because of the fact that the operator of composition is not associative, the order, in which the operators are performed in the expression $\kappa_1 \triangleright \kappa_2 \triangleright \ldots \triangleright \kappa_n$ should be specified by parentheses. To simplify such expressions, we will omit the parentheses if the operators are to be performed from left to right. Therefore

$$\kappa_1 \triangleright \kappa_2 \triangleright \ldots \triangleright \kappa_n = (\ldots((\kappa_1 \triangleright \kappa_2) \triangleright \kappa_3) \triangleright \ldots \triangleright \kappa_{n-1}) \triangleright \kappa_n.$$

Moreover, without loss of generality, in what follows, we will always assume that κ_i is a distribution of variables $\mathbf{K}_i$ and that the composition will be defined in all the formulas wherever the operator appears.

To visualize the structure of a compositional model we use a tool called a *persegram*.

Definition 5 *A* persegram *of a compositional model $\kappa_1 \triangleright \kappa_2 \triangleright \ldots \triangleright \kappa_n$ is a table in which rows correspond to variables from $\mathbf{K}_1 \cup \ldots \cup \mathbf{K}_n$ (in an arbitrary order) and columns correspond to distributions $\kappa_1, \ldots, \kappa_n$ in the respective ordering. A position in the table is marked if the variable is among the arguments of the respective distribution. Markers for the first occurrence of each variable (i.e., the leftmost markers in rows) are* box-markers, *and for other occurrences there are* bullets.

[4] Define $\frac{0 \cdot 0}{0} = 0$.

For an example, the reader is referred to Figure 2a in Section 5, in which the persegram of $\kappa_1(D,N) \triangleright \kappa_2(B,R) \triangleright \kappa_3(R,W) \triangleright \kappa_4(N,R) \triangleright \kappa_5(T,W)$ is depicted.

Persegrams were designed mainly for reading conditional independence relations holding for compositional models. For this, we have to learn what are the trails and avoiding trails in a persegram.

Definition 6 *A sequence of markers $\mathfrak{m}_0, \ldots, \mathfrak{m}_t$ in the persegram of a compositional model $\kappa_1 \triangleright \kappa_2 \triangleright \ldots \triangleright \kappa_n$ is called an* $\mathbf{M}$-avoiding trail ($\mathbf{M} \subseteq \mathbf{K}_1 \cup \mathbf{K}_2 \cup \ldots \cup \mathbf{K}_n$) that connects m_0 and m_t *if it meets the following five conditions:*

1. *neither $\mathfrak{m}_0$ nor $\mathfrak{m}_t$ corresponds to a variable from $\mathbf{M}$;*
2. *for each $s = 1, \ldots, t$, the couple $(\mathfrak{m}_{s-1}, \mathfrak{m}_s)$ is either in the same row (i.e., a horizontal connection) or in the same column (a vertical connection);*
3. *each vertical connection must be adjacent to a box-marker (i.e., at least one of the markers in the vertical connection is a box-marker) - the so-called regular vertical connection;*
4. *no horizontal connection corresponds to a variable from $\mathbf{M}$;*
5. *vertical and horizontal connections regularly alternate with the following possible exception: at most, two vertical connections may be in direct succession if their common adjacent marker is a box-marker of a variable from $\mathbf{M}$.*

If an $\mathbf{M}$-avoiding trail connects two markers corresponding to variables X and Y, we say that these variables are connected by an $\mathbf{M}$-avoiding trail. *This situation is denoted by $X \rightsquigarrow_{\mathbf{M}} Y\ [\kappa_1 \triangleright \kappa_2 \triangleright \ldots \triangleright \kappa_n]$. Symbol $X \not\rightsquigarrow_{\mathbf{M}} Y\ [\kappa_1 \triangleright \kappa_2 \triangleright \ldots \triangleright \kappa_n]$ denote the situation when there does not exist an $\mathbf{M}$-avoiding trail connecting variables X and Y in the corresponding persegram. If $\mathbf{M} = \emptyset$ we speak about a simple trail, and use simplified symbol $X \rightsquigarrow Y$.*

For a simple trail (i.e. $\emptyset$-avoiding trail) connecting variables B and T see Figure 2b. A relationship between the existence of avoiding trails and the conditional independence of variables in a compositional model is expressed in the following assertion, which was originally proven in [9] (an alternative proof was published in [10]).

Theorem 7 *Consider a compositional model $\kappa_1, \kappa_2, \ldots, \kappa_n$ and the corresponding persegram. Let X and Y be two different variables from $\mathbf{K}_1 \cup \mathbf{K}_2 \cup \ldots \cup \mathbf{K}_n$, and $\mathbf{M} \subseteq \mathbf{K}_1 \cup \mathbf{K}_2 \cup \ldots \cup \mathbf{K}_n \setminus \{X, Y\}$. Then*

$$X \not\rightsquigarrow_{\mathbf{M}} Y\ [\kappa_1, \kappa_2, \ldots, \kappa_n] \implies X \perp\!\!\!\perp Y | \mathbf{M}\ [\kappa_1, \kappa_2, \ldots, \kappa_n].$$

4 Information-theoretic notions

In this section we consider a probability distribution $\pi(\mathbf{N})$, and three disjoint subsets $\mathbf{K}, \mathbf{L}, \mathbf{M} \subset \mathbf{N}$, such that $\mathbf{K} \cup \mathbf{L} \cup \mathbf{M} = \mathbf{N}$. Moreover, we assume that $\mathbf{K}$ and $\mathbf{L}$ are nonempty.

The basic notion, from which all the remaining ones are derived, is the famous Shannon entropy defined

$$H(\pi) = -\sum_{x \in \mathbb{X}_{\mathbf{N}} : \pi(x) > 0} \pi(x) \log_2 \pi(x).$$

This concept measures an uncertainty connected with the probability distribution. Its value is always nonnegative, less or equal $\log_2 |\mathbb{X}_{\mathbf{N}}|$. It equals zero if and only if the distribution is degenerated and expresses certainty. In other words, $H(\pi)$ equals zero, if and only if there exists a state $x^* \in \mathbb{X}_{\mathbf{N}}$, for which $\pi(x^*) = 1$. The entropy achieves its maximum only for a uniform distribution, i.e.,

$$H(\pi) = \log_2 |\mathbb{X}_{\mathbf{N}}| \iff \pi(x) = \frac{1}{|\mathbb{X}_{\mathbf{N}}|} \quad \text{for all } x \in \mathbb{X}_{\mathbf{N}}.$$

To measure the strength of dependence between the groups of random variables we employ a notion of *mutual information* defined by the formula

$$MI_\pi(\mathbf{K}; \mathbf{L}) = \sum_{x \in \mathbb{X}_{\mathbf{K} \cup \mathbf{L}} : \pi^{\downarrow \mathbf{K} \cup \mathbf{L}}(x) > 0} \pi^{\downarrow \mathbf{K} \cup \mathbf{L}}(x) \log_2 \left(\frac{\pi^{\downarrow \mathbf{K} \cup \mathbf{L}}(x)}{\pi^{\downarrow \mathbf{K}}(x^{\downarrow \mathbf{K}}) \cdot \pi^{\downarrow \mathbf{L}}(x^{\downarrow \mathbf{L}})} \right).$$

The higher this value, the stronger dependence exists between two disjoint groups of variables: $\mathbf{K}$ and $\mathbf{L}$. If the reader likes, this property can also be expressed in another way. The higher this value, the more information about variables $\mathbf{K}$ we get when learning values of variables $\mathbf{L}$ (or equivalently, because $MI_\pi(\mathbf{K};\mathbf{L}) = MI_\pi(\mathbf{L};\mathbf{K})$, the more information about variables $\mathbf{L}$ we get when learning values of variables $\mathbf{K}$).

Let us summarize the most important properties of mutual information supporting the fact that it is used as the measure of the *strength* of the dependence.

- $0 \leq MI_\pi(\mathbf{K};\mathbf{L}) \leq \min(H(\pi^{\downarrow\mathbf{K}}), H(\pi^{\downarrow\mathbf{L}}))$.
- $MI_\pi(\mathbf{K};\mathbf{L}) = 0 \iff \mathbf{K} \perp\!\!\!\perp \mathbf{L}\ [\pi]$.
- $MI_\pi(\mathbf{K};\mathbf{L}) = H(\pi^{\downarrow\mathbf{K}})$ if and only if variables $\mathbf{K}$ are functionally dependent on variables $\mathbf{L}$. It means that in this case for the conditional distribution $\pi^{\mathbf{K}|\mathbf{L}}$ it holds that

$$\forall y \in \mathbb{X}_\mathbf{L}\ \exists x \in \mathbb{X}_\mathbf{K} \text{ such that } \pi^{\mathbf{K}|\mathbf{L}}(x|y) = 1.$$

In many practical situations, it is useful to normalize the measure of mutual information, to get a measure achieving values from the interval $[0, 1]$. This value suggested by A. Perez [11], who called it *information measure of dependence*, is in this text denoted ID:

$$ID_\pi(\mathbf{K};\mathbf{L}) = \frac{MI_\pi(\mathbf{K};\mathbf{L})}{\min(H(\pi^{\downarrow\mathbf{K}}), H(\pi^{\downarrow\mathbf{L}}))}.$$

It may help the reader to understand the notion of mutual information, if we show that it is actually 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 $\pi(\mathbf{N})$ and $\nu(\mathbf{N})$ by the formula

$$Div(\pi \,||\, \nu) = \begin{cases} \sum\limits_{x \in \mathbb{X}_N : \pi(x) > 0} \pi(x) \log_2 \left(\frac{\pi(x)}{\nu(x)}\right), & \text{if } \pi \ll \nu; \\ +\infty, & \text{otherwise.} \end{cases}$$

It is known that Kullback-Leibler divergence is always nonnegative and equals 0 if and only if $\pi = \nu$ (see [12, 13]). Its only disadvantage is that it is not symmetric, i.e., generally $Div(\pi \,||\, \nu) \neq Div(\nu \,||\, \pi)$. Nevertheless, since it is very easy to show that $\pi^{\downarrow\mathbf{K}\cup\mathbf{L}} \ll \pi^{\downarrow\mathbf{K}} \cdot \pi^{\downarrow\mathbf{L}}$ we see that

$$MI_\pi(\mathbf{K};\mathbf{L}) = Div(\pi^{\downarrow\mathbf{K}\cup\mathbf{L}} \,||\, \pi^{\downarrow\mathbf{K}} \cdot \pi^{\downarrow\mathbf{L}})$$

is always finite, and, as we already said above, equals zero if and only if $\pi^{\downarrow\mathbf{K}\cup\mathbf{L}} = \pi^{\downarrow\mathbf{K}} \cdot \pi^{\downarrow\mathbf{L}}$, which is nothing else than $\mathbf{K} \perp\!\!\!\perp \mathbf{L}\ [\pi]$.

As the reader can expect, not only there is a relationship between independence and mutual information, but there is also an analogous relationship between conditional independence and *conditional mutual information*, defined by

$$MI_\pi(\mathbf{K};\mathbf{L}|\mathbf{M}) = \sum_{x \in \mathbb{X}_\mathbf{N} : \pi(x) > 0} \pi(x) \log_2 \left(\frac{\pi^{\mathbf{K}\cup\mathbf{L}|\mathbf{M}}(x)}{\pi^{\mathbf{K}|\mathbf{M}}(x^{\downarrow\mathbf{K}\cup\mathbf{M}}) \cdot \pi^{\mathbf{L}|\mathbf{M}}(x^{\downarrow\mathbf{L}\cup\mathbf{M}})} \right).$$

(Notice that $MI_\pi(\mathbf{K};\mathbf{L}|\emptyset) = MI_\pi(\mathbf{K};\mathbf{L})$.)

Again, the higher the value of conditional mutual information the stronger the conditional dependence between the respective groups of variables. Since we have not introduced the notion of conditional entropy, in this case, we can precisely formulate only a part of the properties holding for conditional mutual information.

- $MI_\pi(\mathbf{K};\mathbf{L}|\mathbf{M}) \geq 0$.
- $MI_\pi(\mathbf{K};\mathbf{L}|\mathbf{M}) = 0 \iff \mathbf{K} \perp\!\!\!\perp \mathbf{L}|\mathbf{M}\ [\pi]$.

5 Data mining example

This is the main section of the paper presenting a supervised model construction, during which we gain a knowledge from data. We consider six variables $\mathbf{M} = \{B, D, N, R, T, W\}$ with $\mathbb{X}_B = \{1, 2, 3\}$ and $\mathbb{X}_D = \mathbb{X}_N = \mathbb{X}_R = \mathbb{X}_T = \mathbb{X}_W = \{1, 2\}$. We are about to construct a compositional model for these variables from a data file containing 1000 records. Taking into account the fact that the cardinality of the considered state space is $|\mathbb{X}_\mathbf{M}| = 3 \times 2^5 = 96$, we can hardly expect to get any reasonable (i.e., interpretable) knowledge from the respective frequency table depicted in Table 1.

Table 1: Frequencies of states from $\mathbb{X}_{\{B,D,N,R,T,W\}}$.

	$R=0$				$R=1$			
	$T=0$		$T=1$		$T=0$		$T=1$	
	$W=0$	$W=1$	$W=0$	$W=1$	$W=0$	$W=1$	$W=0$	$W=1$
$B=1, D=1, N=1$	0	8	4	15	2	9	23	3
$B=1, D=1, N=2$	0	0	0	0	0	1	3	0
$B=1, D=2, N=1$	0	0	0	0	1	3	2	0
$B=1, D=2, N=2$	0	147	12	66	5	9	1	0
$B=2, D=1, N=1$	0	2	0	10	10	34	70	0
$B=2, D=1, N=2$	0	10	0	7	3	8	1	2
$B=2, D=2, N=1$	0	0	0	0	1	6	13	0
$B=2, D=2, N=2$	0	61	4	31	14	45	22	1
$B=3, D=1, N=1$	0	0	0	4	20	40	78	4
$B=3, D=1, N=2$	0	4	0	2	1	9	3	0
$B=3, D=2, N=1$	0	0	1	0	3	5	13	0
$B=3, D=2, N=2$	0	13	1	7	23	57	20	0

It is not a bad idea to start with computing the value of entropy for all considered variables:

$$H(B) = 1.58, \quad H(D) = 0.98, \quad H(N) = 0.96,$$
$$H(R) = 0.99, \quad H(T) = 0.99, \quad H(W) = 0.93.$$

From this, we do not get any knowledge about the relationship among the considered variables, but we get some information as for how to proceed further. Since the entropy of all binary variables is close to 1, it means that a minimum of entropies for any pair of variables is close to one, either. Therefore, when considering a strength of dependence between two variables, the value of mutual information MI and the value of information measure of dependence ID do not significantly differ from each other. Therefore we compute only values of mutual information. But keep in mind that when the considered variables achieve different numbers of values, there may be substantial differences between the values of the entropy of individual variables. In such a case, considering the information measure of dependence is preferable.

From the point of view of model construction, we are interested in couples of variables, which are closely (strongly) connected, and in couples of independent variables. Therefore, when computing values of mutual information for all pairs of variables, we sort the couples according to the value of mutual information. In the present example, we get

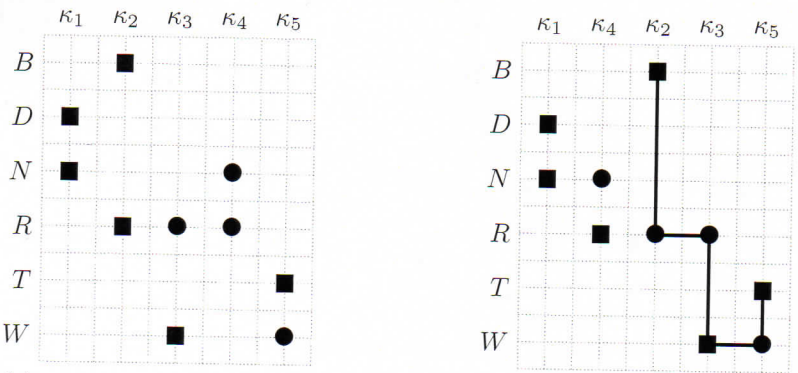
$$\begin{cases} MI(D;N) = 0.4356, \\ MI(B;R) = 0.2871, \\ MI(R;W) = 0.2578, \\ MI(N;R) = 0.2070, \\ MI(T;W) = 0.1813, \end{cases} \qquad \begin{aligned} &MI(N;T) = 0.0709 \\ &MI(D;W) = 0.0627 \\ &MI(B;N) = 0.0619 \\ &MI(D;T) = 0.0421 \\ &MI(B;D) = 0.0342 \end{aligned}$$

$$MI(N;W) = 0.1546$$
$$MI(D;R) = 0.0958$$
$$MI(B;W) = 0.0814$$

$$\begin{cases} MI(R;T) = 0.0019, \\ MI(B;T) = 0.0007. \end{cases}$$

The head of this sequence contains the couples of closely connected variables, the tail of this sequence suggests which pairs of variables may be considered independent. The first five couples are grouped together because these first five couples cover the whole $\mathbf{M}$. Therefore, let us start building compositional models from two-dimensional distributions defined for these couples of

variables. To get their best ordering in a model, the multi-information of the whole model should be taken into account. The higher multi-information, the better model because it incorporates more information from data.



(a) $\pi_1 = \kappa_1 \triangleright \kappa_2 \triangleright \kappa_3 \triangleright \kappa_4 \triangleright \kappa_5$ (b) $\pi_2 = \kappa_1 \triangleright \kappa_4 \triangleright \kappa_2 \triangleright \kappa_3 \triangleright \kappa_5$

Figure 2: Persegrams of models π_1 and π_2.

Consider estimates of the first five two-dimensional distributions and denote them respectively: $\kappa_1(D,N)$, $\kappa_2(B,R)$, $\kappa_3(R,W)$, $\kappa_4(N,R)$, $\kappa_5(T,W)$. If considering model $\pi_1 = \kappa_1 \triangleright \kappa_2 \triangleright \kappa_3 \triangleright \kappa_4 \triangleright \kappa_5$ (see persegram in Figure 2a) we can immediately see that $\pi_1 = \kappa_1 \triangleright \kappa_2 \triangleright \kappa_3 \triangleright \kappa_5$. Distribution κ_4 may be deleted from the model because both the respective markers in the persegram are bullets. This model is decomposable (the reader may easily check RIP) and perfect (data file does not contain missing values, and therefore the estimates of marginals are consistent). Therefore[5],

$$IC(\pi_1) = IC(\kappa_1) + IC(\kappa_2) - IC(\kappa_2^{\downarrow\emptyset}) + IC(\kappa_3) - IC(\kappa_3^{\downarrow R}) + IC(\kappa_5) - IC(\kappa_5^{\downarrow W}) = \sum_{i=1,2,3,5} IC(\kappa_i) = 1.1618,$$

because $IC(\kappa_1) = MI(D;N), IC(\kappa_2) = MI(B;R)$, and so on, and because the multi-information of probability distribution $\kappa(\mathbf{K})$ for $|\mathbf{K}| < 2$ equals zero. However, it is evident that also $\pi_2 = \kappa_1 \triangleright \kappa_4 \triangleright \kappa_2 \triangleright \kappa_3 \triangleright \kappa_5$ is a decomposable model, for which

$$IC(\pi_2) = \sum_{i=1}^{5} IC(\kappa_i) = 1.3687.$$

In fact, this model is the best possible among those assembled from distributions $\kappa_1, \kappa_2, \kappa_3, \kappa_4, \kappa_5$, in case the amount of multi-information is taken as the only criterion of optimality. This is because this model utilizes all the information contained in the distributions from which it is assembled. However, this model does not reflect the other information we obtained from computing the mutual information for all couples of variables: the two smallest values of mutual information suggest that variables T and R, and variables T and B are independent. And, as the reader can deduce from the persegram corresponding to π_2 (see persegram in Figure 2b), one can find simple trails connecting all couples of variables, i.e., also $B \rightsquigarrow T\ [\pi_2]$ and $R \rightsquigarrow T\ [\pi_2]$. Therefore, the independence relations $B \perp\!\!\!\perp T\ [\pi_2]$ and $R \perp\!\!\!\perp T\ [\pi_2]$ are not guaranteed by the model structure.

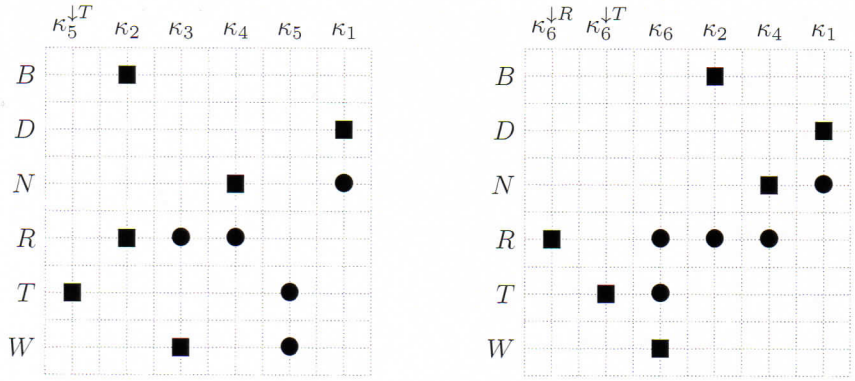
To incorporate this knowledge into the model, one can consider, e.g., model $\pi_3 = \kappa_5^{\downarrow T} \triangleright \kappa_2 \triangleright \kappa_3 \triangleright \kappa_4 \triangleright \kappa_5 \triangleright \kappa_1$. However, as the reader can see from the persegram in Figure 3a, $\pi_3 = \kappa_5^{\downarrow T} \triangleright \kappa_2 \triangleright \kappa_3 \triangleright \kappa_4 \triangleright \kappa_1$, and therefore

$$IC(\pi_3) = \sum_{i=1}^{4} IC(\kappa_i) = 1.1874.$$

The decrease of the multi-information is due to the fact that π_3 dos not incorporate the information from κ_5.

Thus, it may seem that one can incorporate the knowledge about the two independence relations into the model only at the cost of a decrease of multi-information, i.e., at the cost of the loss of

[5]Such a simple formula holds only for decomposable models.



(a) $\pi_3 = \kappa_5^{\downarrow T} \triangleright \kappa_2 \triangleright \kappa_3 \triangleright \kappa_4 \triangleright \kappa_5 \triangleright \kappa_1$ (b) $\pi_5 = \kappa_6^{\downarrow R} \triangleright \kappa_6^{\downarrow T} \triangleright \kappa_6 \triangleright \kappa_2 \triangleright \kappa_4 \triangleright \kappa_1$

Figure 3: Persegrams of models π_3 and π_5.

information. To get out of this trap, let us start studying the way, how variable T is connected with all others. Let us compute from the data the conditional mutual information of T and B given the remaining variables, and similarly, the conditional mutual information of T and R given the remaining variables. We get

$$\begin{aligned} MI(T;B|D) &= 0.002, & MI(T;R|D) &= 0.001, \\ MI(T;B|N) &= 0.006, & MI(T;R|N) &= 0.013, \\ MI(T;B|W) &= 0.024, & MI(T;R|W) &= 0.084. \end{aligned}$$

How to explain the fact that T and R are independent but not conditionally independent? A straightforward explanation is that T and R are independent and jointly influence other variables. In case we know the meaning of the variables, we should choose the one, which is, in our knowledge, directly influenced by T and R (or T and B). Otherwise, we choose the one indicated by the highest value of conditional mutual information: $MI(T;R|W)$. It makes us believe that two independent variables T and R influence W, and the only way how to incorporate this knowledge into the model is to start considering a three-dimensional distribution: let $\kappa_6(R,T,W)$ be the corresponding estimate got from data. Naturally, this three-dimensional distribution is a bearer of all the information expressed by both κ_3 and κ_5, which can be now dropped off from the further consideration. Naturally, κ_6 contains more information than κ_3 and κ_5. It describes the combined influence of T and R on W, which cannot be expressed by two two-dimensional distributions.[6]

After adding κ_6 and deleting κ_3 and κ_5, the remaining distributions $\kappa_1, \kappa_2, \kappa_4, \kappa_6$ can easily be ordered to meet RIP: e.g., $\pi_4 = \kappa_6 \triangleright \kappa_2 \triangleright \kappa_4 \triangleright \kappa_1$ is a decomposable model expressing all the knowledge we consider. Nevertheless, the above discussed independence of variables is not visible from the respective persegram, it is only encoded in the distribution κ_6. Therefore, we can prefer model $\pi_5 = \kappa_6^{\downarrow B} \triangleright \kappa_6^{\downarrow T} \triangleright \kappa_6 \triangleright \kappa_2 \triangleright \kappa_4 \triangleright \kappa_1$, from the persegram of which in Figure 3b the considered independence relations are obvious.

What are the differences between the models π_4 and π_5? Model π_4 is decomposable, and therefore more advantageous when used for computations. On the other hand, model π_5 explicitly manifests the independence $T \perp\!\!\!\perp \{R,B\}|W\ [\pi_5]$. When computing the multi-information of these models we get

$$IC(\pi_4) = \sum_{i=6,2,4,1} IC(\kappa_i) = 0.5234 + 0.2871 + 0.2070 + 0.4356 = 1.4531,$$

and

$$IC(\pi_5) = \sum_{i=6,2,4,1} IC(\kappa_i) - IC(\kappa_6^{\downarrow\{R,T\}}) = IC(\pi_5) - MI(T,R) = 1.4512.$$

[6] To illustrate the fact that a three-dimensional distribution may bear more information than a collection if its two-dimensional marginals, consider the following simple example. Children have usually more fun if the weather is warm. Similarly, they prefer sunny days to days with precipitation. However, in winter, the precipitation in very cold days usually means snowing, which is a great fun for children. And this type of knowledge cannot be expressed just by describing two separate relations: day temperature and children fun, and precipitation and children fun.

The imperceptible decrease of the value of multi-information when transforming π_4 into π_5 is due to small changes necessary for introducing the independence of T and R.

Model π_5 seems to meet all the requirements made for data-based models. Nevertheless, especially when considering supervised approaches, one should not miss the realization of important subsequent steps belonging to a process of model verification rather than to the process of model construction. Let us illustrate these steps by verifying model π_5. Consider the respective persegram in Figure 3b, which enables us to list all (conditional) independence relations holding for the model:

1.	$B \perp\!\!\!\perp D \mid \mathbf{M}$	for $\mathbf{M}$ containing either N or R,
2.	$B \perp\!\!\!\perp N \mid \mathbf{M}$	for $R \in \mathbf{M}$,
3.	$B \perp\!\!\!\perp T \mid \mathbf{M}$	for $R \in \mathbf{M}$, or $W \notin \mathbf{M}$,
4.	$B \perp\!\!\!\perp W \mid \mathbf{M}$	for $R \in \mathbf{M}$,
5.	$D \perp\!\!\!\perp R \mid \mathbf{M}$	for $N \in \mathbf{M}$,
6.	$D \perp\!\!\!\perp T \mid \mathbf{M}$	for $N \in \mathbf{M}$, or $R \in \mathbf{M}$, or $W \notin \mathbf{M}$,
7.	$D \perp\!\!\!\perp W \mid \mathbf{M}$	for $\mathbf{M}$ containing either N or R,
8.	$N \perp\!\!\!\perp T \mid \mathbf{M}$	for $R \in \mathbf{M}$, or $W \notin \mathbf{M}$,
9.	$N \perp\!\!\!\perp W \mid \mathbf{M}$	for $R \in \mathbf{M}$,
10.	$R \perp\!\!\!\perp T \mid \mathbf{M}$	for $W \notin \mathbf{M}$.

From this list, the eights relation covering also the unconditional independence $N \perp\!\!\!\perp T$ is in contradiction with $MI(N;T) = 0.0709$. To set this imperfectness right, we substitute $\kappa_4(N,R)$ by $\kappa_7(N,R,T)$, and consider model $\pi_6 = \kappa_6^{\downarrow B} \triangleright \kappa_6^{\downarrow T} \triangleright \kappa_6 \triangleright \kappa_2 \triangleright \kappa_7 \triangleright \kappa_1$. For this model we have

$$\begin{aligned} IC(\pi_6) &= \sum_{i=6,2,7,1} IC(\kappa_i) - IC(\kappa_6^{\downarrow\{R,T\}}) - IC(\kappa_7^{\downarrow\{R,T\}}) \\ &= 0.5234 + 0.2871 + 0.3236 + 0.4356 - 2 \times 0.0019 = 1.5659. \end{aligned}$$

To accept a model the user should verify that

- the independence relations deduced from the corresponding persegram do not contradict the intuition of the supervising user,
- the independence relations deduced from the corresponding persegram are not in contradiction with the values of (conditional) mutual information values computed from data,
- the marginals from which the resulting model is set up do not differ substantially from the corresponding estimates from data.

To follow these instructions let us transform model π_6 into a form that all the low-dimensional distributions, from which the model is composed, are marginals of the model itself:

$$\begin{aligned} &\nu_1(R) = \kappa_6^{\downarrow R}(R), \\ &\nu_2(T) = \kappa_6^{\downarrow T}(T), \\ &\nu_3(R,T,W) = \nu_1(R) \triangleright \nu_2(T) \triangleright \kappa_6(R,T,W), \\ &\nu_4(B,R) = \nu_1(R) \triangleright \kappa_2(B,R) = \kappa_2(B,R), \\ &\nu_5(N,R,T) = \nu_1(R) \triangleright \nu_2(T) \triangleright \kappa_7(N,R,T), \\ &\nu_6(D,N) = \nu_5^{\downarrow N}(N) \triangleright \kappa_1(D,N). \end{aligned}$$

Thus, $\pi_6 = \nu_1 \triangleright \nu_2 \triangleright \nu_3 \triangleright \nu_4 \triangleright \nu_5 \triangleright \nu_6$, all ν_i (for $i = 1,\ldots,6$) are marginals of π_6. The respective probability distributions generating this model are depicted in Table 2, and the respective persegram is in Figure 4. From this persegram the following list of conditional independence

Table 2: Probability distributions $\nu_1 - \nu_6$.

$\nu_1(R)$			
$\nu_1(1) = .437$	$\nu_1(2) = .563$		
$\nu_2(T)$			
$\nu_2(1) = .566$	$\nu_2(2) = .434$		
$\nu_3(R,T,W)$			
$\nu_3 1,1,1) = .000$	$\nu_3(1,1,2) = .247$	$\nu_3(1,2,1) = .024$	$\nu_3(1,2,2) = .166$
$\nu_3(2,1,1) = .085$	$\nu_3(2,1,2) = .233$	$\nu_3(2,2,1) = .235$	$\nu_3(2,2,2) = .010$
$\nu_4(B,R)$			
$\nu_4(1,1) = .280$	$\nu_4(2,1) = .125$	$\nu_4(3,1) = .032$	
$\nu_4(1,2) = .057$	$\nu_4(2,2) = .230$	$\nu_4(3,2) = .276$	
$\nu_5(N,R,T)$			
$\nu_5(1,1,1) = .010$	$\nu_5(1,1,2) = .036$	$\nu_5(1,2,1) = .136$	$\nu_5(1,2,2) = .197$
$\nu_5(2,1,1) = .238$	$\nu_5(2,1,2) = .153$	$\nu_5(2,2,1) = .182$	$\nu_5(2,2,2) = .048$
$\nu_6(D,N)$			
$\nu_6(1,1) = .45$	$\nu_6(1,2) = .05$	$\nu_6(2,1) = .05$	$\nu_6(2,2) = .45$

relations can be deduced:

$$
\begin{array}{ll}
B \perp\!\!\!\perp D|\mathbf{M} & \text{for } \mathbf{M} \text{ containing either } N \text{ or } R, \\
B \perp\!\!\!\perp N|\mathbf{M} & \text{for } R \in \mathbf{M}, \\
B \perp\!\!\!\perp T|\mathbf{M} & \text{for } \mathbf{M} = \emptyset, \text{ or } R \in \mathbf{M}, \\
B \perp\!\!\!\perp W|\mathbf{M} & \text{for } R \in \mathbf{M}, \\
D \perp\!\!\!\perp R|\mathbf{M} & \text{for } N \in \mathbf{M}, \\
D \perp\!\!\!\perp T|\mathbf{M} & \text{for } N \in \mathbf{M}, \\
D \perp\!\!\!\perp W|\mathbf{M} & \text{for } N \in \mathbf{M}, \text{ or } \{R,T\} \subseteq \mathbf{M}, \\
N \perp\!\!\!\perp W|\mathbf{M} & \text{for } \{R,T\} \subseteq \mathbf{M}, \\
R \perp\!\!\!\perp T|\mathbf{M} & \text{for } \mathbf{M} = \emptyset, \text{ or } \mathbf{M} = \{B\},
\end{array}
$$

neither of which is in contradiction with anything what has been said about the modeled distribution up to now. Distributions ν_1, ν_2 and ν_4 are the original estimates from data. The remaining distributions ν_3, ν_5 and ν_6 are slightly different from the originally estimated distributions. This is due to the modification realized in the process of computation of distributions ν_i. Nevertheless, the deviations from the original data-based estimates are very small, as it can also be seen from the values of Kullback-Leibler divergence

$$
\begin{aligned}
Div(\kappa_6 \,||\, \nu_3) &= 0.00192, \\
Div(\kappa_7 \,||\, \nu_5) &= 0.00192, \\
Div(\kappa_1 \,||\, \nu_6) &= 0.00002.
\end{aligned}
$$

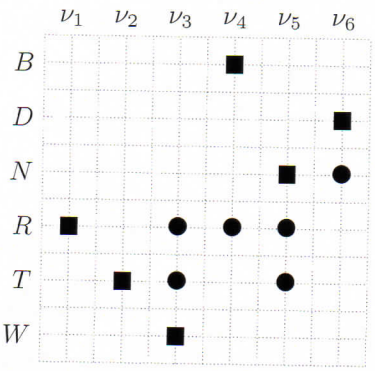


Figure 4: Persegram of model $\pi_6 = \nu_1 \triangleright \nu_2 \triangleright \nu_3 \triangleright \nu_4 \triangleright \nu_5 \triangleright \nu_6$.

(Notice, it is not a pure incidence that $Div(\kappa_6 \amalg \nu_3) = MI(R,T)$; it can be deduced from other properties the information-theoretic characteristics.) Thus we may say that π_6 is a reasonable model of the distribution generating the data.

Acknowledgement
The authors would like to thank the Czech Academy of Sciences and the Taiwanese Ministry of Science and Technology for their financial support of the bilateral project "**Compositional Models for Data Mining**", in the framework of which the described research was performed.

References

[1] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. The weka data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1):10–18, 2009.

[2] David Heckerman. Bayesian networks for data mining. *Data mining and knowledge discovery*, 1(1):79–119, 1997.

[3] David J Hand. Principles of data mining. *Drug safety*, 30(7):621–622, 2007.

[4] Milan Studený. *Probabilistic conditional independence structures*. Springer Science & Business Media, 2006.

[5] Radim Jiroušek. Composition of probability measures on finite spaces. In *Proceedings of the Thirteenth conference on Uncertainty in artificial intelligence*, pages 274–281. Morgan Kaufmann Publishers Inc., 1997.

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

[7] Francesco M Malvestuto. Equivalence of compositional expressions and independence relations in compositional models. *Kybernetika*, 50(3):322–362, 2014.

[8] Francesco M Malvestuto. Marginalization in models generated by compositional expressions. *Kybernetika*, 51(4):541–570, 2015.

[9] Radim Jiroušek. Persegrams of compositional models revisited: conditional independence. In *Proceedings of the 12th International Conference on Information Processing and Management of Uncertainty in Knowledge-based Systems, Malaga*, pages 915–922, 2008.

[10] Radim Jiroušek and Václav Kratochvíl. Foundations of compositional models: structural properties. *International Journal of General Systems*, 44(1):2–25, 2015.

[11] Albert Perez. Personal communication. 1970–1980.

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

[13] Solomon Kullback. An information-theoretic derivation of certain limit relations for a stationary markov chain. *J. SIAM Control*, 4:454–459, 1966.