Czech Academy of Sciences

Institute of Information Theory and Automation

DISCRETE COMPOSITIONAL MODELS FOR DATA MINING

Radim Jiroušek, Václav Kratochvíl, et al.

Prague, 2019

Project team:

Vladislav Bína
Chien Chun-Yu
Radim Jiroušek (Part I Editor)
Václav Kratochvíl (Part II Editor)
Tzong-Ru (Jiun-Shen) Lee
Martin Rod
Jan Švorc
Lucie Váchová
Wang Ching-Yi

This project was supported by Grant No. **MOST-18-04** within the framework of bilateral Czech-Taiwanese scientific cooperation.

Contents

Preface

This brochure has been written with the support of the bilateral Czech-Taiwanese project **Compositional models for data mining** financially supported by the Ministry of Science and Technology, Taiwan, and by the Czech Academy of Sciences under Grant No. **MOST-18-04**. The main output of the project, realized in 2018 and 2019, is a new supervised web system enabling researchers to learn probabilistic (compositional) models (both causal and stochastic) from data. We have opted for the web architecture for two reasons. First, we assume the system will be expanded in subsequent years, and the web application means that the system administrator only has to keep updated one version of program codes. Second, the system is accessible from any place in the world, so it can be applied not only by the members of research teams collaborating within the above-mentioned project but also by all interested researchers from anywhere inthe world.

This booklet should serve as a manual for users of the data mining system. Nevertheless, since the system is based on the theory of compositional models, and no comprehensive text on this theory exists, we decided to set up this text from two parts. The first one describes the theoretical background on which the models constructed from data are based. It also includes chapters showing how the compositional models can be applied to data mining tasks. For this reason, the first part summarizes results scattered in a number of research journal and conference papers, mainly by R. Jiroušek and his coauthors Vl. Bína and V. Kratochvíl [8, 14, 9, 10, 11, 18, 2, 3, 16]. This part, after introducing the notation from general probability theory, puts a special emphasis on the notion of stochastic (conditional) independence, without which one cannot distill knowledge from probability models. Chapters 2-5 sum up excerpts from the original research conference and journal papers. The importance of this part can be seen not only in the fact that it is the first time when these results are surveyed in one comprehensive text but also that it is presented using a new unifying notation, without which it might be difficult to see the links interconnecting individual parts of this theoretical approach.

The text is primarily intended for a user of the web system, who should get familiar with the theory of compositional models. To this end, presenting the proofs of theorems, which are in great majority rather technical, would be unnecessarily boring; we omit them. On the other hand, to avoid the necessity of referring to basic textbooks on probability theory, we include a brief introduction into its main notions and explain thoroughly all the used symbols. Though it might seem admissible, we strongly discourage the reader from using the system without properly studying the underlying theory of compositional systems in Part I.

Part II of this text is a system manual; it advises the reader how to use the system for model construction, how to control the process of model construction, and how to verify the achieved results.

This text is under permanent development. Therefore, the readers are kindly asked to refer all mistakes, errors, and/or suggestions for improvement (e.g. the passages difficult to understand, or notions that should be illustrated by examples) by e-mail to radim@utia.cas.cz (concerning Part I) and velorex@utia.cas.cz (concerning Part II).

Part I

THEORETICAL FOUNDATIONS

Chapter 1

Brief introduction to finite probability theory

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 with mere pieces of local knowledge. Such local knowledge can, within probability theory, be easily represented by low-dimensional distributions. The multidimensional distribution is (in a special way) composed from these local pieces. 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 is, like Bayesian networks, included among the methods of data-mining. For example, constructing compositional models from two data files collected in different cultural environments (in our case in Taiwan and the Czech Republic) enables the user to compare the structures of the two models, thus revealing qualitative differences between the studied societies. Analogously, the comparison of the respective probability tables enables the researchers to describe the quantitative differences.

The goal of a data mining process is not a model itself but its interpretation in the form of distilled knowledge. Nevertheless, as it will be shown in Chapter 6, a greater part of knowledge is already gained during the process of model construction, and only the rest during the model verification – explanation. The supervised approach to model construction enables the researchers to influence the resulting models so that these models are easily comprehensible and interpretable. Both of the above-mentioned processes (model construction and model verification) are supported by the information-theoretic tools introduced in Section 1.4.

The goal of this first Chapter is to introduce basic (and generally well-known) notions from probability theory as well as the above-mentioned concepts from information theory.

1.1 Discrete random variables

Upper-case characters of Latin alphabet (like X, Y, Z, V, W) denote finite valued variables. Finite sets of values of these variables are denoted by $\mathbb{X}_X$, $\mathbb{X}_Y$, $\mathbb{X}_Z$, $\mathbb{X}_V$, $\mathbb{X}_W$. Thus, for example, if variable Y denotes a 'gender' of a respondent, then $\mathbb{X}_Y$ contains just two values corresponding to 'female' and 'male'. Most of the time we will deal with sets of variables denoted by bold-face characters $\mathbf{K}$, $\mathbf{L}$, $\mathbf{M}$, $\mathbf{N}$. Thus, $\mathbf{K}$ may be $\{X,Y,W\}$. By a *state* of variables $\mathbf{K}$ we understand any combination of values of the respective variables, i.e., in the mentioned case $\mathbf{K}=\{X,Y,W\}$, a state is an element of the Cartesian product $\mathbb{X}_X \times \mathbb{X}_Y \times \mathbb{X}_W$. For the sake of simplicity, this Cartesian product is denoted $\mathbb{X}_{\mathbf{K}}$. For a state $y \in \mathbb{X}_{\mathbf{K}}$ and $\mathbf{L} \subset \mathbf{K}$, we denote by $y^{\downarrow \mathbf{L}}$ the *projection* of $y \in \mathbb{X}_{\mathbf{K}}$ into $\mathbb{X}_{\mathbf{L}}$, i.e., $y^{\downarrow \mathbf{L}}$ is the state from $\mathbb{X}_{\mathbf{L}}$ that is obtained from y by disregarding all the values of variables from $\mathbf{K} \setminus \mathbf{L}$.

Example 1.1 *Consider a group of students described by three variables: X – gender, Y – study results, and Z – nationality. Let $\mathbb{X}_X=\{f,m\}$ (female, and male, respectively), $\mathbb{X}_Y=\{e,g,a\}$ (excellent, good, and average, respectively), $\mathbb{X}_Z=\{t,c\}$ (Taiwanese, and Czech, respectively). Denoting $\mathbf{K}=\{X,Y,W\}$, $\mathbb{X}_{\mathbf{K}}$ is the set of all twelve triplets (states): (f,e,t), (f,e,c), ... , (m,a,c). In this case, for example, $(m,g,c)^{\downarrow\{X,Z\}}=(m,c)$ and $\mathbb{X}_{\{X,Z\}}=\{(f,t),(f,c),(m,t),(m,c)\}$.*

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

Having a probability distribution $\kappa(\mathbf{K})$, and a subset of variables $\mathbf{L} \subset \mathbf{K}$, we denote by $\kappa^{\downarrow \mathbf{L}}$ 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 $\kappa^{\downarrow\emptyset}=1$.

For a probability distribution $\kappa(\mathbf{K})$, we introduce a conditional distribution in a standard way. For disjoint $\mathbf{L},\mathbf{M} \subseteq \mathbf{K}$, by a conditional distribution of variables $\mathbf{L}$ given variables $\mathbf{M}$, we understand any function $\kappa^{\mathbf{L}|\mathbf{M}} : \mathbb{X}_{\mathbf{L}\cup\mathbf{M}} \longrightarrow [0,1]$ meeting the following two conditions:

- $\forall x \in \mathbb{X}_{\mathbf{L}\cup\mathbf{M}} \quad \kappa^{\downarrow \mathbf{L}\cup\mathbf{M}}(x)=\kappa^{\mathbf{L}|\mathbf{M}}(x)\cdot\kappa^{\downarrow\mathbf{M}}(x^{\downarrow\mathbf{M}})$,
- $\forall$ fixed $x \in \mathbb{X}_{\mathbf{M}}$ function $\kappa^{\mathbf{L}|\mathbf{M}}$ as a function of variables $\mathbf{L}$ is a probability distribution, i.e., $\sum_{y\in\mathbb{X}_{\mathbf{L}\cup\mathbf{M}};y^{\downarrow\mathbf{M}}=x} \kappa^{\mathbf{L}|\mathbf{M}}(y)=1$.

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

Due to the latter condition, the argument y of the function $\kappa^{\mathbf{L}|\mathbf{M}}$ is often split into two complementary pieces $y^{\downarrow\mathbf{L}}$ and $y^{\downarrow\mathbf{M}}$, and its value is written as $\kappa^{\mathbf{L}|\mathbf{M}}(y^{\downarrow\mathbf{L}}|y^{\downarrow\mathbf{M}})$.

Example 1.2 *Consider three variables $\mathbf{M} = \{X, Y, Z\}$ representing three fair coins, i.e., $|\mathbb{X}_X| = |\mathbb{X}_Y| = |\mathbb{X}_Z| = 2$. Denote $\mathbb{X}_X = \mathbb{X}_Y = \mathbb{X}_Z = \{0, 1\}$, and consider the following random experiment: two coins are randomly tossed and the third one is laid on the table so that the number of '1s' is odd. This experiment is fully described by probability distribution μ, values of which are shown in Table 1.1*

Table 1.1: Probability distribution describing the 3-coin example.

X	0	0	0	0	1	1	1	1
Y	0	0	1	1	0	0	1	1
Z	0	1	0	1	0	1	0	1
μ	0	$\frac{1}{4}$	$\frac{1}{4}$	0	$\frac{1}{4}$	0	0	$\frac{1}{4}$

The reader has certainly noticed that distribution μ shows a kind of symmetry with respect to the considered variables. From Table 1.1, there is no way to say which two coins are tossed randomly and which one is the third one manipulated to have an odd number of '1s'. Namely, it is easy to show that all of its three two-dimensional marginal distributions $\mu^{\downarrow\{X,Y\}}$, $\mu^{\downarrow\{X,Z\}}$,$\mu^{\downarrow\{Y,Z\}}$ are uniform, i.e.,

$$\mu^{\downarrow\{X,Y\}}(x, y) = \mu^{\downarrow\{X,Z\}}(x, z) = \mu^{\downarrow\{Y,Z\}}(y, z) = \frac{1}{4}$$

holds for all $(x, y, z) \in \mathbb{X}_{\{X,Y,Z\}}$. Later we will mention other interesting properties of this distribution. At this moment, we just want to invite the reader to see that[2]

$$\begin{aligned}\mu^{X|Y}(x|y) = \mu^{X|Z}(x|z) = \mu^{Y|X}(y|x) = \mu^{Y|Z}(y|z)& \\ = \mu^{Z|X}(z|x) = \mu^{Z|Y}(z|y) = \frac{1}{2}&\end{aligned}$$

holds for all $(x, y, z) \in \mathbb{X}_{\{X,Y,Z\}}$, and that

$$\mu^{X|\{Y,Z\}}(0|0, 0) = \mu^{X|\{Y,Z\}}(0|1, 1) = \mu^{X|\{Y,Z\}}(1|0, 1) = \mu^{X|\{Y,Z\}}(1|1, 0) = 0,$$

as well as

$$\mu^{X|\{Y,Z\}}(0|0, 1) = \mu^{X|\{Y,Z\}}(0|1, 0) = \mu^{X|\{Y,Z\}}(1|0, 0) = \mu^{X|\{Y,Z\}}(1|1, 1) = 1.$$

[2] When considering a singleton set, we often omit the curly parentheses denoting it as a set, i.e., we use just X instead of more precise (but clumsy) $\{X\}$.

For two probability distributions defined on the same group of variables, say, $\pi(\mathbf{K})$, $\kappa(\mathbf{K})$, we say that κ *dominates* π (writing $\pi \ll \kappa$) if

$$\forall\, x \in \mathbb{X}_{\mathbf{K}} \quad (\kappa(x) = 0 \implies \pi(x) = 0)\,.$$

Notice that this relationship is not too restrictive; e.g., every distribution is dominated by all positive distributions.

Consider two distributions $\kappa(\mathbf{K})$ and $\lambda(\mathbf{L})$. This time we do not assume any restriction on the sets of variables $\mathbf{K}$ and $\mathbf{L}$. They may be disjoint or overlapping. It may even happen that one is a subset of the other. We say that κ and λ are *consistent* if there exists a distribution $\pi(\mathbf{K}\cup\mathbf{L}))$ such that κ and λ are its marginals: $\pi^{\downarrow\mathbf{K}} = \kappa$ and simultaneously $\pi^{\downarrow\mathbf{L}} = \lambda$. In this case we also say that π is a *joint extension* of κ and λ. Notice that it is easy to show that κ and λ are consistent if and only if $\pi^{\downarrow\mathbf{K}\cap\mathbf{L}} = \kappa^{\downarrow\mathbf{K}\cap\mathbf{L}}$.

1.2 Structures of conditional independence

It is clearly impossible to represent a necessary probability distribution by a multidimensional table in a way similar to Table 1.1 when the number of the considered variables exceeds a certain (rather small) limit. However, in problems of practice, one mustconsider collections of features represented by tens or even hundreds of variables. Handling probability distributions of such high dimensionality is made possible by decomposition, which is to be studied in the next Section. Here we are about to introduce a notion of conditional independence, without which such decompositions would not be possible.

Everybody knows that two variables X and Y are *independent* with respect to probability distribution $\pi(X,Y)$ if $\pi(X,Y) = \pi(X)\cdot\pi(Y)$. This is because, in this case[3]:

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

This formula expresses the fact that the knowledge of the value of variable Y does not bear any new information about the value of variable X. The following notion just generalizes this simple idea.

Definition 1.3 *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

[3] Naturally, the following computation is valid for positive $\pi(Y)$.

$\mathbf{M}$ for distribution π *if*[4]

$$\pi^{\downarrow \mathbf{K}\cup\mathbf{L}\cup\mathbf{M}} \cdot \pi^{\downarrow \mathbf{M}} = \pi^{\downarrow \mathbf{K}\cup\mathbf{M}} \cdot \pi^{\downarrow \mathbf{L}\cup\mathbf{M}}. \tag{1.1}$$

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

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

Example 1.4 *Consider, again, distribution μ from Example 1.2. Its marginal distributions $\mu^{\downarrow X}, \mu^{\downarrow Y}$ and $\mu^{\downarrow \{X,Y\}}$ are shown in Table 1.2. Comparing product of $\mu^{\downarrow X}$ and $\mu^{\downarrow Y}$ with $\mu^{\downarrow \{X,Y\}}$ in this Table, one can immediately see that $X \perp\!\!\!\perp Y\ [\mu]$. In the same way one can show that also $X \perp\!\!\!\perp Z\ [\mu]$, and $Y \perp\!\!\!\perp Z\ [\mu]$.*

Table 1.2: Marginals of probability distribution μ from Example 1.2.

X	Y	$\mu^{\downarrow X}$	$\mu^{\downarrow Y}$	$\mu^{\downarrow \{X,Y\}}$
0	0	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{4}$
0	1	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{4}$
1	0	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{4}$
1	1	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{4}$

On the other hand, $X \not\perp\!\!\!\perp Y | Z\ [\mu]$ because

$$\mu^{\downarrow \{X,Y,Z\}}(\{0,0,0\})\mu^{\downarrow Z}(0) = 0 \cdot \frac{1}{2},$$

and

$$\mu^{\downarrow \{X,Z\}}(\{0,0\})\mu^{\downarrow \{Y,Z\}}(\{0,0\}) = \frac{1}{4} \cdot \frac{1}{4}.$$

Similar equalities show that $X \not\perp\!\!\!\perp Z | Y\ [\mu]$, and $Y \not\perp\!\!\!\perp Z | X\ [\mu]$.

To get intuitive insight into the meaning of the conditional independence relationship, it is apposite to realize that

$$X \perp\!\!\!\perp Y | Z\ [\pi] \iff \pi(X|Z) = \pi(X|Y,Z),$$

and, because $X \perp\!\!\!\perp Y | Z\ [\pi] \Leftrightarrow Y \perp\!\!\!\perp X | Z\ [\pi]$,

$$X \perp\!\!\!\perp Y | Z\ [\pi] \iff \pi(Y|Z) = \pi(Y|X,Z).$$

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

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

These formulae hold because, in this case (for a positive distribution π),

$$\begin{aligned}\pi(Y|X,Z) &= \frac{\pi(X,Y,Z)}{\pi(X,Z)} = \frac{\pi(X,Y,Z)\cdot\pi(X)}{\pi(X)\cdot\pi(X,Z)} = \frac{\pi(Y,Z)\cdot\pi(X,Z)}{\pi(X)\cdot\pi(X,Z)} \\ &= \frac{\pi(Y,Z)}{\pi(X)} = \pi(Y|X,Z).\end{aligned}$$

We will thus keep in mind that, generally,

$$\mathbf{K}\perp\!\!\!\perp\mathbf{L}|\mathbf{M}\ [\pi] \iff \pi^{\downarrow\mathbf{K}|\mathbf{M}} = \pi^{\downarrow\mathbf{K}|\mathbf{L}\cup\mathbf{M}}. \tag{1.2}$$

For a probability distribution π, by its *independence structure* we understand the list of all conditional independence relationships holding for π. To describe it in an economical way, the following properties of the notion of conditional independence (for their proofs see, e.g., [32]) may come in handy. For example, using the following Block Independence Property, one can show that the conditional independence structure of distribution π is fully specified by the list of conditional independence relationships of the form $X\perp\!\!\!\perp Y|\mathbf{K}\ [\pi]$ (by the list of conditional independence relationships for singletons).

Theorem 1.5 ***Factorization property.*** *Consider a probability distribution $\pi(\mathbf{J})$, and three disjoint subsets of variables $\mathbf{K}, \mathbf{L}, \mathbf{M}$ ($\mathbf{K}\cup\mathbf{L}\cup\mathbf{M}\subseteq\mathbf{J}$). Let $\mathbf{K}$ and $\mathbf{L}$ be nonempty. Then $\mathbf{K}\perp\!\!\!\perp\mathbf{L}|\mathbf{M}\ [\pi]$ if and only if there exist functions*

$$\begin{aligned}\psi_1 &: \mathbb{X}_{\mathbf{K}\cup\mathbf{M}} \longrightarrow [0,+\infty)\\ \psi_2 &: \mathbb{X}_{\mathbf{L}\cup\mathbf{M}} \longrightarrow [0,+\infty)\end{aligned}$$

such that, for all $x\in\mathbb{X}_{\mathbf{K}\cup\mathbf{L}\cup\mathbf{M}}$,

$$\pi^{\downarrow\mathbf{K}\cup\mathbf{L}\cup\mathbf{M}}(x) = \psi_1(x^{\downarrow\mathbf{K}\cup\mathbf{M}})\cdot\psi_2(x^{\downarrow\mathbf{L}\cup\mathbf{M}}) \tag{1.3}$$

holds.

Though this assertion looks rather abstract, it is just a generalization of the fact that

$$\mathbf{K}\perp\!\!\!\perp\mathbf{L}|\mathbf{M}\ [\pi] \iff \pi^{\downarrow\mathbf{K}\cup\mathbf{L}\cup\mathbf{M}} = \pi^{\downarrow\mathbf{K}\cup\mathbf{M}}\cdot\pi^{\downarrow\mathbf{L}|\mathbf{M}}.$$

The reader can see it immediately after computing the respective marginals from the expression given by Equation 1.3 (for $x\in\mathbb{X}_{\mathbf{K}\cup\mathbf{L}\cup\mathbf{M}}$):

$$\begin{aligned}\pi^{\downarrow\mathbf{K}\cup\mathbf{M}}(x^{\downarrow\mathbf{K}\cup\mathbf{M}}) &= \psi_1(x^{\downarrow\mathbf{K}\cup\mathbf{M}})\cdot\sum_{y\in\mathbb{X}_\mathbf{L}}\psi_2(y,x^{\downarrow\mathbf{M}}),\\ \pi^{\downarrow\mathbf{L}\cup\mathbf{M}}(x^{\downarrow\mathbf{L}\cup\mathbf{M}}) &= \psi_2(x^{\downarrow\mathbf{L}\cup\mathbf{M}})\cdot\sum_{z\in\mathbb{X}_\mathbf{K}}\psi_1(z,x^{\downarrow\mathbf{M}}),\\ \pi^{\downarrow\mathbf{M}}(x^{\downarrow\mathbf{M}}) &= \left(\sum_{z\in\mathbb{X}_\mathbf{K}}\psi_1(z,x^{\downarrow\mathbf{M}})\right)\cdot\left(\sum_{y\in\mathbb{X}_\mathbf{L}}\psi_2(y,x^{\downarrow\mathbf{M}})\right).\end{aligned}$$

Theorem 1.6 ***Block independence property.*** *For any probability distribution $\pi(\mathbf{J})$, and four disjoint subsets of variables $\mathbf{K}, \mathbf{L}, \mathbf{M}, \mathbf{N}$ ($\mathbf{K} \cup \mathbf{L} \cup \mathbf{M} \cup \mathbf{N} \subseteq \mathbf{J}$, and $\mathbf{K}$, $\mathbf{L}$, $\mathbf{M}$ are nonempty) the following two expressions (A) and (B) are equivalent*

(A) $\mathbf{K} \perp\!\!\!\perp \mathbf{L} \cup \mathbf{M} | \mathbf{N}\ [\pi]$,

(B) $\mathbf{K} \perp\!\!\!\perp \mathbf{M} | \mathbf{N}\ [\pi]$ *and* $\mathbf{K} \perp\!\!\!\perp \mathbf{L} | \mathbf{N} \cup \mathbf{M}\ [\pi]$.

Theorem 1.7 ***Block independence property for positive distributions.*** *For any* strictly positive *probability distribution $\pi(\mathbf{J})$, and four disjoint subsets of variables $\mathbf{K}, \mathbf{L}, \mathbf{M}, \mathbf{N}$ ($\mathbf{K} \cup \mathbf{L} \cup \mathbf{M} \cup \mathbf{N} \subseteq \mathbf{J}$, and $\mathbf{K}$, $\mathbf{L}$, $\mathbf{M}$ are nonempty) the following two expressions (A′) and (B′) are equivalent*

(A′) $\mathbf{K} \perp\!\!\!\perp \mathbf{L} \cup \mathbf{M} | \mathbf{N}\ [\pi]$,

(B′) $\mathbf{K} \perp\!\!\!\perp \mathbf{M} | \mathbf{N} \cup \mathbf{L}\ [\pi]$ *and* $\mathbf{K} \perp\!\!\!\perp \mathbf{L} | \mathbf{N} \cup \mathbf{M}\ [\pi]$.

Example 1.8 *This example is included to provide the reader with an illustration of the terms introduced in the previous Sections and to get accustomed to the notation that is commonly used in the field of probabilistic artificial intelligence. Let us point out that the latter may be rather unusual, for example, for statisticians.*

Table 1.3: Four-dimensional probability distribution.

X	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1
Y	0	0	0	0	1	1	1	1	0	0	0	0	1	1	1	1
Z	0	0	1	1	0	0	1	1	0	0	1	1	0	0	1	1
V	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1
μ	0	0	0	$\frac{1}{4}$	$\frac{1}{4}$	0	0	0	$\frac{1}{4}$	0	0	0	0	0	0	$\frac{1}{4}$

Consider four binary variables $\mathbf{N} = \{X.Y.Z, V\}$, again with $\mathbb{X}_X = \mathbb{X}_Y = \mathbb{X}_Z = \mathbb{X}_V = \{0, 1\}$, and a four-dimensional probability distribution $\mu(\mathbf{N})$ from Table 1.3. The observant reader has certainly noticed that its marginal distribution for $\mathbf{M} = \{X.Y.Z\}$ coincides with the distribution from Example 1.2. Moreover, marginalizing the distribution μ for variables Z and V, one can easily show that the added variable V is just a "copy" of the variable Z (the coin corresponding to variable V is laid on the table showing the same face as the coin corresponding to variable Z). Therefore $\mu^{\downarrow\{Z,V\}}(0, 0) = \mu^{\downarrow\{Z,V\}}(1, 1) = \frac{1}{2}$ and $\mu^{\downarrow\{Z,V\}}(0, 1) = \mu^{\downarrow\{Z,V\}}(1, 0) = 0$.

Let us find the independence structure of this distribution. From Example 1.4 we already know that any couple of variables from X, Y, Z are (unconditionally) independent. Therefore also $X \perp\!\!\!\perp V\ [\mu]$, $Y \perp\!\!\!\perp V\ [\mu]$. Intuitively, it is clear that $Z \not\perp\!\!\!\perp V\ [\mu]$; one can see it formally from Table 1.4.

Table 1.4: Marginals of probability distribution μ from Example 1.8.

Z	V	$\mu^{\downarrow Z}$	$\mu^{\downarrow V}$	$\mu^{\downarrow \{Z,V\}}$
0	0	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$
0	1	$\frac{1}{2}$	$\frac{1}{2}$	0
1	0	$\frac{1}{2}$	$\frac{1}{2}$	0
1	1	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$

What about conditional independence relationships? For each pair of variables, one should consider three conditional independence relationships. For X and Y, one should consider $X \perp\!\!\!\perp Y|Z\ [\mu]$, $X \perp\!\!\!\perp Y|V\ [\mu]$ and $X \perp\!\!\!\perp Y|\{Z,V\}\ [\mu]$. None of these three relationships holds true. Regarding the first two of them, it follows from Example 1.4; and it is easy to show it for the last of them as well. Namely, it is a trivial consequence of the fact that V is a "copy" of Z; we leave it to the reader to show it formally (hint: consider state $(0,0,0,0) \in \mathbb{X}_{\mathbf{N}}$).

For the pair X and Z, one should consider $X \perp\!\!\!\perp Z|Y\ [\mu]$, $X \perp\!\!\!\perp Z|V\ [\mu]$ and $X \perp\!\!\!\perp Z|\{Y,V\}\ [\mu]$. From Example 1.4 we know that $X \not\perp\!\!\!\perp Z|Y\ [\mu]$. Contrary to this, $X \perp\!\!\!\perp Z|V\ [\mu]$ and $X \perp\!\!\!\perp Z|\{Y,V\}\ [\mu]$. The validity of the former relationship can be seen from Table 1.5, or from Equivalence (1.2) because it is easy to show that $\mu(X|V) = \mu(X|Z,V)$. The validity of $X \perp\!\!\!\perp Z|\{Y,V\}\ [\mu]$ can be shown in an analogous way, and therefore it is, again, left to the reader.

Table 1.5: Marginals of probability distribution μ from Example 1.8.

X	Z	V	$\mu^{\downarrow V}$	$\mu^{\downarrow \{X,Z,V\}}$	$\mu^{\downarrow \{X,V\}}$	$\mu^{\downarrow \{Z,V\}}$
0	0	0	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{2}$
0	0	1	$\frac{1}{2}$	0	$\frac{1}{4}$	0
0	1	0	$\frac{1}{2}$	0	$\frac{1}{4}$	0
0	1	1	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{2}$
1	0	0	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{2}$
1	0	1	$\frac{1}{2}$	0	$\frac{1}{4}$	0
1	1	0	$\frac{1}{2}$	0	$\frac{1}{4}$	0
1	1	1	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{2}$

It remains to consider conditional independence relationships $Z \perp\!\!\!\perp V|X\ [\mu]$, $Z \perp\!\!\!\perp V|Y\ [\mu]$, and $Z \perp\!\!\!\perp V|\{X,Y\}\ [\mu]$. Neither of the first two relationships *holds because* $\mu(X,Z)\cdot\mu(X,V)$ *(*$\mu(Y,Z)\cdot\mu(Y,V)$*) is positive for all states from* $\mathbb{X}_{\{X,Z,V\}}$ *(*$\mathbb{X}_{\{Y,Z,V\}}$*), which does not hold for* $\mu(X,Z,V)$ *(*$\mu(Y,Z,V)$*). Thus, the only remaining question is that of the validity of* $Z \perp\!\!\!\perp V|\{X,Y\}\ [\mu]$*. It may be a surprise but this conditional independence relationship holds. To show it, the reader can create a table, in a way analogous to Table 1.5, with* 16 *rows corresponding to the states from* $\mathbb{X}_{\{X,Y,Z,V\}}$*, and the columns containing* $\mu^{\downarrow\{X,Y,Z,V\}}$*,* $\mu^{\downarrow\{X,Y\}}$*,* $\mu^{\downarrow\{X,Y,Z\}}$*,* $\mu^{\downarrow\{X,Y,Z\}}$*. At this place, we do not present such a table; our reasoning will be based on the equivalence from Expression* (1.2)*. We know that the values of variables* X *and* Y *uniquely determine the value of variable* Z*. It means that, when knowing the values of these two variables, we know both values of variables* Z *and* V*. Therefore* $\mu(V|\{X,Y,Z\}) = \mu(V|\{X,Y\})$*.*

We can now summarize the preceding paragraphs by presenting the conditional independence structure of the considered probability distribution μ *from Table 1.3:*

$$\begin{array}{lll}
X \perp\!\!\!\perp Y\ [\pi], & X \perp\!\!\!\perp Z|V\ [\pi], & X \perp\!\!\!\perp Z|\{Y,V\}\ [\pi],\\
X \perp\!\!\!\perp Z\ [\pi], & X \perp\!\!\!\perp V|Z\ [\pi], & X \perp\!\!\!\perp V|\{Y,Z\}\ [\pi],\\
X \perp\!\!\!\perp V\ [\pi], & Y \perp\!\!\!\perp Z|V\ [\pi], & Y \perp\!\!\!\perp Z|\{X,V\}\ [\pi],\\
Y \perp\!\!\!\perp Z\ [\pi], & Y \perp\!\!\!\perp V|Z\ [\pi], & Y \perp\!\!\!\perp V|\{X,Z\}\ [\pi],\\
Y \perp\!\!\!\perp V\ [\pi], & & Z \perp\!\!\!\perp V|\{X,Y\}\ [\pi].
\end{array}$$

Recall that, from this list, one can deduce all other conditional independence relationships holding for the considered probability distribution. For example, from $X \perp\!\!\!\perp Z\ [\pi]$ *and* $X \perp\!\!\!\perp V|Z\ [\pi]$ *one can deduce, using the Block Independence Property (Theorem 1.6) that* $X \perp\!\!\!\perp \{Z,V\}\ [\pi]$*.*

1.3 Decomposability

By 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 decompositions 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. In this example, we can see that

- two objects are the result of the decomposition; both the resulting objects are 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 call it a composition) yielding the original object from its decomposed parts – the composition of two integers is their product.

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

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. Under what conditions do we have a chance to reconstruct the original two-dimensional distribution π from its one-dimensional distributins $\pi^{\downarrow X}$ 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) are 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 $\pi^{\downarrow \{Y,Z\}}$ using Equation (1.1).

This consideration leads us to the following general definition.

Definition 1.9 *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. $\mathbf{K} \neq \mathbf{M}$*,* $\mathbf{L} \neq \mathbf{M}$*;*

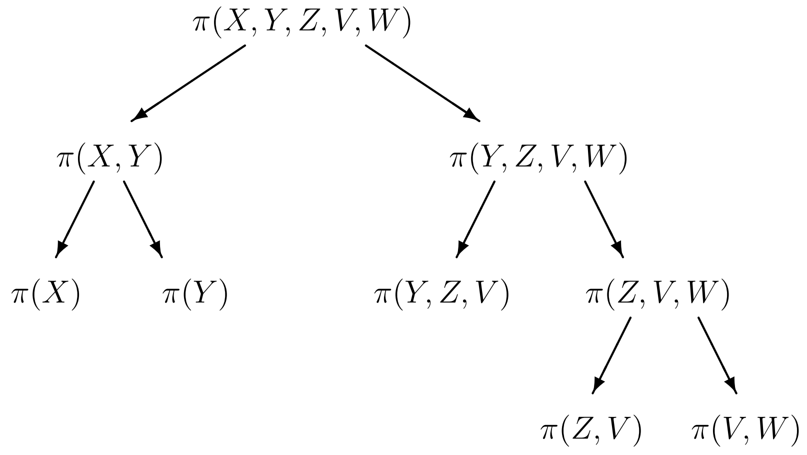


Figure 1.1: Hierarchical decomposition of $\pi(X,Y,Z,V,W)$.

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 other 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 the corresponding tree structure can be seen in Figure 1.1, where distribution $\pi(X,Y,Z,V,W)$ is decomposed into a system of its marginal distributions: $\pi^{\downarrow X}$, $\pi^{\downarrow Y}$, $\pi^{\downarrow\{Y,Z,V\}}$, $\pi^{\downarrow\{Z,V\}}$, and $\pi^{\downarrow\{V,W\}}$. Each decomposition has been made possible by the fact that the respective conditional independence relationship holds for distribution π. The decomposition process from Figure 1.1 has been made possible by the assumption that the following system of conditional independence relationships holds for distribution π (or, in other words, the independence structure of distribution π contains the following relationships):

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

Let us close this Section with a warning about a terminological paradox concerning the notion of *decomposability*. Namely, when speaking about a

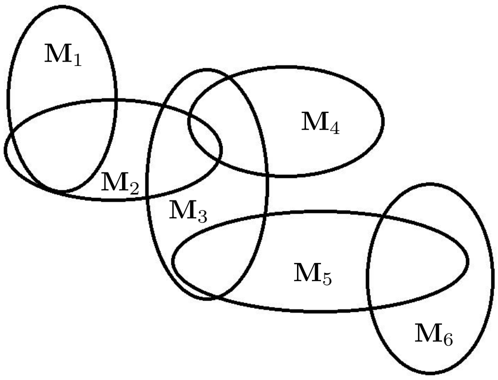


Figure 1.2: Six-dimensional-dimensional flexible model.

probability distribution that can be decomposed, we cannot say that it is decomposable. This term is, as we will see below in Definition 1.12, designated to another property. So, the reader will see that there are distributions that are decomposable but cannot be decomposed, as well as there are distributions that can be decomposed, and simultaneously, they are not decomposable[5].

To define the term of decomposability we need a property that may be met by any sequence of sets.

Definition 1.10 *We say a sequence of variable sets* $\mathbf{M}_1, \mathbf{M}_2, \ldots, \mathbf{M}_m$ *meets the* Running Intersection Property *(RIP) if*

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

At first sight, this property may seem rather technical. However, it has a simple interpretation. Imagine that you are constructing a union of all of the considered sets $\bigcup_{i=1}^{m} \mathbf{M}_i$ step by step. If the sets are ordered to meet the RIP, it means that in the k-th step you are adding set $\mathbf{M}_k$, the intersection of which with all the previously unified sets (i.e. $\mathbf{M}_k \cap \bigcup_{i=1}^{k-1} \mathbf{M}_i$) is covered by one of the previously added sets. In other words, we can say that the newly added set is added “through” one of the preceding sets. Thus, for example, the sequence $\mathbf{M}_1, \mathbf{M}_2$, $\mathbf{M}_3$, $\mathbf{M}_4, \mathbf{M}_5$, $\mathbf{M}_6$ shown in Figure 1.2 meets the RIP. Contrary to this fact, the sequence $\mathbf{M}_6, \mathbf{M}_5$, $\mathbf{M}_4$, $\mathbf{M}_3, \mathbf{M}_2$, $\mathbf{M}_1$ does not meet this property because $\mathbf{M}_3 \cap \bigcup_{i=6,5,4} \mathbf{M}_i$ is covered by none of $\mathbf{M}_4, \mathbf{M}_5$, $\mathbf{M}_6$.

In what follows we will need the following property.

[5] The same paradox also appears in graph theory; decomposable graphs are not those that can be decomposed.

Theorem 1.11 *If* $\mathbf{M}_1, \mathbf{M}_2, \ldots, \mathbf{M}_m$ *meet the RIP then for every* $j = 1, 2, \ldots, m$ *there exists its permutation* $\mathbf{M}_{k_1}, \mathbf{M}_{k_2}, \ldots, \mathbf{M}_{k_m}$ *meeting the RIP such that* $\mathbf{M}_{k_1} = \mathbf{M}_j$.

This assertion guarantees that if a system $\mathbf{M}_1, \mathbf{M}_2, \ldots, \mathbf{M}_m$ can be reordered to meet the RIP, then it can be ordered in many ways to meet this property; there exist at least m such orderings, because each set $\mathbf{M}_i$ may be placed at the beginning of a RIP ordering. Going back to Figure 1.2, the sequence $\mathbf{M}_3$, $\mathbf{M}_4, \mathbf{M}_5$, $\mathbf{M}_2$, $\mathbf{M}_6$, $\mathbf{M}_1$ meets the RIP. We recommend that the reader should find a RIP ordering that also starts with other sets from this example.

Now, we can define what we understand under the terms of a decomposable distribution.

Definition 1.12 *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 ordered so that they meet RIP.*

1.4 Information-theoretic notions

Most of the machine learning methods for probabilistic model construction (whether it is about Bayesian networks or compositional models) are, in a way, supported by notions and theoretical results from information theory. The value of a mutual information helps to find pairs of variables that are tightly connected to each other. The value of a multi-information may be used to select the best model from a considered group of models. Therefore, the user of a supervised system should properly understand these notions and their properties.

In the whole 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 the uncertainty connected with a 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 value 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 two groups of random variables we employ a notion of *mutual information* defined by the formula[6]

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

The higher this value, the stronger the dependence 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:

- $0 \le MI_{\pi}(\mathbf{K};\mathbf{L}) \le \min(H(\pi^{\downarrow \mathbf{K}}), H(\pi^{\downarrow \mathbf{L}}))$.
- $MI_{\pi}(\mathbf{K};\mathbf{L}) = 0 \iff \mathbf{K} \perp\!\!\!\perp \mathbf{L}\ [\pi]$.
- $MI_{\pi}(\mathbf{K};\mathbf{L}) = H(\pi^{\downarrow \mathbf{K}})$ if and only if variables $\mathbf{K}$ are functionally dependent on variables $\mathbf{L}$. It means that, in this case, it holds for the conditional distribution $\pi^{\mathbf{K}|\mathbf{L}}$ 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 other words, for each $y \in \mathbb{X}_{\mathbf{L}}$ there exists one and only one $x \in \mathbb{X}_{\mathbf{K}}$, for which $\pi^{\downarrow \mathbf{K}\cup\mathbf{L}}(x) > 0$.

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 [30], who called it *information measure of dependence*, is denoted by ID in this text:

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

It may help the reader to understand the notion of mutual information if we show that it is actually the measure of similarity between two distributions. In probability theory, several measures of distribution similarity have

[6] The reader has certainly realized that if $\pi^{\downarrow \mathbf{K}\cup\mathbf{L}}(x) > 0$ then also $\pi^{\downarrow \mathbf{K}}(x^{\downarrow \mathbf{K}}) > 0$ and $\pi^{\downarrow \mathbf{L}}(x^{\downarrow \mathbf{L}}) > 0$.

been introduced. One of them, having its origin in information theory, is the Kullback-Leibler divergence defined for $\pi(\mathbf{N})$ and $\nu(\mathbf{N})$ by the formula

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

It is known that the Kullback-Leibler divergence is always nonnegative and equals 0 if and only if $\pi = \nu$ (see [24, 23]). Its only disadvantage is that it is not symmetric, i.e., generally $Div(\pi \shortparallel \nu) \neq Div(\nu \shortparallel \pi)$. 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 can see that

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

is always finite, and, as we have already said above, it 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 other than $\mathbf{K} \perp\!\!\!\perp \mathbf{L}\ [\pi]$.

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

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

Example 1.13 *The notion of conditional mutual information is often employed to control the process of model learning from data. Therefore, it is important to realize that expanding the set of variables contained in the condition may either increase or decrease the value of conditional mutual information. More precisely, for* $\mathbf{M}^* \subseteq \mathbf{M}$,

$$MI_\pi(\mathbf{K}; \mathbf{L}|\mathbf{M}^*) \gtreqless MI_\pi(\mathbf{K}; \mathbf{L}|\mathbf{M}).$$

To illustrate this fact, let us consider three variables X, Y, Z from Example 1.2. For the distribution μ from Table 1.1 it holds:

$$MI_\mu(X; Y) = 0 < MI_\mu(X; Y|Z) = 1.$$

For another distribution of the same variables: distribution π, values of which are given in Table 1.6, we get

$$MI_\pi(X;Y) \doteq 0.046 > MI_\pi(X;Y|Z) = 0.$$

Table 1.6: Probability distribution, for which $X \perp\!\!\!\perp Y|Z\ [\pi]$.

X	0	0	0	0	1	1	1	1
Y	0	0	1	1	0	0	1	1
Z	0	1	0	1	0	1	0	1
π	$\frac{1}{32}$	$\frac{9}{32}$	$\frac{3}{32}$	$\frac{3}{32}$	$\frac{3}{32}$	$\frac{3}{32}$	$\frac{9}{32}$	$\frac{1}{32}$

The last notion to be introduced in this introductory chapter is another generalization of mutual information, for which we will use the term advocated by Studený [32]. By a *multi-information* of a probability distribution $\pi(\mathbf{N})$ we understand the value

$$IC(\pi) = \sum_{x \in \mathbb{X}_\mathbf{N} : \pi(x) > 0} \pi(x) \log_2 \left(\frac{\pi(x)}{\prod_{X \in \mathbf{N}} \pi^{\downarrow\{X\}}(x^{\downarrow\{X\}})} \right).$$

We see that the multi-information is again the Kullback-Leibler divergence between the distribution and a product of all of its one-dimensional marginals. Therefore, for two-dimensional distributions, the value of mutual information coincides with the corresponding value of mutual information. Just from the definition, one can also immediately deduce that it is a nonnegative and finite (because $\pi \ll \prod_{X \in \mathbf{N}} \pi^{\downarrow\{X\}}$), and equals 0 if and only if all variables from N are mutually independent for distribution π. However, notice that this independence is much stronger than pairwise independence between variables. Recall the distribution μ from Example 1.2 (see Table 1.1), for which we have shown that all three variables are pairwise independent, but these variables are not mutually independent (e.g., $X \not\perp\!\!\!\perp Y|Z\ [\mu]$ – see Example 1.4). The reader can easily show that

$$IC(\mu) = 4 \cdot \left(\frac{1}{4} \log_2 \left(\frac{\frac{1}{4}}{\frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2}} \right) \right) = 1.$$

Naturally, the value of multi-information is designed to measure the amount of information represented by multidimensional probability distributions. However, in some computational formulas it can happen that the value of multi-information should be computed from a one-dimensional marginal (or even from $\pi^{\downarrow\emptyset}$); let us point ut that, in this case, the multi-information value equals zero directly from the definition.

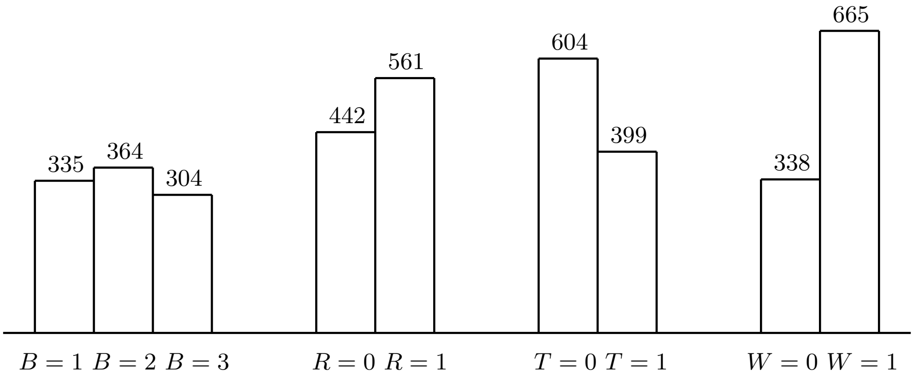


Figure 1.3: Frequencies of values of variables B, R, T, W.

Example 1.14 *This example shows how the introduced information-theoretic characteristics are computed from data. Consider four random variables* B, R, T, W *with* $\mathbb{X}_B = \{1, 2, 3\}$ *and* $\mathbb{X}_R = \mathbb{X}_T = \mathbb{X}_W = \{0, 1\}$*. In this example, we consider a data file with* $1,003$ *records; the frequencies of individual states of* $\mathbb{X}_{\{B,R,T,W\}}$ *are given in Table 1.7.*

Table 1.7: Frequencies of states from $\mathbb{X}_{\{B,R,T,W\}}$.

	$R = 0$				$R = 1$			
	$T = 0$		$T = 1$		$T = 0$		$T = 1$	
	$W = 0$	$W = 1$	$W = 0$	$W = 1$	$W = 0$	$W = 1$	$W = 0$	$W = 1$
$B = 1$	0	168	9	101	11	23	22	1
$B = 2$	0	81	3	49	44	96	89	2
$B = 3$	0	19	1	11	53	109	106	5

After collecting the data, we usually compute the frequencies of values for all variables. An example of such simple statistics is presented in Figure 1.3. From this, we can easily get the entropies of one dimensional marginal tables for all variables (denoted by a slightly nonstandard symbol in this example), e.g.,

$$\begin{aligned} H(B) &= \frac{-1}{1003}\left[\sum_{b\in\{1,2,3\}} f(b)\log_2\left(\frac{f(b)}{1003}\right)\right] \\ &= \frac{-1}{1003}\left[335\log_2\left(\frac{335}{1003}\right) + 364\log_2\left(\frac{364}{1003}\right) + 304\log_2\left(\frac{304}{1003}\right)\right] \\ &= 0.476. \end{aligned}$$

Analogously, $H(R) = 0.298$, $H(T) = 0.292$, $H(W) = 0.278$.

Table 1.8: Two dimensional marginals computed from Table 1.7.

	$R=0$	$R=1$
$T=0$	268	336
$T=1$	174	225

	$B=1$	$B=2$	$B=3$
$W=0$	42	136	160
$W=1$	293	228	144

Now, let us compute the mutual information for variables R and T. Computing the respective two-dimensional marginals (see Table 1.8), we get

$$\begin{aligned} MI(R;T) &= \frac{1}{1003} \sum_{(r,t)\in\mathbb{X}_{\{R,T\}}} f(r,t)\log_2\left(\frac{1003\cdot f(r,t)}{f(r)\cdot f(t)}\right) \\ &= \frac{1}{1003}\left[268\log_2\left(\frac{1003\times 268}{603\times 442}\right) + 336\log_2\left(\frac{1003\times 336}{604\times 561}\right)\right. \\ &\qquad \left. +174\log_2\left(\frac{1003\times 174}{399\times 442}\right) + 225\log_2\left(\frac{1003\times 225}{399\times 561}\right)\right] \\ &= 0.00004, \end{aligned}$$

which, being close to zero, suggests that variables R and T may be considered independent. Let us point out that the information measure of dependence between variables R and T equals

$$ID(R;T) = \frac{MI(R;T)}{\min(H(\pi(R)),(\pi(T)))} = \frac{0.00004}{\min(0.298, 0.292)} = 0.00014.$$

In a similar way, we can compute mutual information and information measure of dependence for variables B and W. From Table 1.8, the reader can check that $MI(B;W) = 0.625$, $ID(B;W) = 2.249$.

What about the conditional mutual information $MI(R;T|W)$? For this value, it is useful to create a three-dimensional frequency table for variables $R,T,W)$ – see Table 1.9

Table 1.9: Frequencies of states from $\mathbb{X}_{\{R,T,W\}}$.

	$R=0$		$R=1$	
	$T=0$	$T=1$	$T=0$	$T=1$
$W=0$	0	0.025	0,629	0.346
$W=1$	0.377	0.070	0.547	0.006

Let us compute[7]

$$\begin{aligned} MI(R;T|W) &= \frac{1}{1003} \sum_{(r,t,w)\in\mathbb{X}_{\{R,T,W\}}} f(r,t,w)\log_2\left(\frac{f(w)\cdot f(r,t,w)}{f(r,w)\cdot f(t.w)}\right) \\ &= \frac{1}{1003}\left[268\log_2\left(\frac{665\times 268}{0.447\times 0.924}\right) + 13\log_2\left(\frac{338\times 13}{0.025\times 0.371}\right)\right. \\ &\quad +161\log_2\left(\frac{665\times 161}{0.447\times 0.076}\right) + 108\log_2\left(\frac{338\times 108}{0.975\times 0.629}\right) \\ &\quad +228\log_2\left(\frac{665\times 228}{0.553\times 0.924}\right) + 217\log_2\left(\frac{338\times 217}{0.975\times 0.371}\right) \\ &\quad \left.+8\log_2\left(\frac{665\times 8}{0.553\times 0.076}\right)\right] = 0.091. \end{aligned}$$

This result is quite interesting from the viewpoint of knowledge discovery: since the conditional mutual information $MI(R;T|W)$ is much higher than the unconditional one, there is a strong chance that variable W may be influenced by both R and T.

1.5 Survey of symbols

Below we provide a list of notations that will be used throughout the entire text and have been introduced in the first Chapter:

X, Y, Z, V, W – finite valued random variables (page 4);

$\mathbb{X}_X$, $\mathbb{X}_Y$, $\mathbb{X}_Z$, $\mathbb{X}_V$, $\mathbb{X}_W$ – – finite sets of values achieved by random variables (page 4);

$\mathbf{K}$, $\mathbf{L}$, $\mathbf{M}$, $\mathbf{N}$ – – sets of random variables (page 4);

$\mathbb{X}_{\mathbf{K}}$, $\mathbb{X}_{\mathbf{L}}$, $\mathbb{X}_{\mathbf{M}}$, $\mathbb{X}_{\mathbf{N}}$ – sets of combinations of random variable values , set of states (page 4);

κ, λ, μ, ν, π – probability distributions (page 4);

$\pi \ll \kappa$ – distribution κ dominates π, (page 6);

$\kappa(\mathbf{K})$ – probability distribution for a set of variables $\mathbf{K}$ (page 4);

$\kappa^{\mathbf{M}|\mathbf{N}}$ – conditional probability distribution for variables $\mathbf{M}$ given variables $\mathbf{N}$ (page 4);

$\kappa^{\downarrow\mathbf{L}}$ – marginal probability distribution for a set of variables $\mathbf{L}$ (page 4);

[7] Recall that we sum over the states with positive frequencies.

$\mathbf{K} \perp\!\!\!\perp \mathbf{L}\ [\nu]$ – independence of variables $\mathbf{K}$ and $\mathbf{L}$ holding for distribution ν (page 7);

$\mathbf{K} \perp\!\!\!\perp \mathbf{L} | \mathbf{M}\ [\nu]$ – conditional independence of variables $\mathbf{K}$ and $\mathbf{L}$ given variables $\mathbf{M}$ holding for distribution ν (Definition 1.3, page 6);

the RIP – Running Intersection Property (Definition 1.10, page 14);

$H(\nu)$ – Shannon entropy of distribution ν (page 15);

$Div(\nu \shortparallel \mu)$ – Kulback-Leibler divergence of distributions ν and μ (page 17);

$MI_{\pi}(\mathbf{K};\mathbf{L})$ – mutual information between groups of variables $\mathbf{K}$ and $\mathbf{L}$ (for distribution π) (page 16);

$MI_{\pi}(\mathbf{K};\mathbf{L}|\mathbf{M})$ – conditional mutual information between groups of variables $\mathbf{K}$ and $\mathbf{L}$ given variables $\mathbf{M}$ (for distribution π) (page 17);

$ID_{\pi}(\mathbf{K};\mathbf{L})$ – information dependence measure between groups of variables $\mathbf{K}$ and $\mathbf{L}$ (for distribution π) (page 16);

$IC(\pi)$ – multi-information (page 18).

Chapter 2

Operator of composition

This chapter introduces and studies the properties of the most important notion on which the entire theory of compositional models is based: the operator of composition. Recall that this operator realizes a process that is an inverse to the process of decomposition discussed in Section 1.3. The first two Sections of this Chapter study the properties necessary for the construction of compositional models and their application to inference, Sections 2.3 and 2.4 are devoted to Csiszár's results concerning what he calls I-geometry of probability distributions. There are two good reasons for their being included in this text. First, they can easily be explained with the help of the operator of composition, second, the iterative process presented in Section 2.4 may advantageously be employed in the process of model learning.

2.1 Basic properties

Recall what we said in Section 1.3: $\pi(\mathbf{N})$ can be decomposed into its marginals $\pi(\mathbf{K})$ and $\pi(\mathbf{L})$ if $\mathbf{K} \cup \mathbf{L} = \mathbf{N}$ and $\pi(\mathbf{N}) \cdot \pi^{\downarrow \mathbf{K} \cap \mathbf{L}} = \pi^{\downarrow \mathbf{K}} \cdot \pi^{\downarrow \mathbf{L}}$. From this, one immediately gets that an inverse operation; namely, the operation of composition is

$$\pi(\mathbf{N}) = \frac{\pi^{\downarrow \mathbf{K}} \cdot \pi^{\downarrow \mathbf{L}}}{\pi^{\downarrow \mathbf{K} \cap \mathbf{L}}}.$$

This is trivial if we compose distributions $\pi(\mathbf{K})$ and $\pi(\mathbf{L})$ that are consistent. The question is whether one can also compose inconsistent distributions, i.e., distributions $\kappa(\mathbf{K})$ and $\lambda(\mathbf{L})$ for which $\kappa^{\downarrow \mathbf{K} \cap \mathbf{L}} \neq \lambda^{\downarrow \mathbf{K} \cap \mathbf{L}}$. For this instance, we accept the following definition that was first introduced in [8].

Definition 2.1 *For two arbitrary 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 fol-*

Table 2.1: Probability distributions κ and λ

κ	$X=0$	$X=1$
$Y=0$	$\frac{1}{4}$	$\frac{1}{4}$
$Y=1$	$\frac{1}{4}$	$\frac{1}{4}$

λ	$Z=0$	$Z=1$
$Y=0$	$\frac{1}{2}$	$\frac{1}{2}$
$Y=1$	0	0

lowing formula[1]

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

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

The presented definition is thus slightly more general than just an inverse operation to decomposition discussed in Section 1.3. In addition to the fact that we do not require the composed distributions to be consistent, we do not require that both $\mathbf{K}$ and $\mathbf{L}$ should be 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 the latter requirement appears advantageous when constructing the compositional models and using the resulting models for inference. For example, it enables the user to specify the required relationships of conditional independence, which would not otherwise be representable in a model.

Example 2.2 *Let us illustrate difficulties which would occur if the formula from Definition 2.1 were applied to situation with $\kappa^{\downarrow \mathbf{K}\cap\mathbf{L}} \not\ll \lambda^{\downarrow \mathbf{K}\cap\mathbf{L}}$.*

Consider distributions $\kappa(X,Y)$ and $\lambda(Y,Z)$ given in Table 2.1, for which $\kappa^{\downarrow Y}(1) > 0$ and $\lambda^{\downarrow Y}(1) = 0$.

The reader can immediately see that while computation $\lambda \triangleright \kappa$ according to the definition makes no problems, the computation of $\kappa \triangleright \lambda$ is impossible, because in this case we multiply a positive probability by an undefined expression $\frac{0}{0}$ – see Table 2.2.

The following assertion summarizes the basic properties of the operator of composition (most of the rule names are due to Prakash Shenoy). To simplify the following exposition, we will always assume that all the expressions in which the operator of composition appears are well defined. That means that the composed distributions meet the requirement on dominance required by Definition 2.1.

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

Table 2.2: Computation of $\lambda \triangleright \kappa$ and $\kappa \triangleright \lambda$

X	Y	Z	$\lambda \triangleright \kappa$	$\kappa \triangleright \lambda$
0	0	0	$\frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$	$\frac{1}{4} \cdot \frac{1}{2} = \frac{1}{8}$
0	0	1	$\frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$	$\frac{1}{4} \cdot \frac{1}{2} = \frac{1}{8}$
0	1	0	$0 \cdot \frac{1}{2} = 0$	$\frac{1}{4} \cdot \frac{0}{0} = ?$
0	1	1	$0 \cdot \frac{1}{2} = 0$	$\frac{1}{4} \cdot \frac{0}{0} = ?$
1	0	0	$\frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$	$\frac{1}{4} \cdot \frac{1}{2} = \frac{1}{8}$
1	0	1	$\frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$	$\frac{1}{4} \cdot \frac{1}{2} = \frac{1}{8}$
1	1	0	$0 \cdot \frac{1}{2} = 0$	$\frac{1}{4} \cdot \frac{0}{0} = ?$
1	1	1	$0 \cdot \frac{1}{2} = 0$	$\frac{1}{4} \cdot \frac{0}{0} = ?$

Theorem 2.3 *Consider three probability distributions $\kappa(\mathbf{K})$, $\lambda(\mathbf{L})$, and $\mu(\mathbf{M})$, for which all the compositions appearing in the following statements are defined. Then the following statements hold:*

1. (Domain): *$\kappa \triangleright \lambda$ is a probability distribution for variables $\mathbf{K} \cup \mathbf{L}$.*
2. (Conditional independence): $\mathbf{K} \setminus \mathbf{L} \perp\!\!\!\perp \mathbf{L} \setminus \mathbf{K} | \mathbf{K} \cap \mathbf{L}\ [\kappa \triangleright \lambda]$.
3. (Composition preserves the first marginal): $(\kappa \triangleright \lambda)^{\downarrow \mathbf{K}} = \kappa$.
4. (Reduction:) *If $\mathbf{L} \subseteq \mathbf{K}$ then, $\kappa \triangleright \lambda = \kappa$.*
5. (Extension:) *If $\mathbf{L} \subseteq \mathbf{K}$ then, $\kappa^{\downarrow \mathbf{L}} \triangleright \kappa = \kappa$.*
6. (Non-commutativity): *In general, $\kappa \triangleright \lambda \neq \lambda \triangleright \kappa$.*
7. (Commutativity under consistency): $\kappa^{\downarrow \mathbf{K} \cap \mathbf{L}} = \lambda^{\downarrow \mathbf{K} \cap \mathbf{L}}$ *if and only if* $\kappa \triangleright \lambda = \lambda \triangleright \kappa$.
8. (Non-associativity): *In general, $(\kappa \triangleright \lambda) \triangleright \mu \neq \kappa \triangleright (\lambda \triangleright \mu)$.*
9. (Associativity under the RIP): *If $\mathbf{K}, \mathbf{L}, \mathbf{M}$ meet the RIP (i.e., if either $\mathbf{K} \supset (\mathbf{L} \cap \mathbf{M})$, or $\mathbf{L} \supset (\mathbf{K} \cap \mathbf{M})$) then, $(\kappa \triangleright \lambda) \triangleright \mu = \kappa \triangleright (\lambda \triangleright \mu)$.*
10. (Stepwise composition): *If $(\mathbf{K} \cap \mathbf{L}) \subseteq \mathbf{M} \subseteq \mathbf{L}$ then, $(\kappa \triangleright \lambda^{\downarrow \mathbf{M}}) \triangleright \lambda = \kappa \triangleright \lambda$.*
11. (Exchangeability): *If $\mathbf{K} \supset (\mathbf{L} \cap \mathbf{M})$ then, $(\kappa \triangleright \lambda) \triangleright \mu = (\kappa \triangleright \mu) \triangleright \lambda$.*

12. (Simple marginalization): *If* $(\mathbf{K} \cap \mathbf{L}) \subseteq \mathbf{M} \subseteq \mathbf{K} \cup \mathbf{L}$ *then,* $(\kappa \triangleright \lambda)^{\downarrow \mathbf{M}} = \kappa^{\downarrow \mathbf{K} \cap \mathbf{M}} \triangleright \lambda^{\downarrow \mathbf{L} \cap \mathbf{M}}$.

13. (Irrelevant composition): *If* $\mathbf{M} \subseteq \mathbf{K} \setminus \mathbf{L}$ *then,* $\kappa \triangleright (\lambda \triangleright \mu) = \kappa \triangleright \lambda$.

The proofs of all the above-presented statements can be found in [14]. We do not repeat them in this text; nevertheless, in the rest of this Section, we comment on some of these assertions and/or illustrate them on simple (counter)examples.

1. *(Domain)* and 2. *(Conditional independence)*: Recall that the composition is a process inverse to decomposition. The decomposition yields, from a distribution, its marginals for smaller groups of variables. Therefore, the expansion of the groups of variables for which the distributions κ and λ are defined is the basic expected property of the composition. In this context, the reader has certainly noticed that, for disjoint $\mathbf{K}$ and $\mathbf{L}$, the composition $\kappa \triangleright \lambda$ is degenerated to a simple product of distributions: $\kappa \triangleright \lambda = \kappa \cdot \lambda$. In a general case, the composition may also be expressed in the following alternative way: $\kappa \triangleright \lambda = \kappa \cdot \lambda^{(\mathbf{L} \setminus \mathbf{K})|(\mathbf{L} \cap \mathbf{K})}$.

3. *(Composition preserves the first marginal):* When decomposing a distribution into its marginals one gets, quite naturally, a pair of consistent distributions. Since the operator is designed for the composition of distributions regardless whether they are consistent or not, the resulting composition must resolve the problem of coping with different information in the case of inconsistent distributions. The introduced operator takes all the information from the first argument and neglects the information from the second distribution if it contradicts the information from the first argument. For the properties of the operator preferring the information from the second argument, the reader is referred to [14]. The only problem remains in the case of the distributions being in total conflict; namely, the first argument equals zero and the second argument assumes a positive value of the respective marginal distribution. This is why the composition remains undefined in this case.

4. *(Reduction:)* Realize that this is the direct consequence of the previous Property 3.

6. *(Non-commutativity):* To illustrate it in the simplest way, consider $\kappa(X) \neq \lambda(X)$. Then, due to Property 4, $\kappa \triangleright \lambda = \kappa \neq \lambda \triangleright \kappa = \lambda$.

7. *(Commutativity under consistency):* This property is obvious from the definition of the operator. Let us only note that since any pair of

distributions defined for disjoint sets of variables are consistent ($\kappa^{\downarrow\emptyset} = \lambda^{\downarrow\emptyset} = 1$), this property also covers the trivial case

$$\mathbf{K} \cap \mathbf{L} = \emptyset \implies \kappa \triangleright \lambda = \kappa \cdot \lambda = \lambda \triangleright \kappa.$$

8. *(Non-associativity):* To illustrate the non-associativity, consider $\mu(X,Y)$ for which $X \not\perp\!\!\!\perp Y\ [\mu]$ and denote $\kappa = \mu^{\downarrow X}$, $\lambda = \mu^{\downarrow Y}$. For this $(\kappa \triangleright \lambda) \triangleright \mu = \mu^{\downarrow X} \cdot \mu^{\downarrow Y}$, and $\kappa \triangleright (\lambda \triangleright \mu) = \mu$ due to Properties 4 and 5.

9. *(Associativity under the RIP):* This property may be formulated as two independent properties: Associativity I, under the condition that $\mathbf{K} \supset (\mathbf{L} \cap \mathbf{M})$, and Associativity II under the condition that $\mathbf{L} \supset (\mathbf{K} \cap \mathbf{M})$. One should realize that these two rules are really independent, neither of them can be proved from the other. Nevertheless, it is not a difficult exercise for the reader to show that Associativity II can be proved from Associativity I and the property of Stepwise composition.

10. *(Stepwise composition):* Notice that it is, in a way, a generalization of Property 5. This property is also the direct consequence of Associativity II mentioned a few lines above, and the Extension property. Namely
$$\kappa \triangleright \lambda^{\mathbf{M}} \triangleright \lambda = \kappa \triangleright \left(\lambda^{\mathbf{M}} \triangleright \lambda\right) = \kappa \cdot \lambda.$$

11. *(Exchangeability):* To show that $(\kappa \triangleright \lambda) \triangleright \mu \neq (\kappa \triangleright \mu) \triangleright \lambda$ holds in general, the reader can consider distributions κ, λ, μ used to show the non-associativity of the operator of composition: $\kappa = \mu^{\downarrow X}$, $\lambda = \mu^{\downarrow Y}$, $\mu(X,Y)$ for which $X \not\perp\!\!\!\perp Y\ [\mu]$. In this case again $(\kappa \triangleright \lambda) \triangleright \mu = \mu^{\downarrow X} \cdot \mu^{\downarrow Y}$, and $(\kappa \triangleright \mu) \triangleright \lambda = \mu$.

12. *(Simple marginalization):* It is important to realize that the assertion holds only if($\mathbf{K} \cap \mathbf{L}) \subseteq \mathbf{M}$. We cannot marginalize out any variable from this intersection because these variables are intermediators of dependence between the variables from $\mathbf{K} \setminus \mathbf{L}$ and the variables from $\mathbf{L} \setminus \mathbf{K}$. Deleting any of these mediator variables could decrease the dependence between $\mathbf{K} \setminus \mathbf{L}$ and $\mathbf{L} \setminus \mathbf{K}$.

2.2 Anticipating operator

The associativity of the operator of composition would be desirable not only to meet the demands of mathematical beauty, but also to make the design of computational algorithms easier. Its lack is, in a way, compensated by the existence of a generalized operator of composition, which is called an anticipating operator and is studied in this Section.

Definition 2.4 *Consider an arbitrary set of variables* $\mathbf{M}$ *and two distributions* $\kappa(\mathbf{K})$, $\lambda(\mathbf{L})$. *Their* anticipating composition *is given by the formula*

$$\kappa \textcircled{$\triangleright$}_{\mathbf{M}}\lambda = (\lambda^{\downarrow(\mathbf{M}\setminus\mathbf{K})\cap\mathbf{L}} \cdot \kappa) \triangleright \lambda.$$

The operator $\textcircled{$\triangleright$}_{\mathbf{M}}$ *is called an* anticipating operator of composition.

Notice that it is a generalization of the operator introduced in Definition 2.1 in the sense that

$$\kappa \textcircled{$\triangleright$}_{\emptyset}\lambda = \kappa \triangleright \lambda.$$

Therefore, it is clear that the result of composition may remain undefined. However, it immediately follows from the respective definitions that if $\kappa\triangleright\lambda$ is defined then so is $\kappa \textcircled{$\triangleright$}_{\mathbf{M}}\lambda$. Both $\kappa \triangleright \lambda$ and $\kappa \textcircled{$\triangleright$}_{\mathbf{M}}\lambda$ are distributions defined for the same set of variables.

Let us also note that the realization of these two operators is of the same computational complexity. So, the main difference between the anticipating operator and the operator $\triangleright$ is that the generalized operator is parameterized by an index set. In the following Theorem (proved in [14]) we articulate the main purpose for which this operator is introduced. Namely, operator $\triangleright$ can be substituted by an anticipating operator while simultaneously changing the ordering of the operations.

Theorem 2.5 *If* $\kappa(\mathbf{K})$, $\lambda(\mathbf{L})$ *and* $\mu(\mathbf{M})$ *are such that* $\mu\triangleright(\kappa \textcircled{$\triangleright$}_{\mathbf{M}}\lambda)$ *is defined, then*

$$(\mu \triangleright \kappa) \triangleright \lambda = \mu \triangleright (\kappa \textcircled{$\triangleright$}_{\mathbf{M}}\lambda).$$

So, the purpose of this operator is to compose the distributions (in our case, distributions κ and λ), but to simultaneously introduce the necessary independence of variables $(\mathbf{M}\setminus\mathbf{K})\cap\mathbf{L}$ and $\mathbf{K}$ that would otherwise be omitted. If we want to compose distributions κ and λ before μ is considered, we have to "anticipate" the independence, which is originally introduced by the realization of the operator when composing μ and κ. This also explains why the operator $\textcircled{$\triangleright$}_{\mathbf{M}}$ is called an anticipating operator.

Example 2.6 *As said above, the specific purpose of the anticipating operator is to introduce the necessary conditional independence that would otherwise be omitted. To illustrate this point, let us consider three distributions* $\mu(X), \kappa(Y), \lambda(X,Y)$ *for which obviously* $(\mu(X) \triangleright \kappa(Y)) \triangleright \lambda(X,Y) = \mu(X)\kappa(Y)$. *If we used the operator* $\triangleright$ *instead of* $\textcircled{$\triangleright$}_{\mathbf{M}}$, *we would get*

$$\mu(X) \triangleright \Big(\kappa(Y) \triangleright \lambda(X,Y)\Big) = \frac{\mu(X)(\kappa(Y)\lambda(X|Y))}{\sum\limits_{x\in\mathbb{X}_X} \kappa(Y)\lambda(x|Y)},$$

which evidently differs from $\mu(X)\kappa(Y)$ because $\mu \triangleright (\kappa \triangleright \lambda)$ inherits the dependence of variables X and Y from λ. Nevertheless, considering

$$\begin{aligned}\mu(X) \triangleright (\kappa(Y) \textcircled{$\triangleright$}_X \lambda(X,Y)) &= \mu(X) \triangleright \Big(\lambda(X)\kappa(Y) \triangleright \lambda(X,Y)\Big) \\ &= \mu(X) \triangleright \lambda(X)\kappa(Y) = \mu(X)\kappa(Y),\end{aligned}$$

we get the desired result.

Perhaps it is also worth mentioning that in this example $\kappa(Y) \textcircled{$\triangleright$}_X \lambda(X,Y) = \lambda(X)\kappa(Y)$ is not a marginal distribution of the resulting $(\mu(X) \triangleright \kappa(Y)) \triangleright \lambda(X,Y) = \mu(X) \cdot \kappa(Y)$.

Example 2.7 *Let us present another, slightly more complex, example illustrating an application of the anticipating operator. This time, we consider distributions $\mu(X,Y,Z,U), \kappa(Y,Z,V), \lambda(Z,U,W)$. In this case, according to Theorem 2.5,*

$$\begin{aligned}&\Big(\mu(X,Y,Z,U) \triangleright \kappa(Y,Z,V)\Big) \triangleright \lambda(Z,U,W) \\ &\qquad = \mu(X,Y,Z,U) \triangleright \Big(\kappa(Y,Z,V) \textcircled{$\triangleright$}_{\{X,Y,Z,U\}} \lambda(Z,U,W)\Big).\end{aligned}$$

According to the definition of the anticipating operator

$$\begin{aligned}\kappa(Y,Z,V) \textcircled{$\triangleright$}_{\{X,Y,Z,U\}} \lambda(Z,U,W) &= \lambda(U)\kappa(Y,Z,V) \triangleright \lambda(Z,U,W) \\ &= \lambda(U)\kappa(Y,Z,V)\lambda(W|Z,U).\end{aligned}$$

The reader has most likely noticed that, thanks to the anticipating operator, $\lambda(W|Z,U)$ appears in this formula, which is exactly the form in which λ occurs in

$$\begin{aligned}&\Big(\mu(X,Y,Z,U) \triangleright \kappa(Y,Z,V)\Big) \triangleright \lambda(Z,U,W) \\ &\qquad = \mu(X,Y,Z,U)\kappa(V|Y,Z)\lambda(W|Z,U).\end{aligned}$$

Moreover,

$$\begin{aligned}&\Big(\kappa(Y,Z,V) \textcircled{$\triangleright$}_{\{X,Y,Z,U\}} \lambda(Z,U,W)\Big)^{\downarrow\{Y,Z,U\}} \\ &\qquad = (\lambda(U)\kappa(Y,Z,V)\lambda(W|Z,U))^{\downarrow\{Y,Z,U\}} \\ &\qquad = \lambda(U)\kappa(Y,Z).\end{aligned}$$

Therefore,

$$\begin{aligned}&\mu(X,Y,Z,U) \triangleright \Big(\kappa(Y,Z,V) \textcircled{$\triangleright$}_{\{X,Y,Z,U\}} \lambda(Z,U,W)\Big) \\ &\qquad = \mu(X,Y,Z,U)\frac{\lambda(U)\kappa(Y,Z,V)\lambda(W|Z,U)}{\lambda(U)\kappa(Y,Z)} \\ &\qquad = \mu(X,Y,Z,U)\kappa(V|Y,Z)\lambda(W|Z,U).\end{aligned}$$

2.3 Projection

In this Section, we study the properties of the operator of composition connected with information-theoretic characteristics of the composed probability distributions. First, let us present an assertion having a close connection to the property

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

that was already presented on page 17.

Theorem 2.8 *Consider a probability distribution $\nu(\mathbf{N})$, and three disjoint subsets of variables $\mathbf{K},\mathbf{L}$, and $\mathbf{M}$, for which $\mathbf{K}\cup\mathbf{L}\cup\mathbf{M}\subseteq\mathbf{N}$, and both $\mathbf{K},\mathbf{L}$ are nonempty. Then $MI_{\nu}(\mathbf{K};\mathbf{L}|\mathbf{M}) = 0$ (which is equivalent with $\mathbf{K}\perp\!\!\!\perp\mathbf{L}|\mathbf{M}\ [\nu]$) if and only if*

$$\nu^{\downarrow\mathbf{K}\cup\mathbf{L}\cup\mathbf{M}} = \nu^{\downarrow\mathbf{K}\cup\mathbf{M}} \triangleright \nu^{\downarrow\mathbf{L}\cup\mathbf{M}}.$$

The rest of this Section describes the results achieved by Imre Csiszár (see [4]), who introduced what he called an *I-geometry of probability distributions.* In the cited paper he discovered an interesting property of the operator of composition presented here in Theorem 2.11 below.

Definition 2.9 *Consider a distribution $\lambda(\mathbf{L})$ and an arbitrary subset of distributions for the same set of variables; denote it by $\Theta(\mathbf{L})$. Distribution*

$$\kappa = \arg\min_{\nu\in\Theta(\mathbf{L})} Div(\nu \shortparallel \lambda)$$

is called a projection[2] *of λ into $\Theta(\mathbf{L})$.*

According to this Definition, the projection is the distribution from $\Theta(\mathbf{L})$ which is, in a sense, closest to λ. As a measure of distance the Kullback-Leibler divergence introduced in Section 1.4 is considered. Recalling that this divergence is not symmetric, one must pay attention to the positions of distributions ν and λ in the Definition. If one minimized the value of $Div(\lambda \shortparallel \nu)$, the result of the projection could be different!

Generally, it may happen that, for given λ and $\Theta(\mathbf{L})$, this projection is not determined uniquely. However, when considering $\Theta(\mathbf{L})$ to be a set of distributions with the given marginal(s), which is always a convex compact set of distributions, the existence of a unique projection is guaranteed just by the existence of (at least) one $\nu \in \Theta(\mathbf{L})$ for which $Div(\nu \shortparallel \lambda)$ is finite. To formulate this property precisely, let us introduce the following notation.

[2] Csiszár calls it I-projection to emphasize the information-theoretic background of this notion.

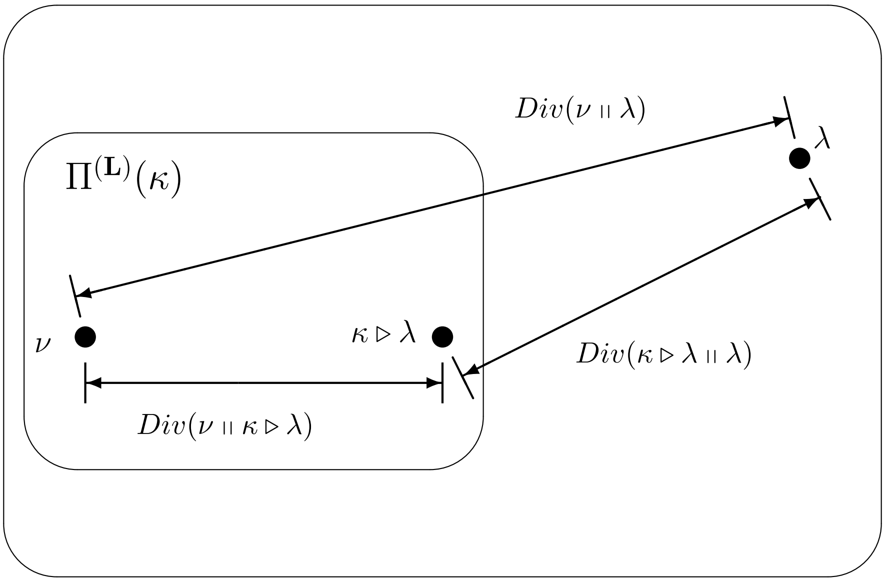

$Div(\nu \shortparallel \lambda)$
λ
$\Pi^{(\mathbf{L})}(\kappa)$
ν
$\kappa \triangleright \lambda$
$Div(\kappa \triangleright \lambda \shortparallel \lambda)$
$Div(\nu \shortparallel \kappa \triangleright \lambda)$

Figure 2.1: Projection of λ into $\Pi^{(\mathbf{L})}(\kappa)$

Definition 2.10 *Consider distribution* $\kappa(\mathbf{K})$ *and* $\mathbf{L} \supset \mathbf{K}$. *The set of all* extensions *of* κ *for variables* L *is denoted*

$$\Pi^{(\mathbf{L})}(\kappa) = \left\{ \nu(\mathbf{L}) : \nu^{\downarrow \mathbf{K}} = \kappa \right\}.$$

Instructions for finding a projection into the sets of extensions are given by the following assertion of Csiszár, which is graphically represented in Figure 2.1.

Theorem 2.11 *Consider* $\kappa(\mathbf{K})$ *and* $\mathbf{L} \supset \mathbf{K}$. *For an arbitrary probability distribution* $\lambda(\mathbf{L})$ *such that* $\kappa \ll \lambda^{\downarrow \mathbf{K}}$, $\kappa \triangleright \lambda$ *is the projection of* λ *into* $\Pi^{(\mathbf{L})}(\kappa)$. *Moreover,*

$$Div(\nu \shortparallel \lambda) = Div(\nu \shortparallel \kappa \triangleright \lambda) + Div(\kappa \triangleright \lambda \shortparallel \lambda),$$

for any $\nu \in \Pi^{(\mathbf{L})}(\kappa)$.

2.4 Iterative Proportional Fitting

Assume that $\kappa_1(\mathbf{K}_1), \kappa_2(\mathbf{K}_2), \ldots, \kappa_n(\mathbf{K}_n)$ are probability distributions. By the *marginal problem* we understand a task to find a multidimensional probability distribution of variables $\mathbf{N} = \mathbf{K}_1 \cup \mathbf{K}_2 \cup \ldots, \cup \mathbf{K}_n$ such that all distributions $\kappa_i(\mathbf{K}_i)$ are its marginals. In other words, we are looking for a joint extension of all distributions $\kappa_1(\mathbf{K}_1), \kappa_2(\mathbf{K}_2), \ldots, \kappa_n(\mathbf{K}_n)$.

Table 2.3: Distributions $\pi_1(X,Y), \pi_2(Y,Z), \pi_3(X,Z)$.

$\pi_1(X,Y)$		
	0	1
0	ε	$\frac{1}{2}--\varepsilon$
1	$\frac{1}{2}--\varepsilon$	ε

$\pi_2(Y,Z)$		
	0	1
0	ε	$\frac{1}{2}--\varepsilon$
1	$\frac{1}{2}--\varepsilon$	ε

$\pi_3(X,Z)$		
	0	1
0	ε	$\frac{1}{2}--\varepsilon$
1	$\frac{1}{2}--\varepsilon$	ε

Generally, it is not an easy task even to decide whether such a solution exists or not. Nevertheless, there are special situations in which one can decide this question almost immediately. The simplest case occurs when there is a couple of inconsistent distributions among the considered set of distributions. Obviously, $\kappa_i(\mathbf{K}_i)$ and $\kappa_j(\mathbf{K}_j)$, such that $\kappa_i^{\downarrow \mathbf{K}_i \cap \mathbf{K}_j} \neq \kappa_j^{\downarrow \mathbf{K}_i \cap \mathbf{K}_j}$, cannot have a joint extension. It means that a pairwise consistence of the distributions $\kappa_1(\mathbf{K}_1), \kappa_2(\mathbf{K}_2), \ldots, \kappa_n(\mathbf{K}_n)$ is a necessary condition for the existence of a joint extension. Unfortunately, this condition is not sufficient, as shown in the following Example. It is known that, in a general case, answering the question whether there exists a solution of a given marginal problem, is as difficult as finding a solution itself to that problem.

Example 2.12 *To see that the pairwise consistency is just a necessary and not a sufficient condition for the existence of a solution of the marginal problem, consider three variables X, Y, Z and three distributions π_1, π_2 and π_3 from Table 2.3. It is obvious that these distributions are pairwise consistent for all $\varepsilon \in [0, \frac{1}{2}]$. Taking $\varepsilon = \frac{1}{4}$, all three distributions π_1, π_2, π_3 are uniform, and the corresponding marginal problem has an infinite number of solutions; a three-dimensional uniform distribution is among them. By contrast, the corresponding marginal problem does not have a solution if we put $\varepsilon = 0$. This can easily be shown in the following way: assuming that $\mu(X,Y,Z)$ is a solution to this marginal problem, one can show that $\mu(x,y,z) = 0$ for each $(x,y,z) \in \mathbb{X} \times \mathbb{Y} \times \mathbb{Z}$. . For $(0,0,0)$ and $(0,0,1)$, this equality follows from the fact that $\mu(0,0,0)+\mu(0,0,1) = \pi_1(0,0)$; for $(0,1,0)$, it follows from $\mu(0,1,0) \leq \pi_3(0,0)$; and so on: $\mu(0,1,1) \leq \pi_2(1,1)$, $\mu(1,0,0) \leq \pi_2(0,0)$, $\mu(1,0,1) \leq \pi_3(1,1)$, $\mu(1,1,0) + \mu(1,1,1) = \pi_1(1,1)$.*

Moreover, the reader can verify that, for $\varepsilon = \frac{1}{6}$, the marginal problem has a unique solution

$$\nu(x,y,z) = \begin{cases} \frac{1}{6}, & \textit{if } x = y = z; \\ \frac{1}{3}, & \textit{otherwise.} \end{cases}$$

(Hint: first show that $\nu(x,y,z) \leq \frac{1}{6}$ for all $(x,y,z) \in \mathbb{X} \times \mathbb{Y} \times \mathbb{Z}$, and then that $\nu(x,y,z) \ = \frac{1}{3}$ if not $x = y = z$.)

The Iterative Proportional Fitting (IPF) procedure studied in this Section is a possible tool to solve the marginal problem. The procedure is ascribed to W. E. Deming and F. F. Stephan [5], who published the cited paper in 1940. But it took until 1975 for Imre Csiszár to prove its convergence [4].

Using the operator of composition, the description of this procedure is very simple. Start with a distribution $\pi_0(\mathbf{N})$ (as we will see later, there are reasons to start with the uniform distribution). Then compute

$$\begin{aligned}
\pi_1 &= \kappa_1 \triangleright \pi_0,\\
\pi_2 &= \kappa_2 \triangleright \pi_1,\\
\pi_3 &= \kappa_3 \triangleright \pi_2,\\
&\vdots\\
\pi_n &= \kappa_n \triangleright \pi_{n-1},\\
\pi_{n+1} &= \kappa_1 \triangleright \pi_n,\\
\pi_{n+2} &= \kappa_1 \triangleright \pi_{n+1},\\
&\vdots\\
\pi_k &= \kappa_{(k-1 \pmod{n})+1} \triangleright \pi_{k-1},\\
&\vdots
\end{aligned}$$

The behavior of this infinite sequence of probability distributions is described by the following Csiszár's theorem.

Theorem 2.13 *Consider probability distributions $\kappa_1(\mathbf{K}_1), \kappa_2(\mathbf{K}_2), \ldots, \kappa_n(\mathbf{K}_n)$ such that $\mathbf{N} = \mathbf{K}_1 \cup \mathbf{K}_2 \cup \ldots, \cup \mathbf{K}_n$. Denote*

$$\Pi^{(\mathbf{N})}(\kappa_1, \ldots, \kappa_n) = \{\nu(\mathbf{N}) : \nu^{\downarrow \mathbf{K}_i} = \kappa_i \textit{ for all } i = 1, \ldots, n\}.$$

Let $\pi_0(\mathbf{N}), \pi_1(\mathbf{N}), \pi_2(\mathbf{N}), \ldots$ be an infinite sequence of distributions computed by IPF procedure from $\kappa_1, \ldots, \kappa_n$. Then there exists $\nu \in \Pi^{(\mathbf{N})}(\kappa_1, \ldots, \kappa_n)$ such that $\nu \ll \pi_0$ if and only if the sequence of probability distributions $\pi_0, \pi_1, \pi_2, \ldots$ converges. In this case this sequence converges to a probability distribution $\pi^* \in \Pi^{(\mathbf{N})}(\kappa_1, \ldots, \kappa_n)$, and the following equality holds for any $\nu \in \Pi^{(\mathbf{N})}(\kappa_1, \ldots, \kappa_n)$

$$Div(\nu \parallel \pi_0) = Div(\nu \parallel \pi^*) + Div(\pi^* \parallel \pi_0).$$

Corollary 2.14 *Under the same assumptions as in Theorem 2.13 and with the same notation, let $\pi_0(\mathbf{N})$ be a product of one-dimensional positive distributions (i.e., $\pi_0(\mathbf{N}) = \prod_{X \in \mathbf{N}} \pi_0^{\downarrow X}$). Then $\Pi^{(\mathbf{N})}(\kappa_1, \ldots, \kappa_n) \neq \emptyset$ if and only if the sequence of probability distributions $\pi_0, \pi_1, \pi_2, \ldots$ converges. In this case this sequence converges to a probability distribution $\pi^* \in \Pi^{(\mathbf{N})}(\kappa_1, \ldots, \kappa_n)$, for which*

$$H(\pi^*) = \max_{\nu \in \Pi^{(\mathbf{N})}(\kappa_1, \ldots, \kappa_n)} (H(\nu)).$$

There are two consequences following from the Csiszár theorem and its Corollary. First, if we choose for π_0 the uniform distribution (which is strictly positive, and it is a product of its one-dimensional marginals), then IPFP converges if and only if the respective marginal problem has a solution, and the process converges to the maximum entropy solution of the marginal problem. Second, if the process with uniform π_0 does not converge, then the respective marginal problem does not have a solution.

Chapter 3

Compositional models

At this point, we start considering *multidimensional compositional models*, i.e., multidimensional probability distributions assembled from sequences of low-dimensional distributions with the help of the operators of composition. The first two Sections, which are based on [14], study properties of two special families of compositional models possessing advantageous properties that will be employed when designing computational procedures for marginalization (Section 3.3) and conditioning (Section 3.4).

To avoid certain technical problems and the necessity of repeating some assumptions too many times, let us formulate the following three important conventions.

- In this and the following Chapters, we consider systems of distributions $\kappa_1(\mathbf{K}_1)$, $\kappa_2(\mathbf{K}_2), \ldots, \kappa_n(\mathbf{K}_n)$. Therefore, whenever we speak about a distribution κ_k, if not explicitly specified otherwise (usually in examples), the distribution κ_k will always be assumed to be defined for variables $\mathbf{K}_k$.

- Based on the convention just stated, the formula $\kappa_1 \triangleright \kappa_2 \triangleright \ldots \triangleright \kappa_n$, if defined, is 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

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

- In what follows, we will always assume that the composition is defined in all the formulas wherever the operator appears. This last convention enables us to omit repeating the assumption regarding the required dominance holding between the composed distributions.

Example 3.1 *In agreement with what has just been said, the compositional model*

$$\kappa_1(X,Z) \triangleright \kappa_2(Z,V) \triangleright \kappa_3(X,U,V,W) \triangleright \kappa_4(Y,V,W)$$

is a distribution

$$\begin{aligned}(\kappa_1 \triangleright \kappa_2 \triangleright \kappa_3 \triangleright \kappa_4)(X,Y,Z,U,V,W)&\\ &= ((\kappa_1(X,Z) \triangleright \kappa_2(Z,V)) \triangleright \kappa_3(X,U,V,W)) \triangleright \kappa_4(Y,V,W)\\ &= \kappa_1(X,Z)\kappa_2(V|Z)\kappa_3(U,W|X,V)\kappa_4(Y|V,W).\end{aligned}$$

When speaking about the multidimensional compositional model $\kappa_1 \triangleright \kappa_2 \triangleright \ldots \triangleright \kappa_n$, we should realize that when computing

$$(\kappa_1 \triangleright \ldots \triangleright \kappa_k) \triangleright \kappa_{k+1} = \frac{(\kappa_1 \triangleright \ldots \triangleright \kappa_k)\, \kappa_{k+1}}{\kappa_{k+1}^{\downarrow \mathbf{K}_{k+1} \cap (\mathbf{K}_1 \cup \ldots \cup \mathbf{K}_k)}},$$

one has to marginalize distribution κ_{k+1}, which is assumed to be a low-dimensional, and therefore this marginalization is easily tractable. Recall the comment on page 26 discussing the property *composition preserves the first marginal*. We mentioned that in [14], an operator preserving the second argument was also defined. If we considered models created with the help of this alternative operator of composition, we would have to marginalize the model assembled from $\kappa_1, \kappa_2, \ldots, \kappa_k$ instead of marginalizing a single distribution κ_{k+1}. And the marginalization of a model would quite often be intractable. This stresses the importance of Property 3 of Theorem 2.3.

Notice that when defining a compositional model we have not imposed any conditions on sets of variables for which the distributions are defined. For example, considering a model, in which one distribution is defined for a subset of variables of another distribution (i.e., $\mathbf{K}_j \subset \mathbf{K}_k$) is fully sensible and may carry information about the distribution. If $\pi(X,Y,Z) = \kappa_1(X) \triangleright \kappa_2(Y) \triangleright \kappa_3(X,Y,Z)$ is a compositional model of a three-dimensional distribution, one can immediately see that variables X and Y are independent. Not knowing the numbers defining the distribution, one cannot say anything similar about distribution $\pi(X,Y,Z)$. (How to read the conditional independence relationships from a compositional model will be discussed in Chapter 4.)

3.1 Perfect models

Not all compositional models are equally efficient when used for the representation of multidimensional distributions. Among them, the so-called perfect models hold an important position.

Definition 3.2 A compositional model $\kappa_1 \triangleright \kappa_2 \triangleright \ldots \triangleright \kappa_n$ is called *perfect* if

$$\begin{aligned}
&\kappa_1 \triangleright \kappa_2 = \kappa_2 \triangleright \kappa_1, \\
&\kappa_1 \triangleright \kappa_2 \triangleright \kappa_3 = \kappa_3 \triangleright (\kappa_1 \triangleright \kappa_2), \\
&\qquad \vdots \\
&\kappa_1 \triangleright \kappa_2 \triangleright \ldots \triangleright \kappa_n = \kappa_n \triangleright (\kappa_1 \triangleright \kappa_2 \triangleright \ldots \triangleright \kappa_{n-1}).
\end{aligned}$$

From this definition one can hardly see the importance of perfect sequences. This importance becomes clearer from the following characterization given in Theorem 3.4. First, however, let us present a technical property, which, being an immediate consequence of an inductive application of Property 7 of Theorem 2.3, can be used for testing the perfectness of compositional models.

Theorem 3.3 *A compositional model $\kappa_1 \triangleright \kappa_2 \triangleright \ldots \triangleright \kappa_n$ is perfect if and only if the pairs of distributions $(\kappa_1 \triangleright \ldots \triangleright \kappa_{i-1})$ and κ_i are consistent for all $i = 2, 3, \ldots, n$.*

Theorem 3.4 *A compositional model $\kappa_1 \triangleright \kappa_2 \triangleright \ldots \triangleright \kappa_n$ is perfect if and only if all the distributions in this sequence are marginals of the distribution $(\kappa_1 \triangleright \kappa_2 \triangleright \ldots \triangleright \kappa_n)$, i.e.,*

$$\forall i = 1, 2, \ldots, n \quad (\kappa_1 \triangleright \kappa_2 \triangleright \ldots \triangleright \kappa_n)^{\downarrow \mathbf{K}_i} = \kappa_i.$$

This Theorem is very important, and its proof is surprisingly simple. It is based on the following idea. Consider a perfect compositional model $(\kappa_1 \triangleright \kappa_2 \triangleright \ldots \triangleright \kappa_n)$. Clearly, κ_1 is its marginal because of Property 3 of Theorem 2.3 Due to the same reason, $\kappa_1 \triangleright \kappa_2$ is also its marginal. For perfect models, fulfilment of the condition $\kappa_1 \triangleright \kappa_2 = \kappa_2 \triangleright \kappa_1$ is required, which means that κ_2 is a marginal of $\kappa_1 \triangleright \kappa_2$ and therefore of the entire model. We can now consider $\kappa_1 \triangleright \kappa_2 \triangleright \kappa_3$, which is a marginal of the entire model. Due to the perfectness of the model, we know that

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

which means that κ_3, being a marginal of $\kappa_1 \triangleright \kappa_2 \triangleright \kappa_3$, is a marginal to the entire model. Repeating this reasoning, we can prove that all elements of a perfect model are its marginal distributions.

Assuming that all $\kappa_1, \kappa_2, \ldots, \kappa_n$ are marginals of a compositional model $\kappa_1 \triangleright \kappa_2 \triangleright \ldots \triangleright \kappa_n$, one knows that all these distributions are pairwise consistent. Therefore $\kappa_1 \triangleright \kappa_2 = \kappa_2 \triangleright \kappa_1$ holds due to Property 7 of Theorem 2.3 (Commutativity under consistency). Because of Property 3 (Composition preserves the first marginal) of the same Theorem, $\kappa_1 \triangleright \kappa_2$ is also a marginal

Table 3.1: Three-dimensional distribution $\pi(X, Y, Z)$

π	$X = 0$		$X = 1$	
	$Y = 0$	$Y = 1$	$Y = 0$	$Y = 1$
$Z = 0$	0.1	0.1	0.2	0.1
$Z = 1$	0.0	0.1	0.0	0.1
$Z = 2$	0.2	0.0	0.0	0.1

of the considered model, and we assume the same condition about κ_3. It means that these two distributions must be consistent, and therefore

$$\kappa_1 \triangleright \kappa_2 \triangleright \kappa_3 = \kappa_3 \triangleright (\kappa_1 \triangleright \kappa_2).$$

Again, this simple idea can be repeated until we prove that the model is perfect, i.e., it meets the requirements of Definition 3.2.

In other words, Theorem 3.4 says that taking low-dimensional distributions κ_i for carriers of local information, the resulting multidimensional distribution, if it is a perfect model, represents global information faithfully reflecting all the local information. This is one of the reasons why we will be very much interested in perfect sequence models.

Example 3.5 *The above-presented Theorem claims that a model preserves all the given marginals if and only if the model is perfect. If the considered model is not perfect than some of the marginal distributions differ from the given ones. In this example we show that non-perfect models need not preserve one-dimensional marginal distributions – even if the given distributions are pairwise consistent.*

Consider a three-dimensional distribution $\pi(X, Y, Z)$ from Table 3.1. Denote $\kappa_1(X) = \pi^{\downarrow X}, \kappa_2(Y) = \pi^{\downarrow Y}$ and $\kappa_3(X, Y, Z) = \pi(X, Y, Z)$. Let us further study the distribution

$$\nu(X, Y, Z) = \kappa_1(X) \triangleright \kappa_2(Y) \triangleright \kappa_3(X, Y, Z)$$

composed from the pairwise consistent distributions (keeping in mind that all the marginals of a given distribution are pairwise consistent). Since both the considered one-dimensional marginal distributions $\kappa_1(X), \kappa_2(Y)$ are uniform, the composition $\kappa_1 \triangleright \kappa_2$ is also uniform. Thus it is an easy task to compute distribution ν, which is done in Table 3.2. Computing one-dimensional marginals $\pi^{\downarrow Z}$ and $\nu^{\downarrow Z}$ from Tables 3.1 and 3.2, we can see that these dis-

Table 3.2: Distribution $\nu(X,Y,Z) = \kappa_1(X) \triangleright \kappa_2(Y) \triangleright \kappa_3(X,Y,Z)$

	$X=0$		$X=1$	
	$Y=0$	$Y=1$	$Y=0$	$Y=1$
$Z=0$	$\frac{2}{24}$	$\frac{3}{24}$	$\frac{6}{24}$	$\frac{2}{24}$
$Z=1$	0.0	$\frac{3}{24}$	0.0	$\frac{2}{24}$
$Z=2$	$\frac{4}{24}$	0.0	0.0	$\frac{2}{24}$

tributions are different:

$$\begin{aligned} \kappa(Z=0) &= 0.5 \qquad & \nu(Z=0) &= \frac{13}{24}, \\ \kappa(Z=1) &= 0.2 \qquad & \nu(Z=1) &= \frac{5}{24}, \\ \kappa(Z=2) &= 0.3 \qquad & \nu(Z=2) &= \frac{6}{24}. \end{aligned}$$

Let us emphasize yet another difference between distributions π and ν. Due to Property 3 of Theorem 2.3, we can see that $\nu^{\downarrow\{X,Y\}} = \kappa_1(X) \cdot \kappa_2(Y)$, which means that $X \perp\!\!\!\perp Y\ [\nu]$. If we want to verify whether variables X and Y are independent even with distribution π, we have to check the equality

$$\pi^{\{X,Y\}}(x,y) = \pi^X(x) \cdot \pi^Y(y)$$

for all states $(x,y) \in \{0,1\} \times \{0,1\}$. In our case for distribution from Table 3.1, we obtain that $X \not\perp\!\!\!\perp Y\ [\pi]$ because (for example) for $x = y = 0$

$$\pi^{\{X,Y\}}(0,0) = \frac{1}{3},$$

and

$$\pi^X(0) = \frac{1}{2}, \quad \pi^Y(0) = \frac{1}{2}.$$

The following assertion is of a great importance. It says that restricting our consideration only to perfect models is not a loss of generality because all compositional models can be transformed into equivalent perfect models.

Theorem 3.6 *For any compositional model $\kappa_1 \triangleright \kappa_2 \triangleright \ldots \triangleright \kappa_n$, the model $\nu_1 \triangleright \nu_2 \triangleright \ldots \triangleright \nu_n$, the distributions of which are computed by the following* *process*

$$\begin{aligned}
\nu_1 &= \kappa_1,\\
\nu_2 &= \nu_1^{\downarrow \mathbf{K}_2 \cap \mathbf{K}_1} \triangleright \kappa_2,\\
\nu_3 &= (\nu_1 \triangleright \nu_2)^{\downarrow \mathbf{K}_3 \cap (\mathbf{K}_1 \cup \mathbf{K}_2)} \triangleright \kappa_3,\\
&\vdots\\
\nu_n &= (\nu_1 \triangleright \ldots \triangleright \nu_{n-1})^{\downarrow \mathbf{K}_n \cap (\mathbf{K}_1 \cup \ldots \cup \mathbf{K}_{n-1})} \triangleright \kappa_n,
\end{aligned}$$

is perfect, and

$$\kappa_1 \triangleright \ldots \triangleright \kappa_n = \nu_1 \triangleright \ldots \triangleright \nu_n.$$

From the theoretical point of view the process of perfectization described in Theorem 3.6 is simple. Unfortunately, its computational complexity is far from being moderate. Namely, the process requires marginalization of compositional models, which may be multidimensional. As we will see in Section 3.3, the marginalization may be computationally hard. Nevertheless, the next Section is devoted to a subclass of models for which the perfectization procedure may be materialized very efficiently.

Now, let us present another important property of perfect models. An arbitrary perfect model $\kappa_1 \triangleright \kappa_2 \triangleright \ldots \triangleright \kappa_n$ (with $n > 1$) can always be reordered (permuted) in such a way way that this permutation $\kappa_{i_1} \triangleright \kappa_{i_2} \triangleright \ldots \triangleright \kappa_{i_n}$ is also perfect. Trivially, if $\kappa_1 \triangleright \kappa_2 \triangleright \ldots \triangleright \kappa_n$ is perfect then $\kappa_2 \triangleright \kappa_1 \triangleright \kappa_3 \triangleright \ldots \triangleright \kappa_n$ is perfect, too (it follows directly from Property 7 of Theorem 2.3 and Theorem 3.4). The following assertion guarantees that if two perfect models are set up from the same system of low-dimensional distributions then these models are equivalent.

Theorem 3.7 *If a compositional model $\kappa_1, \kappa_2, \ldots, \kappa_n$ and its permutation κ_{i_1}, κ_{i_2}, $\ldots, \kappa_{i_n}$ are both perfect then*

$$\kappa_1 \triangleright \kappa_2 \triangleright \ldots \triangleright \kappa_n = \kappa_{i_1} \triangleright \kappa_{i_2} \triangleright \ldots \triangleright \kappa_{i_n}.$$

From the practical point of view, it is important to realize that, for perfect models, one can always compute their information-theoretic characteristics in an efficient way. It is guaranteed by the following two assertions. Theorem 3.8 gives instructions for how to compute entropy of a distribution represented by a perfect sequence model. Theorem 3.9 presents instructions for how to compute the informational content of distributions represented by perfect models. Both these assertions are utilized in the procedures for model learning.

Theorem 3.8 *If the model* $\kappa_1 \triangleright \kappa_2 \triangleright \ldots \triangleright \kappa_n$ *is perfect then*

$$H(\kappa_1 \triangleright \kappa_2 \triangleright \ldots \triangleright \kappa_n) = \sum_{i=1}^{n} H(\kappa_i) - \sum_{i=2}^{n} H\left(\kappa_i^{\downarrow \mathbf{K}_i \cap (\mathbf{K}_1 \cup \ldots \cup \mathbf{K}_{i-1})}\right) \geq H(\mu)$$

for any μ, *which is a joint extension of all distributions*[1] κ_i*:*

$$\mu \in \bigcap_{i=1}^{n} \Pi^{(\mathbf{K}_1 \cup \ldots \cup \mathbf{K}_n)}(\kappa_i) = \left\{\nu(\mathbf{K}_1 \cup \ldots \cup \mathbf{K}_n) \ : \ \forall i = 1, \ldots, n \ \ \nu^{\downarrow \mathbf{K}_i} = \kappa_i\right\}.$$

Theorem 3.9 *If the model* $\kappa_1 \triangleright \kappa_2 \triangleright \ldots \triangleright \kappa_n$ *is perfect, then*

$$IC(\kappa_1 \triangleright \kappa_2 \triangleright \ldots \triangleright \kappa_n) = \sum_{i=1}^{n} IC(\kappa_i) - \sum_{i=2}^{n} IC\left(\kappa_i^{\downarrow \mathbf{K}_i \cap (\mathbf{K}_1 \cup \ldots \cup \mathbf{K}_{i-1})}\right).$$

Let us realize that both the computational formulas expressed in Theorems 3.8 and 3.9 can only be used for perfect models. If a model is not perfect then the computation of entropy and/or informational content of the model is usually more complex (the process usually requires the computation of a system of marginal distributions). Roughly speaking, from the algorithmic point of view, it is as complex as the realization of the perfectization procedure described in Theorem 3.6.

3.2 Decomposable models

Having a compositional model, one can apply Theorem 3.3 to verify whether the model is perfect or not. The perfectness of a model is a strong property and it may happen that its verification is not easy. Nevertheless, in some special situations one can see the answer immediately. As a degenerate example, the reader can consider a model composed from uniform distributions, the perfectness of which is easy to prove. Naturally, this situation is degenerate and therefore uninteresting. From the practical point of view, more important situations are covered by Theorem 3.11 below. In fact, this assertion is just a "translation" of a classical result of Kellerer [21] into the language of this text. To formulate it, let us recall that a sequence of sets $\mathbf{K}_1, \mathbf{K}_2, \ldots, \mathbf{K}_n$ is said to meet the *running intersection property* if

$$\forall j = 2, \ldots, n \ \ \exists k \ \ (1 \leq k < j) \qquad \left(\mathbf{K}_j \cap \left(\bigcup_{i=1}^{j-1} \mathbf{K}_i\right) \subseteq \mathbf{K}_k\right).$$

In agreement with Definition 1.12, in which we defined a notion of a decomposable distribution, we can define a decomposable model in the following way.

[1] Notice that the perfectness of $\kappa_1 \triangleright \kappa_2 \triangleright \ldots \triangleright \kappa_n$ guarantees that $\Pi^{(\mathbf{K}_1 \cup \ldots \cup \mathbf{K}_n)}(\kappa_i) \neq \emptyset$.

Definition 3.10 *We call a compositional model* $\kappa_1(\mathbf{K}_1) \triangleright \kappa_2(\mathbf{K}_2) \triangleright \ldots \triangleright \kappa_n(\mathbf{K}_n)$ decomposable *if the sequence* $\mathbf{K}_1, \ldots, \mathbf{K}_n$ *meets the RIP.*

Let us point out the relationship between Definitions 1.12 and 3.10.

- If a decomposable distribution $\pi(\mathbf{N})$ can be decomposed into its marginals $\pi(\mathbf{M}_1)$, $\pi(\mathbf{M}_2)$, $\ldots$, $\pi(\mathbf{M}_n)$, then (iteratively using Property 2 of Theorem 2.3) one can easily show that $\pi(\mathbf{N}) = \pi(\mathbf{M}_1) \triangleright \pi(\mathbf{M}_2) \triangleright \ldots \triangleright \pi(\mathbf{M}_n)$, for any ordering $\mathbf{M}_1, \mathbf{M}_2, \ldots, \mathbf{M}_n$ that meets the RIP.

- For any decomposable model $\pi(\mathbf{K}_1 \cup \ldots \cup \mathbf{K}_n) = \kappa_1(\mathbf{K}_1) \triangleright \kappa_2(\mathbf{K}_2) \triangleright \ldots \triangleright \kappa_n(\mathbf{K}_n)$, distribution π can be decomposed (again using Property 2 of Theorem 2.3) into a system of its marginals $\pi^{\downarrow \mathbf{K}_1}$, $\pi^{\downarrow \mathbf{K}_2}$, $\ldots$, $\pi^{\downarrow \mathbf{K}_n}$.

Thus, any decomposable probability distribution can be represented in a form of a decomposable model; and any decomposable model represents a decomposable distribution. As we will see below, the importance of decomposable models stems from the fact that, for them, most of the computational procedures appear to be computationally very efficient. As a typical example, let us consider the perfectization procedure described in Theorem 3.6

$$\begin{aligned}
\nu_1 &= \kappa_1, \\
\nu_2 &= \nu_1^{\downarrow \mathbf{K}_2 \cap \mathbf{K}_1} \triangleright \kappa_2, \\
\nu_3 &= (\nu_1 \triangleright \nu_2)^{\downarrow \mathbf{K}_3 \cap (\mathbf{K}_1 \cup \mathbf{K}_2)} \triangleright \kappa_3, \\
&\ \ \vdots \\
\nu_n &= (\nu_1 \triangleright \ldots \triangleright \nu_{n-1})^{\downarrow \mathbf{K}_n \cap (\mathbf{K}_1 \cup \ldots \cup \mathbf{K}_{n-1})} \triangleright \kappa_n,
\end{aligned}$$

and focus on its j-th step: $\nu_j = (\nu_1 \triangleright \ldots \triangleright \nu_{j-1})^{\downarrow \mathbf{K}_j \cap (\mathbf{K}_1 \cup \ldots \cup \mathbf{K}_{j-1})} \triangleright \kappa_j$. Recall that, due to the the RIP, there exists $k < j$ such that (after a trivial reformulation) $\mathbf{K}_j \cap (\mathbf{K}_1 \cup \ldots \cup \mathbf{K}_{j-1}) = \mathbf{K}_j \cap \mathbf{K}_k$. In the j-th step, the submodel $\nu_1(\mathbf{K}_1) \triangleright \nu_2(\mathbf{K}_2) \triangleright \ldots \triangleright \nu_{j-1}(\mathbf{K}_{j-1})$ is already perfect. Therefore ν_k is a marginal of $\nu_1 \triangleright \ldots \triangleright \nu_{j-1}$, and $(\nu_1 \triangleright \ldots \triangleright \nu_{j-1})^{\downarrow \mathbf{K}_j \cap (\mathbf{K}_1 \cup \ldots \mathbf{K}_{j-1})} = \nu_k^{\downarrow \mathbf{K}_j \cap \mathbf{K}_k}$, which means that the computation of $\nu_k^{\downarrow \mathbf{K}_j \cap \mathbf{K}_k}$ is very simple. This is what Lauritzen and Spiegelhater call *local computations* [26]. They use this term for a computational process realized as a sequence of steps in which each step performs computations with only one of the distributions from which the multidimensional model is composed.

The next assertion yields a simple rule to decide whether a decomposable model is perfect or not: it is enough to check whether the distributions are pairwise consistent.

Theorem 3.11 *Let the sequence* $\mathbf{K}_1, \ldots, \mathbf{K}_n$ *meet the RIP. Then the probability distributions* $\kappa_1, \kappa_2, \ldots, \kappa_n$ *are pairwise consistent if and only if* $\kappa_1(\mathbf{K}_1) \triangleright \kappa_2(\mathbf{K}_2) \triangleright \ldots \triangleright \kappa_n(\mathbf{K}_n)$ *is a perfect model.*

Table 3.3: Two-dimensional distributions

$\kappa(X,Z)$	$X=0$	$X=1$
$Z=0$	0.25	0.25
$Z=1$	0.25	0.25

$\lambda(Y,Z)$	$Y=0$	$Y=1$
$Z=0$	0.1	0.1
$Z=1$	0.4	0.4

Before proceeding to other computational procedures studied in the next Sections, let us briefly summarize what we know about perfect and decomposable models.

Theorem 3.7 says that, if a compositional model $\kappa_1 \triangleright \kappa_2 \triangleright \ldots \triangleright \kappa_n$ and its permutation $\kappa_{j_1} \triangleright \kappa_{j_2} \triangleright \ldots \triangleright \kappa_{j_n}$ are both perfect models, then they both represent the same multidimensional distribution

$$\kappa_1 \triangleright \kappa_2 \triangleright \ldots \triangleright \kappa_n = \kappa_{j_1} \triangleright \kappa_{j_2} \triangleright \ldots \triangleright \kappa_{j_n}.$$

It means that perfectness guarantees uniqueness of the model, and all the information is utilized from the low-dimensional distributions from which the model is set up. Perfect models allow for certain reorderings of the distributions in the model, but not more than two different permutations are guaranteed.

We know that a RIP sequence can be reordered in many ways such that all these permutations meet the RIP (Theorem 1.11). It means that decomposable models can be reordered in many ways without violating the decomposability. But we must realize that it does not mean that these models represent the same multidimensional distribution; the resulting models may differ from each other. Therefore, re-ordering the distributions in a model without changing the resulting multidimensional distribution requires that this decomposable model should also be perfect. In such a case, all its RIP permutations are also perfect and all of them define the same multidimensional distribution (due to Theorem 3.7).

Example 3.12 *Consider two two-dimensional distributions from Table 3.3. Since a sequence of two sets always meets the RIP, it is evident that both models $\kappa \triangleright \lambda$ and $\lambda \triangleright \kappa$ are decomposable. The reader can easily verify that these models define different three-dimensional distributions: $\kappa \triangleright \lambda \neq \lambda \triangleright \kappa$ because κ and λ are not consistent. While $\kappa \triangleright \lambda$ is a uniform distribution, $(\lambda \triangleright \kappa)^{\downarrow\{Y,Z\}} = \lambda$, and therefore it cannot be uniform.*

Thus, perfect decomposable models manifest lot of advantageous properties. We said above that it is an easy task to perfectize a decomposable model. Having a perfect model $\kappa_1(\mathbf{K}_1) \triangleright \kappa_2(\mathbf{K}_2) \triangleright \ldots \triangleright \kappa_n(\mathbf{K}_n)$, a natural question arises whether one can find a decomposable model representing the

same multidimensional probability distribution. The precise formulation of this task is the following:
For a perfect compositional model $\kappa_1(\mathbf{K}_1) \triangleright \kappa_2(\mathbf{K}_2) \triangleright \ldots \triangleright \kappa_n(\mathbf{K}_n)$, find a RIP sequence of variable sets $\mathbf{M}_1, \mathbf{M}_2, \ldots, \mathbf{M}_m$, and compute probability distributions $\pi_1(\mathbf{M}_1)$, $\pi_2(\mathbf{M}_2)$, $\ldots$, $\pi_m(\mathbf{M}_m)$, such that

- $\forall\, \mathbf{K}_j \ (j = 1, \ldots, n) \ \exists\, \mathbf{M}_k \quad (\mathbf{K}_j \subseteq \mathbf{M}_k)$;
- $\pi_j(\mathbf{M}_j) = (\kappa_1(\mathbf{K}_1) \triangleright \kappa_2(\mathbf{K}_2) \triangleright \ldots \triangleright \kappa_n(\mathbf{K}_n))^{\downarrow \mathbf{M}_j}$ for all $j = 1, 2, \ldots, m$;
- $\pi(\mathbf{M}_1) \triangleright \pi(\mathbf{M}_2) \triangleright \ldots \triangleright \pi(\mathbf{M}_m) = \kappa_1(\mathbf{K}_1) \triangleright \kappa_2(\mathbf{K}_2) \triangleright \ldots \triangleright \kappa_n(\mathbf{K}_n)$.

If we set certain formal criteria for an optimal solution of this task (e.g., to find – in a way – the smallest possible sets $\mathbf{M}_1, \mathbf{M}_2, \ldots, \mathbf{M}_m$), the problem is known to be NP-hard. Therefore, only heuristic solutions of the task are used in practical situations. One possibility is to use the following algorithm based on the ideas of Tarjan and Yanakakis [33, 34].

Algorithm 3.13 – Transformation into a decomposable model.

INPUT: *Perfect model* $\kappa_1(\mathbf{K}_1) \triangleright \kappa_2(\mathbf{K}_2) \triangleright \ldots \triangleright \kappa_n(\mathbf{K}_n)$.

OUTPUT: *Perfect decomposable model* $\mu_1(\mathbf{M}_1) \triangleright \mu_2(\mathbf{M}_2) \triangleright \ldots \triangleright \mu_m(\mathbf{M}_m)$.

STEP I: *Define a graph* $\mathcal{G} = (\mathbf{V}, \mathcal{E})$*:*
$\mathbf{V} = \mathbf{K}_1 \cup \mathbf{K}_2 \cup \ldots \cup \mathbf{K}_n$*;*
$\mathcal{E} = \{(X, Y) \in \mathbf{V} \times \mathbf{V} : X \neq Y \ \&\ \exists j \in \{1, \ldots, n\} \{X, Y\} \subseteq \mathbf{K}_j\}$.

STEP II: *Enumerate the nodes of* $\mathcal{G}$ *using the* maximum cardinality search*:*
1. *Assign number "1" to an arbitrary node of* $\mathcal{G}$*.*
2. *Repeat assigning the next number to any node with the highest number of already enumerated neighbors*[2] *until all nodes are enumerated.*

STEP III: *For* $k = n, n-1, n-2, \ldots 2$ *do:*
Let X *be the node enumerated by* k *in the previous step.*
Let $\mathbf{L}$ *denote the set of all nodes with assigned numbers smaller than* k*.*
Let $\mathbf{N}$ *be the set of neighbors of* X *in* $\mathcal{G}$*.*
Denote $\mathcal{F} = \{(X, Y) \in \mathbf{L} \cap \mathbf{N} \times \mathbf{L} \cap \mathbf{N} : X \neq Y\}$.
Redefine $\mathcal{E} := \mathcal{E} \cup \mathcal{F}$*.*

STEP IV: *Find all cliques of graph* $\mathcal{G}$ *and order them to meet the RIP*[3]*.*
Denote them by $\mathbf{M}_1, \mathbf{M}_2, \ldots, \mathbf{M}_m$*.*

STEP V: *For all* $j = 1, 2, \ldots, m$*, define* $\mu_j = (\kappa_1 \triangleright \ldots \triangleright \kappa_n)^{\downarrow \mathbf{M}_j}$*.*

[2] If there are more nodes meeting this condition, choose an arbitrary one from among them.

[3] The procedure realized in Step III guarantees that it can always be done.

3.3 Marginalization

The task studied in this Section is the following[4]: for a compositional model $\kappa_1 \triangleright \kappa_2 \triangleright \ldots \triangleright \kappa_n$, and a subset of variables $\mathbf{M} \subset \mathbf{K}_1 \cup \mathbf{K}_2 \cup \ldots \cup \mathbf{K}_n$, find a compositional model $\lambda_1(\mathbf{L}_1) \triangleright \lambda_2(\mathbf{L}_2) \triangleright \ldots \triangleright \lambda_m(\mathbf{L}_m)$ such that

$$(\kappa_1 \triangleright \kappa_2 \triangleright \ldots \triangleright \kappa_n)^{\downarrow \mathbf{M}} = \lambda_1(\mathbf{L}_1) \triangleright \lambda_2(\mathbf{L}_2) \triangleright \ldots \triangleright \lambda_m(\mathbf{L}_m).$$

Unfortunately, there is no general solution to the marginalization task that would be optimal in the sense of computational efficiency. Several heuristic procedures supported by different theoretical results appear in the literature. In this Section, we present some results published in [9, 10, 12, 2], and discuss a heuristic process, which, employing the theorems presented below, solves the problem of marginalization in a way that can be applied to real-size problems.

The simplest rule used for marginalization is a direct application of Property 3 in Theorem 2.3 (Composition preserves first marginal). It says that, in the process of marginalization, we can always cut off the "tail" of the model.

Theorem 3.14 *If there exists $j \in \{1, 2, \ldots, n\}$ such that $\mathbf{M} = \mathbf{K}_1 \cup \ldots \cup \mathbf{K}_j$, then*

$$(\kappa_1 \triangleright \kappa_2 \triangleright \ldots \triangleright \kappa_n)^{\downarrow \mathbf{M}} = \kappa_1 \triangleright \ldots \triangleright \kappa_j.$$

A direct generalization of Property 12 in Theorem 2.3 (Simple marginalization) says that one can easily exclude a variable from a model if the variable is among the arguments of only one distribution.

Theorem 3.15 *Let $X \in \mathbf{N} = \mathbf{K}_1 \cup \mathbf{K}_2 \cup \ldots \cup \mathbf{K}_n$. If there exists $j \in \{1, 2, \ldots, n\}$ such that $X \in \mathbf{K}_j$, and $X \notin \mathbf{K}_1 \cup \ldots \cup \mathbf{K}_{j-1} \cup \mathbf{K}_{j+1} \cup \ldots \cup \mathbf{K}_n$, then*

$$(\kappa_1 \triangleright \kappa_2 \triangleright \ldots \triangleright \kappa_n)^{\downarrow \mathbf{N} \setminus \{X\}} = \kappa_1 \triangleright \ldots \triangleright \kappa_{j-1} \triangleright \kappa_j^{\downarrow \mathbf{K}_j \setminus \{X\}} \triangleright \kappa_{j+1} \triangleright \ldots \triangleright \kappa_n.$$

The following assertion enables us to formulate a rule that eliminates occurrence of a variable among the arguments of individual distributions. Therefore, an iterative application of such a rule eventually makes the deletion of any variable possible.

[4] In [28], F. Malvestuto solves the marginalization problem for a more general class of models called *compositional expressions* introduced in [27]. We do not include his procedure into this text because we exclusively concentrate on the marginalization of compositional models, which, forming a subclass of Malvestuto's sequential expressions, possess special properties that are employed to enhance the process of marginalization.

Theorem 3.16 *Consider $X \in \mathbf{N} = \mathbf{K}_1 \cup \mathbf{K}_2 \cup \ldots \cup \mathbf{K}_n$. Let j and k be indices from $\{1, 2, \ldots, n\}$ such that $j < k$, $X \in \mathbf{K}_j \cap \mathbf{K}_k$, and $X \notin \mathbf{K}_{j+1} \cup \ldots \cup \mathbf{K}_{k-1}$, then*

$$\kappa_1 \triangleright \kappa_2 \triangleright \ldots \triangleright \kappa_k = \kappa_1 \triangleright \ldots \triangleright \kappa_{j-1} \triangleright \kappa_j^{\downarrow \mathbf{K}_j \setminus \{X\}} \triangleright \kappa_{j+1} \triangleright \ldots \triangleright \kappa_{k-1} \triangleright \left(\kappa_j \textcircled{$\triangleright$}_{\mathbf{L}} \kappa_k\right),$$

where $\mathbf{L} = (\mathbf{K}_1 \cup \mathbf{K}_2 \cup \ldots \cup \mathbf{K}_{k-1}) \setminus \{X\}$.

Theorems 3.14 through 3.16 enable us to formulate the following marginalization rules. Their multiple application to a compositional model $\kappa_1 \triangleright \kappa_2 \triangleright \ldots \triangleright \kappa_n$ always lead to the required marginal distribution $(\kappa_1 \triangleright \kappa_2 \triangleright \ldots \triangleright \kappa_n)^{\downarrow \mathbf{M}}$.

Tail deletion rule. Find the smallest j such that $\mathbf{M} \subseteq \mathbf{K}_1 \cup \ldots \cup \mathbf{K}_j$, and consider only $\kappa_1 \triangleright \kappa_2 \triangleright \ldots \triangleright \kappa_j$.

Variable deletion rule. If there is a variable $X \notin \mathbf{M}$ for which there exists $j \in \{1, 2, \ldots, n\}$ such that $X \in \mathbf{K}_j$, and $X \notin \mathbf{K}_i$ for all the remaining $i = 1, \ldots, j-1, j+1, \ldots, n$, then consider model

$$\kappa_1 \triangleright \ldots \triangleright \kappa_{j-1} \triangleright \kappa_j^{\downarrow \mathbf{K}_j \setminus \{X\}} \triangleright \kappa_{j+1} \triangleright \ldots \triangleright \kappa_n.$$

Decrease of a variable appearance rule. For any $X \notin \mathbf{M}$ for which there exist indices $j, k \in \{1, 2, \ldots, n\}$ such that $j < k$, $X \in \mathbf{K}_j \cap \mathbf{K}_k$, and $X \notin \mathbf{K}_{j+1} \cup \ldots \cup \mathbf{K}_{k-1}$, define $\mathbf{L} = (\mathbf{K}_1 \cup \ldots \cup \mathbf{K}_{k-1}) \setminus \{X\}$, and consider model

$$\kappa_1 \triangleright \ldots \triangleright \kappa_{j-1} \triangleright \kappa_j^{\downarrow \mathbf{K}_j \setminus \{X\}} \triangleright \kappa_{j+1} \triangleright \ldots \triangleright \kappa_{k-1} \triangleright \left(\kappa_j \textcircled{$\triangleright$}_{\mathbf{L}} \kappa_k\right) \triangleright \kappa_{k+1} \triangleright \ldots \triangleright \kappa_n.$$

Let us once more emphasize that a suitable iterative application of these rules always leads to the desired marginal distribution. It fully corresponds to Shachter's marginalization procedure [28] (the names of our rules are inspired by Shachter's *node deletion* and *edge reversal* rules). In fact, application of the anticipating operator in a way corresponds to the *inheritance of parents* in his *edge reversal rule*.

In contrast to Shachter's approach, , we still have at our disposal another, more effective rule for the marginalization of compositional models. This rule makes the deletion of a number of variables possible in one computationally simple step. It is applicable under special conditions described in the following Definition.

Definition 3.17 *Tadu to musi byt spatne!!! Let, for a compositional model $\kappa_1 \triangleright \kappa_2 \triangleright \ldots \triangleright \kappa_n$. and index $i \in \mathbf{I} \subsetneq \{1, \ldots, n\}$ be given such that*[5]

$$\left(\bigcup_{k \in \mathbf{I}} \mathbf{K}_k\right) \cap \left(\bigcup_{k \notin \mathbf{I}} \mathbf{K}_k\right) \subseteq \mathbf{K}_i.$$

[5] Symbol $k \notin \mathbf{I}$ stands as an abbreviation for $k \in \{1, 2, \ldots, n\} \setminus \mathbf{I}$.

Then we say that i and $\mathbf{I}$ determine a reduction *of compositional model* $\kappa_1, \ldots, \kappa_n$ *(or simply that* $(i, \mathbf{I})$ is a reduction for model $\kappa_1 \triangleright \kappa_2 \triangleright \ldots \triangleright \kappa_n$*).*

Theorem 3.18 *Let* $(i, \mathbf{I})$ *be a reduction for a compositional model* $\kappa_1 \triangleright \kappa_2 \triangleright \ldots \triangleright \kappa_n$*. Let* $\mathbf{M} = \bigcup_{k \in \mathbf{I}} \mathbf{K}_k$*,* $\mu(\mathbf{K}_i) = (\kappa_1 \triangleright \kappa_2 \triangleright \ldots \triangleright \kappa_n)^{\downarrow \mathbf{K}_i}$*. Define distributions* λ_k

$$\begin{array}{lll} \lambda_k(\mathbf{K}_k) = \kappa_k(\mathbf{K}_k) & \text{for} & k \in \mathbf{I}, \\ \lambda_k(\mathbf{M} \cap \mathbf{L}_k) = \mu^{\downarrow \mathbf{M} \cap \mathbf{L}_k} & \text{for} & k \notin \mathbf{I} \text{ and } \quad \mathbf{L}_k = \bigcup_{\ell \in \{1, \ldots, k\} \setminus \mathbf{I}} \mathbf{K}_\ell. \end{array}$$

Then the marginal distribution $(\kappa_1 \triangleright \kappa_2 \triangleright \ldots \triangleright \kappa_n)^{\downarrow \mathbf{M}}$ *can be expressed as a compositional model*

$$(\kappa_1 \triangleright \kappa_2 \triangleright \ldots \triangleright \kappa_n)^{\downarrow \mathbf{M}} = \lambda_1 \triangleright \lambda_2 \triangleright \ldots \triangleright \lambda_n.$$

The reader should realize that $\mathbf{M} \cap \mathbf{L}_k \subseteq \mathbf{K}_i$ holds. If Theorem 3.18 is applied to a perfect model, then $\mu = \kappa_i$; this fact simplifies the necessary computations because λ_k for all $k \notin \mathbf{I}$ are marginals of κ_i. So we can see that application of this assertion to perfect models is computationally inexpensive. On the other hand, no simple algorithm for seeking a reduction $(i, \mathbf{I})$ of a general compositional model has been proposed. However, such an algorithm becomes very simple for decomposable models because of the following assertion.

Theorem 3.19 *If there exists a reduction* $(i, \mathbf{I})$ *for a decomposable compositional model, then there also exists a reduction* $(j, \mathbf{J})$ *for this model such that* $|\mathbf{J}| = n - 1$*.*

To elaborate, the assertion states that if for a decomposable model there exists a reduction in the sense of Definition 3.17 then the model described in Theorem 3.18 can be obtained by a sequence of simple reductions, each of which reduces the model just by one distribution. This means that the computational complexity of finding and realizing a reduction is linear in the length of a model. And this is also why we formulate a simplified version of the reduction rule valid for perfect decomposable models.

Reduction rule for decomposable models. Let $\kappa_1 \triangleright \kappa_2 \triangleright \ldots \triangleright \kappa_n$ be a perfect decomposable model. If there exists $\mathbf{K}_j$ such that, for a certain $\mathbf{K}_k$ $(k \neq j)$

$$\mathbf{K}_j \cap \Big(\bigcup_{i \neq j} \mathbf{K}_i \Big) \subseteq \mathbf{K}_k,$$

then find the RIP ordering $\mathbf{K}_{j_1}, \mathbf{K}_{j_2}, \ldots, \mathbf{K}_{j_n}$, for which $\mathbf{K}_j = \mathbf{K}_{j_n}$ is true, and consider the model $\kappa_{j_1} \triangleright \kappa_{j_2} \triangleright \ldots \triangleright \kappa_{j_{n-1}}$.

This is the last reduction rule applicable within the following algorithms to be used for the marginalization of compositional models. Their description presented below stresses the principles upon which these algorithms are based. When formulating them, the simplicity of the description is the main criterion. So, if these algorithms are applicable in the way they are formulated, they find the required solution; but to do it efficiently, a lot of programming problems would have to be solved that we do not mention here.

Algorithm 3.20 – Marginalization for decomposable models.

INPUT: *Decomposable model* $\lambda_1(\mathbf{L}_1) \triangleright \lambda(\mathbf{L}_2) \triangleright \ldots \triangleright \lambda_m(\mathbf{L}_m)$*;*
$\mathbf{M} \subset \mathbf{L}_1 \cup \mathbf{L}_2 \cup \ldots \cup \mathbf{L}_m$.

OUTPUT: *Perfect decomposable model* $\kappa_1(\mathbf{K}_1) \triangleright \kappa(\mathbf{K}_2) \triangleright \ldots \triangleright \kappa_n(\mathbf{K}_n)$.

STEP I: *Copy the input model into the output model, which will subsequently be modified:*
For $i = 1, 2, \ldots, m$ *define* $\kappa_i(\mathbf{K}_i) = \lambda_i(\mathbf{L}_i)$*;*
Define $n = m$.

STEP II: *If* $\kappa_1 \triangleright \kappa \triangleright \ldots \triangleright \kappa_n$ *is not perfect, then perfectize it using Theorem 3.6.*

STEP III: *Iteratively use the following rules as long as at least one of them is applicable. Then stop.*
Always use the applicable rule with the lowest number.

1. *If there exist two different indices* $i, j \leq m$ *such that* $\mathbf{K}_i \subseteq \mathbf{K}_j$*, then delete* κ_i *from the model. Renumber the distributions accordingly. Redefine:* $n = n - 1$.

2. *If the* ***Reduction rule for decomposable models*** *is applicable with* $(\mathbf{K}_j \setminus \mathbf{K}_k) \cap \mathbf{M} = \emptyset$*, then apply it. Renumber the distributions accordingly. Redefine:* $n = n - 1$.

3. *If the* ***Variable deletion rule*** *is applicable, then apply it.*

4. *If the* ***Decrease of a variable appearance rule*** *is applicable, then apply it to that variable* X *for which* $|\{j \in \{1, 2, \ldots, n\} : X \in \mathbf{K}_j\}|$ *is minimal.*

Let us complete this Section with the description of a general marginalization algorithm applicable to any compositional model.

Algorithm 3.21 – Marginalization for general models.

INPUT: *General compositional model* $\lambda_1(\mathbf{L}_1) \triangleright \lambda(\mathbf{L}_2) \triangleright \ldots \triangleright \lambda_m(\mathbf{L}_m)$*;*
$\mathbf{M} \subset \mathbf{L}_1 \cup \mathbf{L}_2 \cup \ldots \cup \mathbf{L}_m$.

OUTPUT: *Compositional model* $\kappa_1(\mathbf{K}_1) \triangleright \kappa(\mathbf{K}_2) \triangleright \ldots \triangleright \kappa_n(\mathbf{K}_n)$.

STEP I: *Copy the input model into the output model, which will subsequently be modified:*
For $i = 1, 2, \ldots, m$ define $\kappa_i(\mathbf{K}_i) = \lambda_i(\mathbf{L}_i)$;
Define $n = m$.

STEP III: *Iteratively use the following rules as long as at least one of them is applicable. Then stop.*
Always use the applicable rule with the lowest number.

1. *If the* ***Tail deletion rule*** *is applicable, then apply it.*

2. *If there exist $i \in \{1, 2, \ldots, n\}$ such that $\mathbf{K}_i \subseteq \mathbf{K}_1 \cup \mathbf{K}_2 \cup \ldots \cup \mathbf{K}_{i-1}$, then delete κ_i from the model. Renumber the distributions accordingly. Redefine: $n = n - 1$.*

3. *If the* ***Variable deletion rule*** *is applicable, then apply it.*

4. *If the* ***Decrease of a variable appearance rule*** *is applicable, then apply it to that variable X for which $|\{j \in \{1, 2, \ldots, n\} : X \in \mathbf{K}_j\}|$ is minimal.*

3.4 Conditioning

This Section, which is a survey of results from [3, 17], describes how the theoretical results presented in the previous chapters may be employed for the computation of conditionals. This is a very important part of compositional models theory, and therefore we include it in this text regardless it not being used in the process of model construction, even less for data mining.

When computing conditionals, we need a degenerate one-dimensional distribution expressing certainty. Consider variable X and its value $a \in \mathbb{X}_X$. The probability distribution δ_a^X expressing that variable $X = a$ with certainty, is defined for each $x \in \mathbb{X}_X$ as

$$\delta_a^X(x) = \begin{cases} 1, & \text{if } x = a; \\ 0, & \text{otherwise.} \end{cases}$$

The following assertion gives instructions for how to compute conditional distributions using the operation of composition (for its proof see Theorem 2.3 in [3]).

Theorem 3.22 *Consider a distribution $\kappa(\mathbf{K})$, variable $X \in \mathbf{K}$, its value $a \in \mathbb{X}_X$, and $\mathbf{L} \subseteq \mathbf{K} \setminus \{X\}$. If $\kappa^{\downarrow\{X\}}(a) > 0$, then the corresponding conditional distribution $\kappa^{\mathbf{L}|X=a}$ can be computed as follows:*

$$\kappa^{\mathbf{L}|X=a} = \left(\delta_a^X \triangleright \kappa\right)^{\downarrow\mathbf{L}}.$$

Due to Theorem 3.22, the computation of conditionals from a model $\kappa_1 \triangleright \ldots \triangleright \kappa_n$ means to compute

$$\delta_a^X \triangleright (\kappa_1 \triangleright \kappa_2 \triangleright \ldots \triangleright \kappa_n).$$

This is an easy task if $X \in \mathbf{K}_1$, because, in this special case, the following assertion holds.

Theorem 3.23 *Consider a compositional model $\kappa_1 \triangleright \kappa_2 \triangleright \ldots \triangleright \kappa_n$, variable $X \in \mathbf{K}_1$ and its value $a \in \mathbb{X}_X$. Then,*

$$\delta_{\mathbf{a}}(u) \triangleright (\kappa_1 \triangleright \kappa_2 \triangleright \ldots \triangleright \kappa_n) = (\delta_{\mathbf{a}}(u) \triangleright \kappa_1) \triangleright \kappa_2 \triangleright \ldots \triangleright \kappa_n.$$

This equality can be proved by a multiple application of Property 9 (Associativity under the RIP) of Theorem 2.3. It shows that all the necessary computations are local. If the variable X is not among those for which κ_1 is defined, the computation of $\delta_a^X \triangleright (\kappa_1 \triangleright \kappa_2 \triangleright \ldots \triangleright \kappa_n)$ may be space- and time-demanding. That is why we want to have a conditioning variable among the arguments of the first distribution, and also why we are so interested in models in which the ordering of distributions in the compositional model may be changed without modifying the represented distribution. We already know that, for perfect decomposable models, we can get any variable at the very beginning of the model. The widest class of models that possess the required property is that of flexible models.

Definition 3.24 *A model $\kappa_1 \triangleright \kappa_2 \triangleright \ldots \triangleright \kappa_n$ is called* flexible *if, for each $X \in \mathbf{K}_1 \cup \ldots \cup \mathbf{K}_n$, there exists a permutation $i_1, i_2, \ldots, i_n$ such that $X \in \mathbf{K}_{i_1}$ and*

$$\kappa_{i_1} \triangleright \kappa_{i_2} \triangleright \ldots \triangleright \kappa_{i_n} = \kappa_1 \triangleright \kappa_2 \triangleright \ldots \triangleright \kappa_n.$$

Theorems 1.11 and 3.11 say that a decomposable model consisting of pairwise consistent distributions is flexible. Therefore, any decomposable model can be transformed into a flexible one by applying the perfectization procedure from Theorem 3.6. Nevertheless, it is important to keep in mind that flexibility, in contrast to decomposability, is not a structural property. Decomposability is a property of a sequence of sets $\mathbf{K}_1, \mathbf{K}_2, \ldots, \mathbf{K}_n$, and any compositional model with a sequence of variable sets $\mathbf{K}_1, \mathbf{K}_2, \ldots, \mathbf{K}_n$ meeting the RIP is decomposable. On the contrary, for any sequence of variable sets $\mathbf{K}_1, \mathbf{K}_2, \ldots, \mathbf{K}_n$, one can find a compositional model that is flexible. A trivial example confirming this assertion is a model $\kappa_1 \triangleright \kappa_2 \triangleright \ldots \triangleright \kappa_n$, where all distributions κ_i are uniform. In this case, $\kappa_1 \triangleright \kappa_2 \triangleright \ldots \triangleright \kappa_n$ is a uniform multidimensional distribution regardless of the actual ordering of the distributions in the model.

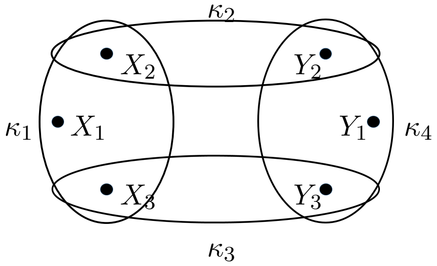


Figure 3.1: Six-dimensional flexible model.

Example 3.25 *Consider the following compositional model (representing two parallel noiseless channels) consisting of four distributions (see Figure 3.1)*

$$\kappa_1(X_1, X_2, X_3) \triangleright \kappa_2(X_2, Y_2) \triangleright \kappa_3(X_3, Y_3) \triangleright \kappa_4(Y_1, Y_2, Y_3).$$

Further assume that $\mathbb{X}_{X_2} = \mathbb{X}_{Y_2}$, and $\mathbb{X}_{X_3} = \mathbb{X}_{Y_3}$, and that the distributions κ_2 and κ_3 realize a noiseless duplex transmission:

$$\begin{aligned} \kappa_2(x, y) &> 0 \text{ if and only if } x = y, \\ \kappa_3(x, y) &> 0 \text{ if and only if } x = y, \end{aligned}$$

which means that $\kappa_2(x, x) = \kappa_2^{\downarrow X_2}(x) = \kappa_2^{\downarrow Y_2}(x)$, and $\kappa_3(x, x) = \kappa_3^{\downarrow X_3}(x) = \kappa_3^{\downarrow Y_3}(x)$. Under the assumption that κ_1 and κ_4 are consistent, i.e., the equality

$$\kappa_1^{\downarrow\{X_2,X_3\}}(x_2, x_3) = \kappa_4^{\downarrow\{Y_2,Y_3\}}(x_2, x_3),$$

holds for all $(x_2, x_3) \in \mathbb{X}_{X_2} \times \mathbb{X}_{X_3} = \mathbb{X}_{Y_2} \times \mathbb{X}_{Y_3}$, it is not difficult to show that

$$\begin{aligned} &\kappa_1(X_1, X_2, X_3) \triangleright \kappa_2(X_2, Y_2) \triangleright \kappa_3(X_3, Y_3) \triangleright \kappa_4(Y_1, Y_2, Y_3) \\ &\qquad = \kappa_4(Y_1, Y_2, Y_3) \triangleright \kappa_2(X_2, Y_2) \triangleright \kappa_3(X_3, Y_3) \triangleright \kappa_1(X_1, X_2, X_3) \end{aligned}$$

holds, which means that the model is flexible. To show the latter, consider an arbitrary state $(x_1, x_2, x_3, y_1, y_2, y_3) \in \mathbb{X}_{X_1} \times \mathbb{X}_{X_2} \times \mathbb{X}_{X_3} \times \mathbb{X}_{Y_1} \times \mathbb{X}_{Y_2} \times \mathbb{X}_{Y_3}$. If either $x_2 \neq y_2$, or $x_3 \neq y_3$ then both $\kappa_1(x_1, x_2, x_3) \triangleright \kappa_2(x_2, y_2) \triangleright \kappa_3(x_3, y_3) \triangleright \kappa_4(y_1, y_2, y_3)$, and $\kappa_4(y_1, y_2, y_3) \triangleright \kappa_2(x_2, y_2) \triangleright \kappa_3(x_3, y_3) \triangleright \kappa_1(x_1, x_2, x_3)$ are *equal to* 0*. In the opposite case,*

$$\begin{aligned}
&\kappa_1(x_1,x_2,x_3) \triangleright \kappa_2(x_2,y_2) \triangleright \kappa_3(x_3,y_3) \triangleright \kappa_4(y_1,y_2,y_3)\\
&\qquad = \kappa_1(x_1,x_2,x_3)\cdot \frac{\kappa_2(x_2,y_2)}{\kappa_2^{\downarrow X_2}(x_2)} \cdot \frac{\kappa_3(x_3,y_3)}{\kappa_3^{\downarrow X_3}(x_3)} \cdot \frac{\kappa_4(y_1,y_2,y_3)}{\kappa_4^{\downarrow Y_2,Y_3}(y_2,y_3)}\\
&\qquad = \kappa_1(x_1,x_2,x_3)\cdot \frac{\kappa_2(x_2,y_2)}{\kappa_2^{\downarrow X_2}(x_2)} \cdot \frac{\kappa_3(x_3,y_3)}{\kappa_3^{\downarrow X_3}(x_3)} \cdot \frac{\kappa_4(y_1,y_2,y_3)}{\kappa_1^{\downarrow X_2,X_3}(x_2,x_3)}\\
&\qquad = \kappa_4(y_1,y_2,y_3)\cdot \frac{\kappa_2(x_2,y_2)}{\kappa_2^{\downarrow Y_2}(y_2)} \cdot \frac{\kappa_3(x_3,y_3)}{\kappa_3^{\downarrow Y_3}(y_3)} \cdot \frac{\kappa_1(x_1,x_2,x_3)}{\kappa_1^{\downarrow X_2,X_3}(x_2,x_3)}\\
&\qquad = \kappa_4(y_1,y_2,y_3) \triangleright \kappa_2(x_2,y_2) \triangleright \kappa_3(x_3,y_3) \triangleright \kappa_1(x_1,x_2,x_3).
\end{aligned}$$

Since $\kappa_1 \triangleright \kappa_2 \triangleright \kappa_3 \triangleright \kappa_4 = \kappa_4 \triangleright \kappa_2 \triangleright \kappa_3 \triangleright \kappa_1$*, and since all variables appear among the arguments of either* κ_1 *or* κ_4*, the introduced model is flexible.*

Let us now turn our attention to the problem of conditioning in the flexible models. We repeat that flexible sequences are those that can be reordered in many ways so that each variable can appear among the arguments of the first distribution. As showed in Example 3.25, it does not mean that each distribution appears at the beginning of the sequence defining the model (this feature is present in perfect decomposable models, but not only in them). As said above, the flexibility is not a structural property (i.e., it is not a property of a sequence $\mathbf{K}_1, \mathbf{K}_2, \ldots, \mathbf{K}_n$). Therefore it is very important that, as expressed in the following assertion, the perfectization procedure from Theorem 3.26 preserves the flexibility.

Theorem 3.26 *If a model* $\kappa_1 \triangleright \kappa_2 \triangleright \ldots \triangleright \kappa_n$ *is flexible then its perfectized form* $\mu_1 \triangleright \mu_2 \triangleright \ldots \triangleright \mu_n$ *defined by the procedure*

$$\begin{aligned}
\mu_1 &= \kappa_1,\\
\mu_2 &= \mu_1^{\downarrow \mathbf{K}_2 \cap \mathbf{K}_1} \triangleright \kappa_2,\\
\mu_3 &= (\mu_1 \triangleright \mu_2)^{\downarrow \mathbf{K}_3 \cap (\mathbf{K}_1 \cup \mathbf{K}_2)} \triangleright \kappa_3,\\
&\ \ \vdots\\
\mu_n &= (\mu_1 \triangleright \ldots \triangleright \mu_{n-1})^{\downarrow \mathbf{K}_n \cap (\mathbf{K}_1 \cup \ldots \cup \mathbf{K}_{n-1})} \triangleright \kappa_n
\end{aligned}$$

is also flexible.

So, having a flexible model, we can easily compute a conditional distribution given a value of a variable for which the probability is positive. In the case of necessity, we can even perfectize the model without violating the flexibility. But, as a rule, when applying a model to inference, we need a conditional distribution given values of several variables. To this end, we need the following Theorem stating that neither does the conditioning spoil the flexibility of a compositional model.

Theorem 3.27 *Consider a flexible model $\pi = \kappa_1 \triangleright \kappa_2 \triangleright \ldots \triangleright \kappa_n$, variable $X \in \mathbf{K}_1 \cup \ldots \cup \mathbf{K}_n$, its value $a \in \mathbb{X}_X$ such that $\pi^{\downarrow\{X\}}(a) > 0$, and the corresponding conditional distribution $\vartheta = \delta_a^X \triangleright (\kappa_1 \triangleright \kappa_2 \triangleright \ldots \triangleright \kappa_n)$. Then*

$$\vartheta = \vartheta^{\downarrow\mathbf{K}_1} \triangleright \vartheta^{\downarrow\mathbf{K}_2} \triangleright \ldots \triangleright \vartheta^{\downarrow\mathbf{K}_n},$$

is a flexible model.

Let us mention that that model $\vartheta^{\downarrow\mathbf{K}_1} \triangleright \vartheta^{\downarrow\mathbf{K}_2} \triangleright \ldots \triangleright \vartheta^{\downarrow\mathbf{K}_n}$ can be computed in the following three steps:

- If $X \notin \mathbf{K}_1$, find a permutation of the model so that $X \in \mathbf{K}_1$.
- Perfectize the model $\left(\delta_a^X \triangleright \kappa_1\right) \triangleright \kappa_2 \triangleright \ldots \triangleright \kappa_n$ getting $\vartheta^{\downarrow\mathbf{K}_1} \triangleright \vartheta^{\downarrow\mathbf{K}_2} \triangleright \ldots \triangleright \vartheta^{\downarrow\mathbf{K}_n}$.
- If necessary, reorder the model to the original permutation.

Thanks to Theorem 3.27, the process can be repeated as many times as necessary to get a compositional model representing the required conditional distribution given an arbitrary number of values of different variables.

Example 3.28 *Consider a compositional model from Example 3.1*

$$\pi(X,Y,Z,U,V,W) = \kappa_1(X,Z) \triangleright \kappa_2(Z,V) \triangleright \kappa_3(X,U,V,W) \triangleright \kappa_4(Y,V,W)$$

with the goal to compute conditional probability distribution $\pi(U,V|X = a, Y = b)$. Since

$$\{X, U.V, W\} \cap (\{X.Z\} \cup \{Z, V\}) = \{X, V\},$$

we can immediately see that this model is not decomposable (not knowing the respective probability distributions (tables), we cannot decide about the flexibility of this model). Therefore, to make available the possibility of performing the computations locally, we first transform the model into a decomposable one.

It is a good exercise for the reader to show that, in this simple case, Algorithm 3.13 yields a decomposable model $\lambda_1(X, Z.V) \triangleright \lambda_2(X,U,V,W) \triangleright \lambda_3(Y,V,W)$ with

$$\begin{aligned} &\lambda_1(X, Z.V) = \kappa_1(X,Z) \triangleright \kappa_2(Z,V),\\ &\lambda_2(X,U,V,W) = \kappa_3(X,U,V,W),\\ &\lambda_3(Y,V,W) = \kappa_4(Y,V,W). \end{aligned}$$

Without loss of generality, we can assume the model is perfect (otherwise we would first perfectize it using Theorem 3.6). Using this model, we can compute

$$\delta_a^X \triangleright \pi = (\delta_a^X \triangleright \lambda_1(X, Z.V)) \triangleright \lambda_2(X,U,V,W) \triangleright \lambda_3(Y,V,W),$$

which is a decomposable model, but evidently not perfect. The application of the perfectization process

$$\begin{aligned}
&\bar{\lambda}_1(X, Z.V) = \delta_a^X \triangleright \lambda_1(X, Z.V),\\
&\bar{\lambda}_2(X, U, V, W) = \bar{\lambda}_1^{\downarrow\{X,V\}} \triangleright \lambda_2(X, U, V, W)\\
&\bar{\lambda}_3(Y, V, W) = \bar{\lambda}_2^{\downarrow\{V,W\}} \triangleright \lambda_3(Y, V, W)
\end{aligned}$$

yields a perfect decomposable model

$$\begin{aligned}
\delta_a^X \triangleright \pi &= \bar{\lambda}_1(X, Z.V)) \triangleright \bar{\lambda}_2(X, U, V, W) \triangleright \bar{\lambda}_3(Y, V, W)\\
&= \bar{\lambda}_3(Y, V, W) \triangleright \bar{\lambda}_2(X, U, V, W) \triangleright \bar{\lambda}_1(X, Z.V)).
\end{aligned}$$

The latter form, starting with $\bar{\lambda}_3(Y, V, W)$, is convenient for the conditioning by variable Y. Thus,

$$\delta_b^Y \triangleright (\delta_a^X \triangleright \pi) = (\delta_b^Y \triangleright \bar{\lambda}_3(Y, V, W)) \triangleright \bar{\lambda}_2(X, U, V, W) \triangleright \bar{\lambda}_1(X, Z.V)$$

is again a decomposable model. Let us perfectize it

$$\begin{aligned}
&\bar{\bar{\lambda}}_1(Y, V, W) = \delta_b^Y \triangleright \bar{\lambda}_3(Y, V, W),\\
&\bar{\bar{\lambda}}_2(X, U, V, W) = \bar{\bar{\lambda}}_1^{\downarrow\{V,W\}} \triangleright \bar{\lambda}_2(X, U, V, W),\\
&\bar{\bar{\lambda}}_3(X, Z.V) = \bar{\bar{\lambda}}_2^{\downarrow\{X,V\}} \triangleright \bar{\lambda}_1(X, Z.V).
\end{aligned}$$

Since this model is perfect, the required probability distribution can be obtained by simply marginalizing the second distribution of this model

$$\pi(U, V|X = a, Y = b) = \bar{\bar{\lambda}}_2^{\downarrow\{U,V\}}.$$

3.5 Causal models

In this Section we show that, if the users want and if they have enough expert knowledge, they can construct *causal* compositional models. Let us stress at the very beginning that to interpret a model as causal, one has to have enough expert knowledge about causal relationships among the considered variables. One can statistically reveal that there is a relationship between two variables. But there is principally no statistical way of finding out whether this dependence is causal. Determining which variable is a cause of the other goes beyond the power of statistical approaches as well.

Analogously to Pearl [29], who assumes to have expert knowledge that allows him to interpret the arrows in a Bayesian network as causal relationships, we have to assume that, for each of the considered variables, we know a group of other variables that are the causes for the variable in question. And it is this additional knowledge that allows us to infer more information from causal models than from the stochastic models described in previous

Sections. In addition to conditioning, we can also compute the impact of an *intervention.* Let us explain the difference between conditioning and intervention by the following trivial example from [3].

Example 3.29 *Consider two (for simplicity binary) variables describing whether there is smoke in a room (variable S) and whether a fire alarm is on or off (variable F). Naturally, since we assume that smoke in the room switches the alarm on, there is an evident causal relationship between these variables: S is a cause for F. Clearly, these variables are mutually dependent, and therefore, denoting the respective probability distribution $\pi(S, F)$, we quite naturally expect that $\pi(S = +|F = +) > \pi(S = +)$ and $\pi(F = +|S = +) > \pi(F = +)$.*

The situation changes when, instead of conditioning, we consider intervention. Using Pearl's "do" notation [29] we denote by $do(S = +)$ the situation when we bring the smoke into the room (for example, we smoke a cigar in it). Analogously, $do(F = +)$ denotes the situation when we switch the alarm on, for example, by pushing a test push-button. In this setup, it is natural to expect that bringing smoke into the room switches the alarm on, but switching the alarm on does not bring any smoke into the room. Therefore, while $\pi(F = +|do(S = +)) > \pi(F = +)$, the equality $\pi(S = +|do(F = +)) = \pi(S = +)$ holds. We have thus obtained that $\pi(S = +|do(F = +)) < \pi(S = +|F = +)$, meaning that if there is alarm on we can expect that there is smoke in the room (stochastic conditioning), but by switching an alarm on we do not increase the probability that there is smoke in the room (intervention).

Let us describe the notion of intervention (and the difference between conditioning and intervention) formally. Consider variable X and a set of variables $\mathfrak{C}(X)$ that are *causes* for X. It means that the behavior of variable X is fully described by a certain distribution $\pi(\mathbf{M})$, for $\mathbf{M} = \{X\} \cup \mathfrak{C}(X)$. Using Properties 4 and 7 from Theorem 2.3, one can immediately see that $\pi(\mathbf{M}) = \pi(\mathfrak{C}(X)) \triangleright \pi(\mathbf{M})$, which seems to be unnecessarily complex, but which is, as we are now going to show, very useful for causal models.

Theorem 3.22 says that

$$\pi(\mathfrak{C}(X)|X = a) = (\delta_a(X) \triangleright \pi(\mathbf{M}))^{\downarrow \mathfrak{C}(X)}, \tag{3.1}$$

and taking into consideration that $\pi(\mathbf{M}) = \pi(\mathfrak{C}(X)) \triangleright \pi(\mathbf{M})$, we get

$$\pi(\mathfrak{C}(X)|X = a) = (\delta_a(X) \triangleright (\pi(\mathfrak{C}(X)) \triangleright \pi(\mathbf{M})))^{\downarrow \mathfrak{C}(X)}. \tag{3.2}$$

Realize that until now we have not utilized the assumption that the model is causal. It comes into our consideration at this moment. Under this assumption, we can also compute the result of intervention $\pi(\mathfrak{C}(X)|do(X = a))$.

From Theorem 3.30 (below), we will learn that it can be computed by a formula that differs from Equality (3.2) in just a pair of parentheses:

$$\pi(\mathfrak{C}(X)|do(X=a)) = (\delta_a(X) \triangleright \pi(\mathfrak{C}(X)) \triangleright \pi(\mathbf{M}))^{\downarrow\mathfrak{C}(X)}. \tag{3.3}$$

Knowing that the operator of composition is not associative, we know that, in general,

$$(\delta_a(X) \triangleright (\pi(\mathfrak{C}(X)) \triangleright \pi(\mathbf{M})))^{\downarrow\mathfrak{C}(X)} \neq ((\delta_a(X) \triangleright \pi(\mathfrak{C}(X))) \triangleright \pi(\mathbf{M}))^{\downarrow\mathfrak{C}(X)}.$$

The difference between these two expressions follows immediately from the application of Properties 4 and 12 (see Theorem 2.3) to the right hand side of the above-stated inequality

$$\begin{aligned}((\delta_a(X) \triangleright \pi(\mathfrak{C}(X))) \triangleright \pi(\mathbf{M}))^{\downarrow\mathfrak{C}(X)} &= (\delta_a(X) \triangleright \pi(\mathfrak{C}(X)))^{\downarrow\mathfrak{C}(X)} \\ &= (\delta_a(X)^{\downarrow\emptyset} \triangleright \pi(\mathfrak{C}(X)) = \pi(\mathfrak{C}(X)).\end{aligned}$$

So, computing $((\delta_a(X) \triangleright \pi(\mathfrak{C}(X))) \triangleright \pi(\mathbf{M}))^{\downarrow\mathfrak{C}(X)}$, we get the marginal $\pi(\mathfrak{C}(X))$. It is in full correspondence with what was illustrated by the example above. The intervention that changes (fixes) the value of the effect variable does not influence the behavior of its causes:

$$\pi(\mathfrak{C}(X)|do(X=a)) = ((\delta_a(X) \triangleright \pi(\mathfrak{C}(X))) \triangleright \pi(\mathbf{M}))^{\downarrow\mathfrak{C}(X)} = \pi(\mathfrak{C}(X)).$$

After explaining the difference between conditioning and intervention in causal models, we still should find answers to the following two questions: What do we understand by a notion of a causal compositional model? How can we compute conditioning and the result of intervention from a causal compositional model? To answer these questions, we present selected results from [3].

Let us consider a set of variables $\mathbf{N}$, and, for each variable $X \in \mathbf{N}$, let $\mathfrak{C}(X) \subset \mathbf{N}\setminus\{X\}$ be its causes. Here we only consider *Markovian* models [29], i. e., the models in which variables can be ordered (without loss of generality, we assume it is the ordering $\{X_1, X_2, \ldots, X_n\} = \mathbf{N}$), which is such that the causes always precede their effects. So, we assume that

$$X_k \in \mathfrak{C}(X_i) \implies k < i,$$

which, as the reader has certainly noticed, means that $\mathfrak{C}(X_1) = \emptyset$, and excludes feedback models from our consideration.

In keeping with the notation introduced above, we denote $\mathbf{K}_i = \mathfrak{C}(X_i) \cup \{X_i\}$, and let $\kappa_i(\mathbf{K}_i)$ stand for the distribution describing the local behavior of X_i. This means that we consider a causal model

$$\pi(X_1, X_2, \ldots, X_n) = \kappa_1(\mathbf{K}_1) \triangleright \kappa_2(\mathbf{K}_2) \triangleright \ldots \triangleright \kappa_n(\mathbf{K}_n). \tag{3.4}$$

In [3], the following Theorem is proved.

Theorem 3.30 *For the causal compositional model π given by Formula* (3.4)*, and for arbitrary $X \in \mathbf{K}_1 \cup \ldots \cup \mathbf{K}_n$, $a \in \mathbb{X}_X$, $\mathbf{L} \subseteq \mathbf{K}_1 \cup \ldots \cup \mathbf{K}_n \setminus \{X\}$,*

$$\kappa(\mathbf{L}|do(X = a)) = \Big(\delta_a(X) \triangleright \kappa_1(\mathbf{K}_1) \triangleright \kappa_2(\mathbf{K}_2) \triangleright \ldots \triangleright \kappa_n(\mathbf{K}_n)\Big)^{\downarrow \mathbf{L}}. \tag{3.5}$$

We can thus see that the difference between conditioning and intervention is given just by a pair of parentheses (see Theorem 3.22):

$$\kappa(\mathbf{L}|X = a) = \Big(\delta_a(X) \triangleright \big(\kappa_1(\mathbf{K}_1) \triangleright \kappa_2(\mathbf{K}_2) \triangleright \ldots \triangleright \kappa_n(\mathbf{K}_n)\big)\Big)^{\downarrow \mathbf{L}}. \tag{3.6}$$

Before we conclude this Section, let us mention two points of great importance from the application point of view.

First, we show that the computation of the effect of multiple interventions is as simple as when considering just one intervention. Consider a causal compositional model $\pi(\mathbf{N}) = \kappa_1(\mathbf{K}_1) \triangleright \ldots \triangleright \kappa_n(\mathbf{K}_n)$, and (for simplicity) two variables $V, X \in \mathbf{N}$. Let $a \in \mathbb{X}_X, b \in \mathbb{X}_V$, and denote $\mathbf{L} = \mathbf{N} \setminus \{V, X\}$. Then

$$\begin{aligned} \kappa(\mathbf{L}|do((V,X) = (b,a)) &= \kappa(\mathbf{L}|do(V = b), do(X = a)) \\ &= (\delta_b(V) \triangleright \delta_a(X) \triangleright \kappa_1(\mathbf{K}_1) \triangleright \kappa_2(\mathbf{K}_2) \triangleright \ldots \triangleright \kappa_n(\mathbf{K}_n))^{\downarrow \mathbf{L}} \\ &= \big(\delta_{(b,a)}(V,X) \triangleright \kappa_1(\mathbf{K}_1) \triangleright \kappa_2(\mathbf{K}_2) \triangleright \ldots \triangleright \kappa_n(\mathbf{K}_n)\big)^{\downarrow \mathbf{L}}. \end{aligned}$$

Second, we want to turn the reader's attention to the fact that one can consider causal compositional models with hidden (unobservable) variables, and yet, under certain conditions, the result of the intervention can be computed. We do not want to go into rather complicated theoretical results in the present text, and therefore we present just a toy example taken from [16], showing that a lack of distinction between conditioning and intervention may lead to incorrect conclusions (and when done deliberately, it may be a way of cheating customers).

Example 3.31 *Assume that there was a statistical survey showing that a disease $\mathbf{d}^+$ has a lower incidence among* New-Drink *consumers than among those who do not make use of* New-Drink*. Since not all people like* New-Drink*, we assume that tendency to drink this beverage is influenced by another cause, say* genetic disposition*, which influences also development of disease $\mathbf{d}^+$. So, let us start considering a simplest possible causal model with three variables*

b – –*drinking* New-Drink	$\mathbb{X}_b = \{\mathbf{b}^+, \mathbf{b}^-\}$	$\mathfrak{C}(b) = \{g\}$,
d – –*disease*	$\mathbb{X}_d = \{\mathbf{d}^+, \mathbf{d}^-\}$	$\mathfrak{C}(d) = \{b, g\}$,
g – –*genetic disposition*	$\mathbb{X}_g$ *unknown*	$\mathfrak{C}(g) = \emptyset$,

i.e., the causal compositional model

$$\pi(b, d, g) = \kappa_1(g) \triangleright \kappa_2(b, g) \triangleright \kappa_3(b, d, g).$$

In this case, unfortunately, the computation of

$$\pi(d|do(b=\mathbf{b}^+)) = \Big(\delta_{\mathbf{b}^+}(b) \triangleright \kappa_1(g) \triangleright \kappa_2(b,g) \triangleright \kappa_3(b,d,g)\Big)^{\downarrow\{d\}}$$

is obviously impossible; because we can estimate only $\kappa_2(b)$, and $\kappa_3(b,d)$. The only way of overcoming this problem is to introduce an additional observable variable. Since the New-Drink *producer claims that, say, the positive impact of drinking their beverage is based on the fact that it decreases the level of cholesterol, let us add the result of the respective laboratory test into the model, and consider, now, a new causal model*

$$\pi(b,c,d,g) = \kappa_1(g) \triangleright \kappa_2(b,g) \triangleright \kappa_3(b,c) \triangleright \kappa_4(c,d,g),$$

which means that now we are considering variables

b − −*drinking* New-Drink	$\mathbb{X}_b = \{\mathbf{b}^+, \mathbf{b}^-\}$	$\mathfrak{C}(b) = \{g\}$,
c − −*cholesterol*	$\mathbb{X}_c = \{\mathbf{c}^{high}, \mathbf{c}^{low}\}$	$\mathfrak{C}(c) = \{b\}$,
d − −*disease*	$\mathbb{X}_d = \{\mathbf{d}^+, \mathbf{d}^-\}$	$\mathfrak{C}(d) = \{c, g\}$,
g − −*genetic disposition*	$\mathbb{X}_g$ *unknown*	$\mathfrak{C}(g) = \emptyset$.

Now, though not simple, the computation of

$$\pi(d|do(b=\mathbf{b}^+)) = \Big(\delta_{\mathbf{b}^+}(b) \triangleright \kappa_1(g) \triangleright \kappa_2(b,g) \triangleright \kappa_3(b,c) \triangleright \kappa_4(c,d,g)\Big)^{\downarrow\{d\}}$$

is possible, regardless of the fact that, from the available data, we can estimate neither κ_1, nor κ_2, nor κ_4, but only κ_3. The computations[6] *take advantage of the fact that the available data also allows for the estimation of the three-dimensional distribution of variables c, d, g which do not appear in the definition of the model. Denoting the estimate of this three-dimensional distribution κ_4, we get*

$$\pi(d|do(b=\mathbf{b}^+)) = \Big(\delta_{\mathbf{b}^+}(b) \triangleright \kappa_3(b,c) \triangleright \big(\kappa_3(b) \cdot \kappa_3(c) \triangleright \kappa_4(b,c,d)\big)^{\downarrow\{c,d\}}\Big)^{\downarrow\{d\}},$$

which is quite different from the conditional distribution that can, for this example, be computed

$$\pi(d|b=\mathbf{b}^+) = \Big(\delta_{\mathbf{b}^+}(b) \triangleright \kappa_4(b,c,d)\Big)^{\downarrow\{d\}}.$$

So, it may easily happen that $\pi(d|b=\mathbf{b}^+) < \pi(d)$, and simultaneously $\pi(d|do(b=\mathbf{b}^+)) > \pi(d)$. The reader can check these inequalities with the data from Table 3.4, for which we get

$$\begin{aligned}
&\pi(d=\mathbf{d}^+) = 0.033,\\
&\pi(d=\mathbf{d}^+|b=\mathbf{b}^+) = 0.027,\\
&\pi(d=\mathbf{d}^+|do(b=\mathbf{b}^+)) = 0.044.
\end{aligned}$$

[6] For a full page of computations, which is not repeated here, the interested reader is referred either to [15], or to [3]; in the latter source the computations are performed in a more general form.

Table 3.4: Frequency table for *New-Drink Example*

$b = \mathbf{b}^{+}$				$b = \mathbf{b}^{-}$			
$c = \mathbf{c}^{high}$		$c = \mathbf{c}^{low}$		$c = \mathbf{c}^{high}$		$c = \mathbf{c}^{low}$	
$d = \mathbf{d}^{+}$	$d = \mathbf{d}^{-}$	$d = \mathbf{d}^{+}$	$d = \mathbf{d}^{-}$	$d = \mathbf{d}^{+}$	$d = \mathbf{d}^{-}$	$d = \mathbf{d}^{+}$	$d = \mathbf{d}^{-}$
$0,010$	$0,122$	$0,008$	$0,520$	$0,009$	$0,263$	$0,006$	$0,062$

From these values one can see that, regardless of the fact that the value of conditional probability $\pi(d = \mathbf{d}^{+} | b = \mathbf{b}^{+}) = 0.027$ *may seem promising, the impact of intervention* $\pi(d = \mathbf{d}^{+} | do(b = \mathbf{b}^{+})) = 0.044$ *is, in fact, negative.*

Chapter 4

Independence structure of models

As defined in Section 1.2 (page 8) *Independence structure* of a probability distribution is a system of all conditional independence relationships holding for the distribution in question [32]. When speaking about the independence structure of a compositional model $\kappa_1 \triangleright \kappa_2 \triangleright \ldots \triangleright \kappa_n$, some of the independence relationships are already encoded in the structure of the model, i.e., in the sequence $\mathbf{K}_1, \mathbf{K}_2, \ldots \mathbf{K}_n$. These relationships can be read from this sequence using Property 2 in Theorem 2.3. Others can be deduced from them using the Block Independence Property (Theorem 1.6). The conditional independence relationships that hold for all probability distributions that can be represented in a form of a compositional model formed by distributions for variable sets $\mathbf{K}_1, \mathbf{K}_2, \ldots \mathbf{K}_n$ (in the given order) are called *structural independence relationships*. Keep in mind that, for a specific compositional model, some other conditional independence relationships may hold. To identify them, one has to check whether the respective equations (from Definition 1.3) hold. In this Chapter we will be interested only in the structural conditional relationships that can be read from the structure of the respective compositional model.

In this Section, which is a brief summarization of the main results from [11, 13, 18], we develop a special tool representing a compositional model structure that serves for communication of ideas between the user and the computer. Naturally, it is also convenient for communication among users. But the main goal, for which it was originally designed in [11], was reading all of the structural conditional independence relationships that hold in the respective model.

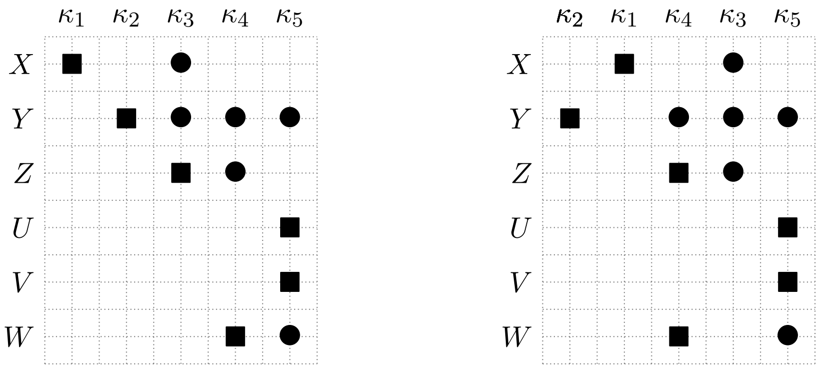
4.1 Persegrams

To visualize the structure of a compositional model, we use a tool called *persegram*[1]. This tool was originally designed in [13] in a slightly different way.

Definition 4.1 *The* persegram *of a compositional model $\kappa_1 \triangleright \kappa_2 \triangleright \ldots \triangleright \kappa_n$ is a table in which the rows correspond to variables from $\mathbf{K}_1 \cup \ldots \cup \mathbf{K}_n$ (in an arbitrary order) and the 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*.*

When speaking about the markers, we represent each of them them as a couple $[\kappa_i, X]$, where X is always a variable from $\mathbf{K}_i$.

Example 4.2 *In Figure 4.1, we can see persegrams of two compositional models: $\kappa_1(X) \triangleright \kappa_2(Y) \triangleright \kappa_3(X, Y, Z) \triangleright \kappa_4(Y, Z, W) \triangleright \kappa_5(Y, U, V, W)$ and $\kappa_2(Y) \triangleright \kappa_1(X) \triangleright \kappa_4(Y, Z, W) \triangleright \kappa_3(X, Y, Z) \triangleright \kappa_5(Y, U, V, W)$ (i.e., the model in Figure 4.1b is a simple permutation of the model from Figure 4.1a).*



	κ_1	κ_2	κ_3	κ_4	κ_5
X	■		●		
Y		■	●	●	●
Z			■	●	
U					■
V					■
W				■	●

(a) $\kappa_1(X) \triangleright \kappa_2(Y) \triangleright \kappa_3(X, Y, Z) \triangleright \kappa_4(Y, Z, W) \triangleright \kappa_5(Y, U, V, W)$

	κ_2	κ_1	κ_4	κ_3	κ_5
X		■		●	
Y	■		●	●	●
Z			■	●	
U					■
V					■
W			■		●

(b) $\kappa_2(Y) \triangleright \kappa_1(X) \triangleright \kappa_4(Y, Z, W) \triangleright \kappa_3(X, Y, Z) \triangleright \kappa_5(Y, U, V, W)$

Figure 4.1: Persegrams of a compositional model and its permutation.

Notice the difference between these persegrams. By reordering the columns – distributions in the model – two markers have changed their shapes: marker $[\kappa_3, Z]$ is a box-marker in the persegram of $\kappa_1(X) \triangleright \kappa_2(Y) \triangleright \kappa_3(X, Y, Z) \triangleright \kappa_4(Y, Z, W) \triangleright \kappa_5(Y, U, V, W)$ but a bullet in the persegram of $\kappa_2(Y) \triangleright \kappa_1(X) \triangleright \kappa_4(Y, Z, W) \triangleright \kappa_3(X, Y, Z) \triangleright \kappa_5(Y, U, V, W)$. Conversely, $[\kappa_4, Z]$ is a bullet in the persegram in Figure 4.1a but a box-marker in Figure 4.1b.

[1] This artificial word refers to the fact that it is a graphical representation of perfect sequences.

Having a closer look at the persegram in Figure 4.1b, we can also notice that the fourth column (corresponding to κ_3) does not contain any box-markers – all the corresponding markers are bullets. This means that, due to Property 4 (Reduction) in Theorem 2.3, this distribution does not influence the model; hence it is completely unnecessary and may be deleted. This is a general property that can be deduced from persegrams: distributions represented in a persegram by columns with no box-markers may be deleted from the model. In the following Sections we start studying the possibility of reading other properties of compositional models from persegrams.

4.2 Simple trails

Originally, as already said, persegrams were designed for reading conditional independence relationships holding in compositional models. Let us start with a simpler task: reading unconditional independence relationships [11].

Definition 4.3 *Consider a compositional model $\kappa_1 \triangleright \kappa_2 \triangleright \ldots \triangleright \kappa_n$ and the corresponding persegram. A sequence of markers $\mathfrak{m}_0, \ldots, \mathfrak{m}_t$ in the persegram of a compositional model is called a* simple trail connecting markers $\mathfrak{m}_0$ and $\mathfrak{m}_t$ *if it meets the following three conditions:*

1. *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);*
2. *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;*
3. *vertical and horizontal connections regularly alternate.*

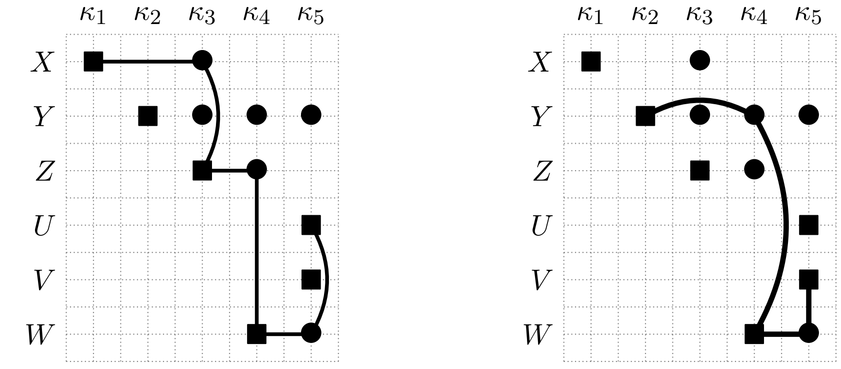
If a simple trail connects two markers corresponding to variables X and Y, we say that these variables are connected by a simple trail. *This situation is denoted by $X \leftrightsquigarrow Y$ $[\kappa_1 \triangleright \kappa_2 \triangleright \ldots \triangleright \kappa_n]$. Symbol $X \not\leftrightsquigarrow Y$ $[\kappa_1 \triangleright \kappa_2 \triangleright \ldots \triangleright \kappa_n]$ denote the situation when there does not exist a simple trail connecting variables X and Y in the corresponding persegram.*

The following theorem reveals the relationship between the existence of simple trails in a persegram and the relationship of (unconditional) independence between variables.

Theorem 4.4 *Consider a compositional model $\kappa_1, \kappa_2, \ldots, \kappa_n$ and the corresponding persegram. Let X and Y be two different variables from $\mathbf{K}_1 \cup \mathbf{K}_2 \cup \ldots \cup \mathbf{K}_n$. Then*

$$X \not\leftrightsquigarrow Y \; [\kappa_1, \kappa_2, \ldots, \kappa_n] \implies X \perp\!\!\!\perp Y \; [\kappa_1, \kappa_2, \ldots, \kappa_n].$$

Example 4.5 *To illustrate the notion of a simple trail, consider compositional model $\kappa_1(X) \triangleright \kappa_2(Y) \triangleright \kappa_3(X,Y,Z) \triangleright \kappa_4(Y,Z,W) \triangleright \kappa_5(Y,U,V,W)$ from Example 4.2 and its persegram, shown in Figure 4.1a. The shortest simple trails connecting two different variables consist from a sole vertical connection: e.g., trail $[\kappa_3, X], [\kappa_3, Z]$ connects X and Z; trail $[\kappa_5, U], [\kappa_5, W]$ connects U and W. Therefore $X \leftrightsquigarrow Z$ $[\kappa_1 \triangleright \kappa_2 \triangleright \ldots \triangleright \kappa_5]$ and $U \leftrightsquigarrow W$ $[\kappa_1 \triangleright \kappa_2 \triangleright \ldots \triangleright \kappa_5]$. However, notice that $[\kappa_5, Y], [\kappa_5, W]$ is not a simple trail, because this vertical connection is not regular – it is not adjacent to a box marker (both $[\kappa_5, Y]$ and $[\kappa_5, W]$ are bullet markers). For longer, simple trails, see Figure 4.2. Simple trail $[\kappa_1, X], [\kappa_3, X]$, $[\kappa_3, Z]$, $[\kappa_4, Z]$, $[\kappa_4, W], [\kappa_5, W], [\kappa_5, U]$ connects variables X and U, and the simple trail $[\kappa_2, Y], [\kappa_4, Y], [\kappa_4, W], [\kappa_5, W], [\kappa_5, V]$ connects variables Y and V.*



(a) $[\kappa_1, X], [\kappa_3, X], [\kappa_3, Z]$, $[\kappa_4, Z], [\kappa_4, W], [\kappa_5, W], [\kappa_5, U]$

(b) $[\kappa_2, Y], [\kappa_4, Y], [\kappa_4, W]$, $[\kappa_5, W], [\kappa_5, V]$

Figure 4.2: Persegram from Figure 4.1a with simple trails.

Let us once more emphasize that $[\kappa_1, X], [\kappa_3, X], [\kappa_3, Y], [\kappa_2, Y]$ is not a simple trail because the vertical connection $[\kappa_3, X], [\kappa_3, Y]$ is not regular (not being adjacent to a box marker). The reader can easily verify that there does not exist a simple trail in this persegram connecting variables X and Y: $X \not\leftrightsquigarrow Y$ $[\kappa_1 \triangleright \kappa_2 \triangleright \ldots \triangleright \kappa_5]$. To do it, notice (from Figure 4.2a) that there is the only connection adjacent to marker $[\kappa_1, X]$, and it is $([\kappa_1, X], [\kappa_3, X])$. It is a horizontal connection, which must be (in any simple trail) followed by a regular vertical connection: the only one is $([\kappa_3, X], [\kappa_3, Z])$. Then, there is no other possibility to continue until we get the trail $[\kappa_1, X], [\kappa_3, X], [\kappa_3, Z]$, $[\kappa_4, Z], [\kappa_4, W], [\kappa_5, W]$. This trail (ending with a horizontal connection) may be extended either by a regular vertical connection to $[\kappa_5, U]$ (as depicted in Figure 4.2a) or to $[\kappa_5, V]$. However, in either of these cases the trail cannot be extended any more, and thus $X \not\leftrightsquigarrow Y$ $[\kappa_1 \triangleright \kappa_2 \triangleright \ldots \triangleright \kappa_5]$. Therefore, due to Theorem 4.4, it holds $X \perp\!\!\!\perp Y$ $[\kappa_1 \triangleright \kappa_2 \triangleright \ldots \triangleright \kappa_5]$.

4.3 Avoiding trails

Definition 4.6 *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* $\mathbf{M}$*.*

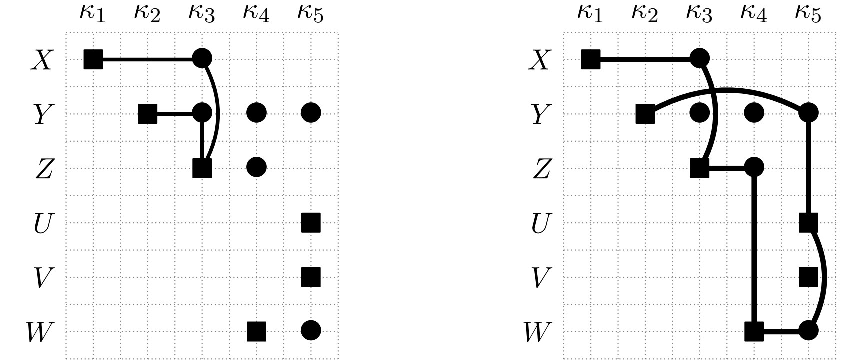
If an $\mathbf{M}$*-avoiding trail connects two markers corresponding to variables* X *and* Y*, we say that these* variables are connected by an $\mathbf{M}$-avoiding trail. *This situation is denoted by* $X \leftrightsquigarrow_{\mathbf{M}} Y$ $[\kappa_1 \triangleright \kappa_2 \triangleright \ldots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

The reader is right to expect that, in analogy to Theorem 4.4, there is a connection between the existence of avoiding trails and the conditional independence of variables in a compositional model. This property is formally expressed in the next important theorem, which was originally proved in [13] (an alternative proof was published in [18]). However, first we should realize that an $\emptyset$-avoiding trail is nothing else but a simple trail. This corresponds with the fact that the conditional independence coincides with the unconditional independence , when the conditioning set of variables is empty.

Theorem 4.7 *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*

$$X \not\leftrightsquigarrow_{\mathbf{M}} Y \; [\kappa_1, \kappa_2, \ldots, \kappa_n] \quad \Longrightarrow \quad X \perp\!\!\!\perp Y | \mathbf{M} \; [\kappa_1, \kappa_2, \ldots, \kappa_n].$$

Example 4.8 *Extend considering the compositional model* $\kappa_1(X) \triangleright \kappa_2(Y) \triangleright \kappa_3(X,Y,Z) \triangleright \kappa_4(Y,Z,W) \triangleright \kappa_5(Y,U,V,W)$ *from Example 4.2 and its persegram in Figure 4.1a. Notice that the simple trail in Figure 4.2a is also an* $\mathbf{M}$*-avoiding trail for any* $\mathbf{M} \subseteq \{Y,V\}$*.* X *and* U *must not be in* $\mathbf{M}$*, because it would violate the first condition in the definition of an avoiding trail (the first and the last markers of an* $\mathbf{M}$*-avoiding trail do not correspond to variables from* $\mathbf{M}$*).* Z *and* W *must not be in* $\mathbf{M}$*, because it would violate the fourth condition of this definition (no horizontal connection may correspond to a variable from* $\mathbf{M}$*). Analogously, the simple trail in Figure 4.2b is also an* $\mathbf{M}$*-avoiding trail for any* $\mathbf{M} \subseteq \{X,Z,U\}$*.*


κ_1 κ_2 κ_3 κ_4 κ_5
X
Y
Z
U
V
W
κ_1 κ_2 κ_3 κ_4 κ_5
X
Y
Z
U
V
W

(a) $\{Z\}$-avoiding trail: $[\kappa_1,X]$, $[\kappa_3,X],[\kappa_3,Z],[\kappa_3,Y],[\kappa_2,Y]$

(b) $[\kappa_2,Y],[\kappa_4,Y],[\kappa_4,W]$, $[\kappa_5,W],[\kappa_5,V]$

Figure 4.3: Persegram from Figure 4.1a with avoiding trails.

In contrast to Example 4.5, where we show that there cannot exist a simple trail connecting X *and* Y*, let us show that there are avoiding trails connecting this couple of variables. The shortest avoiding trail connecting* X *and* Y *can be seen in Figure 4.3a. It is the trail* $[\kappa_2,Y]$*,* $[\kappa_3,Y]$*,* $[\kappa_3,Z]$*,* $[\kappa_3,X]$*,* $[\kappa_1,X]$*, which is an* $\mathbf{M}$*-avoiding trail connecting* X *and* Y *for any* $\mathbf{M}$ *such that* $Z \in \mathbf{M} \subseteq \{Z,U,V,W\}$*. Another, substantially longer is a* $\{U\}$*-avoiding trail in Figure 4.3b:* $[\kappa_2,Y]$*,* $[\kappa_5,Y]$*,* $[\kappa_5,U]$*,* $[\kappa_5,W]$*,* $[\kappa_4,W]$*,* $[\kappa_4,Z]$*,* $[\kappa_3,Z]$*,* $[\kappa_3,X]$*,* $[\kappa_1,X]$*. The reader certainly understands that it is also a* $\{U,V\}$*-avoiding trail connecting* X *and* Y*.*

Chapter 5

Avoiding model overfitting

A *model overfitting* is a well known phenomenon both in statistics [31] and machine learning (artificial intelligence) [1]. It is used to describe a situation when a constructed model reflects noninformative properties of the source data file (like noise and other misleading properties that each randomly generated data file possesses). Let us illustrate this phenomenon on two stochastically dependent variables, the dependence of which is known to be linear. Because the dependence is stochastic, if randomly generated data are plotted in a graph, the respective dots are concentrated along a straight line describing the dependence. Naturally, only a certain proportion of the dots lie directly on the line. If one tries to find a curve that connects all the dots in the plot (see Fig. 5.1), the model becomes useless since it cannot be used for inference – (whether for interpolation or for extrapolation). From the point of view of this Chapter, it is important to realize that such a complex curve is described (defined) by a much larger number of parameters than the straight line, which can be determined just by two points.

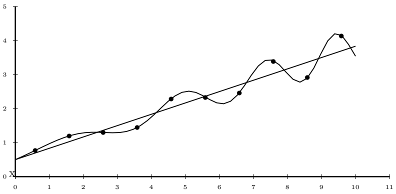


Figure 5.1: Overfitted linear dependence

In agreement with the preceding parts of this text, this Chapter studies the problem from the perspective of information theory. It is compiled from papers [19] and [20], which introduced the original idea that data-based model learning can be viewed as a transformation process, transforming the information contained in the data into the information represented by the model. Thus, using one of the basic laws of information theory, which states that no transformation can increase the amount of information, we get the basic restriction laid on the models constructed from the data: A model is *acceptable* if it does not contain more information than the input data file.

However, the application of this idea leads to a problem: how should we measure the information in a data file, and the information contained in a model. For this purpose, we must go more than a half a century back to seminal papers by Kolmogorov and von Mises, whose ideas are described in the next Section.

5.1 Information and complexity

Both von Mises [35] and Kolmogorov [22] explored relationships interconnecting randomness, complexity and information. They came with the idea that the amount of information in a sequence of 0's and 1's is increasing with the complexity of the sequence, and that the complexity of such a sequence can be measured by the length of the shortest program[1] generating the sequence. We accept that idea here, but instead of the length of an (abstract) program we consider the length of a lossless encoding (one can always generate the sequence from its lossless encoding).

Kolmogorov and Von Mises were also (among others) interested in the question what "the quantity of information conveyed by an individual object 'x' about another individual object 'y' " is [22]. To explain the connection of this question with the aim of this Chapter, take for one of the objects an arbitrary model $\mathfrak{M}$, and for the other one a sequence of 0's and 1's. Considering a lossless encoding $\mathcal{S}$ of the considered model $\mathfrak{M}$, we know that $\mathcal{S}$ does not contain less information than model $\mathfrak{M}$ because it contains all the information of model $\mathfrak{M}$. Now, having two sequences $\mathcal{S}_1$, $\mathcal{S}_2$ of 0's and 1's, which are both lossless encodings of a considered model $\mathfrak{M}$, we can say that both these sequences $\mathcal{S}_1$, $\mathcal{S}_2$ convey the same amount of information about model $\mathfrak{M}$. The same also holds true for an optimum lossless encoding $\mathcal{S}^*$ of model $\mathfrak{M}$. Since the mutual information between two objects is always smaller than or equal to the amount of the information contained in any of these objects, the length of the encoding $\mathcal{S}^*$ can be taken as the best upper estimate of the amount of information contained in model $\mathfrak{M}$. Assuming further that there is no (relative) redundancy in sequence $\mathcal{S}^*$, we can take its length as an estimate of the amount of information (measured in bits) contained in model $\mathfrak{M}$ (in what follows, we will omit the word "estimate", and will speak about the information, or amount of information contained in $\mathfrak{M}$).

The above-presented idea is independent of the type of models considered. We know that the best model for any object is that object itself.

[1] An abstract program for the universal Turing machine.

Therefore, the best model containing all the information contained in the data is the respective data file itself. Therefore, using the above-presented idea, the amount of information in the data equals the number of bits necessary to store an optimum lossless encoding of the respective data file. This enables us to compare the amount of information contained in the data and that contained in a data-learned model. If we get a model with a greater amount of information than that in the data, we can be sure that some undesirable information has been added into the model. Moreover, we know that regardless of the way the data has been collected, it always contains a specific part of the information, employment of which results in the overfitting of the model. Therefore, it should not be included in the model. In other words, all the considered models should contain less information than the input data. Thus we enforce a principle under which **models with the amount of information greater than or equal to that of the input data file are *unacceptable*.** In fact, we accept only models containing *substantially* less information than the input data file. The meaning of the word *substantially* is usually left to the user's discretion.

Notice, that the above-mentioned principle is also fully compatible with the famous *Minimum Description Length* (MDL) principle, which is often used in the process of model learning. For example, it was proposed for Bayesian network learning by Lam and Bacchus [25] (for general sources of this principle see, e.g., [6]).

The considered approach is also fully sensible from the statistical point of view. The less data we have, the smaller amount of bits we are allowed to use to encode the model. It means, among other things, that for small data files we cannot consider probability values specified with a high precision. This fully corresponds to the fact that, having a small amount of data, the confidence intervals for the estimates of probability parameters are rather wide. Therefore it does not make a sense to specify these estimates with a high precision (i.e., , with a large number of digits).

Nevertheless, let us realize that looking for the optimum lossless encoding is intractable, and we must thus use some heuristics. Instead of the optimum lossless encoding we will consider the best from those achieved by a battery of encoding procedures described in Sections 5.3 – 5.5. Though this approach can be considered only suboptimal, it serves well to the purpose of evaluating compositional models.

The next Section is devoted to the famous Huffman's encoding [7], which is utilized by some of these procedures.

5.2 Huffman code

Two of the encoding procedures described in the next Sections take advantage the famous Huffman's procedure [7], which is known to produce (in a sense) an optimum code. The procedure is rather simple and belongs to the fundamental parts of information theory, and therefore we decided to include its description in this text.

In the next Sections we will face the following problem: having a list of nonnegative integers (frequencies) $f_1, f_2, \ldots, f_d$, how many bits do we need to encode this sequence? Huffman's encoding is particularly useful for this purpose in the case that the integers appear in the sequence independently but with different probabilities. The sequences considered in the next Sections will usually consist of small numbers of different integers repeated, and quite often one of them will be prevailing. The Huffman code assigns short codes to frequent numbers, and rare numbers have longer codes. It is exactly the idea already employed by Samuel F. B. Morse, who assigned short codes to (in English) frequent characters (E and T) and long codes to characters that are not so frequent (e.g., J P, and Q).

Thus, the algorithm is based on two simple ideas:

- The frequencies appearing in the sequence more often are encoded by a smaller number of binary digits than those appearing less often.

- The two most sparse frequencies are encoded by codes differing from each other only in the last bit.

Algorithm 5.1 – Huffman algorithm (adapted).

INPUT: *A list of nonnegative integers (frequencies)* $f_1, f_2, \ldots, f_d$.

OUTPUT: *The number* NB *of bits necessary to encode the input sequence.*

STEP I: *Create a list of all values appearing in the input sequence. Let* k *denote the length of this list. Denote by* $p_1, p_2, \ldots, p_k$ *the numbers indicating how many times the respective frequencies appear in the input sequence.*
Define $NB := 0$

STEP II: *If* $k = 1$ *then stop. Otherwise repeat STEP III until* $k = 1$.

STEP III: *Order the values* $p_1, p_2, \ldots, p_k$ *in a descending order.*
Redefine $NB := NB + p_{k-1} + p_k$; $p_{k-1} := p_{k-1} + p_k$; $k = k - 1$.

The reader may have noticed that, for a sequence in which all the frequencies are the same (i.e., $k = 1$), the resulting value $NB = 0$ reflects the fact that all the frequencies are the same and therefore we need not encode them.

Table 5.1: Numbers of frequency occurrences in a sequence.

frequency	number of occurrences	resulting code
1	$p_1 = 46$	0
2	$p_2 = 16$	110
3	$p_3 = 13$	100
4	$p_4 = 11$	101
5	$p_5 = 8$	1110
7	$p_6 = 6$	1111

Example 5.2 *Consider a sequence of frequencies* $1, 1, 3, 2, 1, 5, 1, 1, 2, 1, 1, 4, 3, 1, \ldots, 1$, *the length of which is* 100. *There are only six different frequencies in this sequence, and their occurrences are given in Table 5.1. The behavior of the Huffman algorithm for this sequence is well depicted in Figure 5.2, which, perhaps, does not need a detailed explanation.*

In the leftmost part of this picture we see that $p_1, \ldots, p_6$ *are ordered in descending order, and therefore* p_5 *and* p_6 *are joined together during the first realization of STEP III. New descending ordering of the five numbers are then depicted in the next column of this figure. Therefore, in the second realization of STEP III, numbers* p_3 *and* p_4 *are joined together. In this way we can see that the output value indicating the number of necessary bits to encode this sequence of frequencies is*

$$\begin{aligned} NB &= (p_5 + p_6) + (p_3 + p_4) + (p_2 + p_5 + p_6) + (p_2 + p_3 + p_4 + p_5 + p_6) \\ &\quad + (p_1 + p_2 + p_3 + p_4 + p_5 + p_6) \\ &= p_1 + 3p_2 + +3p_3 + 3p_4 + 4p_5 + 4p_6 = 222. \end{aligned}$$

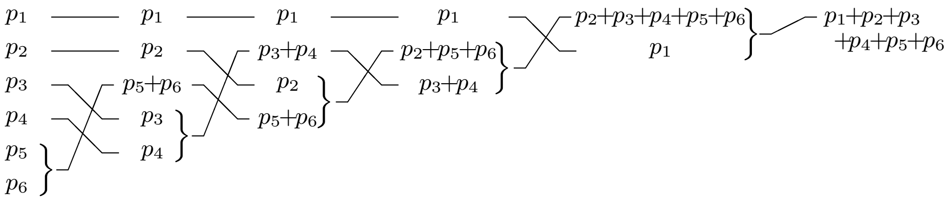


Figure 5.2: Huffman algorithm applied to numbers from Table 5.1.

As stated earlier, we do not need the encoding, we just need the number of bits necessary to encode the data or the model. But we must keep in mind that, when using the Huffman code, we must also encode the coding table, otherwise nobody would be able to reconstruct the original object (data, model). To realize how much space we need for this purpose, see a possible code for this example, which is the content of the right-handpart of

Table 5.1. For each row in this Table, we have to encode the frequency and the corresponding code word.

5.3 Coding Data

In reference to what was said above, the goal of this and the next Sections is to assign numbers to compositional models and/or data files that will be considered to be estimates of their complexity measured in the number of bits necessary for their optimum encoding. These numbers are then used for comparison of the complexity between the two models, or the complexity of the model and that of the data. This is why we neglect encoding the information describing the problem, such as the number of variables, variable names and the cardinalities of their value sets. Encoding this information would just increase all the derived complexity measures by a constant. Therefore, without loss of generality we can assume in this Chapter that variables X_i are identified by their indices i, and their values are $X_i = \{0, \ldots, h_i - 1\}$.

Under the above-stated assumption, when encoding a data file $\mathfrak{D}$ we have to encode a matrix of nonnegative integers with d rows (the records of the data file) and m columns (the variables). To this end, we will consider the following five simple procedures. Let us once more repeat that we are aware of the fact that using more sophisticated types of codes, such as, e.g., arithmetic codes [36], we could achieve even more economical encodings. The following codes realized in the system for compositional model handling are selected as a trade-off between precision and simplicity.

Direct Encoding. For a binary variable, we need just one bit for each entry of the matrix. If the respective $h_i > 2$ then we need[2] $\lceil \log_2 h_i \rceil$ bits to encode the value of variable X_i. To encode a row from a data matrix, we can first transform the respective state (recall that by this term we understand the combination of values of all variables) $(x_1, x_2, \ldots, x_m)$ into the number

$$rep(x_1, x_2, \ldots, x_m) = x_1 * \left(\prod_{k=2}^{m} h_k\right) + x_2 * \left(\prod_{k=3}^{m} h_k\right) + \ldots + x_m.$$

It is obvious that the number rep is unique for each state, and $0 \leq rep < \prod_{k=1}^{m} h_k$. Therefore, for its encoding into a binary sequence we need

$$\left\lceil \log_2 \prod_{k=1}^{m} h_k \right\rceil = \left\lceil \sum_{k=1}^{m} \log_2 h_k \right\rceil$$

[2] $\lceil r \rceil$ denotes the smallest integer that is not smaller than r.

bits[3]. It means that, for direct encoding of the data file, we need

$$c_d(\mathfrak{D}) = d \times \left(\left\lceil \sum_{k=1}^{m} \log_2 h_k \right\rceil \right) + c$$

bits, where c denotes the number of bits necessary to encode the number of records d (that is, the number of rows in the matrix).

Frequency Encoding. For this coding we take advantage of the fact that we can disregard the ordering of records in the data file. We increase the data matrix by one column into which we insert the number of repetitions of each state in the data file. It enables us to keep each state only once in the matrix. Thus, denoting by d_{red} the number of different states appearing in the original data file, and denoting by f_{max} the maximum number of occurrences of the same state in the data file, then we need

$$c_f(\mathfrak{D}) = d_{red} \times \left(\left\lceil \sum_{k=1}^{m} \log_2 h_k \right\rceil + \lceil \log_2(f_{max}) \rceil \right) + 2 \times c$$

bits for this type of encoding. Realize that $2 \times c$ bits are necessary to encode d_{red} and f_{max}.

Huffman Frequency Encoding. In the previous frequency encoding, we paid attention to economical coding of states; but for the number of repetitions we need $\lceil \log_2(f_{max} - -1) \rceil$ regardless of the fact that in many practical situations the number of repetitions is often 1. So, it may be advantageous to encode the numbers of occurrences using the famous Huffman code, the algorithm of which was described in the previous Section. Recall that it means that the states are encoded similar to the previous case, but for encoding the numbers of occurrences we use the code requiring a different number of bits for each different number of occurrences. The more often a number occurs, the fewer bits are required for its encoding. But we must not forget that in this case we do not encode just the data matrix itself but also the code used. The total number of necessary bits for this code (i.e., the number of bits necessary to encode all states plus the number of bits yielded by the Algorithm 5.1 plus the number of bits necessary to encode the coding table) will be denoted by $c_{fH}(\mathfrak{D})$.

Lexicographic Encoding. Analogously to frequency encoding, consider an extended data matrix in which each state appears no more than once, and the $(m+1)$th column contains the number expressing how many times the

[3] Notice that encoding the representative number of the state *rep* is more efficient than encoding each value separately. The latter encoding would require $\sum_{k=1}^{m} \lceil \log_2 h_k \rceil$ bits.

state appears in the data file $\mathfrak{D}$. If the number of variables is rather small, it may happen that the following encoding of the considered matrix is more economical than that by frequency encoding: add to the matrix all the states that do not appear in data (with the number of repetition equal to 0), sort all the states in lexicographic order, and then encode only the numbers from the $(m+1)$th column. Because of the introduced lexicographic ordering, one can easily assign the numbers of occurrences to the respective states. This encoding requires

$$c_l(\mathfrak{D}) = \left(\prod_{k=1}^{m} h_k \right) \times \lceil \log_2(f_{max}+1) \rceil + c$$

bits. $\prod_{k=1}^{m} h_k$ is the number of all states for the considered variables, $(f_{max}+1)$ appears in the formula because we must encode numbers of occurrences that are from $\{0, 1, 2, \ldots, f_{max}\}$, and the last c bits are used to encode f_{max}.

Huffman Lexicographic Encoding. As in the previous case, we only encode frequencies for all $\prod\limits_{k=1}^{m} h_k$ states (regardless of whether they appear in the data file or not). To encode these frequencies, we use the Huffman encoding algorithm. Recall that in this case we need the number of bits yielded by the Algorithm 5.1 and we must not forget to add the number of bits necessary to encode the coding table. The total number of bits necessary for this type of encoding will be denoted by $c_{lH}(\mathfrak{D})$.

Naturally, the reader can imagine that the list of the considered data-encoding procedures is extended by many other approaches (see, e.g., [36]). The web system computes the complexity measure for a data file just as

$$c(\mathfrak{D}) = \min\{c_d(\mathfrak{D}), c_f(\mathfrak{D}), c_{fH}(\mathfrak{D}), c_l(\mathfrak{D}), c_{lH}(\mathfrak{D})\}.$$

Example 5.3 *For the sake of simplicity, we consider just eight binary variables (with values* $0, 1$*), and an artificially generated data file* $\mathfrak{D}_{100}$ *with* 100 *records (binary vectors). To be able to encode much larger data files, let us fix the number of necessary bits to encode the length of the data file to* $c = 32$*. Recall that we omit encoding the information about the considered problem, like the number of variables, their values, etc.*

To apply the direct encoding approach we need to encode the following table

$$d = 100 \left\{ \begin{array}{cccccccc} 0 & 0 & 1 & 1 & 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 1 \\ & & \vdots & & & & & \\ 1 & 1 & 0 & 0 & 0 & 0 & 1 & 0 \end{array} \right.$$

for which $c_d(\mathfrak{D}_{100}) = 100 \times 8 + 32 = 832$ *bits are necessary.*

To encode the same data file with frequency encoding, first transform the data file into the form in which all the rows (states) are unique and the last column contains the numbers of occurrences of the respective state in the original data file. For the considered data file, we get the following table

$$d_{red} = 38 \left\{ \begin{array}{cccccccccc} 0 & 0 & 1 & 1 & 0 & 1 & 0 & 1 & & 27 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & & 10 \\ & & \vdots & & & & & & & \\ 1 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & & 1 \end{array} \right.$$

We thus obtain $c_f(\mathfrak{D}_{100}) = 38 \times (8+5) + 2 \times 32 = 558$ *bits.*

To get what we call the Huffman version of frequency encoding, we need to find the Huffman code for the numbers of occurrences. In our case, such a code is (the numbers in parentheses – the last column – read how many times the respective frequency number appears in the table above)

27	11111	(1×)
10	11110	(1×)
5	1110	(2×)
3	110	(5×)
2	10	(9×)
1	0	(20×)

Using the Huffman version of frequency encoding, we must encode the coding table shown above (which can easily be performed with $6 \times (5+5) = 60$ *bits, and for encoding the numbers of occurrences we only need*[4] $NB = 2 \times 5 + 2 \times 4 + 5 \times 3 + 9 \times 2 + 20 \times 1 = 71$ *bits (instead of* $38 \times 5 = 190$*, which is needed for the frequency encodings in the previous case). So, we get* $c_{fH}(\mathfrak{D}_{100}) = 38 \times (8) + 60 + 71 + 2 \times 32 = 499$ *bits.*

To get the lexicographic encoding, we must consider all 2^8 *states lexicographically ordered:*

$$256 \left\{ \begin{array}{cccccccc|c|} \cline{9-9} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ & & \vdots & & & & & & \\ 0 & 0 & 1 & 1 & 0 & 1 & 0 & 1 & 27 \\ & & \vdots & & & & & & \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 \\ \cline{9-9} \end{array} \right.$$

So, the lexicographic encoding of the framed frequencies requires $c_l(\mathfrak{D}_{100}) = 256 \times 5 + 32 = 1312$ *bits. However, if we use the Huffman approach to*

[4] Recall that NB is the number yielded by the Huffman algorithm described in Section 5.2.

encoding all the numbers appearing in the frame, i.e., if we use the following code

27	111111	(1×)
10	111110	(1×)
5	11110	(2×)
3	1110	(5×)
2	110	(9×)
1	10	(20×)
0	0	(218×)

we will only need $7 \times (5+6) = 77$ *bits to encode this coding table, and* $c_{lH}(\mathfrak{D}_{100}) = 77 + 2 \times 6 + 2 \times 5 + 5 \times 4 + 9 \times 3 + 20 \times 2 + 218 \times 1 = 404$ *bits, which decidedly shows the power of the Huffman codes.*

To illustrate the way in which these complexity measures increase with the amount of the considered data, we have generated (using the same generator) another three data files with 1 000, 10 000 *and* 100 000 *records. A summary of the bit requirements to encode all these data files is given in Table 5.2. From this Table, we can immediately conclude: there is no encoding method among the introduced five approaches that would be generally better than all the others.*

Table 5.2: Requirements for coding the data files

	c_d	c_f	c_{fH}	c_l	c_{lH}
$\mathfrak{D}_{100}$	832	558	499	1,312	**404**
$\mathfrak{D}_{1000}$	8,032	1,680	**976**	2.080	992
$\mathfrak{D}_{10000}$	80,032	4,084	2,713	3,104	**2,362**
$\mathfrak{D}_{100000}$	800,032	5,676	5,800	**3,872**	4,775

5.4 Coding Models

To encode compositional model $\kappa_1(\mathbf{K}_1) \triangleright \kappa_2(\mathbf{K}_2) \triangleright \ldots \triangleright \kappa_n(\mathbf{K}_n)$, we have to encode the sequence of distributions $\kappa_1, \kappa_2, \ldots, \kappa_n$ in the proper order. Each of these distributions κ_i is described by (a) the dimension of the distribution $|\mathbf{K}_i|$ (i.e., the number of variables for which it is defined); (b) the list of variables $\mathbf{K}_i$ for which it is defined; and finally (c) the respective probabilities. It means that we need to encode

(a)	number of variables $\lvert K_i \rvert$	$\lceil \log_2 m \rceil$ bits,
(b)	list of variables	$\lvert K_i \rvert \times \lceil \log_2 m \rceil$ bits.

In addition to these $(|K_i|+1) \times \lceil \log_2 n \rceil$ bits, the encoding of κ_i requires the space necessary for the respective probabilities. The total number of probabilities to be encoded for distribution κ_i is $\prod\limits_{X_j \in K_i} h_j$. Obviously, encoding the probabilities is, as a rule, much more space-demanding than encoding the information mentioned above under (a) and (b). So, let us turn our attention to a simple and efficient way of encoding the list of probabilities.

Naturally, the space requirements for probability encoding is closely connected with the precision with which the respective probabilities are specified. A simple way, which is used in the described web system, is the following:

Select a positive integer, denote it *base*, and (approximately) express all the considered probabilities, each as a ratio of two nonnegative integers

$$\frac{a}{base}.$$

This means that, instead of the respective probability, we need to encode just the respective numerator a. (Notice that we do not need to encode the integer *base* since it is a sum of all the numerators from the considered distribution.) For obvious reasons, it does not make sense to choose the integer *base* larger than the number of records d in the input data file. However, the *base* may be much smaller than d and can be defined with respect to the size of the confidence intervals computed for the probability estimates, or it can be reduced when we want to reduce the complexity of the constructed compositional model.

Employing the idea of representing probabilities by integers we get, in fact, exactly the same situation as in the previous Section: distribution κ_i may be well represented as a list of records, out of which each represents a state $x \in \mathbb{X}_{K_i}$ for which the probability $\kappa_i(x)$ is positive, and by the respective integer a for which $\kappa_i(x) = \frac{a}{base}$ holds. It means that, for encoding the distribution κ_i, we can employ any of the techniques described in the previous Section (an application of the *direct encoding* may come into consideration only in very specific and unusual situations). As a rule, the most economical encoding is yielded by the *Huffman lexicographic encoding*. *Frequency encoding* (both plain and Huffman's) may be applicable only for multi-dimensional distributions, which are positive on a small part of the respective space $\mathbb{X}_{K_i}$.

When encoding compositional models, we will thus face the problem: is it more economical to construct the Huffman code specifically for each distribution (and thus also encode the respective coding table), or to construct a single code for encoding all the marginals from which the model is composed. The reader need not care about this problem because it is solved automatically by the software system.

An analogous problem is connected with the selection of the *base* number. When considering only simple models, we recommend that a sole *base* number should be used for the entire model. However, the reader certainly realizes that in some situations a greater chances to decrease the complexity of the model can be achieved when defining different $base_i$ numbers for different marginals. It makes sense to use a smaller *base* for multi-dimensional distributions for two reasons. First, multi-dimensional distributions require encoding of large amount of parameters (probabilities), therefore an efficient representation of these probabilities is critical. Second, confidence intervals for the estimates of probabilities in multi-dimensional distributions are larger, and therefore it does not make sense to specify the probabilities with great precision. The solution to this problem influences the resulting model, and therefore it is left to the user to decide which of the possible solutions should be preferred.

Example 5.3 (continued) *Let us illustrate the principles described above by encoding a model*

$$\mathfrak{M}_1: \quad \mu_1 = \kappa_1(X_1, X_2) \triangleright \kappa_2(X_3, X_4) \triangleright \kappa_3(X_3, X_5) \triangleright \kappa_4(X_1, X_4, X_5, X_6) \\ \triangleright \kappa_5(X_5, X_6, X_8) \triangleright \kappa_6(X_2, X_5, X_6, X_7, X_8)$$

constructed from the considered data file $\mathfrak{D}_{100}$ with 100 records. As stated above, we do not encode the information about the considered variables because it is the information that should be encoded for all models. Next, we should know how many bits we need to encode the variables and how many bits are necessary to encode the maximum frequencies. Therefore we also assume that we know two numbers encodings of which are not considered in the rest of this Section: the number of variables m, and the size of the data file d, which is the upper limit for the frequencies.

To describe model $\mathfrak{M}_1$ we thus need to specify the number of distributions $n = 6$ (this can be done by $\lceil \log_2 m \rceil$ bits because it does not make sense to compose a model from more distributions than the number of variables).

Let us now turn our attention to efficient encoding of a k-dimensional distribution of binary variables. To encode it by lexicographic encoding, we need:

number of variables k	$\lceil \log_2 m \rceil$ *bits,*
list of variables	$k \times \lceil \log_2 m \rceil$ *bits,*
maximal integer f_{max}	$\lceil \log_2 d \rceil$ *bits,*
frequencies (probabilities)	$2^k \times \lceil \log_2 f_{max} \rceil$ *bits.*

Hence we need

for κ_1:	*3 + 2 × 3 + 7 + 4 × 7 = 44 bits,*
for κ_2:	*3 + 2 × 3 + 7 + 4 × 7 = 44 bits,*
for κ_3:	*3 + 2 × 3 + 7 + 4 × 7 = 44 bits,*
for κ_4:	*3 + 4 × 3 + 7 + 16 × 6 = 118 bits,*
for κ_5:	*3 + 3 × 3 + 7 + 8 × 6 = 67 bits,*
for κ_6:	*3 + 5 × 3 + 7 + 32 × 5 = 185 bits,*

which means that $c_l(\mathfrak{M}_1) = 502$.

Taking into account the fact that, among the 68 frequencies (probabilities) needed to represent the respective six marginals, there appear twenty times "0" *and sixteen times "1", it is not surprising that a more economical encoding is achieved by Huffman's version of the lexicographic encoding, which yields* $c_{lH}(\mathfrak{M}_1) = 423$ *for this model. In any case, whatever type of encoding we may take into consideration, we cannot reach the coding requirements sufficient to encode data* $c_{lH}(\mathfrak{D}_{100}) = 404$. *Therefore, according to the information-theoretic principle described at the beginning of this Section, model* $\mathfrak{M}_1$ *is unacceptable, and we have to simplify it by any of the possibilities described in the next Section.*

5.5 Model Simplification

By simplifying a model we understand the realization of such modifications of models that, with the goal of eliminating a possible model overfitting, decrease the number of bits necessary for the lossless encoding of the model. This is also the reason why the best models are often achieved after heuristic alternative applications of two processes: verification and simplification.

The simplification of models can be achieved in two ways, which can be properly combined: simplification of the model structure, and/or roughening of the probability estimates.

Roughening of the probability estimates. This is perhaps the easiest way to simplify the constructed model. It is realized just by decreasing the constant *base*. Considering model $\mathfrak{M}_1$ with $base = 100$ means that we take all the probability estimates with two digits of precision. Considering $base = 1,000$ means that we take all the probability estimates with three digits of precision. However, rounding these estimates to one decimal digit means to consider $base = 10$. Nevertheless, it is important to realize that we can consider finer roughening by choosing any number. We can set, e.g., $base = 64$.

Example 5.3 (continued) *Denote by* $c_{lH}(\mathfrak{M}_{1:50})$, $c_{lH}(\mathfrak{M}_{1:40})$ *and* $c_{lH}(\mathfrak{M}_{1:32})$ *the complexity values of the Huffman lexicographic encoding for the model* $\mathfrak{M}_1$ *with the base equaling* 50, 40 *and* 32, *respectively. Then, for the probability estimates obtained from the data file* $\mathfrak{D}_{100}$, *we get* $c_{lH}(\mathfrak{M}_{1:50}) = 408$, $c_{lH}(\mathfrak{M}_{1:40}) = 397$, *and* $c_{lH}(\mathfrak{M}_{1:32}) = 284$. *From the information-theoretic viewpoint principle, the last model is acceptable (the difference between* $c(\mathfrak{D}_{100}) = 404$ *and* $c_{lH}(\mathfrak{M}_{1:40}) = 397$ *is too small to recommend model* $\mathfrak{M}_{1:40}$, *but, as said above, the decision is up to the users, their intuition and their expert knowledge of the respective field of application). Let us also note that a greater simplification achieved when changing the base from* 40 *to* 32 *than when changing base from* 50 *to* 40 *is due to the fact that* $\lceil \log_2 40 \rceil > \lceil \log_2 32 \rceil$ *and* $\lceil \log_2 50 \rceil = \lceil \log_2 40 \rceil$. *Therefore, choosing the constant base from among the powers of* 2 *may be recommendable.*

Structure simplification. Another way to simplify the considered model is to simplify its structure. First of all, if there are two distributions in a model such that the set of variables for which one distribution is defined contains the set of variables for which the second distribution is also defined (i.e., $\mathbf{K}_i \subset \mathbf{K}_j$ for some pair if indices), we should try to eliminate the smaller distribution without influencing the distribution represented by the model. If this is not the case for the model in question, then other modifications to change the modeled distribution must be considered. Obviously, in the sense of space requirements, multi-dimensional distributions are more costly. One should therefore try to decrease their dimensionality. Different approaches to realize this task may be applied within different models. The most widespread are the following two. Consider model $\pi = \kappa_1 \triangleright \ldots \triangleright \kappa_i \triangleright \ldots \triangleright \kappa_n$, and assume distribution $\kappa_i(\mathbf{K}_i)$ is the one whose dimension should be decreased. Denote $\mathbf{L} = \mathbf{K}_i \cap (\mathbf{K}_1 \cup \mathbf{K}_2 \cup \ldots \cup \mathbf{K}_{i-1})$.

- Choose a variable $X \in \mathbf{L}$ with the lowest direct influence on variables $\mathbf{K}_i \setminus \mathbf{L}$. Not having an intuition based on expert knowledge about the field of application, choose

  $$X = \arg\min_{Y \in \mathbf{L}} \Big(MI_\pi(Y; \mathbf{K}_i \setminus \mathbf{L} \mid \mathbf{L} \setminus \{Y\}) \Big).$$

  Instead of π, consider its simplification $\bar{\pi} = \kappa_1 \triangleright \ldots \triangleright \kappa_i^{\downarrow K_i \setminus \{X\}} \triangleright \ldots \triangleright \kappa_n$, in which a direct influence of X on $\mathbf{K}_i \setminus \mathbf{L}$ is compensated by a conditional influence through variables $\mathbf{L} \setminus \{Y\}$.

- If $|\mathbf{K}_i \setminus \mathbf{L}| > 1$, consider the possibility of splitting this set into two disjoint parts $\mathbf{M}_1 \cup \mathbf{M}_2$, between which the direct influence can be compensated by conditional independence through variables $\mathbf{L}$. It can be done if

  $$MI_\pi(\mathbf{M}_1; \mathbf{M}_2 \mid \mathbf{L})$$

  is rather low. Then, instead of model π, consider its simplification $\bar{\pi} = \kappa_1 \triangleright \ldots \triangleright \kappa_i^{\downarrow K_i \setminus \mathbf{M}_2} \triangleright \kappa_i^{\downarrow K_i \setminus \mathbf{M}_1} \triangleright \ldots \triangleright \kappa_n$.

Example 5.3 (continued) *Let us, again, consider model*

$$\mathfrak{M}_1: \quad \mu_1 = \kappa_1(X_1, X_2) \triangleright \kappa_2(X_3, X_4) \triangleright \kappa_3(X_3, X_5) \triangleright \kappa_4(X_1, X_4, X_5, X_6) \\ \triangleright \kappa_5(X_5, X_6, X_8) \triangleright \kappa_6(X_2, X_5, X_6, X_7, X_8).$$

Its first simplification is received by a multiple application of the first simplification rule, which suggests deleting variables if their direct influence may be neglected. Based on the fact that $MI(X_6; X_4|\{X_1, X_5\})$ is negligible, we can substitute κ_4 by its marginal $\kappa_4^{\downarrow\{X_1,X_5,X_6\}}$. Similarly, small $MI(X_8; X_6|X_5)$ suggests to substitute κ_5 by its marginal $\kappa_5^{\downarrow\{X_5,X_8\}}$, and small

Table 5.3: Kullback-Leibler divergences

	$\mathfrak{M}_1$	$\mathfrak{M}_{1:50}$	$\mathfrak{M}_{1:40}$	$\mathfrak{M}_{1:32}$	$\mathfrak{M}_2$	$\mathfrak{M}_3$
complexity	423	408	397	284	267	304
K-L divergence	0.2736	0.2795	0.2846	0.2881	0.2964	0.3036

$MI(X_7;\{X_5,X_6\}|\{X_2,X_8\})$ suggests the substitution of κ_6 by its marginal $\kappa_6^{\downarrow\{X_2,X_7,X_8\}}$. In this way, we get a new model

$$\mathfrak{M}_2: \qquad \mu_2 = \kappa_1(X_1,X_2) \triangleright \kappa_2(X_3,X_4) \triangleright \kappa_3(X_3,X_5) \triangleright \kappa_4(X_1,X_5,X_6) \\ \triangleright \kappa_5(X_5,X_8) \triangleright \kappa_6(X_2,X_7,X_8).$$

Let us compare it with another simple model

$$\mathfrak{M}_3: \qquad \mu_3 = \lambda_1(X_3,X_4) \triangleright \lambda_2(X_3,X_5) \triangleright \lambda_3(X_1,X_5,X_6) \triangleright \lambda_4(X_5,X_6,X_8) \\ \triangleright \lambda_5(X_6,X_7,X_8) \triangleright \lambda_6(X_1X_2,X_7).$$

Computing complexities of these models, we get $c_l(\mathfrak{M}_2) = 306$ and $c_l(\mathfrak{M}_3) = 356$ bits, and $c_{lH}(\mathfrak{M}_2) = 267$ and $c_{lH}(\mathfrak{M}_3) = 304$ bits. These values are obtained for models with base = 100. So, comparing these values with $c_{lH}(\mathfrak{D}_{100}) = 404$, we can see that both these models are acceptable from our point of view.

Nevertheless, it is clear that we cannot evaluate a model just on the basis of its complexity (information content), i.e., according to the number of bits necessary for its encoding. We also need a criterion evaluating to what extent each model carries the information contained in the considered data. To this end, we use the Kullback-Leibler divergence between the sample probability distribution defined by the data and the probability distribution defined by the model. So, for each considered model we can compute the Kullback-Leibler divergence between the eight-dimensional sample distribution κ defined by the considered data file with 100 records, and the distribution defined by the respective model. For example, for model $\mathfrak{M}_1$ it is $Div(\kappa;\mu_1)$, where κ is the sample distribution, and μ_1 is the distribution defined by the model[5]. The values of these divergences for all the considered models are shown in Table 5.3. The reader can see that the simplifications leading to models $\mathfrak{M}_2$ and $\mathfrak{M}_3$ are realized at the expense of the accuracy of the resulting models.

[5] Notice that we consider the divergence $Div(\kappa;\mu_1)$ and not $Div(\mu_1;\kappa)$. This is important because of the asymmetry of the Kullback-Leibler divergence. One should realize that, for real size models, the latter divergence is usually $Div(\mu_1;\kappa) = +\infty$ because of $\mu_1 \not\ll \kappa$.

From Table 5.3 we can see that the simplification of a model by decreasing the value of the constant *base*, i.e., by roughening the estimates of probabilities, leads to the decrease in the complexity of the model and a simultaneous increase in the Kullback-Leibler divergence value. The greater this type of simplification, the greater the respective Kullback-Leibler divergence. On the other hand, from the last two columns in Table 5.3, the reader can see that a similar relationship valid for the model simplification by decreasing the complexity of the model would be much more complex. This is based on the fact that though both models $\mathfrak{M}_2$ and $\mathfrak{M}_3$ are the simplification of $\mathfrak{M}_1$, neither of them is a simplification of the other. This means that, for the structure simplification the strength of the simplification cannot be measured by just one parameter, i.e., the amounts of bits necessary for the model encoding, but we also have to introduce a partial order in the set of all potential simplifications, which is a topic for future research.

Chapter 6

Data mining example

The goal of this chapter is to show how the theoretical results presented in the preceding chapters can be utilized during the process of supervised model construction, and what type of knowledge can be gained during this process. A supervised process is used for several reasons. First, no generally accepted method for optimum model construction is known. Second, the user usually has some prior knowledge about the area of application, and this knowledge should be utilized during the process. Further, the user can have some knowledge about data, and the model is constructed on the basis of that knowledge. The user may know that the data is not well stratified and that certain properties should be suppressed, and others emphasized. The user quite often wishes to adapt the constructed model to the purpose for which the model is constructed. Therefore, it is natural that we cannot give general instructions how to proceed when constructing a model. We can present here just a basic simple example. To this end, we consider a slightly extended variant of Example 1.14 presented in Chapter 1.

Consider six random 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 this set of random variables from a data file containing $1,000$ 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 6.1.

It is not a bad idea to start with computing the value of entropy for all considered variables. Using the same notation as in Example 1.14, we get

$$\begin{array}{lll} H(B) = 1.58, & H(D) = 0.98, & H(N) = 0.96, \\ H(R) = 0.99, & H(T) = 0.99, & H(W) = 0.93. \end{array}$$

From this set of numbers, we do not get any knowledge about the relationship among the considered variables, but we get some information how to proceed further. Since the entropy of all binary variables is close to 1, it means

Table 6.1: 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

that the minimum of entropies for any pair of variables is also close to one. In other words, when considering the 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, and we compute only values of mutual information. All the same, we must keep in mind that when the considered variables take on different numbers of values, there may be substantial differences between the values of 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 pairs of variables which are closely (strongly) connected, and in pairs of independent variables. Therefore, when computing values of mutual information for all pairs of variables, we sort the pairs according to the values of mutual information. In the present example we get

$$\left\{\begin{array}{l} MI(D;N) = 0.4356, \\ MI(B;R) = 0.2871, \\ MI(R;W) = 0.2578, \\ MI(N;R) = 0.2070, \\ MI(T;W) = 0.1813, \end{array}\right.$$

$$MI(N;W) = 0.1546$$

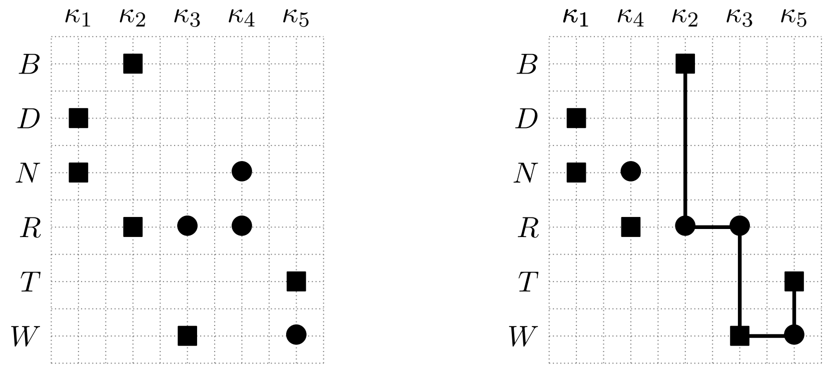
$$MI(D;R) = 0.0958$$

$$MI(B;W) = 0.0814$$

$$\begin{aligned} MI(N;T) &= 0.0709\\ MI(D;W) &= 0.0627\\ MI(B;N) &= 0.0619\\ MI(D;T) &= 0.0421\\ MI(B;D) &= 0.0342 \end{aligned}$$

$$\left\{\begin{aligned} MI(R;T) &= 0.0019,\\ MI(B;T) &= 0.0007. \end{aligned}\right.$$

Since the mutual information can be interpreted as a measure of dependence, the pairs of variables appearing at the head of this sequence are closely connected. On the other hand, if mutual information of a couple of variables is close to zero, the variables can be considered independent. These pairs of variables appear at the tail of the above-presented sequence. The first five pairs are grouped together because they cover the entire **M**. Let us start building the compositional models from the two-dimensional distributions defined for the first five pairs of variables. To get their best ordering in a model, the multi-information of the entire model should be taken into account. The higher the multi-information value, the better the model because it incorporates more information from data. The multi-information of a model can easily be computed using Theorem 3.9 (after the perfectization, if necessary). Nevertheless, let us show that, in this example, we can do without these tedious computations, and instead just utilize the theoretical results presented in the preceding chapters.



(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 6.1: 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 6.1a), 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 of the respective markers in the persegram are bullets. This model is decomposable (the reader may easily check the RIP) and perfect (the data file does not contain missing values, and therefore the estimates of marginals are consistent). Therefore (using Theorem 3.9)

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

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 (see also the comment before Example 1.14). However, it is evident that $\pi_2 = \kappa_1 \triangleright \kappa_4 \triangleright \kappa_2 \triangleright \kappa_3 \triangleright \kappa_5$ is also a perfect decomposable model, for which

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

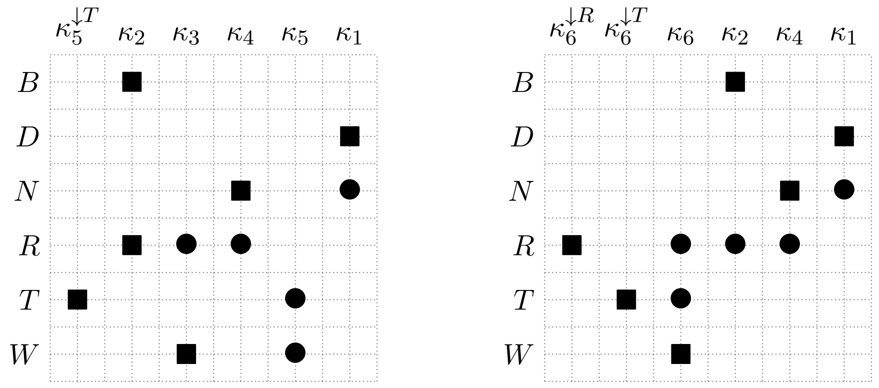
In fact, this model is the best possible among those assembled from distributions $\kappa_1, \kappa_2, \kappa_3, \kappa_4, \kappa_5$, if the amount of multi-information is taken as the only criterion of optimality. This is because this model utilizes all the information contained in the distributions from which it is assembled. More precisely, it is one of the best equivalent models, because, as we know from Theorems 3.11 and 3.7, all the RIP orderings of $\kappa_1, \kappa_2, \kappa_3, \kappa_4, \kappa_5$ yield the same distribution. However, this model does not reflect the other information we obtained from computing the mutual information for all pairs of variables: the two smallest values of mutual information suggest that variables T and R, as well as variables T and B, are independent. The reader can deduce from the persegram corresponding to π_2 (see persegram in Figure 6.1b) that one can find simple trails connecting all couples of variables, i.e., also $B \leftrightsquigarrow T\ [\pi_2]$ and $R \leftrightsquigarrow T\ [\pi_2]$. Therefore, the independence relationships $B \perp\!\!\!\perp T\ [\pi_2]$ and $R \perp\!\!\!\perp T\ [\pi_2]$ are not guaranteed by the model structure.

To incorporate this knowledge into the model, one can, e.g., consider the 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 6.2a, $\pi_3 = \kappa_5^{\downarrow T} \triangleright \kappa_2 \triangleright \kappa_3 \triangleright \kappa_4 \triangleright \kappa_1$, and therefore

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

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

It may thus seem that one can incorporate the knowledge about the two independence relationships into the model only at the cost of decreasing the multi-information value, i.e., at the cost of the loss of information. To get out of this trap, let us start studying the way in which the variable T is connected with all others. Let us compute (from the data) the conditional



(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 6.2: Persegrams of models π_3 and π_5.

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 can we explain the fact that variables T and R can be considered independent ($MI(R;T) = 0.0019$) but not conditionally independent ($MI(T;R|W) = 0.084$)? A straightforward explanation is that T and R are independent and jointly influence other variables. If we know the meanings of these variables, we should choose the one that 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 leads us to believe that two independent variables T and R influence W, and the only way 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 obtained from the data. Naturally, this three-dimensional distribution bears all the information expressed by both κ_3 and κ_5, which can now be dropped from further considerations. 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 not be expressed by two two-dimensional distributions. 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, precipitation in very cold days usually means snowing, which is great fun for children. And this type of knowledge cannot be expressed just by describing two separate relationships: day temperature and children fun, and precipitation and children fun.

Let us turn back to our example. 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 the RIP: e.g., $\pi_4 = \kappa_6 \rhd \kappa_2 \rhd \kappa_4 \rhd \kappa_1$ is a perfect 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 $\pi_5 = \kappa_6^{\downarrow B} \rhd \kappa_6^{\downarrow T} \rhd \kappa_6 \rhd \kappa_2 \rhd \kappa_4 \rhd \kappa_1$, from the persegram of which the considered independence relationships are obvious (see Figure 6.2b).

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

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

The imperceptible decrease in 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. The reader can see that this decrease in the multi-information value exactly corresponds to the amount of the mutual information of variables T and R included in κ_6.

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 the following important subsequent steps belonging to the process of model verification. Let us illustrate these steps by verifying model π_5. Consider the respective persegram in Figure 6.2b, which enables us to list all (conditional) independence relationships holding for the model (regarding the comment before Theorem 1.6 we present here only the relationships for singletons):

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

From this list, the eighth relationship covering also the unconditional independence $N \perp\!\!\!\perp T$ is in contradiction with $MI(N;T) = 0.0709$. To set this serious 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

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

Generally, to accept a model, the user should perfrom the model verification process consisting of the verificationof the following items:

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

To follow these instructions, let us perfectize model π_6 using the procedure described in Theorem 3.6:

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

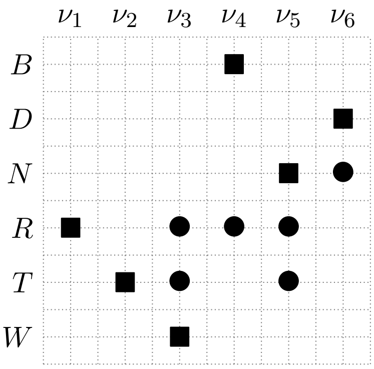
We know that $\pi_6 = \nu_1 \triangleright \nu_2 \triangleright \nu_3 \triangleright \nu_4 \triangleright \nu_5 \triangleright \nu_6$, all ν_i (for $i = 1, \ldots, 6$) are marginals of π_6. The respective probability distributions generating this model are depicted in Table 6.2, and the respective persegram is shown in Figure 6.3. From this persegram, the following list of conditional independence relationships can be deduced:

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

Table 6.2: Probability distributions ν_1 – ν_6.

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

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 the 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 perfectization. Nevertheless, the deviations from the original data-based estimates are very small, as can



	ν_1	ν_2	ν_3	ν_4	ν_5	ν_6
B				■		
D						■
N					■	●
R	■		●	●	●	
T		■	●		●	
W			■			

Figure 6.3: Persegram of model $\pi_6 = \nu_1 \triangleright \nu_2 \triangleright \nu_3 \triangleright \nu_4 \triangleright \nu_5 \triangleright \nu_6$.

also be seen from the values of Kullback-Leibler divergence

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

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

Before completing this example, let us mention another way of improving the constructed model, which has not been employed in this example. Recall that during the process of model construction/verification, we have twice substituted two two-dimensional distributions by one three-dimensional one. In practical situations such an increase of dimensionality of the considered distributions from which the model is constructed must be repeated several times. As a rule, even more than three-dimensional distributions are necessary in real situations. So, it can easily happen that one needs to introduce a distribution defined by too many probabilities. Namely, based on the size of the data file used for the model construction, one can get acceptable estimates of probabilities of limited dimensionality. If the dimension of the distribution to be introduced exceeds the limits corresponding to the data file, we recommend to compute the required distribution by the Iterative Proportional Fitting procedure described in Section 2.4. It should be computed from all those that are to be substituted. For example, if we found that model π_5 does not reflect the relationship between variables N and W well, we could consider an estimate $\kappa_8(N,W)$ from data, and compute distribution $\kappa_9(N,R,T,W)$ by the IPF procedure from the three distributions: $\kappa_7(N,R,T)$, $\kappa_8(N,W)$, and $\kappa_6(R,T,W)$. In this case, we would set up the final model from $\kappa_2(B,R)$, $\kappa_1(D,N)$ and $\kappa_9(N,R,T,W)$ (possibly including again the independence $R \perp\!\!\!\perp T$ by considering the respective one-dimensional marginals).

Part II

SYSTEM MANUAL

Bibliography

[1] P. Berka. *Dobývání znalostí z databází.* Academia, 2003.

[2] V. Bína and R. Jiroušek. Marginalization in multidimensional compositional models. *Kybernetika*, 42(4):405–422, 2006.

[3] V. Bína and R. Jiroušek. On computations with causal compositional models. *Kybernetika*, 51(3):525–539, 2015.

[4] I. Csiszár. I-divergence geometry of probability distributions and minimization problems. *The Annals of Probability*, pages 146–158, 1975.

[5] W. E. Deming and F. F. Stephan. On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *The Annals of Mathematical Statistics*, 11(4):427–444, 1940.

[6] P. Grunwald. A tutorial introduction to the minimum description length principle. *arXiv preprint math/0406077*, 2004.

[7] D. A. Huffman. A method for the construction of minimum-redundancy codes. *Proceedings of the IRE*, 40(9):1098–1101, 1952.

[8] R. 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.

[9] R. Jiroušek. Marginalization in composed probabilistic models. In *Proceedings of the Sixteenth conference on Uncertainty in artificial intelligence*, pages 301–308. Morgan Kaufmann Publishers Inc., 2000.

[10] R. Jiroušek. Decomposition of multidimensional distributions represented by perfect sequences. *Annals of Mathematics and Artificial Intelligence*, 35(1-4):215–226, 2002.

[11] R. Jiroušek. Detection of independence relations from persegrams. *Proceedings of the 9th Information Processing and Management of Uncertainty in Knowledge-based Systems*, pages 1261–1267, 2002.

[12] R. Jiroušek. On computational procedures for probabilistic compositional models. In *Proceedings of the 5th Czech-Japan Seminar on Data Analysis and Decision Making under Uncertainty*, pages 3–10, 2002.

[13] R. 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.

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

[15] R. Jiroušek. On causal compositional models: simple examples. In *International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*, pages 517–526. Springer, 2014.

[16] R. Jiroušek. Brief introduction to causal compositional models. In *Causal Inference in Econometrics*, pages 199–211. Springer, 2016.

[17] R. Jiroušek. On conditioning in multidimensional probabilistic models. In *Robustness in Econometrics*, pages 201–216. Springer, 2017.

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

[19] R. Jiroušek and I. 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.

[20] R. Jiroušek and I. Krejčová. Avoiding overfitting of models: an application to research data on the internet videos. In *Proceedings of the 35th International Conference Mathematical Methods in Economics (MME 2017)*, pages 289–294, 2017.

[21] H. G. Kellerer. Verteilungsfunktionen mit gegebenen marginalverteilungen. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 3(3):247–270, 1964.

[22] A. N. Kolmogorov. Tri podchoda k kvantitativnomu opredeleniju informacii. *Problemy peredachi informacii*, (1):4–7, 1965.

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

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

[25] W. Lam and F. Bacchus. Learning bayesian belief networks: An approach based on the mdl principle. *Computational intelligence*, 10(3):269–293, 1994.

[26] S. L. Lauritzen and D. 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.

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

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

[29] J. Pearl. *Causality: Models, Reasoning, and Inference.* Cambridge university press, 2009.

[30] A. Perez. Personal communication. 1970–1980.

[31] T. P. Ryan. *Modern regression methods*, volume 655. John Wiley & Sons, 2008.

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

[33] R. E. Tarjan and M. Yannakakis. Simple linear-time algorithms to test chordality of graphs, test acyclicity of hypergraphs, and selectively reduce acyclic hypergraphs. *SIAM Journal on computing*, 13(3):566–579, 1984.

[34] R. E. Tarjan and M. Yannakakis. Addendum: Simple linear-time algorithms to test chordality of graphs, test acyclicity of hypergraphs, and selectively reduce acyclic hypergraphs. *SIAM Journal on Computing*, 14(1):254, 1985.

[35] R. Von Mises. *Probability, statistics, and truth.* Courier Corporation, 1981 [Originaly published in German by Springer, 1928].

[36] I. H. Witten, R. M. Neal, and J. G. Cleary. Arithmetic coding for data compression. *Communications of the ACM*, 30(6):520–540, 1987.