INSTITUTE OF INFORMATION THEORY AND AUTOMATION
ACADEMY OF SCIENCES OF THE CZECH REPUBLIC

On mathematical description of probabilistic conditional independence structures

Milan Studený

Prague, May 2001

The thesis for the degree "Doctor of Science" in the field mathematical informatics and theoretical cybernetics (11-07-9).

The thesis was written in the Department of Decision-Making Theory of the Institute of Information Theory and Automation (Academy of Sciences of the Czech Republic) and was supported by the projects K1019101 and GAČR n. 201/01/1482. It is a result of a long-term research performed also within the framework of ESF research program 'Highly Structured Stochastic Systems' for 1997-2000 and the program 'Causal Interpretation and Identification of Conditional Independence Structures' of the Fields Institute for Research in Mathematical Sciences, University of Toronto, Canada (September - November 1999).

Address of the author:

Milan Studený
Institute of Information Theory and Automation
Academy of Sciences of the Czech Republic
Pod vodárenskou věží 4
18208 Prague 8, Czech Republic

`e-mail: studeny@utia.cas.cz`
`web page: http://www.utia.cas.cz/user_data/studeny/studeny_home.html`

Contents

Chapter 1

Introduction

The central topic of this work is how to *describe the structures of probabilistic conditional independence* in a way that the corresponding mathematical model has both relevant interpretation and offers the possibility of computer implementation.

It is a mathematical work which, however, found its motivation in artificial intelligence and statistics. In fact, these two fields are the main areas where the concept of conditional independence was successfully applied. More specifically, graphical models of conditional independence structure are widely used in

- analysis of *contingency tables* which is an area of discrete statistics dealing with categorical data,
- *multivariate analysis* which is a branch of statistics investigating mutual relationships among continuous real-valued variables,
- *probabilistic reasoning* which is an area of artificial intelligence where decision-making under uncertainty is done on basis of probabilistic models.

Moreover, (non-probabilistic) concept of conditional independence was introduced and studied in several other calculi for dealing with knowledge and uncertainty in artificial intelligence (e.g. relational databases, possibility theory, Spohn's kappa-calculus, Dempster-Shafer's theory of evidence). Thus, the presented work has multidisciplinary flavour. Nevertheless, it certainly falls within the scope of *informatics* or *theoretical cybernetics*, and the main emphasis is put on mathematical groundings.

The work uses concepts from several branches of mathematics, in particular measure theory, discrete mathematics, information theory and algebra. Occasional links to further areas of mathematics occur throughout the work, e.g. to probability theory, mathematical statistics, topology and mathematical logic.

1.1 Motivation account

The reader is asked to excuse the following 'methodological' consideration which perhaps explains my motivation. In the sequel I formulate six general questions of interest which may arise in connection with every particular method of description of conditional independence structures. I think that these questions should be answered in order to judge fairly and carefully the quality and suitability of every particular considered method.

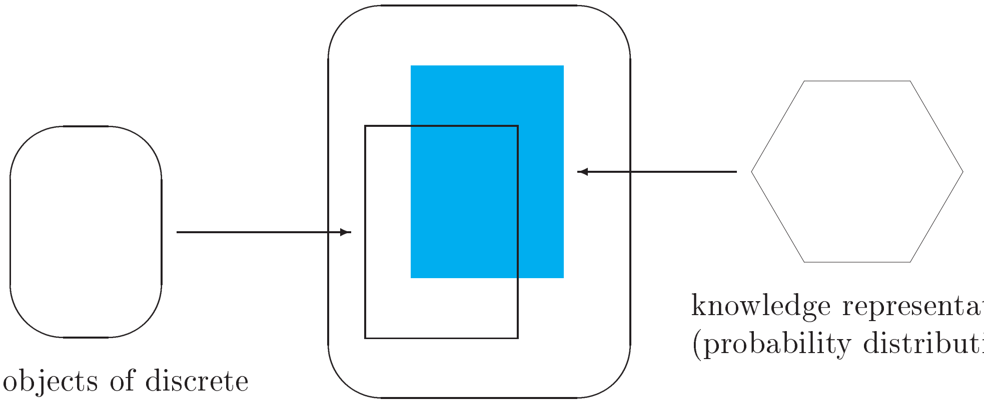


Figure 1.1: Theoretical groundings (informal illustration).

To be more specific one can assume a general situation illustrated by Figure 1.1. One would like to describe conditional independence structures (shortly CI structures) induced by probability distributions from a given fixed class of distributions over a set of variables N. For example, one can consider the class of discrete measures over N (see p. 13), the class of non-degenerate Gaussian measures over N (see p. 28), the class of CG measures over N (see p. 54) or any specific parametrized class of distributions. In probabilistic reasoning every particular discrete probability measure over N represents 'global' knowledge about a (random) system involving variables of N. That means, it serves as a knowledge representative. Thus, one can take even a more general point of view and consider a general class of knowledge representatives within an (alternative) uncertainty calculus of artificial intelligence instead of the class of probability distributions (e.g. a class of possibilistic distributions over N, a class of relational databases over N etc.).

Every knowledge representative of this kind induces a formal independence model over N (for definition see Section 2.2.1 on p. 14). Thus, the class of induced conditional independence models is defined, or in other words, the class of CI structures to be described is specified (the shaded respectively coloured area in Figure 1.1). Well, one has in mind a method of description of CI structures in which objects of discrete mathematics, for example graphs, finite lattices or discrete functions, are used to describe CI structures. Typical examples are classic graphical models widely used in multivariate statistics and probabilistic reasoning (for details see Chapter 3). It is supposed that every object of this type induces a formal independence model over N. Intended interpretation is that the object then 'describes' the induced independence model so that it can possibly describe a conditional independence model, that is one of the CI structures to be described.

The definition of the induced model depends on the type of considered objects. Every class of objects has its specific criterion according to which a formal independence model is ascribed to a particular object. For example, various separation criteria for classic graphical models were obtained as a result of development of miscellaneous Markov properties (see Remark 3.1 in Section 3.1). Evolution ended by the concept of 'global

Markov condition' which establishes a graphical criterion how to determine the maximal set of conditional independence statements represented in a given graph. This set is the induced formal independence model then. The above mentioned implicit assumption is a basic requirement of *consistency*, that is the requirement that every object in the considered class of objects has undoubtedly ascribed a certain formal independence model. Note that some recently developed graphical approaches (see Section 3.5.3) still need to be developed up till the concept of global Markov condition so that they will comply with the basic requirement of consistency.

Under situation above I can formulate first three questions of interest which, in my opinion, are the most important theoretical questions in this general context.

- *Faithfulness* is the question whether every considered object indeed describes one of the CI structures to be described.
- *Completeness* is the question whether every CI structure to be described is described by one of the considered objects. In case this is not the case an advanced subtask occurs, namely to characterize conveniently those formal independence models which can be described by the considered objects.
- *Equivalence* question involves the taks to characterize in a suitable way equivalent objects, that is objects describing the same CI structure. An advanced subquestion is whether one can find a suitable representative for every class of equivalent objects.

The phrase 'faithfulness' was inspired by terminology of [94] where it has similar meaning for graphical objects. Of course, the above notions depend on the considered class of knowledge representatives so that one can differentiate between faithfulness in discrete framework (= relative to the class of discrete measures) and faithfulness in Gaussian framework. Note that in case of classic graphical models faithfulness is usually ensured while completeness is not (see Section 3.6). To avoid misunderstanding let me explain that some authors in the area of (classic) graphical model, including myself, have also used a traditional term "(strong) completeness of a separation graphical criterion" [31, 69, 112, 54]. However, according to the classification above, results of this type belong to the results gathered under label 'faithfulness' (customary reasons of traditional terminology are explained in Remark 3.2 on p. 38). Thus, I distinguish between 'completeness of a criterion' on one hand and 'completeness of a class of objects' (for description of a class of CI structures) one the other hand. Let me remark that not all relevent theoretical questions can be included in the above classification, e.g. the 'inclusion problem' (see p. 129) which can perhaps be regarded as a specific extension of the equivalence question motivated by additional practical questions.

Let me formulate three remaining questions of interest which, in my opinion, are the most important practical questions in this context (for an informal illustrative picture see Figure 1.2).

- *Interpretability* is the question whether considered objects of discrete mathematics can be conveyed to humans in an acceptable way. That usually means whether they can be visualized in a way that they are understood easily and interpreted correctly as CI structures.

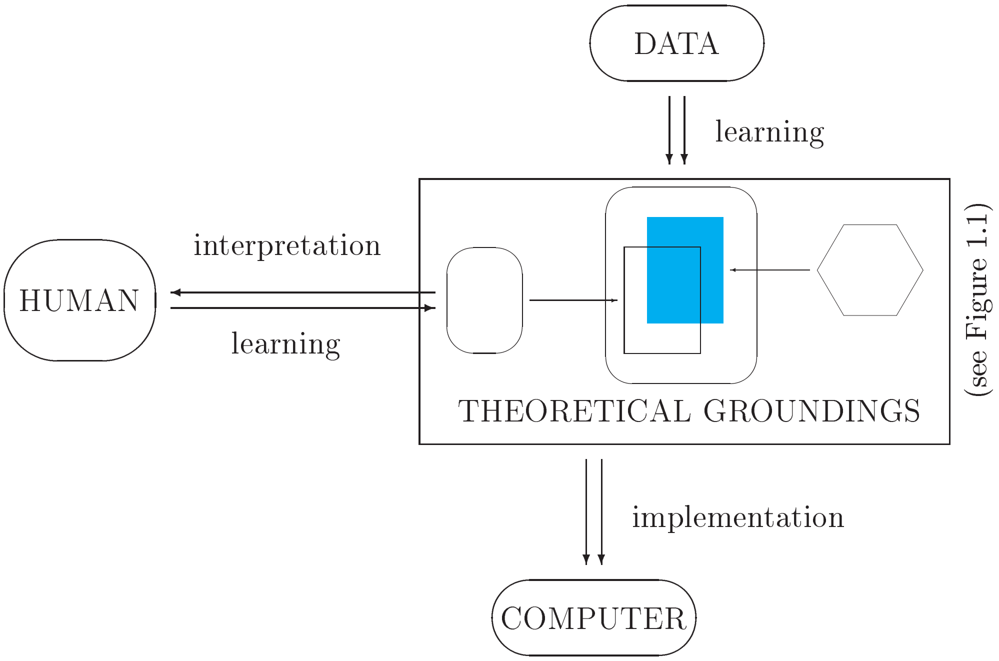


Figure 1.2: Practical questions (informal illustration).

- *Learning* or 'identification' is the question how to determine the most suitable CI structure either on basis of statistical data (= estimation problem) or on basis of expert knowledge provided by human experts. An advanced statistical subtask is to determine even a particular probability distribution inducing the CI structure.

- *Implementation* is the question how to manage the corresponding computational tasks. An advanced subquestion is whether acceptance of a particular CI structure allows one to do respective subsequent calculation with probability distributions effectively, namely whether the describing objects clue in calculation.

Classic graphical models are easily acceptable by humans but their pictorial representation may sometimes lead to another interpretation. For example, acyclic directed graphs can be either interpreted as CI structures or one can prefer 'causal' or 'deterministic' interpretation of their edges [94] which is different. Concerning computational aspects almost ideal framework is provided by the class of *decomposable models* which is a special class of graphical models (see Section 3.4.1). This is a basis of well-known method of 'local computation' [49] which is behind several working probabilistic expert systems [17, 35]. Of course, the presented questions are connected each other. For example, structure learning from experts certainly depends on interpretation while (advanced) distribution learning is closely related to the 'parametrization problem' (see p. 147) which has a strong computational aspect.

The goal of this motivation account is the idea that the practical questions are also strongly connected with theoretical groundings. Thus, in my opinion, before inspection of practical questions one should first solve the related theoretical questions thoroughly. Regretably, some researches in artificial intelligence (marginally in statistics) do not pay enough attention to theoretical groundings and concentrate mainly on practical issues like simplicity of accepted models either from the point of view of computation or visualization. They usually settle in a certain class of 'nice' graphical models (e.g. Bayesian networks - see p. 39) and do not realize that their later technical problems are caused by this limitation.

Even worse, limitation to a small class of models may lead to serious methodological errors! Let me give an example which is my main source of motivation. Consider a hypothetical situation when one is trying to learn CI structure induced by a discrete distribution on basis of statistical data. Suppose, moreover, that one is limited to a certain class of classic graphical models, say Bayesian networks. It is known that this class is not complete in discrete framework (see Chapter 3). Therefore one searches for 'the best approximation'. Well, some of the learning algorithms for graphical models browse thorough the class of possible graphs as follows. One starts with a graph with maximal number of edges, performs certain statistical tests for conditional independence statements and represents the acceptance of these statements by removal of certain edges in the graph. Well, this is a correct procedure in case that the underlying probability distribution indeed induces a CI structure which can be described by a graph within the considered class of graphs. However, in general, this edge removal represents acceptance of a new graphical model together with all other conditional independence statements which are represented in the 'new' graph but which *may not be valid with respect to the underlying distribution*. Let me emphasize once more that this erroneous acceptance of further conditional independence statements is made on basis of a 'correctly recognized' conditional independence statement!

Thus, this error is indeed forced by the limitation to a certain class of graphical models which is not complete. Note that an attitude like this was already criticized several times (see e.g. [125]). In my opinion, these repeated problems in solving practical question of learning are inevitable consequences of omission of theoretical groundings, namely the question of completeness. This maybe motivated several recent attempts to introduce wider and wider classes of graphs which, however, loose easy interpretation and do not achieve completeness. Therefore, in this work, I propose a non-graphical method of description of probabilistic CI structures which primarily solves the completeness and has a potential to take care of practical questions.

1.2 Goals of the work

The aim of the present work is threefold. The first goal is to make an overview of classic methods of description of (probabilistic) CI structures. These methods use mainly graphs whose nodes correspond to variables as a basic tool for visualization and interpretation. The overview involves basic results about conditional independence including those published in my earlier papers.

The second goal is to present a mathematical basis of an alternative method of description of probabilistic CI structures. My alternative method removes certain basic defects of classic methods.

The third goal is an outline of those directions in which the presented method needs to be developed in order to satisfy the requirements of practical applicability. It involves the list of open problems and promising directions of research.

The work is perhaps longer and more detailed than it could be. The reason is that not only experts in the fields and mathematicians are expected audience. My intention was to write a report which can be read and understood by PhD students in computer science and statistics. This was the main stimulus which made me to solve the dilemma 'understandability' versus 'conciseness' in favour of preciseness and understandability.

1.3 Structure of the work

Chapter 2 is an overview of basic definitions, tools and results concerning the concept of *conditional independence*. These notions, including the notion of *imset* which is a certain integer-valued discrete function, are supposed to be a theoretical basis of the rest of the work.

Chapter 3 is an overview of *graphical methods* for description of CI structures. Both classic approaches (undirected graphs, acyclic directed graphs and chain graphs) and recent attempts are included. The chapter makes for a conclusion that a non-graphical method achieving completeness (see Section 1.1, p. 8) is needed.

Chapter 4 introduces a method of this type. The method uses certain imsets, called *structural imsets*, to describe probabilistic CI structures. It is shown that three possible ways of associating probability distributions and structural imsets are equivalent.

Chapter 5 compares two different (but equivalent) *ways of description* CI structures by means of imsets. It is shown that every probabilistic CI structure can be described using this approach and a duality relation between these two ways of description is established.

Chapter 6 is devoted to an advanced question of Markov equivalence (see Section 1.1, p. 8) within the framework of structural imsets. Certain characterization of equivalent imsets is given and related implementation tasks are discussed.

Chapter 7 deals with the problem of choice of a suitable representative of a class of equivalent structural imsets. Possible approaches to this problem are proposed and roughly compared.

Chapter 8 is an overview of open problems to be studied in order to tackle practical question (see Section 1.1 p. 8-9). Chapter 9 (Conclusions) summarizes the presented method.

The Appendix (Chapter 10) is an overview of concepts and facts which are supposed to be elementary and can be omitted by an advanced reader. They are added for several minor reasons: to clarify and unify terminology, to broaden circulation readership and to make reading comfortable as well. It can be used with help of the Index. References conclude the work.

Chapter 2

Basic concepts

Throughout the work the symbol N will denote a non-empty finite set of *variables*. Intended interpretation is that the variables correspond to primitive factors described by random variables. In Chapter 3 variables will be represented by nodes of graphs. The set N will also serve as the basic set for non-graphical tools of discrete mathematics introduced in this work (semi-graphoids, imsets etc.).

The following convention will be used throughout the work: given $A, B \subseteq N$ the juxtaposition AB will denote their union $A \cup B$. Moreover, the following symbols will be reserved for sets of numbers: $\mathbb{R}$ will denote *real numbers*, $\mathbb{Q}$ *rational numbers*, $\mathbb{Z}$ *integers*, $\mathbb{Z}^+$ *non-negative integers* (including 0), $\mathbb{N}$ *natural numbers* (that is positive integers without 0). The symbol $|A|$ will be used to denote the number of elements of a finite set A, that is its *cardinality*. Moreover, the symbol $|x|$ will also denote the *absolute value* of a real number x, that is $|x| = \max\{x, -x\}$.

2.1 Conditional independence

Basic notion of this work is a *probability measure over* N. This phrase will be used to describe the situation when a measurable space $(\mathsf{X}_i, \mathcal{X}_i)$ is given for every $i \in N$ and a probability measure P is defined on the Cartesian product $(\prod_{i\in N} \mathsf{X}_i, \prod_{i\in N} \mathcal{X}_i)$. In this case I will use $(\mathsf{X}_A, \mathcal{X}_A)$ as a shorthand for $(\prod_{i\in A} \mathsf{X}_i, \prod_{i\in A} \mathcal{X}_i)$ for every $\emptyset \neq A \subseteq N$. The *marginal* of P for $\emptyset \neq A \subset N$, denoted by P^A, is defined by the formula

$$P^A(\mathsf{A}) = P(\mathsf{A} \times \mathsf{X}_{N\setminus A}) \quad \text{for} \quad \mathsf{A} \in \mathcal{X}_A \,.$$

Moreover, let us accept two conventions. First, the marginal of P for $A = N$ is P itself, that is $P^N \equiv P$. Second, fully formal convention is that the marginal of P for $A = \emptyset$ is a probability measure on a (fixed appended) measurable space $(\mathsf{X}_\emptyset, \mathcal{X}_\emptyset)$ with trivial σ-algebra $\mathcal{X}_\emptyset = \{\emptyset, \mathsf{X}_\emptyset\}$. Observe that a measurable space of this kind admits only one probability measure $P^\emptyset$.

To give the definition of conditional independence within this framework one needs certain general understanding of the concept of conditional probability. Given a probability measure P over N and disjoint sets $A, C \subseteq N$ by *conditional probability on* X_A *given* C (more specifically given $\mathcal{X}_C$) will be understood a function of two arguments $P_{A|C} : \mathcal{X}_A \times \mathsf{X}_C \to [0, 1]$ which ascribes a $\mathcal{X}_C$-measurable function $P_{A|C}(\mathsf{A}|*)$ to every

$\mathsf{A} \in \mathcal{X}_A$ such that

$$P^{AC}(\mathsf{A} \times \mathsf{C}) = \int_{\mathsf{C}} P_{A|C}(\mathsf{A}|x) \; dP^C(x) \qquad \text{for every} \quad \mathsf{C} \in \mathcal{X}_C \, .$$

Note that no restriction concerning the mappings $\mathsf{A} \mapsto P_{A|C}(\mathsf{A}|x)$, $x \in \mathsf{X}_C$ (often called the regularity requirement - see Section 10.5, Remark on p. 158) is needed within this general approach. Let me emphasize that $P_{A|C}$ depends on the marginal P^{AC} only and that it is defined, for a fixed $\mathsf{A} \in \mathcal{X}_A$, uniquely within the equivalence P^C-almost everywhere. Observe that, owing to the convention above, in case $C = \emptyset$ the conditional probability $P_{A|C}$ coincides in fact with the marginal for A, that means one has $P_{A|\emptyset} \equiv P^A$ (because a constant function can be identified with its value).

Remark 2.1 The conventions above are in concordance with the following unifying perspective. Realize that for every $\emptyset \neq A \subset N$ the measurable space $(X_A, \mathcal{X}_A)$ is isomorphic to the space $(\mathsf{X}_N, \bar{\mathcal{X}}_A)$ where $\bar{\mathcal{X}}_A \subseteq \mathcal{X}_N$ is a certain σ-algebra representing the set A so that inclusion of sets is reflected, namely

$$\bar{\mathcal{X}}_A = \{\mathsf{A} \times \mathsf{X}_{N \setminus A} \, ; \; \mathsf{A} \in \mathcal{X}_A\} = \{\mathsf{B} \in \mathcal{X}_N \, ; \; \mathsf{B} = \mathsf{A} \times \mathsf{X}_{N \setminus A} \quad \text{for} \quad \mathsf{A} \subseteq \mathsf{X}_A\} \, .$$

It is natural to require then that the empty set $\emptyset$ is represented by the trivial σ-algebra $\bar{\mathcal{X}}_\emptyset$ over X_N and N is represented by $\bar{\mathcal{X}}_N = \mathcal{X}_N$. Using this point of view, the marginal P^A corresponds to the restriction of P to $\bar{\mathcal{X}}_A$, and $P_{A|C}$ corresponds to the concept of conditional probability with respect to the σ-algebra $\bar{\mathcal{X}}_C$. Thus, the existence and above mentioned uniqueness of $P_{A|C}$ follows from basic measure-theoretical facts, for details see the Appendix, Section 10.5. $\triangle$

Given a probability measure P over N and pairwise disjoint subsets $A, B, C \subseteq N$ one says that A is *conditionally independent of B given C with respect to P* and writes $A \perp\!\!\!\perp B \,|\, C \; [P]$ if for every $\mathsf{A} \in \mathcal{X}_A$ and $\mathsf{B} \in \mathcal{X}_B$

$$P_{AB|C}(\mathsf{A} \times \mathsf{B}|x) = P_{A|C}(\mathsf{A}|x) \cdot P_{B|C}(\mathsf{B}|x) \qquad \text{for} \quad P^C\text{-a.e.} \;\; x \in \mathsf{X}_C \, . \tag{2.1}$$

Observe that in case $C = \emptyset$ it collapses to a simple equality $P^{AB}(\mathsf{A} \times \mathsf{B}) = P^A(\mathsf{A}) \cdot P^B(\mathsf{B})$, that is to classic independence concept. Note that the validity of (2.1) does not depend on the choice of versions of conditional probabilities given C since these are determined uniquely just within equivalence P^C-almost everywhere.

Remark 2.2 Let me specify the definition for the case of *discrete measures over N*, when X_i is a finite non-empty set and $\mathcal{X}_i = \mathcal{P}(\mathsf{X}_i)$ for every $i \in N$. Then $P_{A|C}$ is determined uniquely exactly on the set $\{x \in \mathsf{X}_C \, ; \; P^C(\{x\}) > 0\}$ by means of the formula

$$P_{A|C}(\mathsf{A}|x) = \frac{P^{AC}(\mathsf{A} \times \{x\})}{P^C(\{x\})} \quad \text{for every} \quad \mathsf{A} \subseteq \mathsf{X}_A \, ,$$

so that $A \perp\!\!\!\perp B \,|\, C \; [P]$ is defined as follows:

$$P_{AB|C}(\mathsf{A} \times \mathsf{B}|x) = P_{A|C}(\mathsf{A}|x) \cdot P_{B|C}(\mathsf{B}|x)$$

for every $\mathsf{A} \subseteq \mathsf{X}_A$, $\mathsf{B} \subseteq \mathsf{X}_B$ and $x \in \mathsf{X}_C$ with $P^C(\{x\}) > 0$. Of course, A and B can be replaced by singletons in this case. Note that the fact that the equality P^C-a.e. concides with the equality on a certain fixed set is a speciality of discrete case. Other common equivalent definitions of conditional independence will be mentioned in Section 2.3. $\triangle$

However, the concept of conditional independence is not exclusively a probabilistic concept. It was introduced in several non-probabilistic frameworks, namely in various calculi for dealing with uncertainty in artificial intelligence - for details and overview see [104, 20, 92]. Formal properties of respective conditional independence concepts may differ in general, but an important fact is that certain basic properties of conditional independence appear to be valid in all these frameworks.

2.2 Semi-graphoid properties

Several authors independently drew attention to these basic formal properties of conditional independence. In modern statistics, they were first accentuated by Dawid [19], then mentioned by Mouchart and Rolin [71], van Putten and van Shuppen [81]. Spohn [96] interpreted them in the context of philosophical logic. Finally, their significance in (probabilistic approach to) artificial intelligence was discerned and highlighted by Pearl and Paz [77]. Their terminology [78] was later widely accepted, so that researchers in artificial intelligence started to call them the *semi-graphoid properties*.

2.2.1 Formal independence models

Formally, a *conditional independence statement over* N is a statement of the form "A is conditionally independent of B given C" where $A, B, C \subseteq N$ are pairwise disjoint subsets of N. A statement of this kind should be always understood with respect to a certain mathematical object $\boldsymbol{o}$ over N, for example a probability measure over N. However, several other objects can occur in place of $\boldsymbol{o}$, for example a graph over N (see Chapter 3), possibility distributions over N [14, 117], relational databases over N [88] or a structural imset over N (see Section 4.4.1). The notation $A \perp\!\!\!\perp B \,|\, C \; [\boldsymbol{o}]$ will be used then; but the symbol $[\boldsymbol{o}]$ can be omitted when it is suitable.

Thus, every conditional independence statement corresponds to a *disjoint triplet over* N, that is a triplet $\langle A, B|C \rangle$ of pairwise disjoint subsets of N. Here, the punctation anticipates the intended role of component sets. The third component, put after the straight line, is the conditioning set while two former components are independent areas, usually interchangeable. Formal difference is that a triplet of this kind can be interpreted either as the corresponding independence statement, or (alternatively) as its negation, that is the corresponding *dependence statement*. Occasionally, I will use the symbol $A \not\perp\!\!\!\perp B \,|\, C \; [\boldsymbol{o}]$ to denote the dependence statement which corresponds to $\langle A, B|C \rangle$. The class of all disjoint triples over N will be denoted by $\mathcal{T}(N)$.

Having established the concept of conditional independence within a certain framework of mathematical objects over N, every object $\boldsymbol{o}$ of this kind defines a certain set of disjoint triplets over N, namely

$$\mathcal{M}_{\boldsymbol{o}} = \{ \langle A, B|C \rangle \in \mathcal{T}(N); \;\; A \perp\!\!\!\perp B \,|\, C \; [\boldsymbol{o}] \}.$$

Let us call it the *conditional independence model induced by* $\boldsymbol{o}$. This phrase is used to indicate that the involved triplets are interpreted as independence statements although from purely mathematical point of view it is nothing but a subset of $\mathcal{T}(N)$. Thus, the conditional independence model induced by a probability measure P over N (according to the definition from Section 2.1) is a special case. Conversely, any class $\mathcal{M} \subseteq \mathcal{T}(N)$ of disjoint triplets over N can be formally interpreted as a conditional independence model if one defines

$$A \perp\!\!\!\perp B \,|\, C\; [\mathcal{M}] \quad \equiv \quad \langle A, B|C\rangle \in \mathcal{M}\,.$$

By *restriction* of a formal independence model $\mathcal{M}$ over N to a set $\emptyset \neq T \subseteq N$ will be undestood the class $\mathcal{M} \cap \mathcal{T}(T)$ denoted by $\mathcal{M}_T$. Evidently, the restriction of a (probabilistic) conditional independence model is again a conditional independence model.

REMARK 2.3 This is to explain my limitation to disjoint triplets over N because some authors [19] do not make this restriction at all. For simplicity of explanation consider discrete probabilistic framework. Indeed, one can introduce, for a discrete probability measure P over N, the statement $A \perp\!\!\!\perp B \,|\, C\; [P]$ even for non-disjoint triplets $A, B, C \subseteq N$ in a reasonable way [27, 61]. However, then the statement $A \perp\!\!\!\perp A \,|\, C\; [P]$ has specific interpretation, namely that the variables in A are functionally dependent on the variables in C (with respect to P), so that it can be interpreted as a *functional dependence statement*. Let us note (cf. Section 2 in [61]) that one can easily derive that

$$A \perp\!\!\!\perp B \,|\, C\; [P] \Leftrightarrow \{\; (A\cap B)\setminus C \perp\!\!\!\perp (A\cap B)\setminus C \,|\, C\cup(B\setminus A)\; [P] \;\;\&\;\; A\setminus C \perp\!\!\!\perp B\setminus AC \,|\, C\; [P] \;\}.$$

Thus, every statement $A \perp\!\!\!\perp B \,|\, C$ of general type can be "reconstructed" from functional dependence statements and from pure conditional independence statements described by disjoint triplets. The topic of this work are pure conditional independence structures; therefore I limit myself to pure conditional independence statements. △

2.2.2 Semi-graphoids

By a *disjoint semi-graphoid over* N is understood any set $\mathcal{M} \subseteq \mathcal{T}(N)$ of disjoint triplets over N (interpreted as independence statements) such that the following conditions hold for every collection of pairwise disjoint sets $A, B, C, D \subseteq N$:

1. triviality $\quad A \perp\!\!\!\perp \emptyset \,|\, C\; [\mathcal{M}]$,
2. symmetry $\quad A \perp\!\!\!\perp B \,|\, C\; [\mathcal{M}]$ implies $B \perp\!\!\!\perp A \,|\, C\; [\mathcal{M}]$,
3. decomposition $\quad A \perp\!\!\!\perp BD \,|\, C\; [\mathcal{M}]$ implies $A \perp\!\!\!\perp D \,|\, C\; [\mathcal{M}]$,
4. weak union $\quad A \perp\!\!\!\perp BD \,|\, C\; [\mathcal{M}]$ implies $A \perp\!\!\!\perp B \,|\, DC\; [\mathcal{M}]$,
5. contraction $\quad A \perp\!\!\!\perp B \,|\, DC\; [\mathcal{M}]$ and $A \perp\!\!\!\perp D \,|\, C\; [\mathcal{M}]$ implies $A \perp\!\!\!\perp BD \,|\, C\; [\mathcal{M}]$.

Note that the terminology above was proposed by Pearl [78] who formulated the formal properties above in the form of inference rules, gave them special names and interpretation and called them the *semi-graphoid axioms*. Of course, the restriction of a semi-graphoid is a semi-graphoid. An important fact is the following one.

LEMMA 2.1 Every conditional independence model induced by a probability measure over N is a disjoint semi-graphoid over N.

Proof: This can be derived easily from Consequence 10.1 proved in the Appendix (see p. 160). Indeed, having a probability measure P over N defined on a measurable space $(\mathsf{X}_N, \mathcal{X}_N)$ one can identify every subset $A \subseteq N$ with a σ-algebra $\bar{\mathcal{X}}_A \subseteq \mathcal{X}_N$ in the way described in Remark 2.1. Then, for a disjoint triplet $\langle A, B|C\rangle$ over N, the statement $A \perp\!\!\!\perp B \,|\, C\; [P]$ is equivalent to the requirement $\bar{\mathcal{X}}_A \perp\!\!\!\perp \bar{\mathcal{X}}_B \,|\, \bar{\mathcal{X}}_C\; [P]$ introduced in Section

10.6. Having in mind that $\bar{\mathcal{X}}_{AB} = \bar{\mathcal{X}}_A \vee \bar{X}_B$ for $A, B \subseteq N$ the rest follows from Consequence 10.1. $\square$

Note that the above mentioned fact is not a special feature of probabilistic framework. Also conditional independence models occuring within other uncertainty calculi mentioned in the end of Section 2.1 are (disjoint) semi-graphoids. Well, even various graphs over N induce semi-graphoids, as explained in Chapter 3.

REMARK 2.4 The limitation to disjoint triplets in the definition of semi-graphoid, is not substantial. One can introduce an *abstract semi-graphoid* on a joint semi-lattice $(\mathcal{S}, \vee)$ as a ternary relation $* \perp\!\!\!\perp * \,|\, *$ over elements A, B, C, D of $\mathcal{S}$ satisfying

- $A \perp\!\!\!\perp B \,|\, C$ whenever $B \vee C = C$,
- $A \perp\!\!\!\perp B \,|\, C$ iff $B \perp\!\!\!\perp A \,|\, C$,
- $A \perp\!\!\!\perp B \vee D \,|\, C$ iff $[\, A \perp\!\!\!\perp B \,|\, D \vee C \;\&\; A \perp\!\!\!\perp D \,|\, C \,]$.

Taking $\mathcal{S} = \mathcal{P}(N)$ one obtains the definition of a non-disjoint semi-graphoid over N. A more complicated example is the semi-lattice of all σ-algebras $\mathcal{A} \subseteq \mathcal{X}$ in a measurable space $(\mathsf{X}, \mathcal{X})$ and the relation $\perp\!\!\!\perp$ of conditional independence of σ-algebras with respect to a probability measure over $(\mathsf{X}, \mathcal{X})$ (see Consequence 10.1). This perspective leads to the general notion of *separoid* introduced in [22] which is a mathematical structure unifying variety of notions of 'irrelevance' arising in probability, statistics, artificial intelligence and other fields. $\triangle$

2.2.3 Elementary independence statements

Well, to store a semi-graphoid over N in memory of a computer one need not allocate all $|\mathcal{T}(N)| = 4^{|N|}$ bits. A more economic way of their representation is feasible. Of course, one can evidently omit *trivial statements* which correspond to triplets $\langle A, B|C \rangle$ over N with $A = \emptyset$ or $B = \emptyset$. Let us denote the class of respective 'trivial' disjoint triplets over N by $\mathcal{T}_{\emptyset}(N)$.

However, principal importance have *elementary statements* or triplets, that is disjoint triplets $\langle A, B|C \rangle$ over N where both A and B are singletons (c.f. [2, 59]). A simplifying convention will be used in that case: braces in singleton notation will be omitted so that $\langle i, j|K \rangle$ or $i \perp\!\!\!\perp j \,|\, K$ will be written only. The class of elementary triplets over N will be denoted by $\mathcal{T}_{\epsilon}(N)$.

LEMMA 2.2 Suppose that $\mathcal{M}$ is a semi-graphoid over N. Then, for every disjoint triplet $\langle A, B|C \rangle$ over N, one has $A \perp\!\!\!\perp B \,|\, C \;[\mathcal{M}]$ iff the following condition holds

$$\forall\, i \in A \quad \forall\, j \in B \quad \forall\, C \subseteq K \subseteq ABC \setminus \{i, j\} \qquad i \perp\!\!\!\perp j \,|\, K \;[\mathcal{M}]. \tag{2.2}$$

In particular, every semi-graphoid is determined by its trace within the class of elementary statements (i.e. by the intersection with $\mathcal{T}_{\epsilon}(N)$).

Proof: (see also [59]) The necessity of the condition (2.2) is easily derivable using decomposition and weak union combined with symmetry.

For converse implication suppose (2.2) and that $\langle A, B|C\rangle$ is not a trivial triplet over N (othervise it is evident). Use induction on $|AB|$; the case $|AB| = 2$ is evident. Supposing $|AB| > 2$ either A of B is not a singleton. Owing to symmetry one can consider without loss of generality $|B| \geq 2$, choose $j \in B$ and put $B' = B \setminus \{j\}$. By induction assumption (2.2) implies both $A \perp\!\!\!\perp j \,|\, B'C$ $[\mathcal{M}]$ and $A \perp\!\!\!\perp B' \,|\, C$ $[\mathcal{M}]$. Hence, by application of the contraction property $A \perp\!\!\!\perp B \,|\, C$ $[\mathcal{M}]$ is derived. □

Sometimes, an *elementary mode of representation* of semi-graphoids (that is by the list of contained elementary statements) is more suitable. The characterization of those collections of elementary triplets which represent semi-graphoids is given in [59].

Remark 2.5 Another reduction of memory demands for semi-graphoid representation follows from symmetry. Instead of keeping a pair of mutually symmetric statements $i \perp\!\!\!\perp j \,|\, K$ and $j \perp\!\!\!\perp i \,|\, K$ one can choose only one of them according to a suitable criterion. In particular, to represent a semi-graphoid over N with $|N| = n$ it suffices to have only $n \cdot (n-1) \cdot 2^{n-3}$ bits. Note that the idea above is also reflected in Section 4.2.1 where just one function corresponds to a 'symmetric' pair of elementary statements.

However, further reduction of the class of considered statements is not possible. The reason is as follows: every elementary triplet $\langle i, j|K\rangle$ over N generates a semi-graphoid over N consisting of $\langle i, j|K\rangle$, its symmetric image $\langle j, i|K\rangle$ and trivial triplets over N (c.f. Lemma 4.5). In fact, these are minimal non-trivial semi-graphoids over N and one has to distinguish them from other semi-graphoids over N. Of course, the above mentioned fact motivated the terminology. △

2.2.4 Problem of axiomatic characterization

Pearl and Paz [77, 78] formulated a conjecture that semi-graphoids coincide with conditional independence models induced by discrete probability measures. However, this conjecture was refuted in [100] by finding a further formal property of these models, not derivable from semi-graphoid properties, namely

$$\begin{aligned} &[\,A \perp\!\!\!\perp B \,|\, CD \quad \& \quad C \perp\!\!\!\perp D \,|\, A \quad \& \quad C \perp\!\!\!\perp D \,|\, B \quad \& \quad A \perp\!\!\!\perp B \,|\, \emptyset\,] \quad \Leftrightarrow \\ &\Leftrightarrow \quad [\,C \perp\!\!\!\perp D \,|\, AB \quad \& \quad A \perp\!\!\!\perp B \,|\, C \quad \& \quad A \perp\!\!\!\perp B \,|\, D \quad \& \quad C \perp\!\!\!\perp D \,|\, \emptyset\,]. \end{aligned}$$

Another formal property of this sort was later derived in [2]. Consequently, a natural question occured. Can conditional independence models arising in discrete probabilistic setting be characterized in terms of a finite number of formal properties of this type? This question is known as the *problem of axiomatic characterization* since a result of this kind would have been a substantial step towards a syntactic description of these models in sense of mathematical logic. Indeed, as explained in Section 5 of [102], then it would have been possible to construct a deductive system which is an analogue of the notion "formal axiomatic theory" from [70]. The wished formal properties then would have played the role of syntactic inference rules of an axiomatic theory of this sort. Unfortunately, the answer to the question above is also negative. It was shown in [102] (for a more didactic proof see [115]) that for every $n \in \mathbb{N}$ there exists a formal property of (discrete) probabilistic conditional independence models which applies on a set of variables N with $|N| = n$ but which cannot be revealed on a set of less cardinality. Note that a basic tool for derivation of these properties was the multiinformation function introduced in Section 2.3.4.

On the other hand, having fixed N, a finite number of possible conditional independence models over N suggests that they can be characterized in terms of a finite number of formal properties. Thus, a related task is, for a small cardinality of N, to characterize them in that way. Well, it makes no problem to verify that in case $|N| = 3$ they coincide with semi-graphoids (see Figure 5.6 for illustration). Discrete probabilistic conditional independence models over N with $|N| = 4$ were characterized in recently completed series of papers [64, 65, 67]. For an overview see [107] where respective formal properties are explicitly formulated (one has 18300 different conditional independence models over N which can be characterized by more than 28 formal properties).

REMARK 2.6 On the other hand several results on relative completeness of semi-graphoid properties were achieved. In [32] and independently in [62] models of "unconditional" stochastic independence (that is submodels consisting of *unconditioned independence statements*, i.e. statements of the form $A \perp\!\!\!\perp B \,|\, \emptyset$) were characterized by means of properties derivable from semi-graphoid properties. Analogous result for the class of *saturated* or *fixed-context conditional independence statements*, that is statements $A \perp\!\!\!\perp B \,|\, C$ with $ABC = N$, was achieved independently in papers [33, 56]. As a specific relative completeness result can be interpreted the result from [109] saying that the semi-graphoid generated by a couple of conditional independence statements is always a conditional independence model induced by discrete probability measure. Note that the problem of axiomatic characterization of CI models mentioned above differs from the problem of axiomatization (in sense of mathematical logic) of a single CI structure over an infinite set of variables N treated in [46]. $\triangle$

2.3 Classes of probability measures

There is no uniform conception of the notion of *probability distribution* in literature. In probability theory authors usually understand by a distribution of a (n-dimensional real) random vector an induced probability measure on the respective sample space ($\mathbb{R}^n$ endowed with the Borel σ-algebra), that is a set function on the sample (measurable) space. On the other hand, authors in artificial intelligence usually identify a distribution of a (finitely-valued) random vector with a pointwise function on the respective (finite) sample space, ascribing probability to every configuration of values (= to every element of the sample space $\prod_{i \in N} \mathsf{X}_i$, where X_i are finite sets). In statistics, either the meaning wavers between these two basic approaches, or authors even avoid the dilemma by describing specific distributions directly by their parameters (e.g. covariance matrix of a Gaussian distribution). Therefore, no exact meaning is assigned to the phrase 'probability distribution' in this work; it is used only in general sense, mainly in vague motivation parts. Moreover, terminological distinction is made between those two above mentioned approaches. The concept of *probability measure* over N from Section 2.1 rather reflects the first approach, which is more general. To relate this to the second approach one has to make an additional assumption on a probability measure P so that it can be also described by a pointwise function, called the *density* of P. Note that many authors simply make an assumption of this type implicitly without mentioning it.

In this section, basic facts about these special probability measures are recalled and several important subclasses of the class of measures having density (called 'marginally

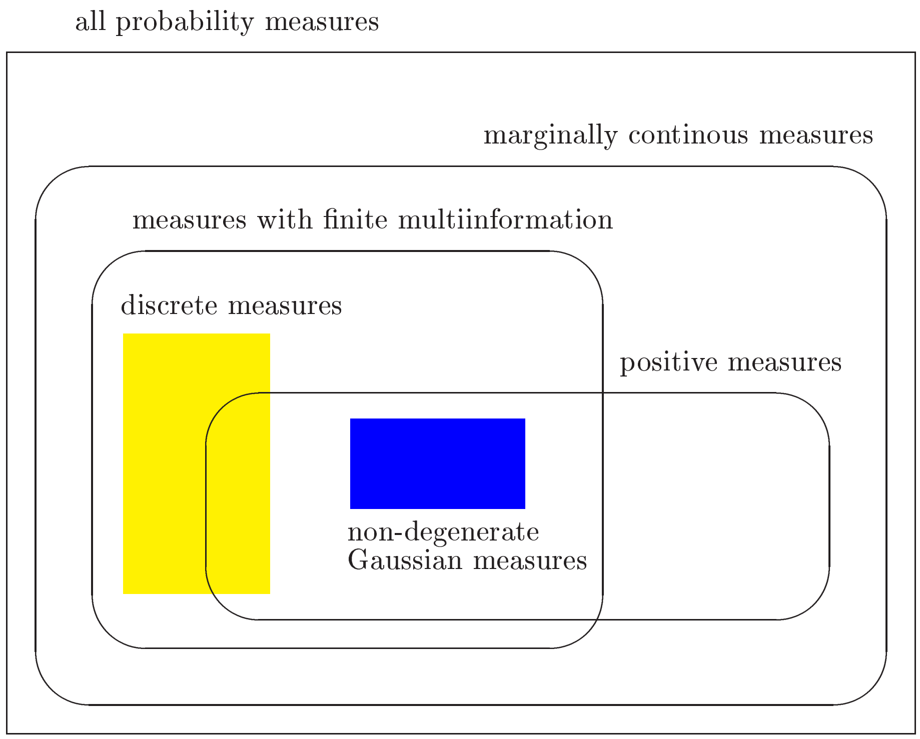


Figure 2.1: The relation of basic classes of probability measures over N.

continuos measures') are introduced. One of them, the class of measures with finite multiinformation is strongly related to the method described in later chapters. The information-theoretical methods are applicable to measures belonging to this class which fortunately involves typical measures used in practice. Mutual relationships among introduced classes of measures are depicted in Figure 2.1.

2.3.1 Marginally continuous measures

A probability measure P over N is *marginally continuous* if it is absolutely continuous with respect to the product of its one-dimensional marginals, that is $P \ll \prod_{i\in N} P^{\{i\}}$. The following lemma contains apparently weaker equivalent definition.

LEMMA 2.3 A probability measure P on $(\mathsf{X}_N, \mathcal{X}_N)$ is marginally continuous iff there exists a collection of σ-finite measures μ_i on $(\mathsf{X}_i, \mathcal{X}_i)$, $i \in N$ such that $P \ll \prod_{i\in N} \mu_i$.

Proof: It was shown in [100], Proposition 1, that in case $|N| = 2$ one has $P \ll \prod_{i\in N} P^{\{i\}}$ iff there are probability measures λ_i on $(\mathsf{X}_i, \mathcal{X}_i)$ with $P \ll \prod_{i\in N} \lambda_i$. One can easily show that for every σ-finite measure μ_i on $(\mathsf{X}_i, \mathcal{X}_i)$ a probability measure λ_i on $(\mathsf{X}_i, \mathcal{X}_i)$ with $\mu_i \ll \lambda_i \ll \mu_i$ exists. Hence, the condition above is equivalent to the requirement of the existence of σ-finite measures μ_i with $P \ll \prod_{i\in N} \mu_i$. Finally, one can use induction on $|N|$ to get the desired conclusion. $\square$

Thus, marginal continuity of P is equivalent to the existence of a *dominating measure μ for P*, that is the product $\mu = \prod_{i\in N} \mu_i$ of some σ-finite measures μ_i on $(\mathsf{X}_i, \mathcal{X}_i)$, $i \in N$ such that $P \ll \mu$. In particular, every discrete measure over N is marginally continuous since the counting measure on X_N can serve a dominating measure. Having fixed a dominating measure μ by a *density of P with respect to μ* will be understood (every version of) the Radon-Nikodym derivative of P with respect to μ.

REMARK 2.7 Let us note without details (see Remark 1 in [100]) that the assumption that a probability measure P over N is marginally continuous also implies that, for every disjoint $A, C \subseteq N$ there exists a regular version of conditional probability $P_{A|C}$ on X_A given $\mathcal{X}_C$ in sense of Loéve [55]. Regularity of conditional probability is usually derived as a consequence of specific topological assumptions on $(\mathsf{X}_i, \mathcal{X}_i)$, $i \in N$ (see the Appendix, Section 10.5). Thus, marginal continuity is a non-topological assumption implying regularity of conditional probabilities. The concept of marginal continuity is closely related to the concept of *dominated experiment* in Bayesian statistics - see §1.2.2 and §1.2.3 in the book [24]. $\triangle$

The next step will be an equivalent definition of conditional independence for marginally continuous measures in terms of densities. To formulate it in an elegant way let us accept the following (notational) convention.

CONVENTION 1 Suppose that a probability measure P on $(\mathsf{X}_N, \mathcal{X}_N)$ together with a fixed dominating measure μ is given. More specifically, $P \ll \mu \equiv \prod_{i \in N} \mu_i$ where μ_i is a σ-finite measure on $(\mathsf{X}_i, \mathcal{X}_i)$ for every $i \in N$.

Then, for every $\emptyset \neq A \subseteq N$, let us put $\mu_A = \prod_{i \in A} \mu_i$, choose a version f_A of Radon-Nikodym derivative $\frac{dP^A}{d\mu_A}$, and fix it. The function f_A will be called a *marginal density of P for A*. It is a $\mathcal{X}_A$-measurable function on X_A.

In order to be able to understand it as a function on X_N as well let us accept the following notation. Given $\emptyset \neq A \subset B \subseteq N$ and $x \in \mathsf{X}_B$, the symbol x_A will denote the *projection of x onto A*, that is $x_A = [x_i]_{i \in A}$ whenever $x = [x_i]_{i \in B}$.

The last formal convention concerns the marginal density $f_\emptyset$ for the empty set. It should be a constant function on (an appended) trivial measurable space $(\mathsf{X}_\emptyset, \mathcal{X}_\emptyset)$. Thus, in formulas below one can simply put $f_\emptyset(x_\emptyset) \equiv 1$ for every $x \in \mathsf{X}_B$, $\emptyset \neq B \subseteq N$. $\diamondsuit$

REMARK 2.8 This is to explain the way of definition of marginal densities in Convention 1. First, let me emphasize that the marginal density is <u>not</u> the Radon-Nikodym derivative of respective marginals since $\mu_A = \prod_{i \in A} \mu_i$ need not coincide with the marginal μ^A of $\mu = \prod_{i \in N} \mu_i$ unless every μ_i is a probability measure.

Indeed, the marginal of a σ-finite measure may not be a σ-finite measure (e.g. $\mu^\emptyset$ in case $\mu(\mathsf{X}_N) = \infty$) so that Radon-Nikodym derivative $\frac{dP^A}{d\mu^A}$ may not exists. Instead, one can take the following point of view. Let us fix a density $f = \frac{dP}{d\mu}$ and introduce, for every $\emptyset \neq A \subset N$, its 'projection' $f^{\downarrow A}$ as a function on X_A defined μ_A-a.e. as follows:

$$f^{\downarrow A}(y) = \int_{\mathsf{X}_{N \setminus A}} f(y, z) \; d\mu_{N \setminus A}(z) \qquad \text{for} \;\; y \in \mathsf{X}_A \,.$$

One can easily conclude using Fubini theorem that $f^{\downarrow A} = \frac{dP^A}{d\mu_A}$ μ_A-a.e., so that there is no substantial difference between $f^{\downarrow A}$ and any version of the marginal density f_A. The convention for the empty set follows this line since one has

$$f^{\downarrow \emptyset}(*) = \int_{\mathsf{X}_N} f(x) \; d\mu(x) = 1 \,.$$

$\triangle$

LEMMA 2.4 Let P be a marginally continuous measure over N. Let us accept Convention 1. Given $\langle A,B|C\rangle \in \mathcal{T}(N)$ one has then $A \perp\!\!\!\perp B\,|\,C\ [P]$ iff the following equality holds

$$f_{ABC}(x_{ABC})\cdot f_C(x_C) = f_{AC}(x_{AC})\cdot f_{BC}(x_{BC}) \qquad \text{for } \mu\text{-a.e. } x \in \mathsf{X}_N. \tag{2.3}$$

Proof: Note that minor omitted details of the proof (e.g. verification of equalities μ-a.e.) can be verified with help of basic measure-theoretical facts gathered in Section 10.5.

As a preparatory step choose and fix a density $f : \mathsf{X}_N \to [0,\infty)$ such that

$$\forall\, \emptyset \neq A \subset N \quad \forall x \in \mathsf{X}_N \qquad f^{\downarrow A}(x_A) \equiv \int_{\mathsf{X}_{N\setminus A}} f(x_A, y)\; d\mu_{N\setminus A}(y) \;<\; \infty\,,$$

and moreover, for every disjoint $A, C \subseteq N$ (with convention $f^{\downarrow N} = f$, $f^{\downarrow\emptyset} \equiv 1$) one has

$$\forall\, x \in \mathsf{X}_N \qquad f^{\downarrow C}(x_C) = 0 \quad\Rightarrow\quad f^{\downarrow AC}(x_{AC}) = 0\,. \tag{2.4}$$

Indeed, these relationships hold μ-a.e. for every version f of $\frac{dP}{d\mu}$ and every version can be overdefined by 0 whenever these relationships do not hold. It makes no problem to verify that $f^{\downarrow A} = \frac{dP^A}{d\mu_A}$ for every $\emptyset \neq A \subseteq N$. Then, for every disjoint $A, C \subseteq N$, one can introduce the function $h_{A|C} : \mathsf{X}_A \times \mathsf{X}_C \to [0,\infty)$ as follows:

$$h_{A|C}(x|z) = \begin{cases} \frac{f^{\downarrow AC}(xz)}{f^{\downarrow C}(z)} & \text{if } f^{\downarrow C}(z) > 0, \\ 0 & \text{if } f^{\downarrow C}(z) = 0, \end{cases} \qquad \text{for } x \in \mathsf{X}_A,\ z \in \mathsf{X}_C.$$

One can verify using Fubini theorem (for $\mu_A \times P^C$), Radon-Nikodym theorem (for $f^{\downarrow C} = \frac{dP^C}{d\mu_C}$) and again Fubini theorem (for $\mu_C \times \mu_A$) that the function

$$(\mathsf{A}, z) \mapsto P_{A|C}(\mathsf{A}|z) \equiv \int_{\mathsf{A}} h_{A|C}(x|z)\; d\mu_A(x) \qquad \text{where } \mathsf{A} \in \mathcal{X}_A,\ z \in \mathsf{X}_C\,,$$

is (a version of) the conditional probability on X_A given $\mathcal{X}_C$.

After this preparatory stage realize that (2.3) can be written as follows:

$$f^{\downarrow ABC}(x_{ABC})\cdot f^{\downarrow C}(x_C) = f^{\downarrow AC}(x_{AC})\cdot f^{\downarrow BC}(x_{BC}) \qquad \text{for } \mu\text{-a.e. } x \in \mathsf{X}_N\,. \tag{2.5}$$

Further, this can be rewritten in the form

$$h_{AB|C}(x_{AB}|x_C)\cdot f^{\downarrow C}(x_C) = h_{A|C}(x_A|x_C)\cdot h_{B|C}(x_B|x_C)\cdot f^{\downarrow C}(x_C) \;\; \text{for } \mu\text{-a.e. } x \in \mathsf{X}_N\,. \tag{2.6}$$

Indeed, owing to (2.4) both (2.5) and (2.6) are trivially valid on $\{x \in \mathsf{X}_N;\ f^{\downarrow C}(x_C) = 0\}$ while they are equivalent on the complement of this set. The next step is to observe that (2.6) is equivalent to the requirement that $\forall\,\mathsf{A} \in \mathcal{X}_A$, $\forall\,\mathsf{B} \in \mathcal{X}_B$, $\forall\,\mathsf{C} \in \mathcal{X}_C$ it holds

$$\int_{\mathsf{C}} \int_{\mathsf{A}\times\mathsf{B}} h_{AB|C}(x_{AB}|x_C)\; d\mu_{AB}(x_{AB})\; dP^C(x_C) =$$

$$= \int_{\mathsf{C}} \int_{\mathsf{A}} h_{A|C}(x_A|x_C)\; d\mu_A(x_A) \cdot \int_{\mathsf{B}} h_{B|C}(x_B|x_C)\; d\mu_B(x_B)\;\; dP^C(x_C)\,.$$

Indeed, as mentioned in Section 10.5 the equality in (2.6) is equivalent to the requirement that their integrals with respect to μ_{ABC} through measurable rectangles $\mathsf{A}\times\mathsf{B}\times\mathsf{C}$ coincide. This can be rewritten using Fubini theorem, Radon-Nikodym theorem and basic properties of integral in the form above. But as explained in the preparatory stage, it can be understood as follows:

$$\int_{\mathsf{C}} P_{AB|C}(\mathsf{A}\times\mathsf{B}|z)\ dP^{C}(z) \ = \int_{\mathsf{C}} P_{A|C}(\mathsf{A}|z)\cdot P_{B|C}(\mathsf{B}|z)\ dP^{C}(z)\,. \tag{2.7}$$

Having fixed $\mathsf{A}\in\mathcal{X}_A$ and $\mathsf{B}\in\mathcal{X}_B$ the equality (2.7) for every $\mathsf{C}\in\mathcal{X}_C$ is equivalent to the requirement that the integrated functions are equal P^C-a.e. Hence, one obtains the condition that (2.1) from p. 13 holds for every $\mathsf{A}\in\mathcal{X}_A$ and $\mathsf{B}\in\mathcal{X}_B$, i.e. $A \perp\!\!\!\perp B\,|\,C\ [P]$. $\square$

Let us observe that in (2.3) one can write 'for μ_{ABC}-a.e. $x\in\mathsf{X}_{ABC}$' instead. Of course, the validity of (2.3) trivially does not depend on the choice of (versions) of densities. The point of Lemma 2.4 is that it even does not depend on the choice of the dominating measure μ since $A \perp\!\!\!\perp B\,|\,C\ [P]$ does depend on it as well. Note that this fact may not be so apparent when one tries to introduce the concept of conditional independence directly by means of densities.

2.3.2 Factorizable measures

Let $\emptyset\neq\mathcal{D}\subseteq\mathcal{P}(N)\setminus\{\emptyset\}$ be a non-empty class of non-empty subsets of N and $D=\bigcup_{T\in\mathcal{D}}$. We say that a marginally continuous measure P over N is *factorizable after* $\mathcal{D}$ (with respect to a dominating measure μ) if the (respective) marginal density of P for D can be expressed in the form

$$f_D\,(x_D)=\prod_{S\in\mathcal{D}} g_S(x_S) \qquad \text{for } \mu\text{-a.e. } x\in\mathsf{X}_N\,, \tag{2.8}$$

where $g_S:\mathsf{X}_S\to[0,\infty)$, $S\in\mathcal{D}$ are $\mathcal{X}_S$-measurable functions, called *potentials*. In fact, factorization does not depend on the choice of a dominating measure μ. One can show that the validity of (2.8) with respect to a general dominating product measure $\mu=\prod_{i\in N}\mu_i$ where μ_i is are σ-finite, is equivalent to the validity of (2.8) with respect to $\prod_{i\in N}P^{\{i\}}$ and with other potentials. Of course, factorization after $\mathcal{D}$ is equivalent to factorization after $\mathcal{D}^{\max}$ and potentials are not unique unless $|\mathcal{D}|=1$.

Further equivalent definition of conditional independence for marginally continuous measures is formulated in terms of factorization (see also [53], § 3.1).

Lemma 2.5 Let P be a marginally continuous measure over N and $\langle A,B|C\rangle\in\mathcal{T}(N)$. Then $A \perp\!\!\!\perp B\,|\,C\ [P]$ if and only if P is factorizable after $\mathcal{D}=\{AC,BC\}$. More specifically, under Convention 1 one has $A \perp\!\!\!\perp B\,|\,C\ [P]$ iff there exist a $\mathcal{X}_{AC}$-measurable function $g:\mathsf{X}_{AC}\to[0,\infty)$ and a $\mathcal{X}_{BC}$-measurable function $h:\mathsf{X}_{BC}\to[0,\infty)$ such that

$$f_{ABC}\,(x_{ABC})=g(x_{AC})\cdot h(x_{BC}) \quad \text{for } \ \mu\text{-a.e. } \ x\in\mathsf{X}_N\,. \tag{2.9}$$

Proof: One can use Lemma 2.4. Clearly, (2.3) $\Rightarrow$ (2.9) where $g=f_{AC}$ and

$$h(x_{BC})=\begin{cases}\frac{f_{BC}(x_{BC})}{f_C(x_C)} & \text{if } \ f_C(x_C)>0\,,\\ 0 & \text{if } \ f_C(x_C)=0\,,\end{cases} \qquad \text{for } x\in\mathsf{X}_N\,,$$

because for μ-a.e. $x\in\mathsf{X}_N$ one has $f_C(x_C)=0 \ \Rightarrow\ f_{BC}(x_{BC})=0$.

For the proof of (2.9) $\Rightarrow$ (2.3) one can first repeat the preparatory step of the proof of Lemma 2.4 (see p. 21), that is to choose a suitable version f of density. Then (2.9) can be rewritten in the form

$$f^{\downarrow ABC}(x_{ABC}) = g\,(x_{AC}) \cdot h\,(x_{BC}) \quad \text{for } \mu\text{-a.e. } x \in \mathsf{X}_N. \tag{2.10}$$

Now, using Fubini theorem and basic properties of integral one can derive from (2.10) by integrating

$$\begin{aligned} f^{\downarrow AC}(x_{AC}) &= g\,(x_{AC}) \cdot h^{\downarrow C}(x_C), \\ f^{\downarrow BC}(x_{BC}) &= g^{\downarrow C}(x_C) \cdot h(x_{BC}), \qquad \text{for } \mu\text{-a.e. } x \in \mathsf{X}_N\,, \\ f^{\downarrow C}(x_C) &= g^{\downarrow C}(x_C) \cdot h^{\downarrow C}(x_C), \end{aligned} \tag{2.11}$$

where the functions

$$g^{\downarrow C}(x_C) = \int\limits_{\mathsf{X}_A} g(y, x_C)\; d\mu_A(y), \quad h^{\downarrow C}(x_C) = \int\limits_{\mathsf{X}_B} h(z, x_C)\, d\mu_B(z) \quad \text{for } x_C \in \mathsf{X}_C\,,$$

are finite μ_C-a.e. (according to Fubini theorem, owing to (2.10) and the fact that $f^{\downarrow ABC}$ is μ_{ABC}-integrable). Thus, (2.10) and (2.11) give together

$$\begin{aligned} f^{\downarrow ABC}(x_{ABC}) \cdot f^{\downarrow C}(x_C) &= g(x_{AC}) \cdot h(x_{BC}) \cdot g^{\downarrow C}(x_C) \cdot h^{\downarrow C}(x_C) = \\ &= f^{\downarrow AC}(x_{AC}) \cdot f^{\downarrow BC}(x_{BC}) \quad \text{for } \mu\text{-a.e. } x \in \mathsf{X}_N\,, \end{aligned}$$

which is equivalent to (2.3). □

As a consequence, one can derive a certain formal property of conditional independence which was already mentioned in discrete case (see [2, 97] and Proposition 4.1 in [61]).

Consequence 2.1 Suppose that P is a marginally continuous measure over N and $A, B, C, D \subseteq N$ are pairwise disjoint sets. Then

$$C \perp\!\!\!\perp D \,|\, AB\; [P],\; A \perp\!\!\!\perp B \,|\, \emptyset\; [P],\; A \perp\!\!\!\perp B \,|\, C\; [P],\; A \perp\!\!\!\perp B \,|\, D\; [P] \;\text{ implies }\; A \perp\!\!\!\perp B \,|\, CD\; [P].$$

Proof: It follows from Lemma 2.4 that the assumption $C \perp\!\!\!\perp D \,|\, AB$ can be rewritted in terms of marginal densities as follows (throughout this proof I write $f(x_S)$ instead of $f_S(x_S)$ for any $S \subseteq N$):

$$f(x_{ABCD}) \cdot f(x_{AB}) \cdot f(x_\emptyset) \cdot f(x_C) \cdot f(x_D) \;=\; f(x_{ABC}) \cdot f(x_{ABD}) \cdot f(x_\emptyset) \cdot f(x_C) \cdot f(x_D)$$

for μ-a.e. $x \in \mathsf{X}_N$. Now, again using Lemma 2.4 the assumptions $A \perp\!\!\!\perp B \,|\, \emptyset$, $A \perp\!\!\!\perp B \,|\, C$ and $A \perp\!\!\!\perp B \,|\, D$ imply that

$$f(x_{ABCD}) \cdot f(x_A) \cdot f(x_B) \cdot f(x_C) \cdot f(x_D) \;=\; f(x_{AC}) \cdot f(x_{BC}) \cdot f(x_{AD}) \cdot f(x_{BD}) \cdot f(x_\emptyset)$$

for μ-a.e. $x \in \mathsf{X}_N$. Since $f(x_A) = 0 \Rightarrow f(x_{ABCD}) = 0$ for μ-a.e. $x \in \mathsf{X}_N$ (and similarly for B, C, D) one can accept the convention $f^{-1}(x_A) = 0$ whenever $f(x_A) = 0$ and obtain

$$\begin{aligned} f(x_{ABCD}) = &\overbrace{f^{-1}(x_A) \cdot f(x_{AC}) \cdot f(x_{AD})}^{g(x_{ACD})} \cdot \\ &\cdot \underbrace{f(x_{BC}) \cdot f(x_{BD}) \cdot f(x_\emptyset) \cdot f^{-1}(x_B) \cdot f^{-1}(x_C) \cdot f^{-1}(x_D)}_{h(x_{BCD})} \quad \text{for } \mu\text{-a.e. } x \in \mathsf{X}_N\,. \end{aligned}$$

Hence, by Lemma 2.5 one has $A \perp\!\!\!\perp B \,|\, CD$. □

2.3.3 Multiinformation and conditional product

Let P be a marginally continuous measure over N. *Multiinformation of* P is the relative entropy $H(P \,|\, \prod_{i\in N} P^{\{i\}})$ of P with respect to the product of its one-dimensional marginals. It is always a value in $[0,+\infty]$ (see Lemma 10.3 in Section 10.7). Common formal convention is that the multiinformation of P is $+\infty$ in case P is not marginally continuous.

REMARK 2.9 The term 'multiinformation' was proposed by my PhD supervisor Albert Perez in late eighties. Note that miscellaneous other terms were used earlier in literature (even by Perez himself); for example 'total correlation' [123], 'dependence tightness' [79] or 'entaxy' [57]. The main reason of Perez's later terminology is that it directly generalizes widely accepted information-theoretical concept of 'mutual information' of two random variables to the case of any finite number of random variables. Indeed, it can serve as a measure of global stochastic dependence among a finite collection of random variables (see § 4 in [115]). Asymptotic behaviour of 'empirical multiinformation' which can be used as a statistical estimate of multiinformation on basis of data was examined in [99]. △

To clarify the significance of multiinformation for study of conditional independence I need the following lemma.

LEMMA 2.6 Let P be a marginally continuous probability measure on $(\mathsf{X}_N, \mathcal{X}_N)$ and $\langle A, B|C\rangle \in \mathcal{T}(N)$. Then there exists unique probability measure Q on $(\mathsf{X}_{ABC}, \mathcal{X}_{ABC})$ such that

$$Q^{AC} = P^{AC}, \quad Q^{BC} = P^{BC} \quad \text{and} \quad A \perp\!\!\!\perp B \,|\, C\; [Q]\,. \tag{2.12}$$

Moreover, $P^{ABC} \ll Q \ll \prod_{i\in ABC} P^{\{i\}}$ and the following equality holds (the symbol H denotes the relative entropy introduced in Section 10.7)

$$\begin{gathered} H(P^{ABC} \,|\, \prod_{i\in ABC} P^{\{i\}}) + H(P^{C} \,|\, \prod_{i\in C} P^{\{i\}}) = \\ H(P^{ABC} \,|\, Q) + H(P^{AC} \,|\, \prod_{i\in AC} P^{\{i\}}) + H(P^{BC} \,|\, \prod_{i\in BC} P^{\{i\}})\,. \end{gathered} \tag{2.13}$$

Proof: Again, omitted technical details can be verified by means of basic measure-theoretical facts from Section 10.5. First, let us verify the uniqueness of Q. Supposing both Q^1 and Q^2 satisfy (2.12) one can observe that $(Q^1)^C = (Q^2)^C$ and then $Q^1_{A|C} \approx Q^2_{A|C}$, $Q^1_{B|C} \approx Q^2_{B|C}$ where $\approx$ indicates the respective equivalence of conditional probabilities (on X_A resp. X_B) given C mentioned in Section 2.1. Because of $A \perp\!\!\!\perp B \,|\, C\; [Q^i]$, $i = 1, 2$ one can derive using (2.1) that $Q^1_{AB|C} \approx Q^2_{AB|C}$ for measurable rectangles which together with $(Q^1)^C = (Q^2)^C$ implies $Q^1 = Q^2$.

For existence proof assume without loss of generality $ABC = N$ and put $\mu \equiv \prod_{i\in N} P^{\{i\}}$. Like in the preparatory step of the proof of Lemma 2.4 (see p. 21) choose a density $f = \frac{dP}{d\mu}$ and respective collection of marginal 'projection' densities $f^{\downarrow A}$, $A \subseteq N$ satisfying (2.4). For brevity, we write $f(x_A)$ instead of $f^{\downarrow A}(x_A)$ in the rest of this proof so that (2.4) has the form

$$\forall\, x \in \mathsf{X}_N \quad \forall\, A, C \subseteq N \text{ such that } A \cap C = \emptyset \qquad f(x_C) = 0 \;\Rightarrow\; f(x_{AC}) = 0\,. \tag{2.14}$$

Let us define a function $g : \mathsf{X}_N \to [0,\infty)$ by

$$g(x) = \begin{cases} \frac{f(x_{AC})\cdot f(x_{BC})}{f(x_C)} & \text{if } f(x_C) > 0, \\ 0 & \text{if } f(x_C) = 0, \end{cases} \qquad \text{for } x \in \mathsf{X}_N = \mathsf{X}_{ABC}\,,$$

and introduce a measure Q on $(\mathsf{X}_N, \mathcal{X}_N)$ as follows:

$$Q(\mathsf{D}) = \int_{\mathsf{D}} g(x)\ d\mu(x) \qquad \text{for} \ \ \mathsf{D} \in \mathcal{X}_N = \mathcal{X}_{ABC}.$$

Now, under the convention $\frac{f(x_{AC})}{f(x_C)} = 0$ in case $f(x_C) = 0$ one can write for every $\mathsf{E} \in \mathcal{X}_{AC}$ using Fubini theorem, (2.14) and Radon-Nikodym theorem

$$\begin{aligned} Q^{AC}(\mathsf{E}) &= \int_{\mathsf{E}\times\mathsf{X}_B} g(x)\ d\mu(x) = \int_{\mathsf{E}} \frac{f(x_{AC})}{f(x_C)} \cdot \int_{\mathsf{X}_B} f(x_B x_C)\ d\mu_B(x_B)\ d\mu_{AC}(x_{AC}) = \\ &= \int_{\mathsf{E}} \frac{f(x_{AC})}{f(x_C)} \cdot f(x_C)\ d\mu_{AC}(x_{AC}) = \int_{\mathsf{E}} f(x_{AC})\ d\mu_{AC}(x_{AC}) = P^{AC}(\mathsf{E}). \end{aligned}$$

Hence, $Q^{AC} = P^{AC}$ and Q is a probability measure. Replace $(\mathsf{X}_A, \mathcal{X}_A)$ by $(\mathsf{X}_B, \mathcal{X}_B)$ in the preceding consideration to obtain $Q^{BC} = P^{BC}$. The way of definition of Q implies $Q \ll \mu$ and $g = \frac{dQ}{d\mu}$. The form of g implies that Q is factorizable after $\{AC, BC\}$ so that $A \perp\!\!\!\perp B\,|\,C\ [Q]$ by Lemma 2.5. To see $P^{ABC} \ll Q$ observe that (2.14) implies $g(x) = 0 \Rightarrow f(x) = 0$ for every $x \in \mathsf{X}_N$, accept the convention $\frac{f(x)}{g(x)} = 0$ in case $g(x) = 0$, and write for every $\mathsf{D} \in \mathcal{X}_N$ using Radon-Nikodym theorem

$$\int_{\mathsf{D}} \frac{f(x)}{g(x)}\ dQ(x) = \int_{\mathsf{D}} \frac{f(x)}{g(x)} \cdot g(x)\ d\mu(x) = \int_{\mathsf{D}} f(x)\ d\mu(x) = P(\mathsf{D})\,.$$

Thus, $P \ll Q$ and $\frac{f}{g} = \frac{dP}{dQ}$. To derive (2.13) realize that it follows from the definition of g (under the convention above) that

$$f(x) \cdot f(x_C) = \frac{f(x)}{g(x)} \cdot f(x_{AC}) \cdot f(x_{BC}) \quad \text{for every} \ \ x \in \mathsf{X}_N\,.$$

Hence, of course

$$\forall\, x \in \mathsf{X}_N \quad \ln f(x) + \ln f(x_C) = \ln \frac{f(x)}{g(x)} + \ln f(x_{AC}) + \ln f(x_{BC}).$$

According to (10.4) and Lemma 10.3 on p. 161 each of five logarithmic terms above is P-quasi-integrable and the integral is a value in $[0, \infty]$ (use $\int_{\mathsf{X}_N} h(x_D)\ dP(x) = \int_{\mathsf{X}_D} h(x_D)\ dP^D(x_D)$ for $D \subseteq N$). Hence, (2.13) was derived. □

Remark 2.10 The measure Q satisfying (2.12) can be interpreted as the *conditional product of P^{AC} and P^{BC}*. Indeed, one can define the conditional product for every pair of *consonant probability measures* (that is measures sharing marginals) in this way. However, in general, some obscurities can occur. First, there exists a pair of consonant measures such that no joint measure having them as marginals exists. Second, even in case joint measures of this type exist, it may happen that none of them complies with the required conditional independence statement. For both examples see [21].

Thus, the assumption of marginal continuity implies the existence of the conditional product. Note that regularity of conditional probabilities $P_{A|C}$ or $P_{B|C}$ is a more general sufficient condition for its existence (see Proposition 2 in [100]). The value of $H(P^{ABC}|Q)$ in (2.13) is known in information theory as the *conditional mutual information of A and B*

given C (*with respect to* P). In case $C = \emptyset$ just the mutual information $H(P^{AB}|P^A \times P^B)$ is obtained, so that it can be viewed as a generalization of mutual information (but from a different perspective than multiinformation). Conditional mutual information is known as a good measure of stochastic dependence between A and B conditional on knowledge of C; for an analysis in discrete case see § 3 in [115]. $\triangle$

2.3.4 Properties of multiinformation function

Supposing P is a probability measure over N the induced *multiinformation function* $m_P : \mathcal{P}(N) \to [0, \infty]$ ascribes the multiinformation of the respective marginal P^S to every non-empty set $S \subseteq N$, that is

$$m_P(S) = H(P^S \mid \prod_{i \in S} P^{\{i\}}) \qquad \text{for every } \emptyset \neq S \subseteq N\,.$$

Moreover, a natural convention $m_P(\emptyset) = 0$ is accepted. The significance of this concept is evident from the following consequence of Lemma 2.6.

Consequence 2.2 Suppose that P is a probability measure over N whose multiinformation is finite. Then the induced multiinformation function m_P is a non-negative real function which satisfies

$$m_P(S) = 0 \quad \text{whenever} \quad S \subseteq N,\ |S| \leq 1, \tag{2.15}$$

and is *supermodular*, that is

$$m_P(ABC) + m_P(C) - m_P(AC) - m_P(BC) \geq 0 \quad \text{whenever} \quad \langle A, B|C\rangle \in \mathcal{T}(N)\,. \tag{2.16}$$

These two conditions imply $m_P(S) \leq m_P(T)$ whenever $S \subseteq T \subseteq N$. Moreover, for every $\langle A, B|C\rangle \in \mathcal{T}(N)$ one has

$$m_P(ABC) + m_P(C) - m_P(AC) - m_P(BC) = 0 \quad \text{iff} \quad A \perp\!\!\!\perp B \mid C\ [P]. \tag{2.17}$$

Proof: The relation (2.15) is evident. Given $S \subseteq N$, put $\langle A, B|C\rangle = \langle S, N \setminus S \mid \emptyset\rangle$ in Lemma 2.6 and (2.13) gives

$$m_P(N) = m_P(N) + m_P(\emptyset) = H(P|Q) + m_P(S) + m_P(N \setminus S)\,.$$

Since all terms here are in $[0, +\infty]$ and $m_P(N) < \infty$ it implies $m_P(S) < \infty$. Therefore (2.13) for general $\langle A, B|C\rangle$ can be always written in the form

$$m_P(ABC) + m_P(C) - m_P(AC) - m_P(BC) = H(P^{ABC} \mid Q),$$

where Q is the conditional product of P^{AC} and P^{BC}. By Lemma 10.3 derive (2.16). It suffices to see $m_P(S) \leq m_P(T)$ when $|T \setminus S| = 1$ which follows directly from (2.16) with $\langle A, B|C\rangle = \langle S, T \setminus S \mid \emptyset\rangle$ and (2.15). The uniqueness of the conditional product Q mentioned in Lemma 2.6 implies that $A \perp\!\!\!\perp B \mid C\ [P]$ iff $P^{ABC} = Q$, that is $H(P^{ABC} \mid Q) = 0$ by Lemma 10.3. Hence (2.17) follows. $\square$

Thus, the class of probability measures over N having finite multiinformation is (by definition) a subclass of the class of marginally continuous measures. It will be shown in

Section 4.1 that it is a quite wide class of measures involving several classes of measures used in practice. The relation (2.17) provides very useful equivalent definition of conditional independence for measures with finite multiinformation, namely by means of an algebraic identity. Note that just relations (2.16) and (2.17) establish a basic method for study of conditional independence used in this work. Because these relations originate from information theory (the expresion in (2.16) is nothing but the conditional mutual information mentioned in Remark 2.10) I dare to call them *information-theoretical tools*. For example, all formal properties of conditional independence from Section 2.2.2 and the result mentioned in the beginning of Section 2.2.4 were derived using this method. Consequence 2.2 also implies that the class of measures with finite multiinformation is closed under marginals. Note without details that it is closed under the operation of conditional product as well.

The following observation appears to be useful later.

LEMMA 2.7 Let P be a probability measure on $(\mathsf{X}_N, \mathcal{X}_N)$ and $P \ll \mu \equiv \prod_{i\in N} \mu_i$ where μ_i is a σ-finite measure on $(\mathsf{X}_i, \mathcal{X}_i)$ for every $i \in N$. Let $\emptyset \neq S \subseteq N$ such that $-\infty < H(P^S \mid \prod_{i\in S} \mu_i) < \infty$ and $-\infty < H(P^{\{i\}} \mid \mu_i) < \infty$ for every $i \in S$. Then $0 \leq m_P(S) < \infty$ and

$$m_P(S) = H(P^S \mid \prod_{i\in S} \mu_i) - \sum_{i\in S} H(P^{\{i\}}|\mu_i)\,. \tag{2.18}$$

Proof: This is just a rough sketch (for technical details see Section 10.5). Suppose without loss of generality $S = N$ and put $\nu = \prod_{i\in N} P^{\{i\}}$. By Lemma 2.3 one knows $P \ll \nu$. Since $P^{\{i\}} \ll \mu_i$ for every $i \in N$ choose versions of $\frac{dP}{d\nu}$ and $\frac{dP^{\{i\}}}{d\mu_i}$ and observe that $\frac{dP}{d\nu} \cdot \prod_{i\in N} \frac{dP^{\{i\}}}{d\mu_i}$ is a version of $\frac{dP}{d\mu}$, defined uniquely P-a.e. Hence derive

$$\ln \frac{dP}{d\nu} = \ln \frac{dP}{d\mu} - \sum_{i\in N} \ln \frac{dP^{\{i\}}}{d\mu_i} \qquad \text{for } P\text{-a.e. } x \in \mathsf{X}_N\,.$$

The assumption of the lemma implies that all logarithmic terms on the right-hand side are P-integrable. Hence, by integration with respect to P (2.18) is obtained. □

2.3.5 Positive measures

A marginally continuous measure P over N is *positive* if there exists a dominating measure μ whose density $f = \frac{dP}{d\mu}$ is (strictly) positive, that is $f(x) > 0$ for μ-a.e. $x \in \mathsf{X}_N$. Note that positivity of density may depend on the choice of dominating measure. However, whenever a measure μ of this kind exists one has $\mu \ll P$. Since $P \ll \prod_{i\in N} P^{\{i\}}$ and $\prod_{i\in N} P^{\{i\}} \ll \prod_{i\in N} \mu_i \equiv \mu$ one can always take $\prod_{i\in N} P^{\{i\}}$ in place of μ. In particular, one can equivalently introduce a positive measure P over N by a simple requirement that $P \ll \prod_{i\in N} P^{\{i\}} \ll P$. Typical example is a *discrete positive measure* P on $\mathsf{X}_N = \prod_{i\in N} \mathsf{X}_i$ with $1 \leq |\mathsf{X}_i| < \infty$, $i \in N$ such that $P(\{x\}) > 0$ for every $x \in \mathsf{X}_N$ (or alternatively only for $x \in \prod_{i\in N} \mathsf{Y}_i$ with $\mathsf{Y}_i = \{y \in \mathsf{X}_i;\ P^{\{i\}}(\{y\}) > 0\}$). These measures play an important role in (probabilistic approach to) artificial intelligence. Pearl [78] noticed that conditional independence models induced by these measures satisfy further special formal property (except semi-graphoid properties) and introduced the following terminology.

A disjoint semi-graphoid $\mathcal{M}$ over N is called a (disjoint) *graphoid* over N if, for every collection of pairwise disjoint sets $A, B, C, D \subseteq N$, one has

6. intersection $\quad A \perp\!\!\!\perp B \,|\, DC \; [\mathcal{M}]$ and $A \perp\!\!\!\perp D \,|\, BC \; [\mathcal{M}]$ implies $A \perp\!\!\!\perp BD \,|\, C \; [\mathcal{M}]$.

It follows from Lemma 2.1 and the observation below that every conditional independence model induced by a positive measure is a disjoint graphoid.

OBSERVATION 2.1 Let P be a marginally continuous measure over N; $A, B, C, D \subseteq N$ are pairwise disjoint, and P^{BCD} is a positive measure over BCD. Then

$$A \perp\!\!\!\perp B \,|\, DC \; [P] \;\; \& \;\; A \perp\!\!\!\perp D \,|\, BC \; [P] \;\; \Rightarrow \;\; A \perp\!\!\!\perp BD \,|\, C \; [P] \, .$$

Proof: (see also [53] for alternative proof under additional restrictive assumption) This is a rough hint only. Let μ be a dominating measure and $f = \frac{dP}{d\mu}$ a density with $f(x_{BCD}) > 0$ for μ-a.e. $x \in \mathsf{X}_N$ (I again follow notational convention from the proof of Lemma 2.4, p. 21). The assumptions $A \perp\!\!\!\perp B \,|\, DC \; [P]$ and $A \perp\!\!\!\perp D \,|\, BC \; [P]$ imply by Lemma 2.4 (one can assume $f(x_E) > 0$ for μ-a.e. $x \in \mathsf{X}_N$ whenever $E \subseteq BCD$)

$$\frac{f(x_{ACD}) \cdot f(x_{BCD})}{f(x_{CD})} = f(x_{ABCD}) = \frac{f(x_{ABC}) \cdot f(x_{BCD})}{f(x_{BC})} \qquad \text{for} \;\; \mu\text{-a.e. } x \in \mathsf{X}_N \, .$$

The terms $f(x_{BCD})$ can be cancelled, so that one derives by dividing

$$f(x_{ACD}) \cdot f(x_{BC}) = f(x_{ABC}) \cdot f(x_{CD}) \qquad \text{for} \;\; \mu\text{-a.e. } x \in \mathsf{X}_N \, .$$

One can take integral with respect to μ_B and get by Fubini theorem

$$f(x_{ACD}) \cdot f(x_C) = f(x_{AC}) \cdot f(x_{CD}) \qquad \text{for} \;\; \mu\text{-a.e. } x \in \mathsf{X}_N \, ,$$

that is $A \perp\!\!\!\perp D \,|\, C \; [P]$ by Lemma 2.4. This, together with $A \perp\!\!\!\perp B \,|\, DC \; [P]$ implies the desired conclusion by the contraction property. $\square$

Let us note that there are discrete probability measures whose induced conditional independence model is not a graphoid, i.e. it does not satisfy the intersection property (see Example 2.2 on p. 32). On the other hand, Observation 2.1 holds also under weaker assumptions on P^{BCD}.

2.3.6 Gaussian measures

These measures are usually treated in multivariate statistics, often under alternative name 'normal distributions'. In this work *Gaussian measures over* N are measures on $(\mathsf{X}_N, \mathcal{X}_N)$ where $(\mathsf{X}_i, \mathcal{X}_i) = (\mathbb{R}, \mathcal{B})$ is the set of real numbers endowed with the σ-algebra of Borel sets for every $i \in N$. Every vector $\boldsymbol{e} \in \mathbb{R}^N$ and every positive semi-definite $N \times N$-matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{N \times N}$ defines a certain measure on $(\mathsf{X}_N, \mathcal{X}_N)$ denoted by $\mathcal{N}(\boldsymbol{e}, \boldsymbol{\Sigma})$ whose *expectation* vector is $\boldsymbol{e}$ and whose *covariance matrix* is $\boldsymbol{\Sigma}$. The components of $\boldsymbol{e}$ and $\boldsymbol{\Sigma}$ are then regarded as parameters of the Gaussian measure.

Attention is almost exclusively paid to *non-degenerate* Gaussian measures which are obtained in case that $\boldsymbol{\Sigma}$ is positive definite (equivalently regular). In that case $\mathcal{N}(\boldsymbol{e}, \boldsymbol{\Sigma})$ can be introduced directly by its density with respect to Lebesgue measure on $(\mathsf{X}_N, \mathcal{X}_N)$

$$f_{\boldsymbol{e},\boldsymbol{\Sigma}}(x) = \frac{1}{\sqrt{(2\pi)^{|N|} \cdot \det(\boldsymbol{\Sigma})}} \cdot \exp^{-\frac{(x-\boldsymbol{e})^\top \cdot \boldsymbol{\Sigma}^{-1} \cdot (x-\boldsymbol{e})}{2}} \qquad \text{for } x \in \mathsf{X}_N, \tag{2.19}$$

where π is Ludolf's constant and $\boldsymbol{\Sigma}^{-1}$ the inverse of the covariance matrix $\boldsymbol{\Sigma}$, called the *concentration matrix*. Its elements are sometimes considered as alternative parameters of a non-degenerate Gaussian measure. Since the density $f_{\boldsymbol{e},\boldsymbol{\Sigma}}$ in (2.19) is positive, non-degenerate Gaussian measures are positive in sense of Section 2.3.5.

On the other hand, in case $\boldsymbol{\Sigma}$ is not regular, the respective *degenerate* Gaussian measure $\mathcal{N}(\boldsymbol{e},\boldsymbol{\Sigma})$ (for a detailed definition see Section 10.9.3) is concentrated on a linear subspace in $\mathbb{R}^N = \mathsf{X}_N$ having Lebesgue measure 0. Thus, degenerate Gaussian measures are not marginally continuous except some rare cases (when the subspace has the form $\{y\} \times \mathsf{X}_A, A \subset N$ for $y \in \mathsf{X}_{N\setminus A}$); for illustration see Example 2.2 below.

Given a Gaussian measure $P = \mathcal{N}(\boldsymbol{e},\boldsymbol{\Sigma})$ over N, non-empty disjoint sets $A, C \subseteq N$ a usual implicit convention (used in multivariate statistics and applicable even in degenerate case) identifies the conditional probability $P_{A|C}$ with its unique 'continuous' version

$$P_{A|C}(*|\, z) = \mathcal{N}(\boldsymbol{e}_A + \boldsymbol{\Sigma}_{A\cdot C} \cdot \boldsymbol{\Sigma}^{-}_{C\cdot C} \cdot (z - \boldsymbol{e}_C), \boldsymbol{\Sigma}_{A\cdot A} - \boldsymbol{\Sigma}_{A\cdot C} \cdot \boldsymbol{\Sigma}^{-}_{C\cdot C} \cdot \boldsymbol{\Sigma}_{C\cdot A}) \quad \text{for every } z \in \mathsf{X}_C.$$

The point is that, for every $z \in \mathsf{X}_C$, it is again a Gaussian measure, whose covariance matrix $\boldsymbol{\Sigma}_{A|C} = \boldsymbol{\Sigma}_{A\cdot A} - \boldsymbol{\Sigma}_{A\cdot C} \cdot \boldsymbol{\Sigma}^{-}_{C\cdot C} \cdot \boldsymbol{\Sigma}_{C\cdot A}$ actually does not depend on the choice of z. Therefore, the matrix $\boldsymbol{\Sigma}_{A|C}$ is called a *conditional covariance matrix*. Recall that in case $C = \emptyset$ one has $\boldsymbol{\Sigma}_{A|C} = \boldsymbol{\Sigma}_{A\cdot A}$ by convention. Elements of misceleaneous conditional covariance matrices can serve as convenient parameters of Gaussian measures - e.g. [7].

Important related fact is that the expectation vector of a Gaussian measure is not significant from the point of view of conditional independence. It follows from the following lemma that single-handed covariance matrix contains all information about conditional independence structure. Therefore it is used in practice almost exclusively.

LEMMA 2.8 Let $P = \mathcal{N}(\boldsymbol{e},\boldsymbol{\Sigma})$ be a Gaussian measure over N and $\langle A, B|C\rangle \in \mathcal{T}(N)$ is a non-trivial triplet over N. Then

$$A \perp\!\!\!\perp B\,|\,C\; [P] \quad \text{iff} \quad (\boldsymbol{\Sigma}_{AB|C})_{A\cdot B} = \mathbf{0}.$$

Proof: The key idea is that topological assumptions (see Remark on p. 158) imply the existence of a regular version of conditional probability on X_{AB} given C, that is a version $\bar{P}_{AB|C}$ such that the mapping $\mathsf{D} \mapsto \bar{P}_{AB|C}(\mathsf{D}\,|\,z)$ is a probability measure on X_{AB} for every $z \in \mathsf{X}_C$. Clearly, for every $\mathsf{A} \in \mathcal{X}_A$, the mapping $z \mapsto \bar{P}_{AB|C}(\mathsf{A} \times \mathsf{X}_B|\, z)$, $z \in \mathsf{X}_C$, is a version of conditional probability on X_A given C; analogously for $\mathsf{B} \in \mathcal{X}_B$. Thus, (2.1) can be rewritten in the form $\forall\, \mathsf{A} \in \mathcal{X}_A$, $\forall\, \mathsf{B} \in \mathcal{X}_B$,

$$\bar{P}_{AB|C}(\mathsf{A} \times \mathsf{B}|\, z) = \bar{P}_{AB|C}(\mathsf{A} \times \mathsf{X}_B|\, z) \cdot \bar{P}_{AB|C}(\mathsf{X}_A \times \mathsf{B}|\, z) \quad \text{for} \quad P^C\text{-a.e. } z \in \mathsf{X}_C\,, \tag{2.20}$$

Since all involved versions of conditional probability are probability measures for every $x \in \mathsf{X}_C$ it is equivalent to the requirement that (2.20) hold for every $\mathsf{A} \in \mathcal{Y}_A$, $\mathsf{B} \in \mathcal{Y}_B$ where $\mathcal{Y}_A$ resp. $\mathcal{Y}_B$ are countable classes closed under finite intersection such that $\sigma(\mathcal{Y}_A) = \mathcal{X}_A$ resp. $\sigma(\mathcal{Y}_B) = \mathcal{X}_B$ (use I.1.5 in [98]). These classes exist in case of Borel σ-algebras on $\mathbb{R}^A$ resp. $\mathbb{R}^B$. The set of $z \in \mathsf{X}_C$ for which (2.20) holds for every $\mathsf{A} \in \mathcal{Y}_A$ and $\mathsf{B} \in \mathcal{Y}_B$ has P^C measure 1 (since $\mathcal{Y}_A$ and $\mathcal{Y}_B$ are countable). For these $z \in \mathsf{X}_C$ then (2.20) holds for every $\mathsf{A} \in \mathcal{X}_A$ and $\mathsf{B} \in \mathcal{X}_B$ by the above mentioned consideration. Hence,

$$A \perp\!\!\!\perp B\,|\,C\; [P] \;\Leftrightarrow\; A \perp\!\!\!\perp B\,|\,\emptyset\; [\bar{P}_{AB|C}(*|\, z)] \qquad \text{for} \quad P^C\text{-a.e. } z \in \mathsf{X}_C.$$

However, in this special case one can suppose that $\bar{P}_{AB|C}(*|\, z)$ is a Gaussian measure (see Section 10.9.3) with the same covariance matrix $\boldsymbol{\Sigma}_{AB|C}$ for every $z \in \mathsf{X}_C$ (while the expectation does depend on z). It is well-known fact that regardless the expectation vector one has $A \perp\!\!\!\perp B \,|\, \emptyset$ with respect to a Gaussian measure iff the $A \times B$-submatrix of its covariance matrix vanishes (see again Section 10.9.3). $\square$

The previous lemma involves the following well-known criteria for elementary conditional independence statements (see also Proposition 5.2 in [53] or Corollaries 6.3.3 and 6.3.4 in [124]).

CONSEQUENCE 2.3 Let P be a Gaussian measure over N with a covariance matrix $\mathbf{\Sigma} = (\sigma_{ij})_{i,j\in N}$ and a correlation matrix $\mathbf{\Gamma} = (\varrho_{ij})_{i,j\in N}$. Then for distinct $i,j \in N$

$$i \perp\!\!\!\perp j \,|\, \emptyset \,[P] \;\Leftrightarrow\; \sigma_{ij} = 0 \;\Leftrightarrow\; \varrho_{ij} = 0\,,$$

and for distinct $i,j,k \in N$

$$i \perp\!\!\!\perp j \,|\, \{k\}\; [P] \;\Leftrightarrow\; \sigma_{kk} \cdot \sigma_{ij} = \sigma_{ik} \cdot \sigma_{kj} \;\Leftrightarrow\; \varrho_{ij} = \varrho_{ik} \cdot \varrho_{kj}\,.$$

If $\mathbf{\Sigma}$ is regular and $\mathbf{\Lambda} = (\kappa_{ij})_{i,j\in N}$ is the concentration matrix, then for distinct $i,j \in N$

$$i \perp\!\!\!\perp j \,|\, N\setminus\{ij\}\; [P] \;\Leftrightarrow\; \kappa_{ij} = 0\,.$$

Proof: The first part is an immediate consequence of Lemma 2.8. For the last fact first observe by elementary computation that a non-diagonal element of a regular 2×2-matrix vanishes iff the same element vanishes in its inverse matrix. In particular,

$$i \perp\!\!\!\perp j \,|\, N\setminus\{i,j\}\; [P] \;\Leftrightarrow\; (\mathbf{\Sigma}_{\{ij\}|N\setminus\{i,j\}})_{ij} = 0 \;\Leftrightarrow ((\mathbf{\Sigma}_{\{ij\}|N\setminus\{i,j\}})^{-1})_{ij} = 0\,.$$

The second observation is that $((\mathbf{\Sigma}_{D|N\setminus D})^{-1})_{D\cdot D} = (\mathbf{\Sigma}^{-1})_{D\cdot D} = \mathbf{\Lambda}_{D\cdot D}$ for every set $D \subseteq N$ containing $\{i,j\}$ (see Section 10.9.1). Hence $((\mathbf{\Sigma}_{D|N\setminus D})^{-1})_{ij} = \kappa_{ij}$ for every such D. $\square$

REMARK 2.11 The proof of Lemma 2.8 reveals notable difference between Gaussian and discrete case. While in discrete case a conditional independence statement $A \perp\!\!\!\perp B \,|\, C\,[P]$ is equivalent to the collection of requirements

$$A \perp\!\!\!\perp B \,|\, \emptyset\; [P_{AB|C}(*|z)] \qquad \text{for every} \;\; z \in \mathsf{X}_C \;\; \text{with} \;\; P^C(z) > 0,$$

in Gaussian case it is equivalent to a single requirement

$$A \perp\!\!\!\perp B \,|\, \emptyset\; [P_{AB|C}(*|z)] \qquad \text{for at least one}\; z \in \mathsf{X}_C,$$

which already implies the same fact for all other $z \in \mathsf{X}_C$ (one uses conventional choice of 'continuous' versions of $P_{AB|C}$ in this case). Informally said, the 'same' conditional independence model is, in Gaussian case, specified by 'less' number of requirements than in discrete case. The reason behind this phenomenon is that the actual number of free parameters characterizing a Gaussian measure over N is, in fact, smaller than the number of parameters characterizing a discrete measure (if $|\mathsf{X}_i| \geq 2$ for $i \in N$). Therefore, discrete measures offer wider variety of induced conditional independence models than Gaussian measures. This is maybe a surprising fact for those who anticipate that a continuous framework should be wider than a discrete framework. The point is that the 'Gaussianity' is quite restrictive assumption. $\triangle$

Thus, one can expect many specific formal properties of conditional independence models arising in Gaussian framework. For example, the following property of a disjoint semi-graphoid $\mathcal{M}$ was recognized by Pearl [78] as a typical property of graphical models (see Chapter 3):

7. composition $\quad A \perp\!\!\!\perp B \,|\, C\; [\mathcal{M}]$ and $A \perp\!\!\!\perp D \,|\, C\; [\mathcal{M}]$ implies $A \perp\!\!\!\perp BD \,|\, C\; [\mathcal{M}]$

for every collection of pairwise disjoint sets $A, B, C, D \subseteq N$. It follows easily from Lemma 2.8 that it is also a typical property of Gaussian conditional independence models.

Consequence 2.4 Let P be a Gaussian measure over N and $A, B, C, D \subseteq N$ are pairwise disjoint. Then

$$A \perp\!\!\!\perp B \,|\, C\; [P] \;\;\&\;\; A \perp\!\!\!\perp D \,|\, C\; [P] \;\;\Rightarrow\;\; A \perp\!\!\!\perp BD \,|\, C\; [P].$$

Proof: Observe that $(\mathbf{\Sigma}_{ABD|C})_{AB\cdot AB} = \mathbf{\Sigma}_{AB|C}$ and $(\mathbf{\Sigma}_{ABD|C})_{AD\cdot AD} = \mathbf{\Sigma}_{AD|C}$ for a covariance matrix $\mathbf{\Sigma}$ (see Section 10.9.1). Thus, the assumptions $(\mathbf{\Sigma}_{ABD|C})_{A\cdot B} = \mathbf{0}$ and $(\mathbf{\Sigma}_{ABD|C})_{A\cdot D} = \mathbf{0}$ imply together $(\mathbf{\Sigma}_{ABD|C})_{A\cdot BD} = \mathbf{0}$. $\square$

However, composition is not universally valid property of conditional independence models as the following example shows.

Example 2.1 There exists a discrete (binary) probability measure P over N with $|N| = 3$ such that

$$i \perp\!\!\!\perp j \,|\, \emptyset\; [P] \;\text{ and }\; \neg(\, i \perp\!\!\!\perp j \,|\, \{k\}\; [P]\,) \quad \text{for any distinct} \;\; i, j, k \in N.$$

Indeed, put $\mathsf{X}_i = \{0, 1\}$ for $i \in N$ and ascribe the probability $\frac{1}{4}$ to every of the following configurations of values: $(0, 0, 0)$, $(0, 1, 1)$, $(1, 0, 1)$, $(1, 1, 0)$. $\diamondsuit$

Further important fact is that every non-degenerate Gaussian measure has finite multiinformation. This follows from Lemma 2.7.

Consequence 2.5 Let P be a non-degenerate Gaussian measure with a correlation matrix $\mathbf{\Gamma}$. Then its multiinformation has the value

$$m_P(N) = -\frac{1}{2} \cdot \ln(\det(\mathbf{\Gamma}))\,. \tag{2.21}$$

Proof: Take the Lebesgue measure on $(\mathsf{X}_N, \mathcal{X}_N)$ in place of μ in Lemma 2.7. Substitution of (10.9) from Section 10.9.3 into (2.18) gives

$$-\frac{|N|}{2} \cdot \ln(2\pi) - \frac{|N|}{2} - \frac{1}{2} \cdot \ln(\det(\mathbf{\Sigma})) - \sum_{i \in N} \{\frac{-\ln{(2\pi)}}{2} - \frac{1}{2} - \frac{1}{2} \cdot \ln{(\sigma_{ii})}\} =$$
$$= \frac{1}{2} \sum_{i \in N} \ln \sigma_{ii} - \frac{1}{2} \cdot \ln(\det(\mathbf{\Sigma})) = -\frac{1}{2} \cdot \ln \frac{\det(\mathbf{\Sigma})}{\prod_{i \in N} \sigma_{ii}} = -\frac{1}{2} \cdot \ln(\det(\mathbf{\Gamma}))\,.$$

$\square$

On the other hand, a degenerate Gaussian measure need not be marginally continuous as the following example shows. It also demonstrates that the intersection property mentioned in Section 2.3.5 is not universally valid.

EXAMPLE 2.2 There exists a Gaussian measure P over N with $|N| = 3$ such that

$$i \perp\!\!\!\perp j \,|\, \{k\} \; [P] \quad \text{and} \quad \neg(\, i \perp\!\!\!\perp j \,|\, \emptyset \; [P]\,) \quad \text{for arbitrarily chosen distinct } i,j,k \in N.$$

Put $P = \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$ where $\mathbf{\Sigma} = (\sigma_{ij})_{i,j\in N}$ with $\sigma_{ij} = 1$ for every $i,j \in N$ and apply Consesquence 2.3. It makes no problem to verify (see Section 10.9.3) that P is concentracted on the subspace $\{(x,x,x)\,;x \in \mathbb{R}\}$ while $P^{\{i\}} = \mathcal{N}(0,1)$ for every $i \in N$. Since $\prod_{i\in N} P^{\{i\}}$ is absolutely continuous with respect to Lebesgue measure, P is not marginally continuous.

Note that the same conditional independence model can be induced by a discrete (binary) measure; put $\mathsf{X}_i = \{0,1\}$ for $i \in N$ and ascribe the probability $\frac{1}{2}$ to configurations $(0,0,0)$ and $(1,1,1)$. ♢

2.3.7 Basic construction

The following lemma provides a basic method of construction of probability measures with prescribed CI structure.

LEMMA 2.9 Let P, Q are probability measures over N. Then there exists a probability measure R over N such that $\mathcal{M}_R = \mathcal{M}_P \cap \mathcal{M}_Q$. Moreover, if P and Q have finite multiinformation then a probability measure R over N with finite multiinformation and $\mathcal{M}_R = \mathcal{M}_P \cap \mathcal{M}_Q$ exists. The same statement holds for the class of discrete measures over N, respectively for the class of positive discrete measures over N.

Proof: Let P be a measure on a space $(\mathsf{X}_N, \mathcal{X}_N) = (\prod_{i\in N} \mathsf{X}_i, \prod_{i\in N} \mathcal{X}_i)$ and Q be a measure on $(\mathsf{Y}_N, \mathcal{Y}_N) = (\prod_{i\in N} \mathsf{Y}_i, \prod_{i\in N} \mathcal{Y}_i)$. Let us put $(\mathsf{Z}_i, \mathcal{Z}_i) = (\mathsf{X}_i \times \mathsf{Y}_i, \mathcal{X}_i \times \mathcal{Y}_i)$ for $i \in N$, introduce $(\mathsf{Z}_N, \mathcal{Z}_N) = \prod_{i\in N}(\mathsf{Z}_i, \mathcal{Z}_i)$ which can be understood as $(\mathsf{X}_N \times \mathsf{Y}_N, \mathcal{X}_N \times \mathcal{Y}_N)$ and define a probability measure R on $(\mathsf{Z}_N, \mathcal{Z}_N)$ as the product of P and Q. The goal is to show that for every $\langle A,B|C\rangle \in \mathcal{T}(N)$

$$A \perp\!\!\!\perp B \,|\, C \; [R] \quad \Leftrightarrow \quad \{\, A \perp\!\!\!\perp B \,|\, C \; [P] \;\; \text{and} \;\; A \perp\!\!\!\perp B \,|\, C \; [Q]\,\}\,. \tag{2.22}$$

Let us take unifying perspective indicated in Remark 2.1: $(\mathsf{Z}_N, \mathcal{Z}_N)$ and R are fixed and respective coordinate σ-algebras $\bar{\mathcal{X}}_A, \bar{\mathcal{Y}}_A, \bar{\mathcal{Z}}_A \subseteq \mathcal{Z}_N$ are ascribed to every $A \subseteq N$. Then P corresponds to the restriction of R to $\bar{\mathcal{X}}_N$, Q to the restriction of R to $\bar{\mathcal{Y}}_N$ and (2.22) takes the form (see Section 10.6 for related concepts)

$$\bar{\mathcal{Z}}_A \perp\!\!\!\perp \bar{\mathcal{Z}}_B \,|\, \bar{\mathcal{Z}}_C \; [R] \quad \Leftrightarrow \quad \bar{\mathcal{X}}_A \perp\!\!\!\perp \bar{\mathcal{X}}_B \,|\, \bar{\mathcal{X}}_C \; [R] \;\; \text{and} \;\; \bar{\mathcal{Y}}_A \perp\!\!\!\perp \bar{\mathcal{Y}}_B \,|\, \bar{\mathcal{Y}}_C \; [R]\,. \tag{2.23}$$

As $\mathcal{X}_A \times \mathcal{Y}_A$-measurable rectangles generate $\mathcal{Z}_A$ for every $A \subseteq N$ by the 'weaker' formulation of the definition of conditional independence for σ-algebras observe that the fact $\bar{\mathcal{Z}}_A \perp\!\!\!\perp \bar{\mathcal{Z}}_B \,|\, \bar{\mathcal{Z}}_C \; [R]$ is equivalent to the requirement: $\forall\, \mathsf{A}^x \in \bar{\mathcal{X}}_A,\, \mathsf{A}^y \in \bar{\mathcal{Y}}_A,\, \mathsf{B}^x \in \bar{\mathcal{X}}_B,\, \mathsf{B}^y \in \bar{\mathcal{Y}}_B$

$$R(\mathsf{A}^x \cap \mathsf{A}^y \cap \mathsf{B}^x \cap \mathsf{B}^y \,|\, \bar{\mathcal{Z}}_C)(z) = R(\mathsf{A}^x \cap \mathsf{A}^y \,|\, \bar{\mathcal{Z}}_C)(z) \cdot R(\mathsf{B}^x \cap \mathsf{B}^y \,|\, \bar{\mathcal{Z}}_C)(z) \text{ for } R\text{-a.e. } z \in \mathsf{Z}_N\,. \tag{2.24}$$

On the other hand $\bar{\mathcal{X}}_A \perp\!\!\!\perp \bar{\mathcal{X}}_B \,|\, \bar{\mathcal{X}}_C \; [R]$ is equivalent to the requirement by usual definition of conditional indepencence for σ-algebras: $\forall\, \mathsf{A}^x \in \bar{\mathcal{X}}_A,\, \mathsf{B}^x \in \bar{\mathcal{X}}_B$

$$P(\mathsf{A}^x \cap \mathsf{B}^x \,|\, \bar{\mathcal{X}}_C)(x) = P(\mathsf{A}^x \,|\, \bar{\mathcal{X}}_C)(x) \cdot P(\mathsf{B}^x \,|\, \bar{\mathcal{X}}_C)(x) \;\; \text{for } R\text{-a.e. } z = (x,y) \in \mathsf{Z}_N, \tag{2.25}$$

and $\bar{\mathcal{Y}}_A \perp\!\!\!\perp \bar{\mathcal{Y}}_B \,|\, \bar{\mathcal{Y}}_C \; [R]$ is equivalent to the requirement: $\forall\, \mathsf{A}^y \in \bar{\mathcal{Y}}_A,\, \mathsf{B}^y \in \bar{\mathcal{Y}}_B$

$$Q(\mathsf{A}^y \cap \mathsf{B}^y \,|\, \bar{\mathcal{Y}}_C)(y) = Q(\mathsf{A}^y \,|\, \bar{\mathcal{Y}}_C)(y) \cdot Q(\mathsf{B}^y \,|\, \bar{\mathcal{Y}}_C)(y) \;\; \text{for } R\text{-a.e. } z = (x,y) \in \mathsf{Z}_N\,. \tag{2.26}$$

Moreover, using the (weaker) definition of conditional probability (see Section 10.5, p. 158) and by the definition of R verify that

$$R(\mathsf{A}^x\cap\mathsf{A}^y\cap\mathsf{B}^x\cap\mathsf{B}^y\,|\,\bar{\mathcal{Z}}_C)(z) \;=\; P(\mathsf{A}^x\cap\mathsf{B}^x\,|\,\bar{\mathcal{X}}_C)(x)\cdot Q(\mathsf{A}^y\cap\mathsf{B}^y\,|\,\bar{\mathcal{Y}}_C)(y) \;\text{ for } R\text{-a.e. } z=(x,y)\in\mathsf{Z}_N. \quad (2.27)$$

Thus, to evidence (2.24)⇒(2.25) put $\mathsf{A}^y=\mathsf{B}^y=\mathsf{Z}_N$; to evidence (2.24)⇒(2.26) put $\mathsf{A}^x=\mathsf{B}^x=\mathsf{Z}_N$. Conversely, (2.25),(2.26)⇒(2.24) by (2.27) which means (2.23) was verified.

In both P and Q have finite multiinformation then $R^{\{i\}}=P^{\{i\}}\times Q^{\{i\}}$ are marginals of R on $(\mathsf{Z}_i,\mathcal{Z}_i)$ for $i\in N$ and $R\ll\prod_{i\in N}P^{\{i\}}\times\prod_{j\in N}Q^{\{j\}}=\prod_{k\in N}P^{\{k\}}\times Q^{\{k\}}$. Thus, R is marginally continuous measure over N and one can apply Lemma 2.6 to R with 'doubled' N to see that

$$H(R\,|\prod_{i\in N}P^{\{i\}}\times\prod_{j\in N}Q^{\{i\}})=H(P\,|\prod_{i\in N}P^{\{i\}})+H(Q\,|\prod_{j\in N}Q^{\{j\}})\,.$$

Note for explanation that in the considered case R is the conditional product of P and Q and therefore the term $H(P^{ABC}|Q)$ in (2.13) vanishes by Lemma 10.3 from Section 10.7. In particular, the multiinformation of R is the sum of the multiinformations P and Q and therefore it is finite. The statement concerning discrete and positive discrete measures easily follows from the given construction. □

Elementary constructions of probability measures are needed to utilize the method from Lemma 2.9. One of them is the product of one-dimensional probability measures.

Observation 2.2 There exists a discrete (binary) probability measure P over N such that

$$A\perp\!\!\!\perp B\,|\,C\;[P]\quad\text{for every }\;\langle A,B|C\rangle\in\mathcal{T}(N).$$

Observation 2.3 Suppose that $|N|\geq 2$ and $A\subseteq N$ with $|A|\geq 2$. Then there exists a discrete (binary) probability measure P over N such that

$$m_P(S)=\begin{cases}\ln 2 & \text{if } A\subseteq S,\\ 0 & \text{otherwise.}\end{cases}$$

Proof: Put $\mathsf{X}_i=\{0,1\}$ for $i\in N$ and ascribe the probability $2^{1-|N|}$ to every configuration of values $[x_i]_{i\in N}$ with even $\sum_{i\in A}x_i$ (remaining configurations have zero probability). □

Lemma 2.10 Suppose that $|N|\geq 3$, $2\leq l\leq|N|$ and $\mathcal{L}\subseteq\{S\subseteq N;|S|=l\}$. Then there exists a discrete probability measure P over N such that

$$\forall\,\langle i,j|K\rangle\in\mathcal{T}_\epsilon(N)\;\text{ with }|ijK|=l\qquad i\perp\!\!\!\perp j\,|\,K\;[P]\;\Leftrightarrow\;ijK\notin\mathcal{L}\,. \qquad (2.28)$$

Proof: If $\mathcal{L}=\emptyset$ then use Observation 2.2. If $\mathcal{L}\neq\emptyset$ then apply Observation 2.3 to every $A\in\mathcal{L}$ and Consequence 2.2 to get a binary probability measure $P_{[A]}$ such that

$$\forall\text{ elementary triplet }\langle i,j|K\rangle\text{ with }|ijK|=l\qquad i\perp\!\!\!\perp j\,|\,K\;[P_{[A]}]\Leftrightarrow ijK\neq A.$$

Then Lemma 2.9 can be applied repeatedly to get a measure over N satisfying (2.28). □

This gives a lower estimate of the number of 'discrete' probabilistic CI structures.

Consequence 2.6 If $n=|N|\geq 3$ then the number of distinct CI structures induced by discrete probability measures over N exceeds the number $2^{2^{\lfloor\frac{n}{2}\rfloor}}$ where $\lfloor\frac{n}{2}\rfloor$ denotes the lower integer part of $\frac{n}{2}$.

Proof: Let us put $l = \frac{n}{2}$ for even n, respectively $l = \frac{n+1}{2}$ for odd n. By Lemma 2.10 for every subclass $\mathcal{L}$ of $\{S \subseteq N; |S| = l\}$ a respective probability measure $P_{[\mathcal{L}]}$ exists. By (2.28) these measures induce distinct CI models over N. Therefore, the number of distinct induced CI models exceeds 2^s where s is the number of elements of $\{S \subseteq N; |S| = l\}$. Find suitable lower estimates for s. If $l = \frac{n}{2}$ then write

$$s = \binom{2l}{l} = \frac{1 \cdot 2 \cdot \ldots \cdot 2l}{(1 \cdot \ldots \cdot l) \cdot (1 \cdot \ldots \cdot l)} = \frac{1 \cdot 3 \cdot \ldots \cdot (2l-1)}{1 \cdot 2 \cdot \ldots \cdot l} \cdot \frac{2 \cdot 4 \cdot \ldots \cdot 2l}{1 \cdot 2 \cdot \ldots \cdot l} \geq 2^l = 2^{\lfloor \frac{n}{2} \rfloor}.$$

Similarly, in case $l = \frac{n+1}{2}$ write

$$s = \binom{2l-1}{l} = \frac{1 \cdot 3 \cdot \ldots \cdot (2l-1)}{1 \cdot 2 \cdot \ldots \cdot l} \cdot \frac{2 \cdot 4 \cdot \ldots \cdot (2l-2)}{1 \cdot 2 \cdot \ldots \cdot (l-1)} \geq 2^{l-1} = 2^{\lfloor \frac{n}{2} \rfloor}.$$

which implies the desired conclusion $2^s \geq 2^{2^{\lfloor \frac{n}{2} \rfloor}}$ in both cases. □

2.4 Imsets

By an *imset over* N is understood an integer-valued function on the power set of N, that is any function $u : \mathcal{P}(N) \to \mathbb{Z}$, or alternatively an element of $\mathbb{Z}^{\mathcal{P}(N)}$. Basic operation with imsets, namely summation, subtraction, multiplication by an integer are defined coordinatewisely. Analogously, we write $u \leq v$ for imsets u, v over N if $u(S) \leq v(S)$ for every $S \subseteq N$. *Multiset* is an imset with non-negative values, that is any function $m : \mathcal{P}(N) \to \mathbb{Z}^+$. Any imset u over N can be written as a difference $u = u^+ - u^-$ of two multisets over N where u^+ is the *positive part* of u and u^- is the *negative part* of u, defined as follows:

$$u^+(S) = \max\{u(S), 0\}, \quad u^-(S) = \max\{-u(S), 0\} \qquad \text{for } S \subseteq N.$$

By *positive domain* of u will be undestood the class of sets $\mathcal{D}_u^+ = \{S \subseteq N;\, u(S) > 0\}$, by *negative domain* of u the class $\mathcal{D}_u^- = \{S \subseteq N;\, u(S) < 0\}$.

Remark 2.12 The word ‘multiset’ is taken from combinatorial theory [1] while the word ‘imset’ is an abbreviation for **i**nteger-valued **multiset**. Later in this work certain special imsets will be used to describe probabilistic conditional independence structures (see Section 4.2.3). △

Trivial example of an imset is the *zero imset* denoted by 0 which ascribes zero value to every $S \subseteq N$. Another simple example is the *identificator* of a set $A \subseteq N$ denoted by δ_A and defined as follows:

$$\delta_A(S) = \begin{cases} 1 & \text{in case } S = A, \\ 0 & \text{in case } S \subseteq N,\ S \neq A. \end{cases}$$

Special notation $m^{A\uparrow}$ respectively $m^{A\downarrow}$ will be used for multisets which serve as identificators of classes of subsets respectively classes of supersets of a set $A \subseteq N$:

$$m^{A\downarrow}(S) = \begin{cases} 1 & \text{if } A \supseteq S, \\ 0 & \text{otherwise}, \end{cases} \qquad \text{and} \quad m^{A\uparrow}(S) = \begin{cases} 1 & \text{if } A \subseteq S, \\ 0 & \text{otherwise}. \end{cases}$$

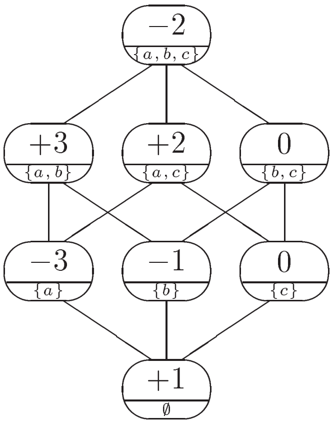


Figure 2.2: Hasse diagram of an imset over $N = \{a, b, c\}$.

It is clear how to represent an imset over N in memory of a computer, namely by a vector with $2^{|N|}$ integral components which correspond to subsets of N. However, for a small number of variables, one can also visualize imsets in a more telling way, using special pictures. The power set $\mathcal{P}(N)$ is a distributive lattice and can be represented in the form of *Hasse diagram* (see p. 6 in [8]). Nodes of this diagram correspond to elements of $\mathcal{P}(N)$, that is to subsets of N, and a link is made between two nodes if the symmetric difference of the represented sets is a singleton. A function on $\mathcal{P}(N)$ can be visualized by writing assigned values into respective nodes. For example, the imset u over $N = \{a, b, c\}$ defined by the table

S	$\emptyset$	$\{a\}$	$\{b\}$	$\{c\}$	$\{a,b\}$	$\{a,c\}$	$\{b,c\}$	$\{a,b,c\}$
$u(S)$	$+1$	-3	-1	0	$+3$	$+2$	0	-2

can be visualized in the form of the diagram from Figure 2.2. The third possible way of description of an imset (used in this work) is to write it as a combination of more elementary imsets with integral coefficients. For example, the imset u from Figure 2.2 can be written as follows:

$$u \;=\; -2 \cdot \delta_N + 3 \cdot \delta_{\{a,b\}} + 2 \cdot \delta_{\{a,c\}} - 3 \cdot \delta_{\{a\}} - \delta_{\{b\}} + \delta_{\emptyset}\,.$$

In this work, certain special imsets over N will be used. Effective dimension of these imsets, that is the actual number of free values is not $2^{|N|}$ but $2^{|N|} - |N| - 1$ only. There are several ways of standardization of imsets of this kind. I will distinguish three basic ways of standardization (for justification of terminology see Remark 5.3 in Section 5.1.2). An imset u over N is *o-standardized* if

$$\sum_{S \subseteq N} u(S) = 0 \quad \text{and} \qquad \forall\, i \in N \sum_{S \subseteq N,\, i \in S} u(S) = 0\,.$$

Alternatively, the second requirement can be formulated in the form $\sum_{S \subseteq N \setminus \{j\}} u(S) = 0$ for every $j \in N$. An imset u is *ℓ-standardized* if

$$u(S) = 0 \qquad \text{whenever} \;\; S \subseteq N,\; |S| \leq 1\,,$$

and *u-standardized* if

$$u(S) = 0 \qquad \text{whenever } S \subseteq N,\ |S| \geq |N| - 1\,.$$

An imset u over N is called *normalized* if the collection of values $\{u(S);\, S \ \subseteq N\}$ has no common prime divisor. Except basic operations with imsets the operation of *scalar product* of a real function $m : \mathcal{P}(N) \rightarrow \mathbb{R}$ and an imset u over N denoted by $\langle m, u\rangle$ and defined by

$$\langle m, u\rangle = \sum_{S\subseteq N} m(S)\cdot u(S)$$

will be used. Indeed, it is a scalar product on the Eucledian space $\mathbb{R}^{\mathcal{P}(N)}$. Note that the function m can be an imset as well, it will be often a multiset.

Chapter 3

Graphical methods

Graphs whose nodes correspond to random variables are traditional tools for description of CI structures. One can distinguish three classic approaches: using *undirected graphs*, using *acyclic directed graphs* and using *chain graphs*. This chapter is an overview of graphical methods of description of CI structures with the main emphasis put on theoretical questions mentioned in Section 1.1. Both classic and advanced approaches are included. Note that elementary graphical concepts are introduced in Section 10.3.

3.1 Undirected graphs

Graphical models based on undirected graphs are also known as *Markov networks* [78]. Given an undirected graph G over N one says that a disjoint triplet $\langle A,B|C\rangle \in \mathcal{T}(N)$ is *represented in* G and writes $A \perp\!\!\!\perp B \,|\, C\ [G]$ if every route (equivalently every path) in G between a node in A and a node in B contains a node in C, that is C separates between A and B in G. For illustration see Figure 3.1. Thus, every undirected graph G over N induces a formal independence model over N by means of the *separation criterion* (*for undirected graphs*):

$$\mathcal{M}_G = \{\, \langle A,B|C\rangle \in \mathcal{T}(N)\,;\ A \perp\!\!\!\perp B \,|\, C\ [G]\,\}\,.$$

Let us call every independence model obtained in this way an *UG model*. These models were characterized in [77] in terms of a finite number of formal properties:

1. triviality $A \perp\!\!\!\perp \emptyset \,|\, C\ [G]$,
2. symmetry $A \perp\!\!\!\perp B \,|\, C\ [G]$ implies $B \perp\!\!\!\perp A \,|\, C\ [G]$,
3. decomposition $A \perp\!\!\!\perp BD \,|\, C\ [G]$ implies $A \perp\!\!\!\perp D \,|\, C\ [G]$,
4. strong union $A \perp\!\!\!\perp B \,|\, C\ [G]$ implies $A \perp\!\!\!\perp B \,|\, DC\ [G]$,
5. intersection $A \perp\!\!\!\perp B \,|\, DC\ [G]$ and $A \perp\!\!\!\perp D \,|\, BC\ [G]$ implies $A \perp\!\!\!\perp BD \,|\, C\ [G]$,
6. transitivity $A \perp\!\!\!\perp B \,|\, C\ [G]$ implies $A \perp\!\!\!\perp \{d\} \,|\, C\ [G]$ or $\{d\} \perp\!\!\!\perp B \,|\, C\ [G]$.

This axiomatic characterization implies that every UG model is a graphoid satisfying the composition property.

Remark 3.1 Let me note that the above mentioned separation criterion was a result of certain development. Theory of Markov fields stems from statistical physics [73] where undirected graphs were used to model geometric arrangement in space. Several types of

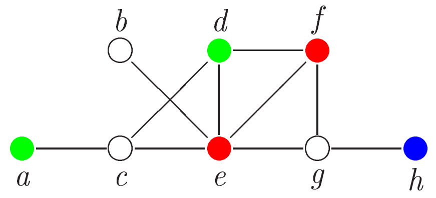


Figure 3.1: The set $C = \{e, f\}$ separates between $A = \{a, d\}$ and $B = \{h\}$.

Markov conditions were later introduced [52] in order to associate these graphs and probabilistic CI structures. The original 'pairwise Markov property' was strengthened to the 'local Markov property' and this was strengthened to the 'global Markov property'. The latter property corresponds to the separation criterion and appeared to be the strongest possible Markov condition in a certain sense (see Remark 3.2). The Markov conditions differ in general (e.g. [60]) but coincide in case of positive measures [52].

Note that similar story was observed in case of acyclic directed graphs and chain graphs (for an overview see Chapter 3.2 of [53]) and has been repeated recently in case of advanced graphical models (see Section 3.5). However, in this work attention is paid only to the result of this development, that is to graphical criteria which correspond to respective global Markov conditions. △

A probability measure P over N is *Markovian with respect to an undirected graph* G over N if

$$A \perp\!\!\!\perp B \,|\, C \; [G] \quad \text{implies} \quad A \perp\!\!\!\perp B \,|\, C \; [P] \quad \text{for every } \langle A, B|C\rangle \in \mathcal{T}(N)$$

and *perfectly Markovian* if the converse implication holds as well. It was shown in [33] (Theorem 11) that a perfectly Markovian discrete probability measure exists for every undirected graph over N. In other words, every UG model is a (probabilistic) CI model and faithfulness of UG models (in sense of Section 1.1) is ensured.

Remark 3.2 This is to explain certain habitual terminology used sometimes in literature. The remark holds also in case of acyclic directed graphs and chain graphs (see Sections 3.2, 3.3, 3.5.4, 3.5.5). The existence of a perfectly Markovian measure which belongs to a class of measures Ψ implies the following weaker result. Whenever a disjoint triplet t is not represented in a graph G then there exists a measure $P \in \Psi$ which is Markovian with respect to G and t is <u>not</u> valid conditional independence statement with respect to P. Some authors [25, 44, 39] say then that *the class of measures* Ψ *is perfect* with respect to G. Thus, Theorem 2.3 from [26] says that the class of CG measures with prescibed layout od discrete and continuous variables is perfect with respect to every undirected graph. However, the claim about perfectness of a class Ψ is also referred in literature [31, 112, 54] as the *completeness* (*of the respective graphical criterion* relative to Ψ) since it says that the criterion cannot be strengthened within Ψ any more (unlike the pairwise and local Markov conditions in case of the class of positive measures - see Remark 3.1). By *strong completeness* is then meant the existence of a perfectly Markovian measure over N with prescribed non-trivial sample space X_N [69, 54]. △

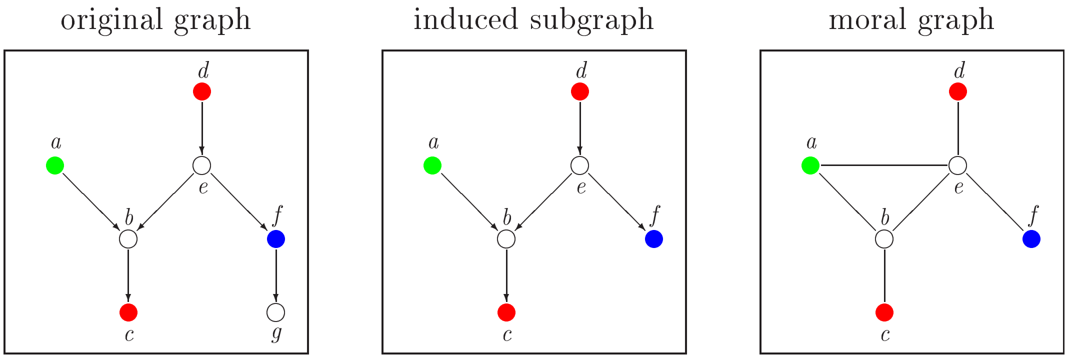


Figure 3.2: Testing $\langle a, f \,|\, \{c, d\} \rangle$ according to the moralization criterion.

One can say that two undirected graphs G and H over N are *Markov equivalent* if the classes of Markovian measures with respect to G and H coincide. The result about the existence of perfectly Markovian measures implies that it occurs iff $\mathcal{M}_G = \mathcal{M}_H$. Moreover, the observation that $a - b$ in G iff $\neg(a \perp\!\!\!\perp b \,|\, N \setminus ab\ [G]\,)$ implies that $\mathcal{M}_G = \mathcal{M}_H$ iff $G = H$. Thus, the equivalence question (in sense of Section 1.1) has a simple solution in case of undirected graphs.

Remark 3.3 A marginaly continuous probability measure over N is called *factorizable with respect to an undirected graph* G over N if it factorizes after the class (see p. 22) of its cliques. It is known that every factorizable measure is Markovian [52], the converse is true for positive measures [43] but not for all (discrete) measures [60].

One can say that two graphs are *factorizably equivalent* if the corresponding classes of factorizable measures coincide. However, this notion is not very sensible in the frameowrk of undirected graphs since it reduces to identity of graphs in this case (one can use the same reasoning like in case of Markov equivalence). $\triangle$

The restriction of an UG model to a set $\emptyset \neq T \subseteq N$ is an UG model [111]. However, the corresponding *marginal graph* G^T differs from the usual induced subgraph G_T. For $a, b \in T$ one has $a - b$ in G^T iff there exists a path in G between a and b consisting of nodes of $\{a, b\} \cup (N \setminus T)$.

3.2 Acyclic directed graphs

These graphical models are also known under name *Bayesian networks* [78]. Note that the majority of authors became accustomed to the phrase 'directed acyclic graphs' which is not accurate from grammatical point of view (since the adjectives do not commute). The respective abbreviation DAG is therefore commonly used.

Two basic criteria to determine whether a triplet $\langle A, B|C \rangle \in \mathcal{T}(N)$ is represented in an acyclic directed graph G were developed. Lauritzen *et. al.* [52] proposed the *moralization criterion* while the group around Pearl [30] used the *d-separation criterion* (d means 'directional').

The moralization criterion has three stages. First, one takes the set $T = an_G(ABC)$ and considers the induced subgraph G_T. Second, G_T is changed into its *moral graph* H, that is the underlying graph of the graph K (with mixed edges) over T which is obtained from the graph G_T by adding a line $a - b$ in K whenever there exists $c \in T$ having both a and b as parents in G_T. The name 'moral graph' was motivated by the fact that the nodes having a common child are 'married'. The third step is to decide whether C separates between A and B in H. If yes, one says that $\langle A, B|C\rangle$ is *represented in G according to the moralization criterion*. For illustration see Figure 3.2 where the tested triplet is not represented in the original graph.

To formulate d-separation criterion one needs some auxiliary concepts as well. Let $\omega : c_1, \ldots, c_n$, $n \geq 1$ be a route in a directed graph G. By a *collider node* with respect to ω is understood every node c_i, $1 < i < n$ such that $c_{i-1} \to c_i \leftarrow c_{i+1}$ in ω. One says that ω is *active with respect to a set $C \subseteq N$* if

- every collider node with respect to ω belongs to $an_G(C)$,
- every other node of ω is outside C.

Route which is not active with respect to C is *blocked by C*. A triplet $\langle A, B|C\rangle$ is *represented in G according to the d-separation criterion* if every route (equivalently every path) in G from A to B is blocked by C. For illustration of d-separation criterion see Figure 3.3. It was shown in [52] that the moralization and d-separation criteria for acyclic directed graphs are equivalent. Note that the moralization criterion is effective if $\langle A, B|C\rangle$ is represented in G while d-separation is suitable for the opposite case. The third possible equivalent criterion (a compromise between those two criteria) appeared in [58].

One writes $A \perp\!\!\!\perp B \,|\, C\ [G]$ whenever $\langle A, B|C\rangle \in \mathcal{T}(N)$ is represented in an acyclic directed graph G according to one of the criteria. Thus, every acyclic directed graph G induces a formal independence model

$$\mathcal{M}_G = \{\, \langle A, B|C\rangle \in \mathcal{T}(N)\, ;\ A \perp\!\!\!\perp B \,|\, C\ [G]\, \}\,.$$

Following common practice let me call every independence model obtained in this way a *DAG model*. These models were not characterized like UG models, just several formal properties of DAG models were given in [78]. They imply that every DAG model is a graphoid satisfying the composition property. The problem of axiomatic characterization of DAG models seems to be more complicated - see Remark 3.5.

The definition of *Markovian* and *perfectly Markovian measure with respect to an acyclic directed graph* is analogous to the case of undirected graphs. It was shown in [31] that a perfectly Markovian discrete probability measure exists for every acyclic directed graph. Hence, the existence of a perfectly Markovian measure with prescribed non-trivial discrete sample space was derived [69]. Thus DAG models are also probabilistic CI models.

Two acyclic directed graphs are *Markov equivalent* if their classes of Markovian measures coincide. The problem of graphical characterization of this equivalence was probably first solved in [120] but the result can be found also in other publications [5, 94] and follows from an analogous result for chain graphs [25] as well. Let's call by an *immorality* in an acyclic directed graph G every induced subgraph of G for a set $T = \{a, b, c\}$ such that $a \to c$ in G, $b \to c$ in G and $[a, b]$ is not an edge in G. Two acyclic directed graphs are Markov equivalent iff they have the same underlying graph and the same immoralities.

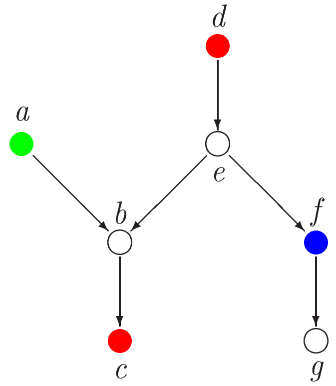


Figure 3.3: The path $a \to b \leftarrow e \to f$ is active with respect to $C = \{c, d\}$.

Note that the word 'immorality' has the same justification like 'moralization criterion'; other authors used various alternative names like 'unshield colliders', 'v-structures' and 'uncoupled head-to-head nodes'.

However, the question of choice of a suitable representative of equivalence class has no natural solution in the framework of acyclic directed graphs. There is no distinguished representative in every class of equivalent graphs. Thus, hybrid graphs like *essential graphs* [5] or (*completed*) *pattern* [120] were used in literature to represent uniquely equivalence classes of acyclic directed graphs. The problem of estimation of DAG models from data, more exactly estimation of an essential graph on basis of the induced independence model (which could be obtained as a result of statistical tests based on data) was treated in [121, 68, 16].

Remark 3.4 It is a speciality of the case of acyclic directed graphs that for marginally continuous probability measures the respective concept of (*recursively*) *factorizable* measure coincides with the concept of Markovian measure [52]. Another specific feature of this case is that an analogue of the 'local Markov property' is equivalent to the 'global Markov property' [52]. This fact can be also derived from the result in [119] saying that the least semi-graphoid containing the following collection of independence statements

$$a_i \perp\!\!\!\perp \{a_1, \ldots, a_{i-1}\} \setminus pa_G(a_i) \,|\, pa_G(a_i) \quad \text{for } i = 1, \ldots, n$$

where $a_1, \ldots, a_n$, $n \geq 1$ is an ordering of nodes of G consonant with direction of arrows, is nothing but the induced model $\mathcal{M}_G$. The above collection of independence statement is called often an *causal input list* or 'stratified protocol'. △

Unlike the case of UG models the restriction of a DAG model need not be a DAG model as the following example shows.

Example 3.1 There exists a DAG model over $N = \{a, b, c, d, e\}$ whose restriction to $T = \{a, b, c, d\}$ is not a DAG model over T. Consider the independence model induced by the graph in Figure 3.4. It was shown in [21] (Lemma 5.1) that its restriction to T is not a DAG model. This unpleasant property of DAG models probably motivated attempts to extend study to DAG models with hidden variables, that is restriction of DAG models - see Section 3.5.7. ♢

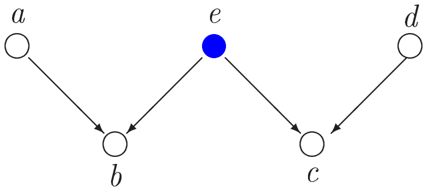


Figure 3.4: Acyclic directed graph with hidden variable e.

Remark 3.5 An indirect consequence of the preceding example is that DAG models cannot be characterized in terms of properties of 'semi-graphoid' type (unlike UG models). To evidence it take the following perspective. Let us call by a *relevance statement* over N any independence or dependence statement which corresponds to a disjoint triplet over N. By a *full-consistent set* of relevance statements is understood a set of these statements over N such that for every $\langle A, B|C\rangle \in \mathcal{T}(N)$ exclusively either the corresponding independence statement or the corresponding dependence statements belongs to the set. Every independence model can be easily identified with a set of relevance statements of this kind: every 'missing' independence statement is automatically regarded as a dependence statement.

Consider special formal properties of full-consistent sets of relevance statements where a finite conjunction of relevance statements (which may be empty) implies another relevance statement. These formal properties are general enough since every requirement that a finite conjuction implies a finite disjunction (of relevance statements) can be equivalently described in this way (because of full-consistency). To be more specific, I have in mind properties expressed in the form of 'syntactic inference rules' (e.g. semi-graphoid properties on p. 15 or the properties characterizing UG models on p. 37). The interpretation for a given set of variables N is this: a rule of this sort is applicable only when the substitution of subsets of N for capital letters $A, B, C, \ldots$ (resp. elements of N for lower case letters like d; the symbol $\emptyset$ has specific meaning) leads to relevance statements over N which correspond to disjoint triplets over N (for all involved statements!).

Basic observation is that the restriction of any full-consistent set of relevance statements over N satisfying a formal property of this kind to a set $\emptyset \neq T \subseteq N$ is a full-constistent set of relevance statements over T satisfying the same formal property. This hold for any (even infinite) collection of formal properties of this type. Therefore, because of Example 3.1, DAG models can never be characterized by means of any collection of these properties.

However, perhaps DAG models can be characterized by means of more general formal properties where 'elementary clauses' are more complex and represent sets of relevance statements. For example, one symbol $A \top\!\!\!\top B \mid C^{\uparrow}$ could represent a class of all dependence statements $A \top\!\!\!\top B \mid D$ where $C \subseteq D \subseteq N \setminus AB$. According to e-mail communication by T. Verma such a characterization is possible but very complex. $\triangle$

3.3 Classic chain graphs

A *chain graph* is a hybrid graph without directed cycles or equivalently a hybrid graph which admits a chain (see Section 10.3, p. 154). The class of chain graphs was introduced

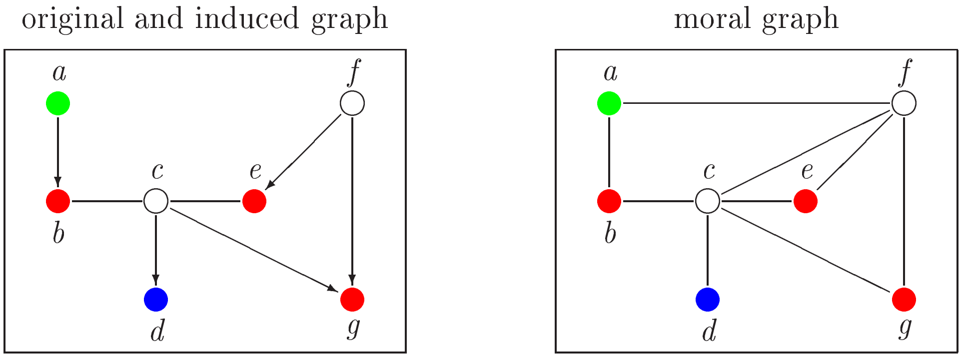


Figure 3.5: Testing $\langle a, d \,|\, \{b, e, g\} \rangle$ according to the moralization criterion for chain graphs.

by Lauritzen and Wermuth in middle eighties in the report [47] which became later a basis of a journal paper [50].

Classic interpretation of chain graphs is based on the *moralization criterion for chain graphs* established by Lauritzen [51] and Frydenberg [25]. The main distinction between the moralization criterion for chain graphs and for acyclic directed graphs (see p. 40) is a more general definition of the moral graph in case of chain graphs. Supposing G_T is a hybrid graph over $\emptyset \neq T \subseteq N$ one defines a graph K with mixed edges over T by adding lines $a — b$ in K whenever there exist $c, d \in T$ belonging to the same connectivity component of G_T (possibly $c = d$) such that $a \to c$ in G_T and $b \to d$ in G_T. The *moral graph* H of G_T is then the underlying graph of K. A triplet $\langle A, B|C \rangle \in \mathcal{T}(N)$ *represented in a chain graph* G *over* N *according to the moralization criterion* if C separates between A and B in the moral graph G_T where $T = \mathrm{an}_G(ABC)$. For illustration see Figure 3.5.

An equivalent *c-separation criterion* (c stands for 'chain') which generalizes the d-separation criterion for acyclic directed graphs was introduced in [11]. This criterion was later simplified as follows [113]. By a *section* of a route $\omega : c_1, \ldots, c_n$, $n \geq 1$ in a hybrid graph G is understood a maximal undirected subroute $c_i — \ldots — c_j$ of ω (that is either $i = 1$ or $[c_{i-1}, c_i]$ is not a line, analogously for j). By a *collider section* of ω is understood a section $c_i, \ldots, c_j$, $1 < i \leq j < n$ such that $c_{i-1} \to c_i — \ldots — c_j \leftarrow c_{j+1}$ in ω. A route ω is *superactive with respect to a set* $C \subseteq N$ if

- every collider section of ω contains a node of C,
- every other section of ω is outside C.

Route which is not superactive with respect to C is *blocked by* C. A triplet $\langle A, B|C \rangle \in \mathcal{T}(N)$ is *represented in* G *according to the c-separation criterion* if every route in G from A to B is blocked by C. The equivalence of the c-separation criterion and the moralization criterion was shown in [112] (Consequence 1). One writes $A \perp\!\!\!\perp B \,|\, C \;[G]$ if $\langle A, B|C \rangle$ is represented in a chain graph G according to one of these criteria. The induced formal independence model is then

$$\mathcal{M}_G = \{ \langle A, B|C \rangle \in \mathcal{T}(N) \, ; \; A \perp\!\!\!\perp B \,|\, C \; [G] \} \, .$$

Thus, the class of *CG models* was introduced. Since c-separation generalizes both the separation criterion for undirected graphs and the d-separation criterion for acyclic directed graphs every UG model and every DAG model is a CG model (for illustration see

Figure 3.6 on p. 46). Every CG model is a graphoid satisfying the composition property [112]. Note that Example 3.1 can serve also as an example that the restriction of a CG model need not be a CG model. Therefore, one can repeat the arguments from Remark 3.5 showing that CG models cannot be characterized by means of formal properties of 'semi-graphoid' type.

REMARK 3.6 Unlike the case of undirected and acyclic directed graphs blocking of all routes required in c-separation criterion is not equivalent to blocking of all paths. Consider the chain graph G in the left-hand picture of Figure 3.5. The only path between $A = \{a\}$ and $B = \{d\}$ is $a \to b - c \to d$ which is blocked by $C = \{b, e, g\}$. However, the route $a \to b - c - e \leftarrow f \to g \leftarrow c \to d$ is active with respect to C. Thus, one has $\neg\{\, a \perp\!\!\!\perp d \,|\, \{b, e, g\}\ [G]\,\}$. Despite the fact that the class of all routes between two sets could be infinite c-separation is finitely implementable for another reason - see Section 5 in [113].

Note that the above mentioned phenomenon was the main reason why the original version of c-separation [11] looked akward. It was formulated for a special finite class of routes called 'trails' and complicated by subsequent inevitable intricacies. $\triangle$

A probability measure P over N is *Markovian with respect to a chain graph* G over N if

$$A \perp\!\!\!\perp B \,|\, C\ [G] \quad \text{implies} \quad A \perp\!\!\!\perp B \,|\, C\ [P] \quad \text{for every } \langle A, B|C\rangle \in \mathcal{T}(N)$$

and *perfectly Markovian* if the converse implication holds as well. The main result of [112] says that a perfectly Markovian positive discrete probability measure exists for every chain graph. In particular, faithfulness of CG models (in sense of Section 1.1) is ensured as well.

Two chain graphs over N are *Markov equivalent* if their classes of Markovian measures coincide. These graphs were characterized in graphical terms by Frydenberg [25]. By a *complex* in a hybrid graph G over N is understood every induced subgraph of G for a set $T = \{d_1, \ldots, d_k\}$, $k \geq 3$ such that $d_1 \to d_2$, $d_i - d_{i+1}$ for $i = 2, \ldots, k-2$, $d_{k-1} \leftarrow d_k$ in G and no additional edge between (distinct) nodes of $\{d_1, \ldots, d_k\}$ exists in G. Two chain graphs over N are Markov equivalent iff they have the same underlying graph and the same complexes.

However, unlike the case of acyclic directed graphs the advanced question of representation of Markov equivalence classes has an elegant solution. Every class of Markov equivalent chain graphs contains a naturally distinguished member! Given two chain graphs G and H over N having the same underlying graph one says that G is *larger* than H if every arrow in G is an arrow in H with the same direction. Frydenberg [25] showed that every class of Markov equivalent chain graphs contains a graph which is larger than every other chain graph within the class (that is, it has the greatest number of lines). This distinguished graph is named the *largest chain graph* of the equivalence class. An elegant graphical characterization of those graphs which are the largest chain graphs was presented in [122]. The paper also describes an algorithm for transformation of every chain graph into the respective largest chain graph. An alternative algorithm is presented in [110] where the problem of finding the largest chain graph on basis of induced formal independence model is solved. This could be utilized for learning CG models.

REMARK 3.7 Lauritzen [53], Section 3.2.3 defined the concept of (marginally continuous) *factorizable measure* with respect to a chain graph. Like in case of undirected graphs every factorizable measure is Markovian and the converse is true for positive measures [25].

Well, having fixed the sample space $(\mathsf{X}_N, \mathcal{X}_N)$ where $\mathcal{X}_i$ is non-trivial for each $i \in N$ one can say that two chain graphs over N are *factorizably equivalent* if the corresponding classes of factorizable measures (on X_N) coincide. However, unlike the case of undirected and acyclic directed graphs the hypothesis that this equivalence coincides with Markov equivalence has not been confirmed untill now - see Question 3. $\triangle$

3.4 Within classic graphical models

This section deals with some methods of description of probabilistic structures which in fact fall within the scope of classic graphical models.

3.4.1 Decomposable models

Very important class of undirected graphs is the class of triangulated graphs. An undirected graph G is called *triangulated* or *chordal* if every cycle $a_1, \ldots, a_n$, $n \geq 5$ in G has a 'chord', that is a line between nodes of $\{a_1, \ldots, a_{n-1}\}$ different from the lines of the cycle. There are several equivalent definitions of a chordal graph; one of them says that the graph can be decomposed in a certain way into its cliques (see [53], Proposition 2.5) which motivated other alternative name *decomposable graph* (see p. 141). For this reason UG models induced by triangulated graphs are named *decomposable models* [78]. Another equivalent definition (see [53], Proposition 2.17) is that all cliques of the graph can be ordered into a sequence $C_1, \ldots, C_m$, $m \geq 1$ satisfying the *running intersection property*

$$\forall\, 2 \leq i \leq m \;\; \exists\, 1 \leq k < i \qquad S_i \equiv C_i \cap (\bigcup_{j<i} C_j) \subseteq C_k \,. \tag{3.1}$$

Note that the phrase *acyclic hypergraph* is sometimes used in literature for a class of sets admitting an ordering of this type. The sets S_i are then called *separators* since S_i separates the 'history' $H_i = \bigcup_{j<i} C_j \setminus S_i$ from the 'residuals' $R_i = C_i \setminus S_i$ in the graph for every $i = 2, \ldots, m$ (see [53], p. 15). Actually, separators and their multiplicity (i.e. the number of indices $i \in \{2, \ldots, m\}$ for which $S = S_i$) do not depend on the choice of the sequence satisfying the running intersection property (see Lemma 7.2 in Section 7.2.2 or [48]). Note that the running intersection property has a close connection to marginal problem within the framework of probabilistic expert systems [37, 40].

One can show (by repeated application of Proposition 3.17 from [53]) that a marginally continuous probability measure P is Markovian with respect to a triangulated undirected graph G over N iff its marginal densities f_A, $A \subseteq N$ satisfy the following product formula

$$f_N(x) \;=\; \frac{\prod_{C \in \mathcal{C}} f_C(x_C)}{\prod_{S \in \mathcal{S}} f_S(x_S)^{w(S)}} \tag{3.2}$$

where $\mathcal{C}$ is the class of cliques, $\mathcal{S}$ the class of separators and $w(S)$ denotes the multiplicity of a separator. Thus, to store a discrete measure P in memory of a computer one needs to store only its clique marginals.

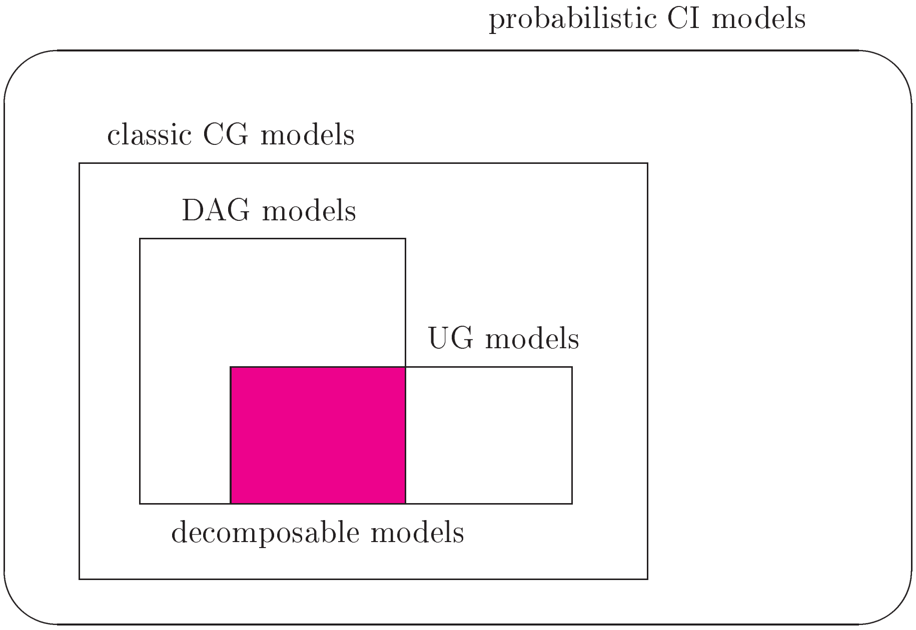


Figure 3.6: Relationships among classic graphical models.

Further related equivalent definition of a triangulated graph is the existence of a *junction tree* ([17], Theorem 4.6) of its cliques (and separators). Junction trees then form a mathematical basis for miscellaneous effective computational methods [89, 17] which originate from the *local computation method* [49]. Thus, decomposable models are very suitable from the point of view of implementation (see Section 1.1, p. 9).

Perhaps another characterization of decomposable models is worthy of mentioning. Decomposable models are just those formal independence models which are simutaneously UG models and DAG models. For illustration see Figure 3.6. A characterization of decomposable models in terms of a finite number of formal properties is given in [15]. It implies that decomposable models are closed under restriction.

3.4.2 Recursive causal graphs

The concept of *recursive causal graph* [41] seems to precede the concept of chain graph. It can be equivalently defined as a chain graph which admits a chain such that all its lines belong to the first block. Thus, both undirected and acyclic directed graphs are special cases of recursive causal graphs. The way of ascribing of an independence model to a recursive graph is consonant with the way used in case of classic chain graphs.

3.4.3 Lattice conditional independence models

Andersson and Perlman [4] came with an idea to describe probabilistic CI structures by finite lattices (of subsets of N). Given a ring $\mathcal{R}$ of subsets of N one says that a probability measure over N satisfies the *lattice conditional independence model* (= LCI model) induced by $\mathcal{R}$ if

$$\forall\, E,F \in \mathcal{R} \qquad (E \setminus F) \perp\!\!\!\perp (F \setminus E) \,|\, E \cap F \;[P]\,.$$

However, it was found later in [6] that LCI models coincide with DAG models induced by *transitive acyclic directed graphs* in which $a \rightarrow b$ and $b \rightarrow c$ implies $a \rightarrow c$. Thus, LCI models also fall within the scope of classic graphical models. Note that these models are advantageous from the point of view of learning. It was shown in [80] that an explicit formula for the maximum likelihood estimate exists even in case of 'non-monotone' pattern of missing data.

3.4.4 Bubble graphs

Shafer in Section 2.3 of [89] defined bubble graphs which are not graphs in standard sense mentioned in Section 10.3. A *bubble graph* over N is specified by an ordered decomposition $B_1, \ldots, B_n$, $n \geq 1$ of N into non-empty subsets called *bubbles* and by a collection of directed links which point to bubbles although they originate from single nodes taken from the preceding bubbles. Every graph of this type describes a class of probability measures over N which satisfy certain factorization formula.

One can associate a chain graph with every bubble graph as follows. Join nodes in each bubble by lines and replace any directed link from a node $a \in N$ to a bubble $B \subset N$ by the collection of arrows from a to every node of B. Then one can show easily that a probability measure over N satisfies the factorization formula corresponding to the bubble graph iff it factorizes with respect to the ascribed chain graph in sense of Remark 3.7. In particular, every bubble graph can be interpreted as a classic chain graph. On the other hand, every DAG model can be described by a bubble graph.

3.5 Advanced graphical models

Various types of graphs have been recently proposed in literature in order to describe probabilistic structures (possibly expressed in terms of structural equations for random variables). Some of these graphs can be viewed as tools for description of CI structures (although this may not be the original aim of respective authors). This section gives an overview of these graphical approaches. Note that the majority of formal independence models ascribed to these graphs are semi-graphoids satisfying the composition property.

3.5.1 General directed graphs

A natural way of generalization is to allow directed cycles. Spirtes, Glymour and Scheines (see Chapter 12 in [94]) mentioned possible use of general directed graphs for description of models allowing feedback. They proposed to use d-separation criterion (see p. 40) to ascribe a formal independence model to a directed graph (even allowing multiple edges). It was shown in [95] that even in case of general directed graphs, d-separation criterion is equivalent to the moralization criterion and the criteria are complete (in sense of Remark 3.2) relative to the class of non-degenerate Gaussian measures. Richardson [83] published a graphical characterization of Markov equivalent directed graphs. It is rather complex in comparison with the case of acyclic directed graphs (six independent conditions are involved).

3.5.2 Reciprocal graphs

Koster [44] introduced very general class of reciprocal graphs. A *reciprocal graph G* over N is a graph with mixed edges over N (multiple edges are allowed) such that there is no arrow in G between nodes belonging to the same connectivity component of G. Thus, every classic chain graph is a reciprocal graph and every (general) directed graph is a reciprocal graph as well. The moralization criterion for chain graphs (see p. 43) can be used to ascribe a formal independence model to every reciprocal graph. Note that in case of directed graphs it reduces to the moralization criterion treated by Spirtes [95].

Thus, consistency of reciprocal graphs (see p. 8) is ensured. The question of their faithfulness remains open but the related question of existence of a perfect class of measures (see Remark 3.2) was answered positively. Koster's aim was to apply these graphs to simultaneous equation systems (*LISREL models* [38]). A certain reciprocal graph can be ascribed to every LISREL model so that the class of non-degenerate Gaussian measures satisfying the LISREL model is perfect with respect to the assigned reciprocal graph (in sense of Remark 3.2).

3.5.3 Joint-response chain graphs

Cox and Wermuth [18] generalized the concept of chain graph by introducing two additional types of edges. A *joint-response chain graph G* is a chain graph (in sense of Section 10.3) in which, however, every arrow is either a *solid arrow* or a *dashed arrow* and every line is either a *solid line* or a *dashed line*. Thus, even four types of edges are allowed in a graph of this type. Moreover, two technical conditions are required for every connectivity component C of a joint-response chain graph, namely

- all lines within C are of the same type (i.e. either solid or dashed),
- all arrows directed to nodes of C are of the same type.

The interpretation of these graphs (see [18], Section 2.3) is more likely in terms of what is known as pairwise Markov property (see Remark 3.1). Namely, the absence of an edge between nodes a and b is interpreted as a CI statement $a \perp\!\!\!\perp b \,|\, C$ where the set $C \subseteq N \setminus ab$ depends on the type of 'absent' edge. Note that technical conditions above allow one to deduce implicitly what is the type of the 'absent' edge.

The resulting interpretation of joint-response chain graphs with solid lines and arrows only is then in concordance with the original interpretation of chain graphs (see Section 3.3) so that they generalize classic chain graphs. An analogue of global Markov property was established in two other special cases (see Sections 3.5.4 and 3.5.5).

Remark 3.8 Following an analogy with development of classic graphical models (see Remark 3.1) observe that in order to determine the strongest possible Markov condition (on basis of pairwise Markov condition) one needs to know what is the respective class of probability measures. This class of measures was traditionally closely connected with the considered class of graphs. It was the class of positive measures in case of CG models and UG models (which are called *concentration graphs* by Cox and Wermuth [18]), the class of Gaussian measures in case of covariance graphs (see Section 3.5.4) and the class of all probability measures in case of acyclic directed graphs (see Remark 3.4). Since Cox and Wermuth did not explicate the class of measures which should correspond to general joint-response chain graphs one cannot derive 'automatically' the respective global Markov condition. Well, I can only speculate that they have probably in mind the class of non-degenerate Gaussian measures. In particular, global Markov condition for general joint-response chain graph was not established so far (see Section 2.4.5 of [18]) and the question of consistency (see Section 1.1) remains to be solved. $\triangle$

Thus, other theoretical questions mentioned in Section 1.1 do not have sense for joint-response chain graphs untill consistency is established for them.

3.5.4 Covariance graphs

However, consistency was ensured in a special case of undirected graphs made of dashed lines and named *covariance graphs*. Kauermann [39] formulated a global Markov property for covariance graphs which is equivalent to the above mentioned condition of Cox and Wermuth for every probability measure whose induced independence model satisfies the composition property. A triplet $\langle A, B|C\rangle \in \mathcal{T}(N)$ is represented in a covariance graph G if $N \setminus ABC$ separates between A and B. Thus, every covariance graph induces a graphoid satisfying the composition property. Kauerman [39] also showed that the class of Gaussian measures is perfect (in sense of Remark 3.2) with respect to every covariance graph. In particular, his criterion is the strongest possible one for the considered class of measures.

3.5.5 Alternative chain graphs

Another class of joint-response chain graphs for which the global Markov property was established are chain graphs with solid lines and dashed arrows only. Lead by a specific way of parametrization of non-degenerate Gaussian measures Andersson, Madigan and Perlman [7] introduced 'alternative Markov property' (AMP) for chain graphs. Their *alternative chain graphs* are chain graphs in sense of Section 10.3 but their interpretation is different from the interpretation of classic chain graphs (see Section 3.3) so that they correspond to the above mentioned joint-response chain graphs (see [7], §1 for details).

The corresponding *augmentation criterion* is analogous to the moralization criterion for classic chain graphs but it is more complex. Testing whether a triplet $\langle A, B|C\rangle \in \mathcal{T}(N)$ is represented in an alternative chain graph G over N consists of 3 steps. The first step is a specific restriction of G to an 'extended graph" over a set $T \subseteq N$ involving ABC (= an analogue of the induced graph G_T in the moralization criterion). The second step is transformation of the extended graph into an undirected 'augmented graph'. This is done by adding some edges and taking the underlying graph (= an analogue of the moralization procedure). The third step is testing whether C separates between A and B in the augmented graph.

Like in case of classic chain graphs an equivalent *p-separation criterion* (p stands for 'path') was introduced [54]. The main result of [54] is the existence of perfectly Markovian non-degenerate Gaussian measure for every alternative chain graph. Thus, the faithfulness of these models is ensured. Moreover, Markov equivalent alternative chain graphs were characterized in graphical terms as well [7]. Every class of Markov quivalence can be represented by respective *essential graph* (for details see [7], §7).

3.5.6 Annotated graphs

Paz proposed in [75] a special fast implementation (modification) of the moralization criterion for acyclic directed graphs. In the preparatory stage of that procedure the original directed graph G is changed into its moral graph and every its immoralitity $a \rightarrow c \leftarrow b$ in G is recorded by annotation of the edge $a - b$ of the moral graph by the set C of all descendants of c in G. Thus, the original graph over N is changed into an *annotated graph* over N, that is an undirected graph supplemented by a collection of 'elements' $[\,\{a,b\}\,|\,C\,]$ where $a,b \in N$, $a \neq b$ and $\emptyset \neq C \subseteq N \setminus ab$ which represents annotated edges. Testing whether a triplet $\langle A,B|C\rangle \in \mathcal{T}(N)$ is represented in G then consists in application of a special *membership algorithm* for annotated graphs. This algorithm consists in successive restriction of the graph, removal of (respective) annotated edges and final checking whether C separates between A and B in the resulting graph. All this procedure is equivalent to the moralization algorithm [75].

The point is that this approach has much wider applicability. In [76] the class of *regular annotated graphs* was introduced together with the corresponding general membership algorithm. Formal independence model induced in this way by a regular annotated graph was shown to be a graphoid. Regular annotated graph can serve as a condensed record for the least graphoid containing the unions of UG models (= their graphoid closure). Given a sequence of undirected graphs $G_i = (N_i, \mathcal{L}_i)$, $i = 1,\ldots,k$ ($k \geq 1$) such that $N_i \subseteq N_{i+1}$ and $\mathcal{L}_i \subseteq \mathcal{L}_{i+1}$ for $i = 1,\ldots,k-1$ a specific *annotation algorithm* described in [76] allows one to construct a regular annotated graph over $N = N_k$ such that the independence model induced by it is just the graphoid closure of all UG models induced by G_i, $i = 1,\ldots,k$. Since every (classic) CG model can be obtained in this way regular annotated graphs generalize classic graphical models.

3.5.7 Hidden variables

Example 3.1 shows that the restriction of a DAG model need not be a DAG model. This maybe led to an idea to describe restrictions of DAG models by means of graphical diagrams. These models are usually named the *models with hidded variables* since except 'observed' variables in N one anticipates other 'unobserved' hidden variables K and a DAG model over NK.

Geiger, Paz and Pearl [34] introduced the concept of *embedded Bayesian network*. It is a graph over (observed variables) N allowing both directed and bidirected edges (without multiple edges) such that purely directed cycles (that is directed cycles made exclusively of arrows) are not present in the graph. A generalized d-separation criterion was used to ascribe a formal independence model over N to a graph of this type. It is mentioned in [34] that one can always find a DAG model over a set $M \supseteq N$ whose restriction to N is the ascribed independence model. Moreover, according to Pearl's oral communication, Verma showed that the restriction of every DAG model can be described in this way. Note that faithfulness of embedded Bayesian networks is an easy consequence of faithfullness of DAG models and the above mentioned claims.

However, there are other graphical methods for description of models with hidden variables. For example, *summary graphs* from [18], §8.5 or ancestral graphs mentioned below.

3.5.8 Ancestral graphs

Motivated by the need to describe classes of Markov equivalent (general) directed graphs Richardson [84] proposed to use special graphical objects called *partial ancestral graphs* (PAGs) for this purpose. PAGs are graphs whose edges have 3 possible endings for both end-nodes and where the endings of different edges near a common end-node may be connected by two possible 'connections'. Every mark of this type in a PAG express certain graphical property shared by all graphs within the Markov equivalence class, for example that a node is not an ancestor of another node in all equivalent graphs.

The idea of graphical representation of common features of classes of Markov equivalent graphs was later substantially simplified. In a recent paper [85] Richardson and Spirtes introduced *ancestral graphs*. These graphs admit three types of edges, namely lines, arrows and bi-directed edges (e.g. $a \leftrightarrow b$) and satisfy some additional requirements. These requirements imply that multiple edges and loops are not present in ancestral graphs. A formal independence model over N is ascribed to an ancestral graph over N by means of the *m-separation criterion* which generalizes the d-separation criterion for acyclic directed graphs.

Additional standardization of ancestral graphs is suitable. *Maximal ancestral graph* (MAG) is an ancestral graph G such that $[a, b]$ is an edge in G iff $\neg\{\, a \perp\!\!\!\perp b \,|\, C\ [G]\,\}$ for every $C \subseteq N \setminus ab$. MAGs exhibit some elegant mathematical properties. One can define graphical operation of marginalizing and conditioning of MAGs which corresponds to the respective operation with induced formal independence models (c.f. Section 8.2.1). Edges of a MAG G correspond to single real parameters in a certain parametrization of the class of non-degenerate Gaussian measures which are Markovian with respect to G. Moreover, there exists a perfectly Markovian Gaussian measure with respect to G. Thus, the question of faithfulness (see Section 1.1) has positive solution in this framework. Note that MAG models involve both UG models and DAG models and coincide with the class of models induced by summary graphs - see §9.3.1 in [85].

3.5.9 MC graphs

Koster [45] introduced a certain class of graphs which admit the same three types of edges as ancestral graphs. However, in these graphs, called *MC graphs*, multiple edges and some loops are allowed. The abbreviation MC means that graphical operations of 'marginalizing and conditioning' can be applied to these graphs (like in case of MAGs). However, unlike m-separation the respective separation criterion for MC graphs requires blocking of all routes (like in the c-separation criterion for classic chain graphs - c.f. Remark 3.6). As mentioned in §9.2 of [85] the separation criterion for MC graphs generalizes m-separation criterion. Thus, the class of formal independence models induced by MC graphs involves MAG models. On the other hand, although MC graphs include chain graphs the respective separation criterion in case of chain graphs differs both from the c-separation criterion and from the p-separation criterion.

3.6 Incompleteness of graphical approaches

Let me raise the question how many probabilistic CI models can be described by graphs (cf. the question of completeness in Section 1.1, p. 8). Expressiveness of graphical methods varies. For example, in case $|N| = 3$ one has 8 UG models and 11 DAG models (= CG models). But in case $|N| = 4$ one has 64 UG models, 185 DAG models and 200 CG models, while in case $|N| = 5$ there exist 1024 UG models, 8782 DAG models [5] and 11519 CG models [122]. However, this is not enough for description of CI structures induced by discrete probability measures. Well, in case $|N| = 3$ one has 22 discrete CI models but in case $|N| = 4$ already 18300 CI models! [107] So, there is a tremendeous gap between the number of classic graphical models and the number of discrete probabilistic CI structures in case $|N| = 4$ and this gap increases with $|N|$. In particular, classic graphical models cannot describe all CI structures.

The reader may object that a sufficiently wide class of graphs could possibly cure the problem. Let me give an argument against it. Having fixed a class of graphs over N in which only finitely many types of edges are allowed the number of these graphs is bounded by the cardinality of the power set of the set of possible edges which grows polynomially with $n = |N|$. On the contrary, as shown in Consequence 2.6 on p. 33 the number of discrete probabilistic CI structures grows with n at least as rapidly as the power set of power of n.

Thus, in my opinion, one can hardly achieve completeness of a graphical approach (see p. 8) relative to the class of discrete measures and this may result in serious methodological errors (see Section 1.1 p. 10). Well, perhaps one can think about a class of advanced complex graphs which allow exponentially many '(hyper)edges' (e.g. annotated graphs) and which has a chance to achieve completeness. But complex graphs of this sort loose their easy interpretability for humans.

The conclusion above is the reason for an attempt to develop a non-graphical approach to the description of probabilistic CI structures. The approach described in subsequent chapters achieves completeness in discrete framework and had minor chance to be acceptable by humans. On the other hand, the mathematical objects which are used describe more than necesary in sense that some induced formal independence models are not probabilistic CI models. The loss of faithfulness is a natural price for the possibility of interpretation and relatively good solution of the equivalence question. Nevertheless, I consider these two gains more valuable than faithfulness.

Chapter 4

Structural imsets: fundamentals

The moral of the proceding chapter is that the main drawback of graphical models is their inability to describe all probabilistic conditional independence structures. This motivated an attempt to develop an alternative method of their description which overcomes this drawback and keeps some assets of graphical methods. The central notion of this method is the concept of *structural imset* introduced in this chapter. Note that basic ideas of the theory were presented earlier [108] but (later recognized) superfluous details worsened understanding of the message of the original series of papers. This work brings (in the next four chapters) much simpler presentation supplemented by facts and perspectives revealed later.

4.1 Basic class of distributions

The class of probability measures for which this approach is applicable, that is whose induced conditional independence models can be described by structural imsets, is relatively wide. It is the class of *measures over N with finite multiinformation* mentioned in Section 2.3.4. The aim of this section is to show that this class involves three basic classes of measures used in practice in artificial intelligence and multivariate statistics.

4.1.1 Discrete measures

These simple probability measures (see Remark 2.2, p. 13) are mainly used in probabilistic reasoning [78] which is an area of artificial intelligence. Positive discrete probability measures are behind the models used in analysis of contingency tables (see [53], Chapter 4) which is an area of statistics. The fact that every discrete probability measure over N has finite multiinformation is trivial.

4.1.2 Non-degenerate Gaussian measures

These measures (see Section 2.3.6 for their basic properties) are widely used in mathematical statistics, in particular in multivariate statistics [18]. Consequence 2.5 says that every non-degenerate Gaussian measure over N has finite multiinformation.

4.1.3 Non-degenerate conditional Gaussian measures

This class of measures was proposed by Lauritzen and Wermuth [50] with the aim to unify discrete and continuous graphical models. Non-degenerate conditional Gaussian measure P over N, in the sequel called shortly *CG-measure over* N, will be specified as follows. The set N is partitioned into the set Δ of *discrete variables* and the set Γ of *continuous variables*. For every $i \in \Delta$, X_i is a finite non-empty set and $\mathcal{X}_i = \mathcal{P}(\mathsf{X}_i)$. For every $i \in \Gamma$, $\mathsf{X}_i = \mathbb{R}$ and $\mathcal{X}_i$ is the class of Borel sets in $\mathbb{R}$. A (discrete) probability measure P_Δ on $(\mathsf{X}_\Delta, \mathcal{X}_\Delta)$ is given and a vector $\boldsymbol{e}(x) \in \mathbb{R}^\Gamma$ and a positive definite $\Gamma \times \Gamma$-matrix $\boldsymbol{\Sigma}(x) \in \mathbb{R}^{\Gamma\times\Gamma}$ is ascribed to every $x \in \mathsf{X}_\Delta$ with $P_\Delta(x) > 0$ if $\Gamma \neq \emptyset$. Then P is simply specified by its marginal for Δ and the conditional probability on X_Γ given Δ:

$$P^\Delta \equiv P_\Delta, \quad P_{\Gamma|\Delta}(*\,|\,x) \equiv \mathcal{N}(\boldsymbol{e}(x), \boldsymbol{\Sigma}(x)) \quad \text{for every } \ x \in \mathsf{X}_\Delta \ \text{ with } \ P_\Delta(x) > 0\,.$$

Of course, these requirements determine unique probability measure on $(\mathsf{X}_N, \mathcal{X}_N)$. The above definition collapses in case $\Gamma = \emptyset$ to a discrete measure over N and in case $\Delta = \emptyset$ to a non-degenerate Gaussian measure over N.

REMARK 4.1 Note that *positive CG-measures* (when $P_\Delta(x) > 0$ for every $x \in \mathsf{X}_\Delta$) are mainly used in practice. A CG-measure of this type can be defined directly (see [53] §6.1.1) by its density f with respect to the product of the counting measure on X_Δ and the Lebesgue measure on X_Γ

$$f(x,y) = \exp^{\,\boldsymbol{g}(x)+\boldsymbol{h}(x)^\top\cdot y-\frac{1}{2}\cdot y^\top\cdot\boldsymbol{\Gamma}(x)\cdot y} \qquad \text{for } \ x \in \mathsf{X}_\Delta,\ y \in \mathsf{X}_\Gamma\,,$$

where $\boldsymbol{g}(x), \boldsymbol{h}(x) \in \mathbb{R}^\Gamma$ and positive definite matrices $\boldsymbol{\Gamma}(x) \in \mathbb{R}^{\Gamma\times\Gamma}$ are named *canonical characteristics of* P. One can compute them directly from parameters $P_\Delta(x), \boldsymbol{e}(x), \boldsymbol{\Sigma}(x)$ which are named *moment characteristics* of the CG-distribution as follows (see [53], p. 159):

$$\boldsymbol{\Gamma}(x) = \boldsymbol{\Sigma}(x)^{-1}, \qquad \boldsymbol{h}(x) = \boldsymbol{\Sigma}(x)^{-1}\cdot\boldsymbol{e}(x),$$

$$\boldsymbol{g}(x) = \ln P_\Delta(x) - \frac{|\Gamma|}{2}\cdot\ln(2\pi) - \frac{1}{2}\cdot\ln(\det(\boldsymbol{\Sigma}(x))) - \frac{1}{2}\cdot\boldsymbol{e}(x)^\top\cdot\boldsymbol{\Sigma}(x)^{-1}\cdot\boldsymbol{e}(x).$$

These measures are positive in sense of Section 2.3.5 but they do not involve all discrete measures. Therefore, the class of CG-measures was slightly enlarged in this work. $\triangle$

To evidence that every CG-measure has finite multiinformation (and thus it is marginally continuous) I use auxiliary estimates with relative entropies modified in a certain way.

Supposing $(\mathsf{X}, \mathcal{X})$ is a measurable space, P and Q are probability measures and μ is a σ-finite measure on $(\mathsf{X}, \mathcal{X})$ such that $P, Q \ll \mu$ by *Q-perturbated relative entropy of P with respect to μ* will be understood the integral

$$H(P\,|\,\mu : Q) = \int_{\mathsf{X}} \ln\frac{dP}{d\mu}(x)\ \ dQ(x) \equiv \int_{\mathsf{X}} \frac{dQ}{d\mu}(x)\cdot\ln\frac{dP}{d\mu}(x)\ \ d\mu(x)$$

provided that the function $\ln\frac{dP}{d\mu}$ is Q-quasi-integrable. Of course, the value does not depend on the choice of versions of Radon-Nikodym derivatives $\frac{dP}{d\mu}$ or $\frac{dQ}{d\mu}$. In case $Q = P$ it coincides with $H(P\,|\,\mu)$ mentioned in Section 10.7. Note that a discrete version of this concept is known in information theory as Kerridge's inaccuracy [118] p. 322-323.

LEMMA 4.1 Let $(\mathsf{X}, \mathcal{X})$ be a measurable space and μ a σ-finite measure on $(\mathsf{X}, \mathcal{X})$. Suppose that $P_1, \ldots, P_r,\ r \geq 1$ is a finite collection of probability measures on $(\mathsf{X}, \mathcal{X})$ such that $-\infty < H(P_k|\mu : P_l) < +\infty$ for every $k, l \in \{1, \ldots, r\}$. Then every convex combination of $P_1, \ldots, P_r$ has finite relative entropy with respect to μ, that is

$$-\infty < H(\sum_{k=1}^{r} \alpha_k \cdot P_k |\, \mu) < +\infty \quad \text{whenever } \ \alpha_1, \ldots, \alpha_r \geq 0,\ \sum_{k=1}^{r} \alpha_k = 1.$$

Proof: Put $P = \sum_{k=1}^{r} \alpha_k \cdot P_k$, choose and fix a version of $\frac{dP_k}{d\mu}$ for every k and fix the version of $\frac{dP}{d\mu} = \sum_{l=1}^{r} \alpha_l \cdot \frac{dP_l}{d\mu}$. The assumption says

$$\forall\, k, l \in \{1, \ldots, r\} \qquad \int_{\mathsf{X}} \frac{dP_l}{d\mu}(x) \cdot |\, \ln \frac{dP_k}{d\mu}(x)\,|\, d\mu(x) < \infty.$$

One has to show that

$$\int_{\mathsf{X}} |\, \ln \frac{dP}{d\mu}(x)\,|\; dP(x) = \int_{\mathsf{X}} (\, \ln \frac{dP}{d\mu}(x)\,)^{+}\; dP(x) + \int_{\mathsf{X}} (\, \ln \frac{dP}{d\mu}(x)\,)^{-}\; dP(x) < \infty\,.$$

To estimate the first term above use Radon-Nikodym theorem, the observation that the function $y \mapsto (y \cdot \ln y)^{+}$ is convex and the inequality $y^{+} \leq |y|$:

$$\int_{\mathsf{X}} (\, \ln \frac{dP}{d\mu}(x)\,)^{+}\; dP(x) \leq \int_{\mathsf{X}} (\, \frac{dP}{d\mu}(x) \cdot \ln \frac{dP}{d\mu}(x)\,)^{+}\; d\mu(x) \leq$$

$$\leq \sum_{k=1}^{r} \alpha_k \cdot \int_{\mathsf{X}} (\, \frac{dP_k}{d\mu}(x) \cdot \ln \frac{dP_k}{d\mu}(x)\,)^{+}\; d\mu(x) \leq \sum_{k=1}^{r} \alpha_k \cdot \int_{\mathsf{X}} \frac{dP_k}{d\mu}(x) \cdot |\ln \frac{dP_k}{d\mu}(x)|\, d\mu(x) < \infty\,.$$

To estimate the second use the fact that the function $y \mapsto (\ln y)^{-}$ is convex, the inequality $y^{-} \leq |y|$, Radon-Nikodym theorem and the form of $\frac{dP}{d\mu}$

$$\int_{\mathsf{X}} (\, \ln \frac{dP}{d\mu}(x)\,)^{-}\; dP(x) \leq \sum_{k=1}^{r} \alpha_k \cdot \int_{\mathsf{X}} (\, \ln \frac{dP_k}{d\mu}(x)\,)^{-}\; dP(x) \leq \sum_{k=1}^{r} \alpha_k \cdot \int_{\mathsf{X}} |\, \ln \frac{dP_k}{d\mu}(x)\,|\; dP(x) =$$

$$= \sum_{k=1}^{r} \alpha_k \cdot \int_{\mathsf{X}} \sum_{l=1}^{r} \alpha_l \cdot \frac{dP_l}{d\mu}(x) \cdot |\ln \frac{dP_k}{d\mu}(x)|\; d\mu(x) = \sum_{k,l=1}^{r} \alpha_k \cdot \alpha_l \cdot \int_{\mathsf{X}} \frac{dP_l}{d\mu}(x) \cdot |\ln \frac{dP_k}{d\mu}(x)|\, d\mu(x) < \infty\,.$$

$\square$

LEMMA 4.2 Let P be a CG-measure over $N = \Delta \cup \Gamma$ and $\mu = \prod_{i \in N} \mu_i$ where $\mu_i = \upsilon$ for $i \in \Delta$ and $\mu_i = \lambda$ for $i \in \Gamma$. Then

$$-\infty < H(P \,|\, \mu) < \infty \quad \text{and} \quad -\infty < H(P^{\{i\}}|\mu_i) < \infty \quad \text{for every } i \in N\,.$$

Proof: A direct formula for $H(P|\mu)$ is easy to derive. Indeed, write

$$\frac{dP}{d\mu}(x, y) = P_{\Delta}(x) \cdot f_{\boldsymbol{e}(x), \boldsymbol{\Sigma}(x)}(y) \quad \text{for } \ x \in \mathsf{X}_{\Delta} \ \text{ with } \ P_{\Delta}(x) > 0 \ \text{ and } \ y \in \mathsf{X}_{\Gamma},$$

apply logarithm, integrate it with respect to P and obtain using standard properties of integral

$$\begin{aligned} H(P|\mu) &= \int_{\mathsf{X}_N} \ln\, P_\Delta(x)\, dP(x,y) + \int_{\mathsf{X}_N} \ln\, f_{\boldsymbol{e}(x),\boldsymbol{\Sigma}(x)}(y)\ dP(x,y) = \\ &= H(P_\Delta|\mu_\Delta) + \sum_{x\in\mathsf{X}_\Delta, P_\Delta(x)>0} P_\Delta(x)\cdot H(P(*|x)|\mu_\Gamma)\,. \end{aligned}$$

The fact $-\infty < H(P^{\{i\}} \,|\, \mu_i) < \infty$ for $i \in \Delta$ is trivial. For fixed $i \in \Gamma$ first realize that $P^{\Delta\cup\{i\}}$ is again a CG-measure where

$$P_{\{i\}|\Delta}(*|x) = \mathcal{N}(\boldsymbol{e}(x)_i, \boldsymbol{\Sigma}(x)_{i\cdot i}) \qquad \text{for } \ i \in \mathsf{X}_\Delta \ \text{ with } \ P_\Delta(x) > 0.$$

Therefore, the marginal $P^{\{i\}}$ is nothing but a convex combination of non-degenerate Gaussian measures. To verify $-\infty < H(P^{\{i\}}|\lambda) < \infty$ one can use Lemma 4.1. Indeed, suppose $P_k = \mathcal{N}(e,\beta)$ and $P_l = \mathcal{N}(f,\gamma)$ where $e, f \in \mathbb{R}$, $\beta, \gamma > 0$ are the corresponding parameters. Because expectation and variance of P_l are known one can compute easily

$$\begin{aligned} H(P_k|\lambda : P_l) = \int_{\mathbb{R}} -\frac{1}{2}\cdot\ln(2\pi\beta) - \frac{(x-e)^2}{2\beta}\, dP_l(x) = -\frac{1}{2}\cdot\ln(2\pi\beta) - \frac{1}{2\beta}\cdot\int_{\mathbb{R}} (x-e)^2\, dP_l(x) = \\ = -\frac{1}{2}\cdot\ln(2\pi\beta) - \frac{1}{2\beta}\cdot\int_{\mathbb{R}} (x-f)^2 + 2\cdot(f-e)\cdot x + (e^2-f^2)\, dP_l(x) = \\ = -\frac{1}{2}\ln(2\pi\beta) - \frac{1}{2\beta}[\gamma + 2\cdot(f-e)\cdot f + (e^2-f^2)] = -\frac{1}{2}\cdot\ln(2\pi\beta) - \frac{\gamma+(e-f)^2}{2\cdot\beta}\,. \end{aligned}$$

The result is evidently a finite number. □

Consequence 4.1 Every CG-measure over N has finite multiinformation.

Proof: Owing to Lemma 4.2 the assumptions of Lemma 2.7 on p. 27 for $S = N$ are fulfilled. □

The fact above was verified by finding finite lower and upper estimates for multiinformation. The question whether there exists a suitable exact formula for values of multiinformation function in terms of parameters of CG-measure remains open (see Theme 1 in Chapter 8).

Remark 4.2 The class of CG-meassures is not closed under marginalizing which may lead to problems when one tries to study CI within this context. However, it was shown that this class can be embedded into a wider class of measures with finite multiinformation which is already closed under marginalizing (see Consequence 2.2). △

4.2 Classes of structural imsets

Definitions and elementary facts concerning structural imsets are gathered in this section.

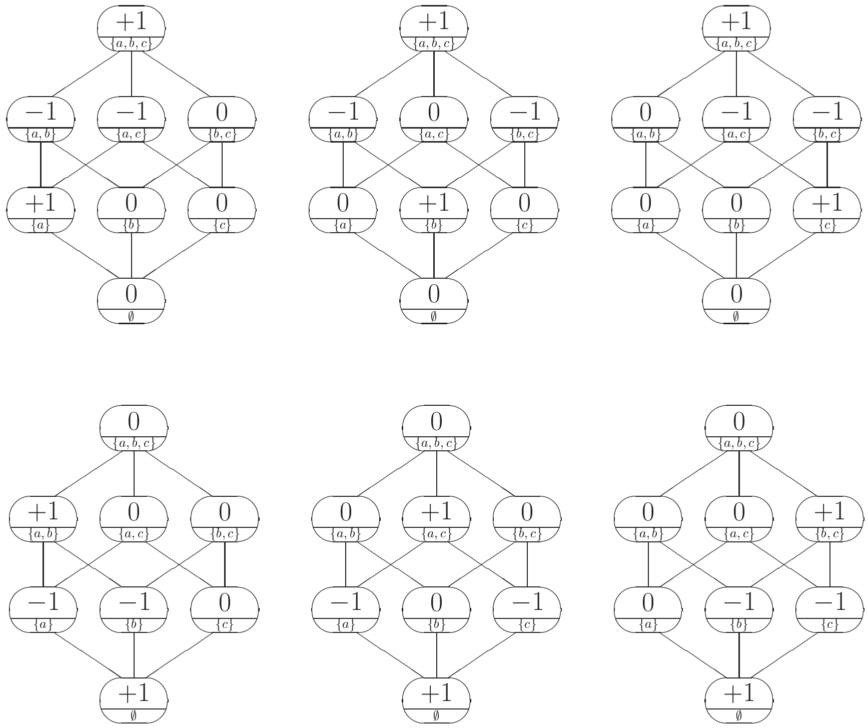


Figure 4.1: Elementary imsets over $N = \{a, b, c\}$.

4.2.1 Elementary imsets

Elementary imset over a set of variables N with $|N| \geq 2$ is an imset of a specific form. Given (an elementary) triplet $\langle i, j|K\rangle$ where $K \subseteq N$ and $i, j \in N \setminus K$ are distinct ($i \neq j$) the corresponding *elementary imset* $u_{\langle i,j|K\rangle}$ over N is defined by the formula

$$u_{\langle i,j|K\rangle} = \delta_{\{i,j\}\cup K} + \delta_K - \delta_{\{i\}\cup K} - \delta_{\{j\}\cup K}\,.$$

The class of elementary imsets over N will be denoted by $\mathcal{E}(N)$. In case $|N| = 1$ is the class $\mathcal{E}(N)$ empty by convention. By *level* of an elementary imset $u_{\langle i,j|K\rangle}$ is understood the number $|K|$. For every $l = 0, \ldots, |N| - 2$, the class of elementary imsets of level l will be denoted by $\mathcal{E}_l(N)$. Supposing $|N| = n \geq 2$ it is easy to see that $|\mathcal{E}_l(N)| = \binom{n}{2} \cdot \binom{n-2}{l}$ and $|\mathcal{E}(N)| = \binom{n}{2} \cdot 2^{n-2}$. Thus, in case $N = \{a, b, c\}$ one has 6 elementary imsets of 2 possible levels. They are shown in Figure 4.1.

The following observation is a basis of later results.

Observation 4.1 Supposing $n = |N| \geq 2$ and $l \in \{0, \ldots, n-2\}$ let us introduce a multiset m_l over N by means of the formula

$$m_l(S) = \max\{\,|S| - l - 1, 0\}\quad \text{for every } S \subseteq N,$$

and a multiset m_* over N by means of the formula

$$m_*(S) = \frac{1}{2}\cdot|S|\cdot(|S|-1)\quad \text{for every } S \subseteq N.$$

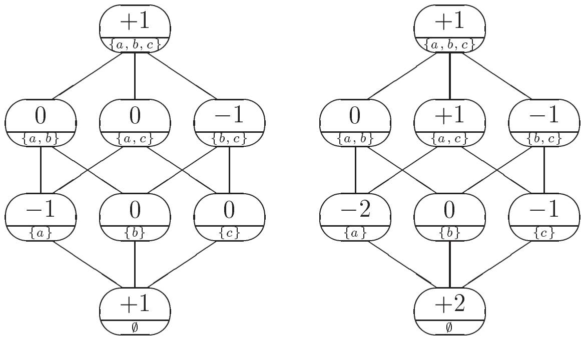


Figure 4.2: Two combinatorial imsets over $N = \{a, b, c\}$.

Then one can observe the following facts

$$\begin{array}{lll} \text{(a)} & \forall\, u \in \mathcal{E}_l(N) & \langle m_l, u\rangle = 1, \\ \text{(b)} & \forall\, u \in \mathcal{E}(N) \setminus \mathcal{E}_l(N) & \langle m_l, u\rangle = 0, \\ \text{(c)} & \forall\, u \in \mathcal{E}(N) & \langle m_*, u\rangle = 1. \end{array}$$

Proof: The first two facts are easy to evidence, the third fact follows from the identity $m_* = \sum_{l=0}^{n-2} m_l$ and the previous facts. Indeed, this identity can be verified for $S \subseteq N$, $|S| \geq 2$ as follows:

$$\sum_{l=0}^{n-2} m_l(S) = \sum_{l=0}^{|S|-2} m_l(S) = \sum_{l=0}^{|S|-2} |S| - \sum_{l=0}^{|S|-2} l - \sum_{l=0}^{|S|-2} 1 = |S| \cdot (|S|-1) - \frac{1}{2}(|S|-1) \cdot |S| = m_*(S).$$

□

4.2.2 Semi-elementary and combinatorial imsets

Given $\langle A, B|C\rangle \in \mathcal{T}(N)$ the corresponding *semi-elementary imset* $u_{\langle A,B|C\rangle}$ is defined by the formula

$$u_{\langle A,B|C\rangle} \;=\; \delta_{ABC} + \delta_C - \delta_{AC} - \delta_{BC}\,.$$

Evidently, zero imset is semi-elementary as $u_{\langle A,B|C\rangle} = 0$ for any $\langle A, B|C\rangle \in \mathcal{T}_\emptyset(N)$. Every elementary imset is semi-elementary as well. An example of non-zero semi-elementary imset which is not elementary is the imset $u_{\langle a,bc|\emptyset\rangle}$ shown in the left-hand picture of Figure 4.2. Provided one accepts the convention that zero imset is a combination of the empty set of imsets one can observe the following fact.

Observation 4.2 Every semi-elementary imset is a combination of elementary imsets with non-negative integral coefficients.

Proof: A non-zero semi-elementary imset has the form $u_{\langle A,B|C\rangle}$ where $\langle A, B|C\rangle \in \mathcal{T}(N) \setminus \mathcal{T}_\emptyset(N)$. The formulas $u_{\langle A,BD|C\rangle} = u_{\langle A,B|DC\rangle} + u_{\langle A,D|C\rangle}$ and $u_{\langle AD,B|C\rangle} = u_{\langle A,B|DC\rangle} + u_{\langle D,B|C\rangle}$ can be applied repeatedly. □

By a *combinatorial imset* over N will be understood every imset u which is a combination of elementary imsets with non-negative integral coefficients, that is

$$u = \sum_{v \in \mathcal{E}(N)} k_v \cdot v \quad \text{where} \quad k_v \in \mathbb{Z}^+. \tag{4.1}$$

The class of combinatorial imsets over N will be denoted by $\mathcal{C}(N)$. By Observation 4.2, every semi-elementary imset is a combinatorial imset. The converse is not true: the imset $u_{\langle a,b|c\rangle} + 2 \cdot u_{\langle a,c|\emptyset\rangle}$ in the right-hand picture of Figure 4.2 is not semi-elementary. Clearly, every combination of combinatorial imsets with coefficients from $\mathbb{Z}^+$ is again a combinatorial imset. In particular, combinatorial imsets can be equivalently introduced as combinations of semi-elementary imsets with non-negative integral coefficients.

Of course, a particular combinatorial imset can be sometimes expressed in several different ways. For example, the imset u from the left-hand picture of Figure 4.2 can be written either as $u_{\langle a,b|c\rangle} + u_{\langle a,c|\emptyset\rangle}$ or as $u_{\langle a,c|b\rangle} + u_{\langle a,b|\emptyset\rangle}$. On the other hand, there are characteristics which do not depend on a particular way of combination. Supposing (4.1) one can introduce the *degree* of a combinatorial imset u, denoted by $deg(u)$, as follows

$$deg(u) = \sum_{v \in \mathcal{E}(N)} k_v \, .$$

Similarly, if $|N| \geq 2$ then introduce the *level-degree* of u for every $l = 0, \ldots, |N| - 2$, denoted by $deg(u, l)$, as the number

$$deg(u, l) = \sum_{v \in \mathcal{E}_l(N)} k_v \, .$$

The following lemma implies that these numbers do not depend on the choice of coefficients k_v for $v \in \mathcal{E}(N)$.

Observation 4.3 Supposing $u \in \mathcal{C}(N)$ and $l \in \{0, \ldots, |N| - 2\}$ with $|N| \geq 2$

$$deg(u, l) = \langle m_l, u\rangle, \qquad deg(u) = \langle m_*, u\rangle \, ,$$

where the multisets m_l, m_* are introduced in Observation 4.1 on p. 57.

Proof: Substitute (4.1) in $\langle m_l, u\rangle$ and $\langle m_*, u\rangle$ and use Observation 4.1. □

4.2.3 Structural imsets

An imset u over N will be called *structural* if there exists $n \in \mathbb{N}$ such that the multiple $n \cdot u$ is a combinatorial imset, that is

$$n \cdot u = \sum_{v \in \mathcal{E}(N)} k_v \cdot v \quad \text{for some } n \in \mathbb{N},\ k_v \in \mathbb{Z}^+. \tag{4.2}$$

In other words, an imset is structural if it is a combination of elementary imsets, respectively semi-elementary imsets, with non-negative rational coefficients. The class of structural imsets over N will be denoted by $\mathcal{S}(N)$. By definition, every combinatorial imset is structural. In case $|N| \leq 4$ the converse is true [101]. However, the question whether this is true in general remains open (see Question 7 on p. 142).

OBSERVATION 4.4 Every structural imset u over N is o-standardized, $\langle m^{A\uparrow}, u\rangle \geq 0$ and $\langle m^{A\downarrow}, u\rangle \geq 0$ (see pp. 34-35). The only imset $w \in \mathcal{S}(N)$ with $-w \in \mathcal{S}(N)$ is the zero imset $w = 0$.

Proof: All three properties hold for zero and elementary imsets and can be extended to combinatorial imsets and then to structural imsets. Given $w \in \mathbb{Z}^{\mathcal{P}(N)}$ with $\langle m^{A\downarrow}, w\rangle = 0$ for every $A \subseteq N$ the condition $w(S) = 0$ for $S \subseteq N$ can be verified by induction on $|S|$. $\square$

Given a structural imset u let us introduce the *lower class* of u, denoted by $\mathcal{L}_u$, as the descending class induced by the negative domain of u, that is

$$\mathcal{L}_u = \{\, T \subseteq N;\ \exists\, S \subseteq N \text{ such that } T \subseteq S \ \text{ and } u(S) < 0 \,\} \equiv (\mathcal{D}_u^-)^{\downarrow}.$$

Similarly, one can introduce the *upper class* of u, denoted by $\mathcal{U}_u$, as the descending class induced by the positive domain of u

$$\mathcal{U}_u = \{\, T \subseteq N;\ \exists\, S \subseteq N \text{ such that } T \subseteq S \ \text{ and } u(S) > 0 \,\} \equiv (\mathcal{D}_u^+)^{\downarrow}.$$

Terminology is motivated by the next fact and later results (Consequence 4.3 on p. 66).

OBSERVATION 4.5 Whenever u is a structural imset one has $\mathcal{L}_u \subseteq \mathcal{U}_u$. Moreover

$$\bigcup_{S \in \mathcal{L}_u} S \ = \ \bigcup_{S \in \mathcal{U}_u} S\,. \tag{4.3}$$

Proof: Supposing $T \in \mathcal{L}_u$ find $T \subseteq S \subseteq N$ with $u(S) < 0$. The fact $\langle m^{S\uparrow}, u\rangle \geq 0$ from Observation 4.4 implies the existence of $S \subseteq K \subseteq N$ with $u(K) > 0$. The fact that u is o-standardized says $\langle m^{\{i\}\uparrow}, u\rangle = 0$ for every $i \in N$ which implies (4.3) then. $\square$

Given a structural imset u over N, by the *range* of u, denoted by R_u, will be understood the set union from (4.3). The following lemma is a basis of a later result.

LEMMA 4.3 Supposing u is a non-zero combinatorial imset over N let us consider a fixed particular combination

$$u = \sum_{v \in \mathcal{E}(N)} k_v \cdot v \quad \text{where } k_v \in \mathbb{Z}^+,\ \ u \neq 0\,.$$

Then there exists $v \in \mathcal{E}(N)$ such that $k_v > 0$ and $\mathcal{L}_v \subseteq \mathcal{L}_u$.

Proof: Since $u \neq 0$ necessarily $\mathcal{L}_u \cup \mathcal{U}_u \neq \emptyset$. Because u is structural, by Observation 4.5 $\mathcal{L}_u \subseteq \mathcal{U}_u$, and therefore $\mathcal{U}_u \neq \emptyset$. Take maximal $K \in \mathcal{U}_u$ and again using $\mathcal{L}_u \subseteq \mathcal{U}_u$ observe that $u(K) > 0$ and $\forall\, L \supset K \ \ u(L) = 0$. Introduce

$$\mathcal{S}_u = \mathcal{P}(N) \setminus \mathcal{L}_u = \{\, T \subseteq N;\ \forall\, T \subseteq S \subseteq N \ \ u(S) \geq 0 \}\,.$$

Clearly, $\mathcal{S}_u$ is an ascending class and $K \in \mathcal{S}_u$; let us consider a multiset $s = \sum_{S \in \mathcal{S}_u} \delta_S$. It follows from the definition of $\mathcal{S}_u$ that $\langle s, u\rangle \geq u(K) > 0$. Thus, one can write

$$0 < \langle s, u\rangle = \sum_{v \in \mathcal{E}(N)} k_v \cdot \langle s, v\rangle \leq \sum_{v \in \mathcal{E}(N),\, \langle s,v\rangle > 0} k_v \cdot \langle s, v\rangle,$$

which implies the existence $v \in \mathcal{E}(N)$ with $k_v > 0$ and $\langle s, v\rangle > 0$. Well, since $\mathcal{S}_u$ is ascending, an elementary imset $v = u_{\langle i,j|K\rangle}$ satisfies $\langle s, v\rangle > 0$ iff $\{i,j\} \cup K \ \in \mathcal{S}_u$ and $\{i\} \cup K, \{j\} \cup K \ \notin \mathcal{S}_u$ (see Section 4.2.1). However, this implies $\mathcal{L}_v \cap \mathcal{S}_u = \emptyset$, which means $\mathcal{L}_v \subseteq \mathcal{L}_u$. $\square$

4.3 Product formula induced by a structural imset

This formula provides a direct way of associating a structural imset with a probability measure. It can be viewed as a generalization of the concept of factorization into marginal densities. To give a sensible definition I need the following auxiliary concept whose sense becomes evident later (see Section 4.5). Suppose that P is a probability measure on $(\mathsf{X}_N, \mathcal{X}_N)$ which has finite multiinformation. By a *reference system of measures for P* will be understood any collection $\{\mu_i;\ i \in N\}$ of σ-finite measures on $(\mathsf{X}_i, \mathcal{X}_i)$, $i \in N$ such that

$$P^{\{i\}} \ll \mu_i \quad \text{and} \quad -\infty < H(P^{\{i\}}|\,\mu_i) < +\infty \qquad \text{for every} \;\; i \in N\,.$$

Having fixed a reference system $\{\mu_i;\ i \in N\}$ one can put $\mu = \prod_{i \in N} \mu_i$ and observe $P \ll \mu$, that is μ is a dominating measure for P. Thus, one can repeat what is done in Convention 1 (p. 20), that is to choose marginal density f_S (= a version of $\frac{dP^S}{d\mu_S}$) for every $S \subseteq N$.

Given a structural imset u over N one says that P satisfies the *product formula induced by u* if

$$\prod_{S \subseteq N} f_S(x_S)^{u^+(S)} = \prod_{S \subseteq N} f_S(x_S)^{u^-(S)} \qquad \text{for} \;\; \mu\text{-a.e.} \;\; x \in X_N. \tag{4.4}$$

Of course, the validity of this formula does not depend on the choice (of versions) of marginal densities. The influence of the choice of a reference system of measures will appear to be seeming as well (see Section 4.5). On the other hand, flexibility in its choice is advantageous since miscellaneous special cases can be described in more details.

4.3.1 Examples of reference systems of measures

Let me illustrate this concept by four basic examples. The first one shows that one can always find a reference system for a probability measure with finite multiinformation. The other examples correspond to important special cases mentioned already in Section 4.1.

Universal reference system

Given a probability measure P over N with $H(P\,|\prod_{i \in N} P^{\{i\}}) < \infty$ one can simply put $\mu_i = P^{\{i\}}$ for every $i \in N$. It is evidently a reference system of measures since $H(P^{\{i\}}|\mu_i) = 0$ for every $i \in N$. Let us call it the *universal reference system* because it can be established for any measure with finite multiinformation.

Discrete case

Supposing P is a discrete measure on $(\mathsf{X}_N, \mathcal{X}_N)$ with $1 \leq |\mathsf{X}_i| < \infty$, $i \in N$ one can consider the counting measure υ on X_i in place of μ_i for every $i \in N$. This is evidently a reference system for P leading to the following system of marginal densities:

$$f_S(x_S) = P^S(\{x_S\}) \qquad \text{for every} \;\; S \subseteq N,\; x \in \mathsf{X}_N.$$

REMARK 4.3 An alternative choice of a reference system in discrete case is possible. One can take uniformly distributed probability measure $\hat{\mu}_i = \frac{\upsilon}{|\mathsf{X}_i|}$ on X_i for every $i \in N$. This leads to alternative marginal densities

$$\hat{f}_S(x_S) = \frac{P^S(\{x_S\})}{|X_S|} \quad \text{for every} \;\; S \subseteq N, \;\; x \in \mathsf{X}_N,$$

with convention $|\mathsf{X}_\emptyset| = 1$. $\triangle$

Gaussian case

Supposing $P = \mathcal{N}(\boldsymbol{e}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma} = (\sigma_{ij})_{i,j\in N}$ is a non-degenerate Gaussian measure over N one consider the Lebesgue measure λ on $\mathbb{R}$ is place of μ_i for every $i \in N$. It is a reference system for P because $H(P^{\{i\}}|\lambda) = -\frac{1}{2} - \frac{1}{2} \cdot \ln(2\pi\sigma_{ii})$ for every $i \in N$ by (10.9) in Section 10.9.3. Owing to the fact that the marginal of a Gaussian measure is again Gaussian and (2.19) one can choose marginal densities f_S for $\emptyset \neq S \subseteq N$ in the form

$$f_S(y) = \frac{1}{\sqrt{(2\pi)^{|S|} \cdot \det(\boldsymbol{\Sigma}_{S\cdot S})}} \cdot \exp^{-\frac{1}{2}\cdot(y-\boldsymbol{e}_S)^\top \cdot (\boldsymbol{\Sigma}_{S\cdot S})^{-1}\cdot(y-\boldsymbol{e}_S)} \quad \text{for } y \in \mathbb{R}^S.$$

CG-measures

Let P be a non-degenerate CG-measure over N partitioned into the set Δ of discrete variables and the set Γ of continuous variables. By a *standard reference system for P* will be understood the system $\{\mu_i;\, i \in N\}$ where $\mu_i = \upsilon$ is the counting measure on finite X_i for $i \in \Delta$ and $\mu_i = \lambda$ is the Lebesgue measure on $X_i = \mathbb{R}$ for $i \in \Gamma$. By Lemma 4.2 it is indeed a reference system of measures for P. In purely discrete or Gaussian case it coicides with two above mentioned reference systems which I recalled explicitly in order to emphasize the importance of these two classic cases.

One can choose the following versions of marginal densities f_S for $\emptyset \neq S \subseteq N$ (of course, the formula is more complex than in purely discrete or purely Gaussian case)

$$f_S(x,y) = \sum_{\substack{z\in \mathsf{X}_{\Delta\setminus S} \\ P_\Delta(x,z)>0}} P_\Delta(x,z) \cdot f_{\boldsymbol{e}(x,z)_{S\cap\Gamma}, \boldsymbol{\Sigma}(x,z)_{S\cap\Gamma\cdot S\cap\Gamma}}(y) \quad \text{for } x \in \mathsf{X}_{S\cap\Delta},\ y \in \mathsf{X}_{S\cap\Gamma},$$

where a detailed formula for $f_{\boldsymbol{e}(*),\boldsymbol{\Sigma}(*)}(*)$ is in (2.19).

4.3.2 Topological assumptions

The reader can object that the product formula (4.4) is not elegant enough since it is dimmed by non-uniqueness of marginal densities and the equality is understood in 'almost everywhere' sense. However, under certain topological assumptions usually valid in practice and additional natural convention it turns into a fair equality 'everywhere'.

A reference system $\{\mu_i; i \in N\}$ for a probability measure P on $(\mathsf{X}_N, \mathcal{X}_N)$ with finite multiinformation will be called *continuous* if the following three conditions are fulfilled.

(a) X_i is a separable metric space and $\mathcal{X}_i$ is the class of Borel sets in X_i for every $i \in N$.

(b) Every open ball in X_i has positive measure μ_i for every $i \in N$, that is

$$\forall\, i \in N \quad \forall\, x \in \mathsf{X}_i \quad \forall\, \varepsilon > 0 \quad \mu_i(U(x,\varepsilon)) > 0.$$

(c) For every $\emptyset \neq S \subseteq N$ there exists a version f_S of $\frac{dP^S}{d\mu_S}$ (where $\mu_S = \prod_{i\in S}\mu_i$) which is continuous with respect to the product topology on $\mathsf{X}_S = \prod_{i\in S}\mathsf{X}_i$.

The following observation is easy to evidence (see the Appendix, Sections 10.4, 10.5 and 10.9 for relevant facts).

OBSERVATION 4.6 The standard reference system of measures for a non-degenerate CG-measure over N is continuous.

In case of a continuous reference system Convention 1 can be explicated as follows.

CONVENTION 2 Suppose that P is a probability measure on $(\mathsf{X}_N, \mathcal{X}_N)$ with finite multiinformation and $\{\mu_i; i \in N\}$ is a continuous reference system for P. Then (a) implies that X_S is a separable metric space and $\mathcal{X}_S$ is Borel σ-algebra on X_S for every $\emptyset \neq S \subseteq N$. Put $\mu_S = \prod_{i\in S} \mu_i$, choose a version f_S of Radon-Nikodym derivative $\frac{dP^S}{d\mu_S}$ which is continuous with respect to respective topology on X_S and fix it. Note that it is possible owing to (c). Let us call it the *continuous marginal density of P for S*. Note that it follows from (b) that it is determined uniquely (use arguments from the proof of the next lemma).

Other notational habits from Convention 1 remain valid. In particular, every f_S can be viewed as a continuous function on X_N endowed with the product topology. △

LEMMA 4.4 Let P be a probability measure over N with finite multiinformation and $\{\mu_i; i \in N\}$ a continuous reference system of measures for P. Let's accept Convention 2. Then (4.4) is equivalent to the requirement

$$\prod_{S\subseteq N} f_S(x_S)^{u^+(S)} = \prod_{S\subseteq N} f_S(x_S)^{u^-(S)} \qquad \text{for every } x \in \mathsf{X}_N\,. \tag{4.5}$$

Proof: By (a) assume that $(\mathsf{X}_i, \varrho_i)$ is a separable metric space for every $i \in N$. Observe that X_N endowed with the distance

$$\varrho(x, y) = \max_{i\in N}\ \varrho_i(x_i, y_i) \qquad \text{for } x, y \in \mathsf{X}_N$$

is a separable metric space inducing product topology which generates $\mathcal{X}_S$ (see e.g. [98] Theorem I.2.3). This definition implies that open balls in X_N are Cartesian products of open balls in X_i and therefore one derives from (b)

$$\forall\, x \in \mathsf{X}_N \quad \forall\, \varepsilon > 0 \quad \mu_N(U_\varrho(x, \varepsilon)) = \mu_N(\prod_{i\in N} U_{\varrho_i}(x_i, \varepsilon)) > 0\,.$$

Now, both the left-hand side and the right-hand side of (4.4) are continuous functions on X_N by (c) (see Convention 2) and (4.4) says that their difference g (which is also a continuous function on X_N) vanishes μ_N-a.e. Hence, $\int_{\mathsf{X}_N} |g(y)|\, d\mu_N(y) = 0$.

Suppose for contradiction that $g(x) \neq 0$ for some $x \in \mathsf{X}_N$. Then there exists $\varepsilon > 0$ such that $\forall\, y \in U(x, \varepsilon)$ one has $|g(y)| \geq \frac{|g(x)|}{2}$ and therefore

$$\int_{\mathsf{X}_N} |g(y)|\, d\mu_N(y) \geq \int_{U(x,\varepsilon)} |g(y)|\, d\mu_N(y) \geq \frac{|g(x)|}{2} \cdot \mu_N(U(x, \varepsilon)) > 0$$

which contradicts the fact above. Therefore $g(x) = 0$ for every $x \in \mathsf{X}_N$. □

Thus, by Observation 4.6 one can interpret the product formula induced by a structural imset as a real identity of uniquely determined marginal densities in three basic cases used in practice: for discrete measures, for non-degenerate Gaussian measures and for non-degenerate CG-measures. Of course, this need not hold for arbitrary measure with finite multiinformation and respective universal reference system of measures.

4.4 Markov condition

The second basic way of associating a structural imset with a probability measure is an analogue of Markov condition used in graphical models. That means, one requires that some conditional independence statements determined by an imset through a certain criterion are valid conditional independence statements with respect to the measure.

4.4.1 Semi-graphoid induced by a structural imset

One says that a disjoint triplet $\langle A,B|C\rangle \in \mathcal{T}(N)$ is *represented in a structural imset* u over N and writes $A \perp\!\!\!\perp B \,|\, C\ [u]$ if there exists $k \in \mathbb{N}$ such that $k \cdot u - u_{\langle A,B|C\rangle}$ is a structural imset over N as well. An equivalent requirement is that there exists $l \in \mathbb{N}$ such that $l \cdot u - u_{\langle A,B|C\rangle}$ is a combinatorial imset over N. The class of represented triplets then defines the (conditional independence) *model induced by* u

$$\mathcal{M}_u = \{\, \langle A,B|C\rangle \in \mathcal{T}(N);\quad A \perp\!\!\!\perp B \,|\, C\ [u]\,\}\,.$$

Trivial example is the model induced by zero imset.

Observation 4.7 $\mathcal{M}_u = \mathcal{T}_\emptyset(N)$ for $u = 0$.

Proof: Inclusion $\mathcal{T}_\emptyset(N) \subseteq \mathcal{M}_u$ is trivial. Suppose for contradiction $\langle A,B|C\rangle \in \mathcal{T}(N) \setminus \mathcal{T}_\emptyset(N)$ which means that $-u_{\langle A,B|C\rangle}$ is a structural imset. This contradicts Observation 4.4. $\square$

Further example is the model induced by an elementary imset.

Lemma 4.5 Supposing $v = u_{\langle i,j|K\rangle} \in \mathcal{E}(N)$ one has

$$\mathcal{M}_v = \{\, \langle i,j|K\rangle\,,\ \langle j,i|K\rangle\,\} \cup \mathcal{T}_\emptyset(N)\,.$$

Proof: The facts $\langle i,j|K\rangle\,,\ \langle j,i|K\rangle \in \mathcal{M}_v$ and $\mathcal{T}_\emptyset(N) \subseteq \mathcal{M}_v$ are evident. Suppose that $\langle A,B|C\rangle \in \mathcal{M}_v \setminus \mathcal{T}_\emptyset(N)$ and $k \cdot v = u_{\langle A,B|C\rangle} + w$ for $k \in \mathbb{N}$ and a structural imset w. To evidence $ABC \subseteq ijK$ use Observation 4.4 to derive

$$k \cdot \langle m^{ABC\uparrow}, v\rangle = \langle m^{ABC\uparrow}, k\cdot v\rangle = \langle m^{ABC\uparrow}, u_{\langle A,B|C\rangle}\rangle + \langle m^{ABC\uparrow}, w\rangle > 0\,. \qquad (4.6)$$

The fact $\langle m^{ABC\uparrow}, u_{\langle i,j|K\rangle}\rangle > 0$ then implies $ABC \subseteq ijK$. Analogously, to evidence $C \supseteq K$ use also Observation 4.4 with $m^{C\downarrow}$ in (4.6) instead of $m^{ABC\uparrow}$. The fact that $\langle A,B|C\rangle$ is a disjoint triplet and $K \subseteq C \subset ABC \subseteq ijK$ then implies that $\langle A,B|C\rangle$ coincides either with $\langle i,j|K\rangle$ or with $\langle j,i|K\rangle$. $\square$

A basic fact is this.

Lemma 4.6 Every structural imset over N induces a disjoint semi-graphoid over N.

Proof: Semi-graphoid properties (see Section 2.2.2, p. 15) easily follow from the definition above and the fact that the sum of structural imsets is a structural imset. Indeed, for triviality property realize $\mu_{\langle A,\emptyset|C\rangle} = 0$, for symmetry $\mu_{\langle A,B|C\rangle} = \mu_{\langle B,A|C\rangle}$ and for remaining properties $\mu_{\langle A,BD|C\rangle} = \mu_{\langle A,B|DC\rangle} + \mu_{\langle A,D|C\rangle}$. $\square$

For the proof of the equivalence result in Section 4.5 I need a technical lemma. In its proof the following simple observation concerning upper classes (see p. 60) is used.

OBSERVATION 4.8 Supposing $u = w + v$ where w, v are structural imsets one has $\mathcal{U}_u = \mathcal{U}_w \cup \mathcal{U}_v$.

Proof: Inclusion $\mathcal{U}_u \subseteq \mathcal{U}_w \cup \mathcal{U}_v$ is trivial. To show $\mathcal{U}_w \subseteq \mathcal{U}_u$ take $S \in \mathcal{U}_w^{\max}$. By Observation 4.5 $w(S) > 0$ and $w(T) = 0$ whenever $S \subset T \subseteq N$. Hence, by Observation 4.4

$$0 < \langle m^{S\uparrow}, w\rangle + \langle m^{S\uparrow}, v\rangle = \langle m^{S\uparrow}, u\rangle = \sum_{T, S\subseteq T} u(T)$$

which implies that $S \in \mathcal{U}_u$. The inclusion $\mathcal{U}_v \subseteq \mathcal{U}_u$ is analogous. □

LEMMA 4.7 Suppose that u is a structural imset over N. Then there exists a sequence $\mathcal{U}_u = \mathcal{D}_0, \ldots, \mathcal{D}_r = \mathcal{L}_u,\ r \geq 0$ of descending classes of subsets of N and a sequence $\langle a_1, b_1|C_1\rangle, \ldots, \langle a_r, b_r|C_r\rangle$ of elementary triplets over N (which is empty in case $r = 0$) such that for every $i = 1, \ldots, r$

(a) $a_i \perp\!\!\!\perp b_i \,|\, C_i\ [u]$,

(b) $a_iC_i,\ b_iC_i \in \mathcal{D}_i$ and $\mathcal{D}_{i-1} = \mathcal{D}_i \cup \{S;\ S \subseteq a_ib_iC_i\}$.

Proof: Observe that for every combinatorial imset and every $n \in \mathbb{N}$ one has $\mathcal{U}_{n\cdot u} = \mathcal{U}_u$, $\mathcal{L}_{n\cdot u} = \mathcal{L}_u$ and $A \perp\!\!\!\perp B \,|\, C\,[n\cdot u]$ iff $A \perp\!\!\!\perp B \,|\, C\,[u]$ for each $\langle A, B|C\rangle \in \mathcal{T}(N)$. Therefore, it suffices to assume that u is a combinatorial imset and prove the proposition by induction on $deg(u)$.

In case $deg(u) = 0$ necessarily $u = 0$ and one can put $r = 0$ and $\mathcal{D}_0 = \mathcal{U}_u = \mathcal{L}_u = \emptyset$. In case $deg(u) \geq 1$ one has $u \neq 0$ by Observation 4.3 and can apply Lemma 4.3 to find $v = u_{\langle a,b|C\rangle} \in \mathcal{E}(N)$ with $\mathcal{L}_v \subseteq \mathcal{L}_u$ such that $w = u - v$ is a combinatorial imset. Of course $\{aC, bC\} \subseteq \mathcal{L}_u,\ a \perp\!\!\!\perp b \,|\, C\,[u]$ and one can observe that $\mathcal{L}_w \subseteq \mathcal{L}_u \cup \{S; S \subseteq abC\}$. Moreover, by Observations 4.3 and 4.1

$$deg(w) = \langle m_*, w\rangle = \langle m_*, u\rangle - \langle m_*, v\rangle = deg(u) - 1\,.$$

In particular, one can apply the induction hypothesis to w an conclude that there exists a sequence $\mathcal{U}_w = \mathcal{F}_0, \ldots, \mathcal{F}_{r-1} = \mathcal{L}_w,\ r - 1 \geq 0$ of descending classes and a sequence $\langle a_i, b_i|C_i\rangle,\ i = 1, \ldots, r-1$ of elementary triplets with $a_i \perp\!\!\!\perp b_i \,|\, C_i\,[w]$ and

$$a_iC_i, b_iC_i \in \mathcal{F}_i, \quad \mathcal{F}_{i-1} = \mathcal{F}_i \cup \{S; S \subseteq a_ib_iC_i\}.$$

Let us put $\mathcal{D}_i = \mathcal{F}_i \cup \mathcal{U}_v \cup \mathcal{L}_u$ for $i = 0, \ldots, r-1$ and for $i = r$ define $\mathcal{D}_r = \mathcal{L}_u$ and $\langle a_r, b_r|C_r\rangle = \langle a, b|C\rangle$. By Observations 4.8 and 4.5 $\mathcal{D}_0 = \mathcal{F}_0 \cup \mathcal{U}_v \cup \mathcal{L}_u = (\mathcal{U}_w \cup \mathcal{U}_v) \cup \mathcal{L}_u = \mathcal{U}_u$. It makes no problem to evidence that $\mathcal{D}_0, \ldots, \mathcal{D}_r$ satisfies the required conditions. Indeed, $a_i \perp\!\!\!\perp b_i \,|\, C_i\,[w]$ implies $a_i \perp\!\!\!\perp b_i \,|\, C_i\,[u]$ for $i \leq r$ and since $\mathcal{L}_w \subseteq \mathcal{L}_u \cup \{S;\ S \subseteq abC\}$ by $u = w + v$ one has $\mathcal{D}_{r-1} = \mathcal{L}_w \cup \mathcal{U}_v \cup \mathcal{L}_u = \mathcal{L}_u \cup \{S; S \subseteq abC\} = \mathcal{D}_r \cup \{S; S \subseteq a_rb_rC_r\}$. □

The significance of the preceding lemma (summarized in the consequence below) is that one can always 'reach' the upper class of a structural imset from its lower class with help of its induced conditional independence statements. Note that 'reverse order' in formulation of Lemma 4.7 (going from the upper class to the lover class) is used because it is more suitable from the point of view of the proof(s).

CONSEQUENCE 4.2 Let u be a structural imset over N. Then every descending system $\mathcal{E} \subseteq \mathcal{T}(N)$ containing $\mathcal{L}_u$ and satisfying

$$\forall\ \langle A, B|C\rangle \in \mathcal{T}(N) \quad A \perp\!\!\!\perp B \,|\, C\ [u] \text{ and } AC, BC \in \mathcal{E} \quad \text{implies} \quad ABC \in \mathcal{E}\,, \tag{4.7}$$

necessarily contains $\mathcal{U}_u$.

Proof: Apply Lemma 4.7 and prove $\mathcal{D}_r \subseteq \mathcal{E}$ by reverse induction on $i = r, \ldots, 0$. □

4.4.2 Markovian measures

Suppose that u is a structural imset over N and P is a probability measure over N. One says that P is *Markovian with respect to* u if

$$A \perp\!\!\!\perp B \,|\, C \,[u] \quad \text{implies} \quad A \perp\!\!\!\perp B \,|\, C \,[P] \quad \text{whenever } \langle A, B|C\rangle \in \mathcal{T}(N)\,.$$

Thus, statistical meaning of an 'imsetal' model is completely analogous to statistical meaning of a graphical model. Every structural imset u over N represents a class of probability measures over N (within the respective framework of measures, e.g. discrete measures, Gaussian measures etc.), namely the class of measures which are Markovian with respect to u. In fact, 'imsetal' models generalize graphical models: given a classic graph there exists a structural imset having the same class of Markovian distributions (for DAG models see Lemma 7.1).

One says that P is *perfectly Markovian with respect to a structural imset* u over N if u induces exactly the conditional independence model induced by P, that is for every $\langle A, B|C\rangle \in \mathcal{T}(N)$ one has

$$A \perp\!\!\!\perp B \,|\, C \,[u] \quad \text{if and only if} \quad A \perp\!\!\!\perp B \,|\, C \,[P]\,.$$

One of the results of this work (Theorem 5.2) is that every probability measure with finite multiinformation is perfectly Markovian with respect to a structural imset. On the other hand, there are 'superfluous' structural imsets whose induced semi-graphoid is not a model induced by any probability measure with finite multiinformation.

EXAMPLE 4.1 There exists a structural imset u over $N = \{a, b, c, d\}$ such that no marginally continuous measure over N is perfectly Markovian with respect to u. Put

$$u = u_{\langle c,d|\{a,b\}\rangle} + u_{\langle a,b|\emptyset\rangle} + u_{\langle a,b|\{c\}\rangle} + u_{\langle a,b|\{d\}\rangle}\,.$$

Evidently $c \perp\!\!\!\perp d \,|\, \{a, b\}\ [u]$, $a \perp\!\!\!\perp b \,|\, \emptyset\ [u]$, $a \perp\!\!\!\perp b \,|\, \{c\}\ [u]$ and $a \perp\!\!\!\perp b \,|\, \{d\}\ [u]$. To show that $a \not\perp\!\!\!\perp b \,|\, \{c, d\}\ [u]$ consider the multiset $m_\dagger$ in Figure 4.3 and observe that $\langle m_\dagger, v\rangle \geq 0$ for every $v \in \mathcal{E}(N)$. Hence, $\langle m_\dagger, w\rangle \geq 0$ for every structural imset w over N. Because $\langle m_\dagger,\ k\ \cdot u - u_{\langle a,b|\{c,d\}\rangle}\rangle = -1$ for every $k \in \mathbb{N}$ the imset $k \cdot u - u_{\langle a,b|\{c,d\}\rangle}$ is not structural. However, by Consequence 2.1 there is no marginally continuous probability measure over N which is perfectly Markovian with respect to u. ♢

Another important consequence of Lemma 4.7 is that marginals of a Markovian measure with respect to a structural imset u for sets in $\mathcal{L}_u$ determine uniquely its marginals for sets in $\mathcal{U}_u$. This motivated the terminology lower and upper class of u introduced in Section 4.2.3. Note that one often has $\mathcal{U}_u = \mathcal{P}(N)$ in which case whole Markovian measure is determined by its marginals on the lower class.

CONSEQUENCE 4.3 Suppose that both P and Q are probability measures on $(\mathsf{X}_N, \mathcal{X}_N)$ which are Markovian with respect to a structural imset u. Then

$$[\, P^S = Q^S \ \text{ for every } S\ \in \mathcal{L}_u\,] \;\Rightarrow\; [\, P^S = Q^S \ \text{ for every } S\ \in \mathcal{U}_u\,]\,.$$

Proof: One can repeat the arguments used in the beginning of the proof of Lemma 2.6 (p. 24) to verify the following 'uniqueness principle'. For every $\langle A, B|C\rangle \in \mathcal{T}(N)$

$$A \perp\!\!\!\perp B \,|\, C\ [P],\ A \perp\!\!\!\perp B \,|\, C\ [Q],\ P^{AC} = Q^{BC},\ P^{BC} = Q^{BC} \;\Rightarrow\; P^{ABC} = Q^{ABC}.$$

Then, owing to the fact that $S \subseteq T$, $P^T = Q^T$ implies $P^S = Q^S$, one can apply Lemma 4.7 and show by reverse induction on $i = r, \ldots, 0$ that $[\, P^S = Q^S\ \text{ for every } S\ \in \mathcal{D}_i\,]$. □

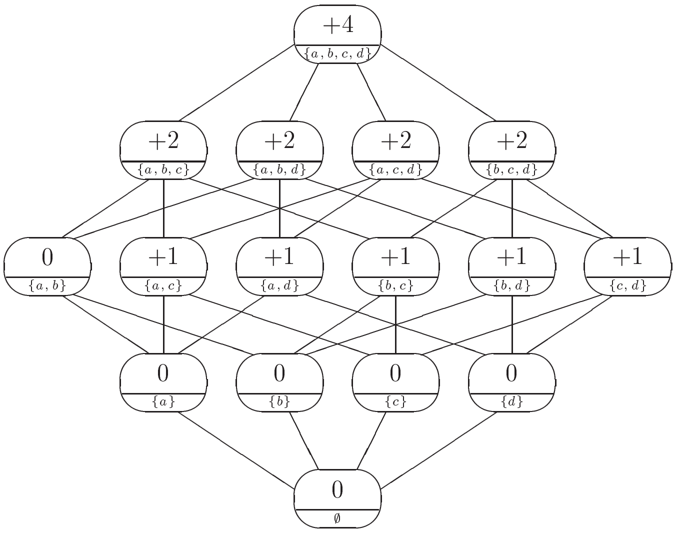


Figure 4.3: The multiset $m_\dagger$ from Example 4.1.

4.5 Equivalence result

The third way of associating a structural imset with a probability measure is an algebraic identity in which the measure is represented by its multiinformation function.

One says that a probability measure P over N with finite multiinformation *complies with a structural imset* u over N if $\langle m_P, u\rangle = 0$ where m_P denotes the multiinformation function defined in Section 2.3.4.

Remark 4.4 The concept above can be introduced alternatively in the following way. Suppose that P is a probability measure over N which has a dominating measure (see p. 19) $\mu = \prod_{i\in N} \mu_i$ such that $-\infty < H(P^S \,|\, \prod_{i\in S} \mu_i) < +\infty$ for each $S \subseteq N$. Note that by Lemma 2.7 P has finite multiinformation then. Thus, one can introduce the *entropy function of P relative to μ* as follows:

$$h_{P,\mu}(S) = -H(P^S \,|\, \prod_{i\in S} \mu_i) \qquad \text{for} \quad \emptyset \neq S \subseteq N$$

and $h_{P,\mu}(\emptyset) = 0$ by convention. Then P complies with a structural imset u over N iff $\langle h_{P,\mu}, u\rangle = 0$. Indeed, (2.18) implies together with the fact that u is o-standardized

$$\sum_{S\subseteq N} m_P(S)\cdot u(S) = -\sum_{S\subseteq N} h_{P,\mu}(S)\cdot u(S) + \sum_{j\in N} h_{P,\mu}(\{j\})\cdot \underbrace{\sum_{S\subseteq N, j\in S} u(S)}_{0}\ ,$$

that is $\langle m_P, u\rangle = -\langle h_{P,\mu}, u\rangle$. Note that in case of a discrete probability measure one can always take the counting measure υ in place of μ. The corresponding entropy function is then non-negative and has very pleasant properties which enable one to characterize functional dependence statements (p. 15) with respect to P [61] (in addition to pure conditional independence statements). Namely, $h_{P,\upsilon}(A) \leq h_{P,\upsilon}(AC)$ for $A, C \subseteq N$ while the equality occurs iff $A \perp\!\!\!\perp A \,|\, C\,[P]$ (c.f. Remark 2.3). However, this pleasant phenomenon seems to be more or less limited to discrete case. It is not clear which dominating measures in general produce entropy functions with behaviour of this type towards functional dependence (except measures concentrated on a countable set). For example, in Gaussian case the entropy function relative to Lebesgue measure need not be non-negative or even monotone. This is the main reason why I prefer multiinformation function to entropy function. The second one is that entropy function does depend on the choice of a suitable dominating measure unlike multiinformation function. $\triangle$

The main result of this chapter says that all three ways of associating structural imsets with probability measures are equivalent. In words, a probability measure complies with a structural imset iff it is Markovian with respect to it or iff the product formula induced by it holds. In the proof below the following simple observation is used.

OBSERVATION 4.9 Suppose the situation from Convention 1 (p. 20). Then

$$\forall\, S \subseteq T \subseteq N \quad f_S(x_S) = 0 \;\Rightarrow\; f_T(x_T) = 0 \quad \text{for } \mu\text{-a.e. } x \in \mathsf{X}_N\,. \tag{4.8}$$

Proof: Combine the arguments used in Remark 2.8 with the formula (2.4) in the proof of Lemma 2.4. $\square$

THEOREM 4.1 *Let u be a structural imset over N, P a probability measure on $(\mathsf{X}_N, \mathcal{X}_N)$ with finite multiinformation. Suppose that $\{\mu_i;\ i \in N\}$ is a reference system of measures for P (p. 61); let us accept Convention 1 on p. 20. Then the following four conditions are equivalent.*

(i) $\prod_{S\subseteq N} f_S(x_S)^{u^+(S)} = \prod_{S\subseteq N} f_S(x_S)^{u^-(S)}$ *for μ-a.e. $x \in \mathsf{X}_N$,*

(ii) $\prod_{S\subseteq N} f_S(x_S)^{u(S)} = 1$ *for P-a.e. $x \in \mathsf{X}_N$,*

(iii) $\langle m_P, u\rangle = 0$,

(iv) *$A \perp\!\!\!\perp B\,|\,C\ [u]$ implies $A \perp\!\!\!\perp B\,|\,C\ [P]$ for every $\langle A, B|C\rangle \in \mathcal{T}(N)$.*

Proof: Implication (i) $\Rightarrow$ (ii) is trivial since $P \ll \mu$ and $f_S(x_S) > 0$ for P-a.e. $x \in \mathsf{X}_N$ and every $S \subseteq N$. To show (ii) $\Rightarrow$ (iii) apply logarithm to the assumed equality first and get

$$\sum_{S\subseteq N} u(S)\cdot \ln f_S(x_S) = \ln\big(\prod_{S\subseteq N} f_S(x_S)^{u(S)}\big) = 0 \qquad \text{for } P\text{-a.e. } x \in \mathsf{X}_N\,.$$

Then by integrating with respect to P (notation is from Convention 1)

$$\sum_{S\subseteq N} u(S)\cdot H(P^S\,|\,\mu_S) = \int_{\mathsf{X}_N} \sum_{S\subseteq N} u(S)\cdot \ln f_S(x_S)\; dP(x) = 0\,.$$

As explained in Remark 4.4 this is equivalent to $\langle m_P, u\rangle = 0$.
To see (iii) $\Rightarrow$ (iv) consider a structural imset $w = k\ \cdot u - u_{\langle A,B|C\rangle}$ with $k \in \mathbb{N}$ and write

$$0 = \langle m_P, k\ \cdot u\rangle = \langle m_P, u_{\langle A,B|C\rangle}\rangle + \langle m_P, w\rangle.$$

By Consequence 2.2, the inequality (2.16), both terms on the right-hand side are non-negative and therefore they vanish. Thus, by (2.17) one has $A \perp\!\!\!\perp B \,|\, C\,[P]$.

Supposing (iv) one already knows that $f_S(x_S) \geq 0$ for every $x \in \mathsf{X}_N$, $S \subseteq N$. Thus, the condition (i) can be proved separately on the set $\mathsf{Y} = \{y \in \mathsf{X}_N;\ \prod_{S\subseteq N} f_S(y_S)^{u^-(S)} = 0\}$ and on the set $\mathsf{Z} = \{z \ \in \mathsf{X}_N;\ \prod_{S\subseteq N} f_S(z_S)^{u^-(S)} > 0\}$. Because of $\mathcal{L}_u \subseteq \mathcal{U}_u$ (Observation 4.5) it follows from (4.8) in Observation 4.9 that

$$\prod_{S\subseteq N} f_S(y_S)^{u^+(S)} = 0 \quad \text{for } \mu\text{-a.e. } \ y \in \mathsf{Y}\,,$$

and both sides of the expression in (i) vanish μ-a.e. on Y.

Suppose now $z \in \mathsf{Z}$ and put $\mathcal{E}_z = \{S \subseteq N;\ f_S(z_S) > 0\}$. Observe that $\mathcal{L}_u \subseteq \mathcal{E}_z$ for every $z \in \mathsf{Z}$ and that $\mathcal{E}_z$ is a descending class for μ-a.e. $z \in \mathsf{Z}$ by (4.8). Having fixed $\langle A, B|C\rangle \in \mathcal{T}(N)$ the assumption $A \perp\!\!\!\perp B \,|\, C\ [u]$ implies by (iv) $A \perp\!\!\!\perp B \,|\, C\ [P]$ and hence by Lemma 2.4 derive that

$$f_{AC}(x_{AC}) \cdot f_{BC}(x_{BC}) > 0 \ \Rightarrow \ f_{ABC}(x_{ABC}) > 0 \qquad \text{for } \mu\text{-a.e. } \ x \in \mathsf{X}_N\,.$$

In particular, for μ-a.e. $z \in \mathsf{Z}$ the fact $AC, BC \in \mathcal{E}_z$ implies $ABC \in \mathcal{E}_z$. Altogether, for μ-a.e. $z \in \mathsf{Z}$ the assumptions of Consequence 4.2 with $\mathcal{E} = \mathcal{E}_z$ are fulfilled and therefore $\mathcal{U}_u \subseteq \mathcal{E}_z$, that is

$$\forall\ S \subseteq N \quad u(S) > 0 \Rightarrow f_S(z_S) > 0 \qquad \text{for } \mu\text{-a.e. } \ z \in \mathsf{Z}\,. \tag{4.9}$$

Since u is a structural imset one has $n \cdot u = \sum_{v\in\mathcal{E}(N)} k_v \cdot v$ for $n \in \mathbb{N}$ and $k_v \in \mathbb{Z}^+$ (see Section 4.2.3). For every $v = u_{\langle i,j|K\rangle} \in \mathcal{E}(N)$ with $k_v > 0$ one has $i \perp\!\!\!\perp j \,|\, K\ [u]$ and therefore by (iv) and (2.3) on p. 21 derives

$$\prod_{S\subseteq N} f_S(x_S)^{v^+(S)} = \prod_{S\subseteq N} f_S(x_S)^{v^-(S)} \qquad \text{for } \mu\text{-a.e. } \ x \in \mathsf{X}_N\,.$$

These equalities can be multiplied each other so that one gets

$$\prod_{S\subseteq N} f_S(z_S)^{\sum_{v\in\mathcal{E}(N)} k_v\cdot v^+(S)} = \prod_{S\subseteq N} f_S(z_S)^{\sum_{v\in\mathcal{E}(N)} k_v\cdot v^-(S)} \quad \text{for } \mu\text{-a.e. } z \in \mathsf{Z}\,. \tag{4.10}$$

Let us introduce the multiset $w = \sum_{v\in\mathcal{E}(N)} k_v \cdot v^+ - n \cdot u^+ = \sum_{v\in\mathcal{E}(N)} k_v \cdot v^- - n \cdot u^-$. For every $S \subseteq N$ the fact $w(S) > 0$ implies $v(S) > 0$ for some $v \in \mathcal{E}(N)$ with $k_v > 0$. Hence, $S \in \mathcal{U}_v \subseteq \mathcal{U}_{n\cdot u} = \mathcal{U}_u$ by Observation 4.8. By application of (4.9) to some $T \supseteq S$ and (4.8) one derives $f_S(z_S) > 0$ for μ-a.e. $z \in \mathsf{Z}$. This consideration implies

$$\prod_{S\subseteq N} f_S(z_S)^{w(S)} > 0 \qquad \text{for } \mu\text{-a.e. } \ z \in \mathsf{Z}\,.$$

Thus, one can divide (4.10) by this non-zero expression for μ-a.e. $z \in \mathsf{Z}$ and conclude that

$$\prod_{S\subseteq N} f_S(z_S)^{n\cdot u^+(S)} = \prod_{S\subseteq N} f_S(z_S)^{n\cdot u^-(S)} \qquad \text{for } \mu\text{-a.e. } z \in \mathsf{Z}\,.$$

Take the n-th root of it and obtain what is desired. □

Let me note that one can always take the universal reference system (p. 61) in Theorem 4.1 which implies that the conditions (iii) and (iv) which are not dependent on the choice of a reference system are always equivalent (for a probability measure with finite multiinformation).

Further comment is that another equivalent definition of conditional independence can be derived from Theorem 4.1. Suppose that P is a probability measure over N with finite multiinformation, $\langle A, B|C\rangle \in \mathcal{T}(N)$ and accept Convention 1. It suffices to put $u = u_{\langle A,B|C\rangle}$ and use (ii) in Theorem 4.1 to see that $A \perp\!\!\!\perp B\,|\,C\;[P]$ iff

$$f_{ABC}(x_{ABC}) = \frac{f_{AC}(x_{AC}) \cdot f_{BC}(x_{BC})}{f_C(x_C)} \qquad \text{for } P\text{-a.e. } \; x \in X_N\,. \tag{4.11}$$

Note that $f_S(x_S) > 0$ for P-a.e. $x \in X_N$.

Chapter 5

Description of probabilistic models

Two basic approaches to description of probabilistic CI structures are dealt with in this chapter. The first one which uses structural imsets was already mentioned in Section 4.4.1. The second one which uses supermodular functions is closely related to the first one. It can also use imsets over N to describe CI models over N but the respective class of imsets and their interpretation are completely different. However, despite formal difference, the approaches are equivalent. In fact, there exists a certain duality relation between these two methods: one approach is complementary to the other (see Section 5.4). The main result of the chapter says that every CI model induced by a probability measure with finite multiinformation can be described both by a structural imset and by a supermodular function.

5.1 Supermodular set functions

A real set function $m : \mathcal{P}(N) \to \mathbb{R}$ is called a *supermodular function over* N if

$$m(U \cup V) + m(U \cap V) \geq m(U) + m(V) \qquad \text{for every} \quad U, V \subseteq N. \tag{5.1}$$

The class of all supermodular functions on $\mathcal{P}(N)$ will be denoted by $\mathcal{K}(N)$. The definition can be formulated in several equivalent ways.

Observation 5.1 A set function $m : \mathcal{P}(N) \to \mathbb{R}$ is supermodular iff any of the following three conditions holds:

(i) $\langle m, u \rangle \geq 0$ for every structural imset u over N,

(ii) $\langle m, u \rangle \geq 0$ for every semi-elementary imset u over N,

(iii) $\langle m, u \rangle \geq 0$ for every elementary imset $u \in \mathcal{E}(N)$.

Proof: Evidently (i) $\Rightarrow$ (ii) $\Rightarrow$ (iii). The implication (iii) $\Rightarrow$ (i) follows from the definition of structural imset (see Section 4.2.3 p. 59) and linearity of scalar product. The condition (5.1) is equivalent to the requirement $\langle m, u_{\langle A,B|C \rangle} \rangle \geq 0$ for every $\langle A, B|C \rangle \in \mathcal{T}(N)$ which is nothing but (ii). $\square$

Further evident observation is as follows.

Observation 5.2 The class of supermodular functions $\mathcal{K}(N)$ is a cone:

$$\forall\, m_1, m_2 \in \mathcal{K}(N) \quad \forall\, \alpha, \beta \geq 0 \quad \alpha \cdot m_1 + \beta \cdot m_2 \in \mathcal{K}(N)\,. \tag{5.2}$$

Remark 5.1 This is to warn the reader that a different terminology is used in game theory, where supermodular set functions are named either 'convex set functions' [86] or even 'convex games' [91]. I followed that terminology in some of my former reports [101, 108]. However, another common term 'supermodular' is used here in order to avoid confusion with usual meaning of the adjective 'convex' in mathematics. As mentioned in [13] supermodular functions are also named '2-monotone Choquet capacities'. $\triangle$

5.1.1 Semi-graphoid produced by a supermodular function

One says that a disjoint triplet $\langle A, B|C\rangle \in \mathcal{T}(N)$ is *represented in a supermodular function* m over N and writes $A \perp\!\!\!\perp B \,|\, C\ [m]$ if $\langle m, u_{\langle A,B|C\rangle}\rangle = 0$. The class of represented triplets then defines the *model produced by* m

$$\mathcal{M}^m = \{\, \langle A, B|C\rangle \in \mathcal{T}(N)\,; \quad A \perp\!\!\!\perp B \,|\, C\ [m]\,\}\,.$$

Two supermodular functions over N are *model equivalent* if they represent the same class of disjoint triplets over N.

Remark 5.2 This is to explain terminology. I usually say that a model is <u>induced</u> by a mathematical object over N (see Section 2.2.1), for example by a probability measure over N or by a graph over N (see Chapter 3). However, in this chapter and in subsequent chapters I need to distinguish two different ways of inducing formal independence models by imsets. Both ways appear to be equivalent as concerns the class of obtained models (see Consequence 5.4). The problem is that some imsets (e.g. zero imset or $u_{\langle a,b|\emptyset\rangle}$ in case $N = \{a, b\}$) may 'induce' different models depending on the way of 'inducing'. To prevent misunderstanding I decided to emphasize the difference both in terminology (*induced* versus *produced*) and in notation ($\mathcal{M}_u$ versus $\mathcal{M}^m$). Regretably, I have to confess that the adjective 'induced' was used in a former report [116] also for supermodular functions over N. $\triangle$

Basic fact is this.

Lemma 5.1 A supermodular function over N produces a disjoint semi-graphoid over N.

Proof: This follows easily from respective formulas for semi-elementary imsets and linearity of scalar pruduct. Let m be a supermodular function over N. For triviality property realize $\langle m, u_{\langle A,\emptyset|C\rangle}\rangle = \langle m, 0\rangle = 0$, for symmetry $\langle m, u_{\langle A,B|C\rangle}\rangle = \langle m, u_{\langle B,A|C\rangle}\rangle$. The formula

$$\langle m, u_{\langle A,BD|C\rangle}\rangle = \langle m, u_{\langle A,B|DC\rangle}\rangle + \langle m, u_{\langle A,D|C\rangle}\rangle \tag{5.3}$$

implies directly contraction. To verify decomposition and weak union use Observation 5.1 which says that both terms on the right-hand side of (5.3) are non-negative. $\square$

Typical example of a supermodular set function is the multiinformation function introduced in Section 2.3.4. In fact, Consequence 2.2 on p. 26 says the following.

Observation 5.3 Given a probability measure P over N with finite multiinformation the multiinformation function m_P is an ℓ-standardized supermodular function.

One can conclude even more.

Consequence 5.1 Given a probability measure P over N with finite multiinformation there exists an ℓ-standardized supermodular function m such that $\mathcal{M}_P = \mathcal{M}^m$.

Proof: Let us put $m = m_P$. The relation (2.17) from Consequence 2.2 says

$$A \perp\!\!\!\perp B \,|\, C \; [P] \;\Leftrightarrow\; A \perp\!\!\!\perp B \,|\, C \; [m_P] \qquad \text{for every} \;\; \langle A, B|C\rangle \in \mathcal{T}(N)$$

which implies the desired fact. □

Note that the value $\langle m_P, u_{\langle A,B|C\rangle}\rangle$ for a probability measure P and a disjoint triplet $\langle A, B|C\rangle$ is nothing but the relative entropy of P^{ABC} with respect to the conditional product of P^{AC} and P^{BC} (see the proof of Consequence 2.2 on p. 26). This number can be interpreted (in discrete case) as a numerical evaluation of the degree of stochastic conditional dependence between A and B given C with respect to P [115]. Thus, given a supermodular function m over N and $\langle A, B|C\rangle \in \mathcal{T}(N)$ the non-negative value $\langle m, u_{\langle A,B|C\rangle}\rangle$ could be interpreted as a generalized degree of dependence between A and B given C with respect to m. Having in mind this point of view there is no reason to distinguish between two supermodular functions for which scalar products with semi-elementary imsets coincide. This motivated the next definition.

5.1.2 Strong equivalence of supermodular functions

One says that two supermodular functions m_1 and m_2 over N are *strongly equivalent* if

$$\langle m_1, u_{\langle A,B|C\rangle}\rangle = \langle m_2, u_{\langle A,B|C\rangle}\rangle \quad \text{for every} \;\; \langle A, B|C\rangle \in \mathcal{T}(N)\,. \tag{5.4}$$

Obviously, m_1 and m_2 are then model equivalent. Strong equivalence can be equivalently described with help of a special class of functions, namely the class of functions inducing the maximal independence model $\mathcal{T}(N)$. A function $l : \mathcal{P}(N) \to \mathbb{R}$ is called *modular* if

$$l(U \cup V) + l(U \cap V) = l(U) + l(V) \quad \text{for every} \;\; U, V \subseteq N. \tag{5.5}$$

The class of modular functions over N will be denoted by $\mathcal{L}(N)$. Evidently $\mathcal{L}(N) \subseteq \mathcal{K}(N)$.

Observation 5.4 The only ℓ-standardized modular function is the zero function.

Proof: Indeed, supposing that $l : \mathcal{T}(N) \to \mathbb{R}$ is ℓ-standardized modular function one can show by induction on $|S|$ that $l(S) = 0$ for every $S \subseteq N$. This is evident in case $|S| \leq 1$. If $|S| \geq 2$ then take $u_{\langle i,j|K\rangle} \in \mathcal{E}(N)$ such that $S = ijK$. The fact $\langle l, u_{\langle i,j|K\rangle}\rangle = 0$ says

$$l(S) = l(iK) + l(jK) - l(K)$$

and the right-hand side of this equality vanishes by the induction hypothesis. □

Lemma 5.2 A supermodular function m produces $\mathcal{T}(N)$ iff $m \in \mathcal{L}(N)$. Two supermodular functions m_1, m_2 over N are strongly equivalent iff $m_1 - m_2 \in \mathcal{L}(N)$. Every supermodular function is strongly equivalent to an ℓ-standardized supermodular function. The class $\mathcal{L}(N)$ is a linear subspace of dimension $|N| + 1$. The functions $m^{\emptyset\uparrow}$ and $m^{\{i\}\uparrow}$ for $i \in N$ (see p. 34) form its linear base.

Proof: Clearly, $m : \mathcal{P}(N) \to \mathbb{R}$ is modular iff both m and $-m$ are supermodular. Hence, by Observation 5.1(ii) one has $m \in \mathcal{L}(N)$ iff $\langle m, u \rangle = 0$ for every semi-elementary imset u which means $\mathcal{M}^m = \mathcal{T}(N)$. On the other hand because of linearity of scalar product two supermodular functions m_1 and m_2 are strongly equivalent iff $\langle m_1 - m_2, u \rangle = 0$ for every semi-elementary imset u over N.

Let m be a supermodular function over N. The function

$$\widetilde{m} \equiv m - m(\emptyset) \cdot m^{\emptyset\uparrow} - \sum_{i \in N} \{m(\{i\}) - m(\emptyset)\} \cdot m^{\{i\}\uparrow} \tag{5.6}$$

is evidently ℓ-standardized and supermodular as $m^{\emptyset\uparrow} \in \mathcal{L}(N)$ and $m^{\{i\}\uparrow} \in \mathcal{L}(N)$ for $i \in N$.

Of course, $\mathcal{L}(N)$ is a linear subspace. Observe that $m^{\emptyset\uparrow}$ and $m^{\{i\}\uparrow}$ for $i \in N$ are linearly independent. To show that they generate $\mathcal{L}(N)$ take $m \in \mathcal{L}(N)$ and introduce $\widetilde{m}$ by means of the formula (5.6). By Observation 5.4 $\widetilde{m} = 0$. □

Remark 5.3 Thus, to have a clear view on 'quantitative' dependence structures produced by supermodular functions one should choose one representative from every class of strong equivalence in a systematic way. The choice should follow relevant mathematical principles: to have geometric insight one should do the choice 'linearly'. This can be made as follows. Take a linear subspace $\mathcal{S}(N) \subseteq \mathbb{R}^{\mathcal{P}(N)}$ such that $\mathcal{S}(N) \cap \mathcal{L}(N) = \{0\}$ and $\mathcal{S}(N) \oplus \mathcal{L}(N) = \mathbb{R}^{\mathcal{P}(N)}$. Then every $m \in \mathcal{K}(N)$ can be uniquely decomposed: $m = s + l$ where $s \in \mathcal{S}(N)$, $l \in \mathcal{L}(N)$. The fact $-\mathcal{L}(N) \subseteq \mathcal{K}(N)$ and Observation 5.2 implies $s \in \mathcal{K}(N)$. Moreover, s is strongly equivalent to m by Lemma 5.2 and the function $s \in \mathcal{K}(N) \cap \mathcal{S}(N)$ conicides for strongly equivalent functions $m \in \mathcal{K}(N)$.

However, there is flexibility in the choice of $\mathcal{S}(N)$. Fixing on a space $\mathcal{S}(N)$ satisfying the above requirements means that one restricts attention to this linear subspace and represents the class of supermodular functions $\mathcal{K}(N)$ by respective class of standardized supermodular functions $\mathcal{K}(N) \cap \mathcal{S}(N)$. In this work only three ways of standardization are mentioned (they are justified by some theoretical reasons). Preferred standardization using the linear subspace

$$\mathcal{S}_\ell(N) = \{\, m \in \mathbb{R}^{\mathcal{P}(N)} \,;\; m(S) = 0 \;\text{ whenever }\; |S| \leq 1 \,\}$$

is in concordance with the property (2.15) of multiinformation functions (see p. 26). Functions $m \in \mathcal{K}(N) \cap \mathcal{S}_\ell(N)$ are *non-decreasing*: $m(S) \leq m(T)$ whenever $S \subseteq T$ (see Consequence 2.2). In particular, they are non-negative.

However, from purely mathematical point of view another standardization which uses the subspace

$$\mathcal{S}_u(N) = \{\, m \in \mathbb{R}^{\mathcal{P}(N)} \,;\; m(S) = 0 \;\text{ whenever }\; |S| \geq |N| - 1 \,\}$$

is equally entitled. This standardization can be viewed as 'reflection' the of former one since composition with the mapping $S \longmapsto N \setminus S$, $S \subseteq N$ on $\mathcal{P}(N)$ transforms

$\mathcal{K}(N) \cap \mathcal{S}_\ell(N)$ onto $\mathcal{K}(N) \cap \mathcal{S}_u(N)$. Thus, this standardization leads to non-increasing standardized supermodular function, which are non-negative as well.

The third natural option is to take the orthogonal complement of $\mathcal{L}(N)$

$$\mathcal{S}_o(N) = \{\, m \in \mathbb{R}^{\mathcal{P}(N)}\,;\ \sum_{S\subseteq N} m(S) = 0 \ \text{ and } \sum_{S\subseteq N, i\in S} m(S) = 0 \ \text{ for } i \in N \,\}\,.$$

Note that every independence model produced by a supermodular function is even produced by a supermodular imset (see Consequence 5.4 in Section 5.3). Thus, standardizations of imsets mentioned in Section 2.4 are just the standardization of supermodular functions. The letters ℓ, u, o distinguish the types of standardization: ℓ-standardization means that the '*lower*' part of the respective diagram of an imset is 'vanished', u-standardization means that the '*upper*' part is 'vanished' and o-standardization means that the respective linear space is the *orthogonal* complement of $\mathcal{L}(N)$. △

5.2 Skeletal supermodular functions

A supermodular function m over N will be called *skeletal* if $\mathcal{M}^m \subset \mathcal{T}(N)$ but there is no supermodular function r over N such that $\mathcal{M}^m \subset \mathcal{M}^r \subset \mathcal{T}(N)$. Thus, a supermodular function is skeletal iff it produces 'submaximal' independence model. The definition implies that a supermodular function which is model equivalent to a skeletal function is also skeletal. In particular, strong equivalence has the same property. Of course, model equivalence decomposes the collection of skeletal functions into finitely many equivalence classes. The aim of this section is to characterize these equivalence classes. To have a clear geometric view on the problem it is suitable to simplify the situation with help of ℓ-standardization mentioned in Remark 5.3.

Introduce the class of *ℓ-standardized supermodular functions* $\mathcal{K}_\ell(N) = \mathcal{K}(N) \cap \mathcal{S}_\ell(N)$. Basic observation is that $\mathcal{K}(N)$ is a direct sum of $\mathcal{K}_\ell(N)$ and $\mathcal{L}(N)$, in notation $\mathcal{K}(N) = \mathcal{K}_\ell(N) \oplus \mathcal{L}(N)$.

Observation 5.5 $\mathcal{K}_\ell(N) \cap \mathcal{L}(N) = \{0\}$ and every $m \in \mathcal{K}(N)$ has unique decomposition $m = \widetilde{m} + l$ where $\widetilde{m} \in \mathcal{K}_\ell(N)$ and $l \in \mathcal{L}(N)$.

Proof: Put $l = m(\emptyset)\cdot m^{\emptyset\uparrow} + \sum_{i\in N}\{\, m(\{i\}) - m(\emptyset)\,\}\cdot m^{\{i\}\uparrow}$. By Lemma 5.2 $l \in \mathcal{L}(N)$. As $(-l) \in \mathcal{L}(N) \subseteq \mathcal{K}(N)$ by Observation 5.2 $\widetilde{m} \equiv m + (-l) \in \mathcal{K}(N)$. The facts $l(\emptyset) = m(\emptyset)$ and $l(\{i\}) = m(\{i\})$ for $i \in N$ imply that $\widetilde{m}$ is ℓ-standardized. The uniqueness of the decomposition follows from Observation 5.4 since $\mathcal{L}(N) \cap \mathcal{K}_\ell(N) = \mathcal{L}(N) \cap \mathcal{S}_\ell(N)$. □

The following lemma summarizes substantial facts concerning $\mathcal{K}_\ell(N)$ (for related concepts see Section 10.8.2).

Lemma 5.3 The set $\mathcal{K}_\ell(N)$ is a pointed rational polyhedral cone in $\mathbb{R}^{\mathcal{P}(N)}$. In particular, it has finitely many extreme rays and every extreme ray of $\mathcal{K}_\ell(N)$ contains exactly one non-zero normalized imset (see p. 36). The set $\mathcal{K}_\ell(N)$ is a conical closure of this collection of normalized imsets.

Proof: To evidence that $\mathcal{K}_\ell(N)$ is a rational polyhedral cone observe that it is the dual cone $\mathcal{F}^* = \{m \in \mathbb{R}^{\mathcal{P}(N)}\,;\ \langle m, u\rangle \geq 0 \text{ for } u \in \mathcal{F}\}$ to a finite set $\mathcal{F} \subseteq \mathbb{Q}^{\mathcal{P}(N)}$, namely to

$$\mathcal{F} = \mathcal{E}(N) \cup \{\delta_\emptyset, -\delta_\emptyset\} \cup \bigcup_{i\in N}\{\delta_{\{i\}}, -\delta_{\{i\}}\}.$$

The fact that it is pointed, that is $\mathcal{K}_\ell(N) \cap (-\mathcal{K}_\ell(N)) = \mathcal{L}(N) \cap \mathcal{S}_\ell(N) = \{0\}$ follows from Observation 5.4. All remaining statements of Lemma 5.3 follow from well-known properties of pointed rational polyhedral cones gathered in Section 10.8.2. Observe that every (extreme) ray of $\mathcal{K}_\ell(N)$ which contains a non-zero element of $\mathbb{Q}^{\mathcal{P}(N)}$ must contain a non-zero element of $\mathbb{Z}^{\mathcal{P}(N)}$, that is a non-zero imset. But only one non-zero imset within the ray is normalized. □

5.2.1 Skeleton

Let us denote by $\mathcal{K}_\ell^\diamond(N)$ the collection of non-zero normalized imsets belonging to extremal rays of $\mathcal{K}_\ell(N)$ and call this set the *ℓ-skeleton* over N. It is empty in case $|N| = 1$. The first important observation concerning $\mathcal{K}_\ell^\diamond(N)$ is the following one.

LEMMA 5.4 An imset u over N is structural iff it is o-standardized and $\langle m, u\rangle \geq 0$ for every $m \in \mathcal{K}_\ell^\diamond(N)$.

Proof: The necessity of the conditions follows from Observations 4.4 and 5.1(i). For sufficiency suppose that $u \in \mathbb{Z}^{\mathcal{P}(N)}$ is o-standardized and $\langle m, u\rangle \geq 0$ for any $m \in \mathcal{K}_\ell^\diamond(N)$. The fact that u is o-standardized means that $\langle m^{\emptyset\uparrow}, u\rangle = 0$ and $\langle m^{\{i\}\uparrow}, u\rangle = 0$ for $i \in N$. Hence, by Lemma 5.2 derive that $\langle l, u\rangle = 0$ for every $l \in \mathcal{L}(N)$. The fact $\mathcal{K}_\ell(N) = con(\mathcal{K}_\ell^\diamond(N))$ (see Lemma 5.3) implies that $\langle m, u\rangle \geq 0$ for every $m \in \mathcal{K}_\ell(N)$. Hence, by Observation 5.5 $\mathcal{K}(N) = \mathcal{K}_\ell(N) \oplus \mathcal{L}(N)$ implies $\langle m, u\rangle \geq 0$ for every $m \in \mathcal{K}(N)$, i.e. u belongs to the dual cone $\mathcal{K}(N)^*$. However, $\mathcal{K}(N)$ was introduced as the dual cone $\mathcal{E}(N)^*$ in $\mathbb{R}^{\mathcal{P}(N)}$- see Observation 5.1(iii). This says $u \in \mathcal{E}(N)^{**}$, but $\mathcal{E}(N)^{**}$ is nothing but the conical closure $con(\mathcal{E}(N))$ - see Section 10.8.2. Hence $u \in con(\mathcal{E}(N)) \cap \mathbb{Z}^{\mathcal{P}(N)}$ and by Fact from Section 10.8.2 u is a combination of elementary imsets with non-negative rational coefficients. Therefore, it is a structural imset - see Section 4.2.3. □

The following consequence of Lemma 5.4 will be utilized later.

CONSEQUENCE 5.2 Let u be a structural imset over N and $\langle A, B|C\rangle \in \mathcal{T}(N)$. Then $A \perp\!\!\!\perp B \,|\, C\ [u]$ iff $\forall\, r \in \mathcal{K}_\ell^\diamond(N)\ \ \langle r, u_{\langle A,B|C\rangle}\rangle > 0$ implies $\langle r, u\rangle > 0$.

Proof: Since both u and $v \equiv u_{\langle A,B|C\rangle}$ are o-standardized, $w_k \equiv k \cdot u - v$ is o-standardized for every $k \in \mathbb{N}$. By Lemma 5.4 w_k is structural iff $\langle r, k \cdot u - v\rangle \geq 0$ for every $r \in \mathcal{K}_\ell^\diamond(N)$. Thus, by definition of $\mathcal{M}_u$ on p. 64 $A \perp\!\!\!\perp B \,|\, C\ [u]$ iff

$$\exists\, k \in \mathbb{N} \quad \forall\, r \in \mathcal{K}_\ell^\diamond(N) \quad k \cdot \langle r, u\rangle \geq \langle r, v\rangle\,. \tag{5.7}$$

This clearly implies that

$$\forall\, r \in \mathcal{K}_\ell^\diamond(N) \qquad \langle r, v\rangle > 0 \ \ \Rightarrow\ \ \langle r, u\rangle > 0\,. \tag{5.8}$$

Conversely, supposing (5.8) observe that $\forall\, r \in \mathcal{K}_\ell^\diamond(N)$ there exists $k_r \in \mathbb{N}$ such that $k \cdot \langle r, u\rangle \geq \langle r, v\rangle$ for any $k \in \mathbb{N}$, $k \geq k_r$. Indeed, owing to Observation 5.1 $k_r = 1$ in case $\langle r, v\rangle = 0$ and k_r is the least integer greater than $\frac{\langle r,v\rangle}{\langle r,u\rangle}$ in case $\langle r, v\rangle > 0$. As $\mathcal{K}_\ell^\diamond(N)$ finite one can put $k = \max\{k_r; r \in \mathcal{K}_\ell^\diamond(N)\}$ to evidence (5.7). □

An important auxiliary result is the following 'separation' lemma.

LEMMA 5.5 For every $m \in \mathcal{K}_\ell^\diamond(N)$ there exists a structural imset $u \in \mathcal{S}(N)$ such that $\langle m, u\rangle = 0$ and $\langle r, u\rangle > 0$ for any other $r \in \mathcal{K}_\ell^\diamond(N) \setminus \{m\}$. Moreover, for every pair $m, r \in \mathcal{K}_\ell^\diamond(N)$, $m \neq r$ there exists an elementary imset $v \in \mathcal{E}(N)$ such that $\langle m, v\rangle = 0$ and $\langle r, v\rangle > 0$. Consequently, $\mathcal{M}^m \setminus \mathcal{M}^r \neq \emptyset \neq \mathcal{M}^r \setminus \mathcal{M}^m$ for distinct $m, r \in \mathcal{K}_\ell^\diamond(N)$.

Proof: By Lemma 5.3 $\mathcal{K}_\ell(N)$ is a pointed rational polyhedral cone. It can be viewed as a cone in $\mathbb{R}^{\mathcal{P}_*(N)}$ where $\mathcal{P}_*(N) = \{T \subseteq N, |T| \geq 2\}$. Observe that this change of standpoint does not influence the concept of extreme ray and ℓ-skeleton. One can apply Lemma from Section 10.8.2 to the extreme ray generated by m. The respective $q \in \mathbb{Q}^{\mathcal{P}_*(N)}$ can be multiplied by a natural number to get $u \in \mathbb{Z}^{\mathcal{P}_*(N)}$. This integer-valued function on $\mathcal{P}_*(N)$ can be extended to an o-standardized imset over N by means of the formulas

$$u(\{i\}) = -\sum_{S, \{i\} \subseteq S} u(S) \quad \text{for } i \in N, \qquad u(\emptyset) = -\sum_{S, S \neq \emptyset} u(S).$$

As every element of $\mathcal{K}_\ell^\diamond(N)$ is ℓ-standardized the obtained imset u satisfies the required conditions: it is a structural imset by Lemma 5.4. The existence of $v \in \mathcal{E}(N)$ is a clear consequence of the existence of u since $n \cdot u = \sum_{v \in \mathcal{E}(N)} k_v \cdot v$ for some $k_v \in \mathbb{Z}^+$, $n \in \mathbb{N}$. Indeed, linearity of scalar product and the fact $\langle r, u\rangle > 0$ implies that $k_v > 0$ and $\langle r, v\rangle > 0$ for some $v \in \mathcal{E}(N)$. Moreover, $\langle m, v\rangle = 0$ by Observation 5.1. □

However, the main lemma of this section is the following proposition.

LEMMA 5.6 A function $m \in \mathcal{K}_\ell(N)$ is skeletal iff it is non-zero function belonging to an extreme ray of $\mathcal{K}_\ell(N)$.

Proof: For necessity suppose that m is skeletal. Then $m \neq 0$ because $\mathcal{M}^m \neq \mathcal{T}(N)$. By Lemma 5.3 write

$$m = \sum_{r \in \mathcal{K}_\ell^\diamond(N)} \alpha_r \cdot r \qquad \text{for some } \alpha_r \geq 0\,. \tag{5.9}$$

Since $m \neq 0$ there exists $r \in \mathcal{K}_\ell^\diamond(N)$ such that $\alpha_r > 0$. Linearity of scalar product with help of Observation 5.1 says that $\langle m, u\rangle = 0$ implies $\langle r, u\rangle = 0$ for every semi-elementary imset u over N. Thus $\mathcal{M}^m \subseteq \mathcal{M}^r$. The fact $r \in \mathcal{K}_\ell^\diamond(N)$ implies $\mathcal{M}^r \neq \mathcal{T}(N)$ by Lemma 5.2 and Observation 5.4. The assumption that m is skeletal forces $\mathcal{M}^m = \mathcal{M}^r$. By Lemma 5.5 at most one $r \in \mathcal{K}_\ell^\diamond(N)$ with $\mathcal{M}^r = \mathcal{M}^m$ exists. Thus, (5.9) says $m = \alpha_r \cdot r$ for some $r \in \mathcal{K}_\ell^\diamond(N)$ and $\alpha_r > 0$.

For necessity suppose that $m \neq 0$ belongs to an extreme ray R of $\mathcal{K}_\ell(N)$. The fact $m \neq 0$ implies by Lemma 5.2 with help of Observation 5.5 $\mathcal{M}^m \neq \mathcal{T}(N)$. Suppose that r is a supermodular function with $\mathcal{M}^m \subseteq \mathcal{M}^r$. The aim is to show that either $\mathcal{M}^r = \mathcal{M}^m$ or $\mathcal{M}^r = \mathcal{T}(N)$. By Lemma 5.2 r is strongly equivalent an ℓ-standardized supermodular function. Therefore assume without loss of generality $r \in \mathcal{K}_\ell(N)$. The assumption $\mathcal{M}^m \subseteq \mathcal{M}^r$ says $\langle m, u\rangle = 0 \Rightarrow \langle r, u\rangle = 0$ for every semi-elementary imset u over N. Thus, by Observation 5.1 $\langle r, u\rangle > 0$ implies $\langle m, u\rangle > 0$. This means that there exists $k_u \in \mathbb{N}$ with $k_u \cdot \langle m, u\rangle \geq \langle r, u\rangle$. Since the class of semi-elementary imsets is finite there exists $k \in \mathbb{N}$ such that $k \cdot \langle m, u\rangle \geq \langle r, u\rangle$ for every semi-elementary imset u over N. By linearity of scalar product and Observation 5.1 conclude that $k \cdot m - r$ is supermodular. Since both m and r are ℓ-standardized $k \cdot m - r \in \mathcal{K}_\ell(N)$. The assumption that R is extreme ray of $\mathcal{K}_\ell(N)$ and decomposition $k \cdot m = (k \cdot m - r) + r$ implies that

$r \in \mathsf{R}$. Thus, $r = \alpha \cdot m$ for $\alpha \geq 0$. If $\alpha = 0$ then $r = 0$ says $\mathcal{M}^r = \mathcal{T}(N)$, if $\alpha > 0$ then necessarily $\mathcal{M}^r = \mathcal{M}^m$. $\square$

Hence the desired characterization of model equivalence classes of skeletal imsets is obtained.

CONSEQUENCE 5.3 Every class of model equivalence of skeletal supermodular functions over N is characterized by unique element of ℓ-skeleton $\mathcal{K}_\ell^\diamond(N)$ belonging to the class. Given $m \in \mathcal{K}_\ell^\diamond(N)$ the respective equivalence class consists of functions

$$\widetilde{m} = \alpha \cdot m + l \qquad \text{where} \;\; \alpha > 0,\; l \in \mathcal{L}(N)\,.$$

In particular, every skeletal function is model equivalent to a skeletal imset and $\mathcal{K}_\ell^\diamond(N)$ characterizes all skeletal functions.

Proof: Given a skeletal function $r \in \mathcal{K}(N)$ by Observation 5.5 and Lemma 5.2 find unique strongly equivalent skeletal function $\tilde{r} \in \mathcal{K}_\ell(N)$ and apply Lemma 5.6 to find $m \in \mathcal{K}_\ell^\diamond(N)$ and $\alpha > 0$ with $\tilde{r} = \alpha \cdot m$. The fact that m is the unique model equivalent element of the ℓ-skeleton follows from Lemma 5.5. $\square$

5.2.2 Significance of skeletal imsets

One of the main results of this chapter is the following theorem which explains the significance of the concept of ℓ-skeleton.

THEOREM 5.1 *There exists the least finite set of normalized ℓ-standardized imsets $\mathcal{N}(N)$ such that for every imset u over N*

$$u \text{ is structural} \quad \Leftrightarrow \quad u \text{ is o-standardized and } \langle m, u\rangle \geq 0 \text{ for every } m \in \mathcal{N}(N). \tag{5.10}$$

Moreover, $\mathcal{N}(N)$ is nothing but $\mathcal{K}_\ell^\diamond(N)$.

Proof: Lemma 5.4 says that $\mathcal{K}_\ell^\diamond(N)$ is a finite set of normalized ℓ-standardized imsets satisfying (5.10). Let $\mathcal{N}(N)$ be any finite class of this type; the aim is to show that $\mathcal{K}_\ell^\diamond(N) \subseteq \mathcal{N}(N)$. Suppose for contradiction that $m \in \mathcal{K}_\ell^\diamond(N) \setminus \mathcal{N}(N)$. By Lemma 5.5 there exists a structural imset u over N such that $\langle m, u\rangle = 0$ and $\langle r, u\rangle > 0$ for any other $r \in \mathcal{K}_\ell^\diamond(N)$. Basic observation is that $\langle s, u\rangle > 0$ for every $s \in \mathcal{N}(N), s \neq 0$.

Indeed, (5.10) implies with help of Observation 5.1 that s is supermodular and therefore $s \in \mathcal{K}_\ell(N)$. By Lemma 5.3 write $s = \sum_{r \in \mathcal{K}_\ell^\diamond(N)} \alpha_r \cdot r$ for $\alpha_r \geq 0$. Observe that $\alpha_{\bar{r}} > 0$ for some $\bar{r} \in \mathcal{K}_\ell^\diamond(N) \setminus \{m\}$ since otherwise $s = \alpha_m \cdot m$ for $\alpha_m > 0$ (as $s \neq 0$) and the fact that both s and m are normalized imsets implies $s = m$ which contradicts $m \notin \mathcal{N}(N)$. Hence, by Observation 5.1

$$\langle s, u\rangle = \sum_{r \in \mathcal{K}_\ell^\diamond(N)} \alpha_r \cdot \langle r, u\rangle \geq \alpha_{\bar{r}} \cdot \langle \bar{r}, u\rangle > 0\,.$$

Further step is to take an o-standardized imset w over N such that $\langle m, w\rangle < 0$ and put $v_k = k \cdot u + w$ for every $k \in \mathbb{N}$. The inequality $\langle m, v_k\rangle = \langle m, w\rangle < 0$ implies by Lemma 5.4 that v_k is not a structural imset over N. On the other hand, for every $0 \neq s \in \mathcal{N}(N)$ one has $\langle s, u\rangle > 0$ and therefore there exists $k_s \in \mathbb{N}$ with $\langle s, v_{k_s}\rangle = k_s \cdot \langle s, u\rangle + \langle s, w\rangle \geq 0$.

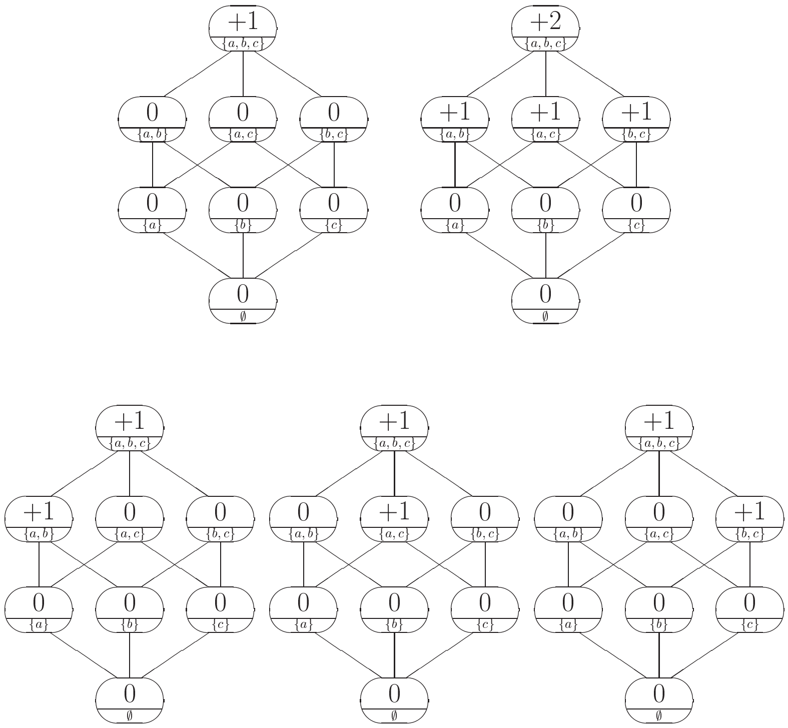


Figure 5.1: ℓ-skeleton over $N = \{a, b, c\}$.

Since $\mathcal{N}(N)$ is finite there exists $k \in \mathbb{N}$ such that $\langle s, v_k \rangle \geq 0$ for every $s \in \mathcal{N}(N)$. By the fact that v_k is o-standardized and (5.10) derive that v_k is a structural imset which contradicts the conclusion above. Thus, there is no $m \in \mathcal{K}_\ell^\diamond(N) \setminus \mathcal{N}(N)$ and the desired inclusion $\mathcal{K}_\ell^\diamond(N) \subseteq \mathcal{N}(N)$ was verified. $\square$

Remark 5.4 The number of elements of ℓ-skeleton $\mathcal{K}_\ell^\diamond(N)$ depends on $|N|$. In trivial case $|N| = 1$ one has $|\mathcal{K}_\ell^\diamond(N)| = 0$. The simplest non-trivial case $|N| = 2$ is not interesting because $|\mathcal{K}_\ell^\diamond(N)| = 1$ then: the cone $\mathcal{K}_\ell^\diamond(N)$ consists of a single ray generated by δ_N in that case. However, in case $N = \{a, b, c\}$ the ℓ-skeleton has already 5 imsets (see Example 1 in [101]). Figure 5.1 gives their list. Thus, in case $|N| = 3$ one needs to check only 5 inequalities to find out whether an o-standardized imset is structural. In case $|N| = 4$ ℓ-skeleton has 37 imsets; Hasse diagrams of ten basic types of skeletal imsets are in the Appendix of [116] - for the proof see [101]. However, the ℓ-skeleton in case $|N| = 5$ was found by means of a computer; it has 117978 imsets - see [116] and

http://www.utia.cas.cz/user_data/studeny/fivevar.htm

for a related virtual catalogue. Note that the problem of suitable characterization of skeletal imsets remains open - see Theme 4 in Section 8.1.2. $\triangle$

Remark 5.5 The name skeleton was inspired by the idea that the collection of extreme rays of $\mathcal{K}_\ell^\diamond(N)$ can be viewed as outer skeleton of the cone $\mathcal{K}_\ell(N)$. I used this word in [108] to name the least finite set of normalized imsets defining a pointed rational polyhedral cone as its dual cone; that is the ℓ-skeleton in case of the conical closure $\mathcal{E}(N)$. Another possible

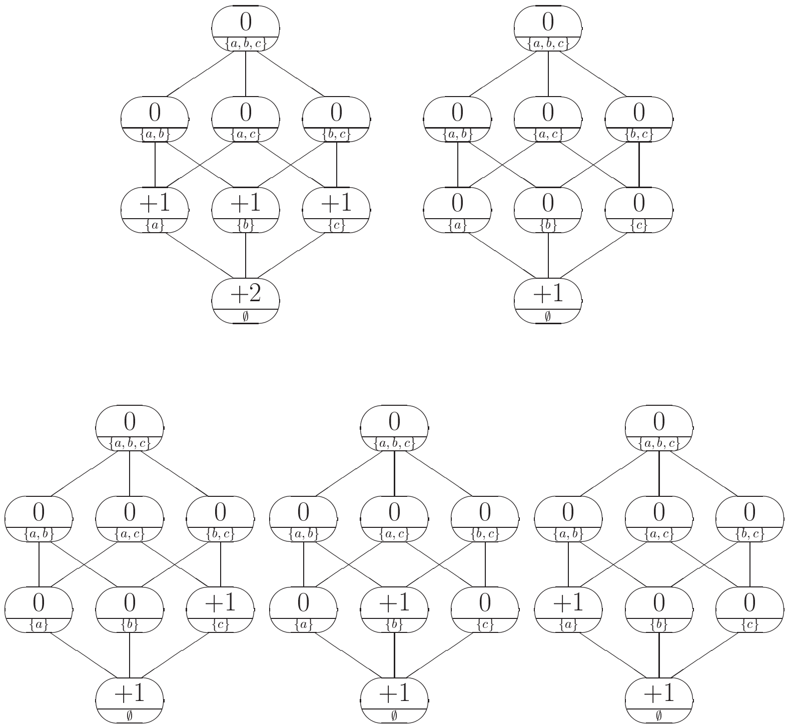


Figure 5.2: u-skeleton over $N = \{a, b, c\}$.

justification is that every independence model produced by a supermodular function is intersection of submaximal independence models of this type (see Theorem 5.3) so that 'skeletal' CI models form a certain 'generator' of the whole lattice of CI models produced by supermodular functions. Note for explanation that some authors interested in graphical models [5, 7] use the word 'skeleton' to name the underlying graph of a chain graph (see Section 10.3, p. 154). $\triangle$

Remark 5.6 As explained in Remark 5.3 ℓ-standardization is not the only possible way of standardization of supermodular functions. An interesting fact is that all results gathered in Section 5.2 can be achieved also for u-standardization and o-standardization. Thus, one can introduce the *u-skeleton* $\mathcal{K}_u^\diamond(N)$ as uniquely determined (finite) collection of non-zero normalized imsets belonging to extreme rays of the cone of u-standardized supermodular functions $\mathcal{K}_u(N)$. Analogously, the *o-skeleton* $\mathcal{K}_o^\diamond(N)$ consists of non-zero normalized imsets from extreme rays of the cone of o-standardized supermodular functions $\mathcal{K}_o(N)$. The point is that extreme rays of $\mathcal{K}_\ell(N)$, $\mathcal{K}_u(N)$ and $\mathcal{K}_o(N)$ correspond to each other; they describe respective classes of model equivalent skeletal imsets.

Thus, given a skeletal imset m the corresponding element of ℓ-skeleton m_ℓ can be obtained as follows. Put

$$\widetilde{m}_\ell = m - m(\emptyset) \cdot m^{\emptyset\uparrow} + \sum_{i \in N} \{m(\emptyset) - m(\{i\})\} \cdot m^{\{i\}\uparrow} \tag{5.11}$$

and then 'normalize' $\widetilde{m}_\ell$, i.e. put $m_\ell = k^{-1} \cdot \widetilde{m}_\ell$ where k is the greatest common prime divisor of $\{\widetilde{m}_\ell(S); S \subseteq N\}$ (see Figure 6.4 for an example of m and the respective element of the ℓ-skeleton). Given a skeletal imset m over N put

$$\nu(i) = m(N) - m(N \setminus \{i\}) \quad \text{for } i \in N\,, \qquad x = -m(N) + \sum_{i \in N} \nu(i)\,,$$

introduce

$$\widetilde{m}_u = m + x \cdot m^{\emptyset\uparrow} - \sum_{i \in N} \nu(i) \cdot m^{\{i\}\uparrow} \tag{5.12}$$

and normalize $\widetilde{m}_u$ to get the respective element of $\mathcal{K}_u^\diamond(N)$. Figure 5.2 shows u-skeleton for $N = \{a, b, c\}$. Finally, given a skeletal imset m over N put

$$\mu(i) = 2 \cdot \sum_{S \subseteq N} m(S) - 4 \cdot \sum_{S, i \in S} m(S) \quad \text{for } i \in N\,,$$

and

$$y = 2 \cdot \sum_{S \subseteq N} |S| \cdot m(S) - (|N| + 1) \cdot \sum_{S \subseteq N} m(S).$$

Then the formula

$$\widetilde{m}_o = 2^{|N|} \cdot m + y \cdot m^{\emptyset\uparrow} + \sum_{i \in N} \mu(i) \cdot m^{\{i\}\uparrow} \tag{5.13}$$

defines an o-standardized imset which after normalization yields the respective element of the o-skeleton $\mathcal{K}_o^\diamond(N)$. Figure 5.3 consists of Hasse diagrams of o-skeletal imsets over $N = \{a, b, c\}$.

Note that the proof of result of Section 5.2 for alternative standardization are analogous. The only noteworthy modification is needed in the proof of Lemma 5.5 in case of o-standardization. The cone $\mathcal{K}_o^\diamond(N)$ is viewed as a cone in $\mathbb{R}^{\mathcal{P}(N)}$ and after application of Lemma from Section 10.8.2 the respective $q \in \mathbb{Q}^{\mathcal{P}(N)}$ is multiplied to get $u \in \mathbb{Z}^{\mathcal{P}(N)}$. Then the formula (5.13) with u in place of m defines the desired o-standardized imset over N. Remaining arguments are analogous. △

5.3 Description of models by structural imsets

Semi-graphoid induced by a structural imset was introduced already in Section 4.4.1 on p. 64. The aim of this section is to relate those semi-graphoids to semi-graphoids produced by supremodular functions introduced in Section 5.1.1. The first observation is this.

OBSERVATION 5.6 Let m be a supermodular function over N and u a structural imset over N. Then $\langle m, u \rangle = 0$ iff $\mathcal{M}_u \subseteq \mathcal{M}^m$.

Proof: Supposing $\langle m, u \rangle = 0$ and $\langle A, B|C \rangle \in \mathcal{M}_u$ there exists $k \in \mathbb{N}$ such that $k \cdot u - u_{\langle A,B|C \rangle} \in \mathcal{S}(N)$. Write

$$0 = k \cdot \langle m, u \rangle = \langle m, k \cdot u \rangle = \langle m, k \cdot u - u_{\langle A,B|C \rangle} \rangle + \langle m, u_{\langle A,B|C \rangle} \rangle.$$

By Observation 5.1 both terms on the right-hand side of this equality are non-negative and must vanish. Thus, $\langle m, u_{\langle A,B|C \rangle} \rangle = 0$ which means $\langle A, B|C \rangle \in \mathcal{M}^m$. Conversely, supposing

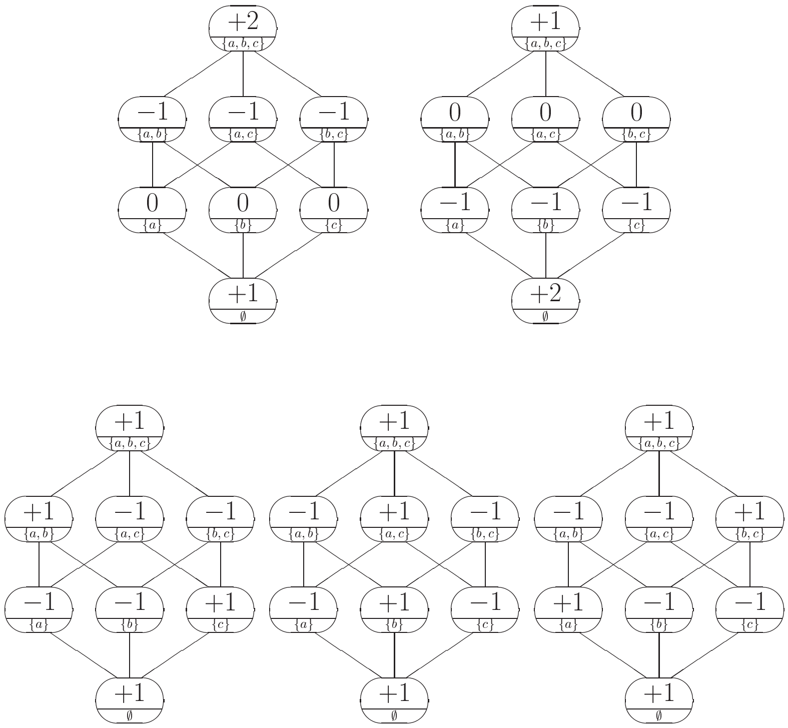


Figure 5.3: o-skeleton over $N = \{a, b, c\}$.

$\mathcal{M}_u \subseteq \mathcal{M}^m$ write by (4.2) $n \cdot u = \sum_{v \in \mathcal{E}(N)} k_v \cdot v$ for $n \in \mathbb{N}$, $k_v \in \mathbb{Z}^+$. For every $v = u_{\langle i,j|K\rangle} \in \mathcal{E}(N)$ such that $k_v > 0$ observe $\langle i, j|K\rangle \in \mathcal{M}_u$ and deduce $\langle m, v\rangle = 0$ by $\mathcal{M}_u \subseteq \mathcal{M}^m$. In particular,

$$n \cdot \langle m, u\rangle = \sum_{v \in \mathcal{E}(N)} k_v \cdot \langle m, v\rangle = 0,$$

which implies $\langle m, u\rangle = 0$. $\square$

An important auxiliary result is the following.

Lemma 5.7 Given a structural imset u over N one has $\mathcal{M}_u = \bigcap_{r \in \mathcal{R}} \mathcal{M}^r = \mathcal{M}^m$ where

$$\mathcal{R} = \{\, r \in \mathcal{K}_\ell^\diamond(N);\ \mathcal{M}_u \subseteq \mathcal{M}^r \,\} \quad \text{and} \quad m = \sum_{r \in \mathcal{R}} r\,.$$

Proof: The fact $\mathcal{M}_u \subseteq \bigcap_{r \in \mathcal{R}} \mathcal{M}^r$ is evident. For converse inclusion use Consequence 5.2: if $\langle A, B|C\rangle \in \mathcal{T}(N) \setminus \mathcal{M}_u$ then there exists $r \in \mathcal{K}_\ell^\diamond(N)$ such that $\langle r, u_{\langle A,B|C\rangle}\rangle > 0$ and $\langle r, u\rangle = 0$. By Observation 5.6 $\mathcal{M}_u \subseteq \mathcal{M}^r$. Thus, $r \in \mathcal{R}$ and $\langle A, B|C\rangle \notin \mathcal{M}^r$.

The inclusion $\bigcap_{r \in \mathcal{R}} \mathcal{M}^r \subseteq \mathcal{M}^m$ follows from the fact $m = \sum_{r \in \mathcal{R}} r$ by linearity of scalar product, the converse inclusion can be derived similarly with help of Observation 5.1. $\square$

Substantial fact is the following proposition.

Consequence 5.4 Given $\mathcal{M} \subseteq \mathcal{T}(N)$ the following four condition are equivalent:

(i) $\mathcal{M} = \mathcal{M}^m$ for a supermodular function m over N,

(ii) $\mathcal{M} = \mathcal{M}_u$ for a combinatorial imset u over N,

(iii) $\mathcal{M} = \mathcal{M}_u$ for a structural imset u over N,

(iv) $\mathcal{M} = \mathcal{M}^m$ for a supermodular ℓ-standardized imset m over N.

Proof: For (i)⇒(ii) put $u = \sum_{\langle A,B|C\rangle \in \mathcal{M}} u_{\langle A,B|C\rangle}$. As a combination of semi-elementary imsets is u a combinatorial imset. For every $\langle A,B|C\rangle \in \mathcal{M}$ observe that $u - u_{\langle A,B|C\rangle}$ is a combinatorial imset and therefore $A \perp\!\!\!\perp B \,|\, C\ [u]$. Thus $\mathcal{M} \subseteq \mathcal{M}_u$. For converse implication observe $\langle m, u\rangle = 0$ and use Observation 5.6. The implication (iii)⇒(iv) is an easy consequence of Lemma 5.7, (ii)⇒(iii) and (iv)⇒(i) are evident. □

Now, the main result of this chapter can be easily derived.

THEOREM 5.2 *Let P be probability measure over N with finite multiinformation. Then there exists a structural imset u over N such that P is perfectly Markovian with respect to u, that is $\mathcal{M}_P = \mathcal{M}_u$.*

Proof: By Consequence 5.1 on p. 73 there exists a supermodular function m over N such that $\mathcal{M}_P = \mathcal{M}^m$. By Consequence 5.4 $\mathcal{M}^m = \mathcal{M}_u$ for a structural imset u over N. □

REMARK 5.7 Going back to the motivation account from Section 1.1 Theorem 5.2 means that structural imsets solve satisfactorily theoretical question of completeness. The answer is affirmative, every CI structure induced by a probability measure with finite multiinformation can be described by a structural imset. On the other hand, natural price for this achievement is that structural imsets describe some 'superfluous' semi-graphoids. That means, there are semi-graphoids induced by structural imsets which are not induced by discrete probability measures as Example 4.1 on p. 66 shows (the left-hand picture of Figure 6.1 depicts the respective structural imset). In particular, another theoretical question of faithfulness from Section 1.1 has negative answer.

However, mathematical objects which 'answer' affirmatively both faithfulness and completeness question are not advisable because they cannot solve satisfactorily practical question of implementation (see Section 1.1, p. 9). These objects must be difficult to handle by a computer as the lattice of probabilistic CI models is quite complicated. For example, in case of 4 variables there exists meet-irreducible models which are not coatoms (= submaximal models) [107] which makes implementation complicated. The asset of structural imsets is that the lattice of models induced by them is fairly elegant and gives a chance of efficient computer implementation. △

5.4 Galois connection

The relation of both methods of description of CI models mentioned in this chapter can be lucidly explained with help of the view of theory of 'formal concept analysis'. This approach, developed in [28], is a specific application of theory of complete lattices on (conceptual) data analysis and knowledge processing. Because of its philosophical roots formal concept analysis is very near to human conceptual thinking. The most important mathematical notion behind this approach is a well-known notion of Galois connection. This view helps one to interpret the relation of structural imsets and supermodular imsets (functions) as a duality relation. I hope that presentation with help of Galois connection will make theory of structural CI models easy understandable for readers.

5.4.1 Formal concept analysis

Let me recall basic ideas of Chapter 1 of [28]. *Formal context* consists of the following items:

- the set of *objects* Œ,
- the set of *attributes* Æ,
- binary *incidence relation* $\Im \subseteq$ Œ $\times$ Æ between objects and attributes.

If $(x, y) \in \Im$ for $x \in$ Œ, $y \in$ Æ then write $x \Im y$ and say that the object x has the attribute y. In general, Galois connection is defined for a pair of posets ([8], Section 6 of Chapter IV). However, in treated special case, *Galois connection* can be introduced as a pair of mappings between power sets of Œ and Æ (which are posets with respect to inclusion):

$$\begin{array}{rcl} X \subseteq \text{Œ} & \longrightarrow & X^{\triangleright} = \{y \in \text{Æ}\,;\ x \Im y \text{ for every } x \in X\}\,, \\ Y \subseteq \text{Æ} & \longrightarrow & Y^{\triangleleft} = \{x \in \text{Œ}\,;\ x \Im y \text{ for every } y \in Y\}\,. \end{array}$$

Thus $X^{\triangleright}$ is the set of attributes common to objects in X while $Y^{\triangleleft}$ is the set of objects which have all attributes in Y. Clearly, $X_1 \subseteq X_2$ implies $X_1^{\triangleright} \supseteq X_2^{\triangleright}$ and $Y_1 \supseteq Y_2$ implies $Y_1^{\triangleleft} \subseteq Y_2^{\triangleleft}$. The consequence is that the mapping $X \mapsto X^{\triangleright\triangleleft}$ is a closure operation on subsets of Œ and the mapping $Y \mapsto Y^{\triangleleft\triangleright}$ is a closure operation on subsets of Æ.

By a *formal concept* of the context (Œ, Æ, $\Im$) is understood a pair (X, Y) with $X \subseteq$ Œ, $Y \subseteq$ Æ, $X^{\triangleright} = Y$ and $Y^{\triangleleft} = X$. The set X is called the *extent* and the set Y the *intent* of the concept. Observe that the concept is uniquely determined either by its extent, that is the list of objects forming the concept or by its intent which is the list of attributes (= properties) which characterize the concept. It reflects two different philophical-methodological ways of defining concepts: constructive and descriptive definitions.

One says that the concept (X_1, Y_1) is a *subconcept* of the concept (X_2, Y_2) and writes $(X_1, Y_1) \preceq (X_2, Y_2)$ if $X_1 \subseteq X_2$. Basic properties of Galois connection and the definition of notion formal concept implies that $X_1 \subseteq X_2$ iff $Y_1 \supseteq Y_2$. Thus, the class of all concepts of a given context (Œ, Æ, $\Im$) is a poset ordered by the relation $\preceq$. In fact, it is a complete lattice (see Theorem 3 in Chapter 1 of [28]) where supermum and infimum (of two concepts) are defined as follows:

$$\begin{array}{rcl} (X_1, Y_1) \vee (X_2, Y_2) & = & (\,(X_1 \cup X_2)^{\triangleright\triangleleft}\,,\ Y_1 \cap Y_2\,), \\ (X_1, Y_1) \wedge (X_2, Y_2) & = & (\,X_1 \cap X_2\,,\ (Y_1 \cup Y_2)^{\triangleleft\triangleright}\,). \end{array}$$

The lattice is called the *concept lattice.*

REMARK 5.8 Note that it follows from the properties of Galois connection that the above mentioned concept lattice is order-isomorphic to the poset $\{X \subseteq \text{Œ}; X = X^{\triangleright\triangleleft}\}$ ordered by inclusion $\subseteq$. Thus, the lattice can be described only in terms of objects with help of the closure operation $X \mapsto X^{\triangleright\triangleleft}$ on subsets of Œ. However, for analogous reason the same concept lattice is order-isomorphic to the poset $\{Y \subseteq \text{Æ}; Y = Y^{\triangleleft\triangleright}\}$ ordered by reversed inclusion $\supseteq$. This means that the lattice can be described dually in terms of attributes and the respective closure operation $Y \mapsto Y^{\triangleleft\triangleright}$ as well: this closure operation induces ordinary inclusion ordering $\subseteq$ on $\{Y \subseteq \text{Æ}; Y = Y^{\triangleleft\triangleright}\}$ (see Section 10.2). Thus, the same

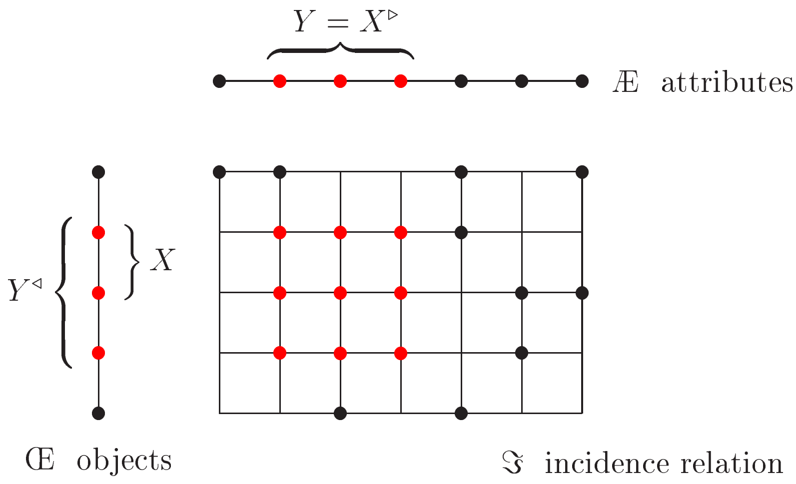


Figure 5.4: Galois connection - informal illustration.

mathematical structure can be described from two different points of view, in terms of objects or in terms of attributes. This again corresponds to two different methodological methods how to describe relations between concepts. The message of Section 5.4 is that the relation between description of CI models in terms of structural imsets and in terms of supermodular functions is just the relation of this kind. On the other hand, the role of objects and attributes in a formal context is evidently exchangeable. See Figure 5.4 for illustration. $\triangle$

5.4.2 Lattice of structural models

Let us introduce the class $\mathcal{U}(N)$ of *structural independence models*

$$\mathcal{U}(N) = \{\mathcal{M} \subseteq \mathcal{T}(N)\,;\ \mathcal{M} = \mathcal{M}_u \text{ for a structural imset } u \text{ over } N\}\,. \tag{5.14}$$

Consequence 5.4 implies that it coincides with the class of formal independence models produced by supermodular functions

$$\mathcal{U}(N) = \{\mathcal{M} \subseteq \mathcal{T}(N)\,;\ \mathcal{M} = \mathcal{M}^m \text{ for a supermodular function } m \text{ over } N\}\,. \tag{5.15}$$

The class $\mathcal{U}(N)$ is naturally ordered by inclusion $\subseteq$. The main result of this section says that $\mathcal{U}(N)$ is a finite concept lattice. Indeed, the respective formal context can be constructed as follows:

$$Œ = \mathcal{E}(N),\ Æ = \mathcal{K}_\ell^\diamond(N) \text{ and } u \,\Im\, m \text{ iff } \langle m, u\rangle = 0 \text{ for } u \in \mathcal{E}(N), m \in \mathcal{K}_\ell^\diamond(N). \tag{5.16}$$

Figure 5.5 gives an example of this formal context in case $N = \{a, b, c\}$. The following theorem summarizes the results.

THEOREM 5.3 *The poset $(\mathcal{U}(N), \subseteq)$ is a finite concept lattice which is, moreover, both atomistic and coatomistic. The null of $\mathcal{U}(N)$ is $\mathcal{T}_\emptyset(N)$, the model induced by zero structural imset. The atoms of $\mathcal{U}(N)$ are just the models induced by elementary imsets $\mathcal{M}_v$, $v \in \mathcal{E}(N)$. The coatoms of $\mathcal{U}(N)$ are just the models produced by skeletal supermodular functions $\mathcal{M}^m$, $m \in \mathcal{K}_\ell^\diamond(N)$. The unit of $\mathcal{U}(N)$ is $\mathcal{T}(N)$, the model produced by any modular set function $l \in \mathcal{L}(N)$.*

$\text{Æ} = \mathcal{K}_\ell^\diamond(N)$

	δ_N	$2\cdot\delta_N+\delta_{ab}+\delta_{ac}+\delta_{bc}$	$\delta_N+\delta_{ab}$	$\delta_N+\delta_{ac}$	$\delta_N+\delta_{bc}$
$u_{\langle b,c\|a\rangle}$		•	•	•	
$u_{\langle a,c\|b\rangle}$		•	•		•
$u_{\langle a,b\|c\rangle}$		•		•	•
$u_{\langle a,b\|\emptyset\rangle}$	•			•	•
$u_{\langle a,c\|\emptyset\rangle}$	•		•		•
$u_{\langle b,c\|\emptyset\rangle}$	•		•	•	

$\text{Œ} = \mathcal{E}(N)$

Figure 5.5: Formal context (5.16) in case $N = \{a,b,c\}$.

Proof: The first observation is that $\mathcal{U}(N)$ is a complete lattice. Indeed, it suffices to show that every subset of $\mathcal{U}(N)$ has infimum (see Section 10.2). Let us use (5.15) for this purpose. Given supermodular functions $m_1,\ldots,m_n, n \geq 1$ the function $m = \sum_{i=1}^n m_i$ defines a supermodular function such that $\mathcal{M}^m = \bigcap_{i=1}^n \mathcal{M}^{m_i}$. This follows from Observation 5.1. Infimum of the empty subset of $\mathcal{U}(N)$ is $\mathcal{T}(N)$ (see Lemma 5.2).

The second observation is that $\{\mathcal{M}_v; v \in \mathcal{E}(N)\}$ is supremum-dense in $\mathcal{U}(N)$:

$$\forall \mathcal{M} \in \mathcal{U}(N) \quad \mathcal{M} = \sup\{\mathcal{M}_u : u \in \mathcal{Q}\} \text{ where } \mathcal{Q} = \{u_{\langle i,j|K\rangle} \in \mathcal{E}(N); \langle i,j|K\rangle \in \mathcal{M}\}. \quad (5.17)$$

Indeed, evidently $\mathcal{M}_v \subseteq \mathcal{M}$ for every $v \in \mathcal{Q}$ (c.f. Lemma 4.5). On the other hand, supposing $\mathcal{K} \in \mathcal{U}(N)$ satisfies $\mathcal{M}_v \subseteq \mathcal{K}$ for every $v \in \mathcal{Q}$ every elementary independence statement from $\mathcal{M}$ belongs to $\mathcal{K}$. Then by Lemma 2.2 conclude $\mathcal{M} \subseteq \mathcal{K}$ which implies (5.17).

The third observation is that $\{\mathcal{M}^m; m \in \mathcal{K}_\ell^\diamond(N)\}$ is infimum-dense in $\mathcal{U}(N)$:

$$\forall \mathcal{M} \in \mathcal{U}(N) \quad \mathcal{M} = \inf\{\mathcal{M}^m; m \in \mathcal{R}\} \text{ where } \mathcal{R} = \{r \in \mathcal{K}_\ell^\diamond(N); \mathcal{M} \subseteq \mathcal{M}^r\}. \quad (5.18)$$

Indeed, by (5.14) one can apply Lemma 5.7 to $\mathcal{M}$ and observe that $\mathcal{M} = \bigcap_{m\in\mathcal{R}} \mathcal{M}^m$. This clearly implies (5.18). To show that $(\mathcal{U}(N), \subseteq)$ is even a concept lattice one can use Theorem 3 in Chapter 1 of [28]. The theorem says that to show that $\mathcal{U}(N)$ is order-isomorphic to the concept lattice of a formal context $(\text{Œ}, \text{Æ}, \Im)$ it suffices to show that there exist mappings $\gamma : \text{Œ} \to \mathcal{U}(N)$ and $\delta : \text{Æ} \to \mathcal{U}(N)$ such that

(a) $\gamma(\text{Œ})$ is supremum-dense in $\mathcal{U}(N)$,

(b) $\delta(\text{Æ})$ is infimum-dense in $\mathcal{U}(N)$,

(c) $u \,\Im\, m \Leftrightarrow \gamma(u) \subseteq \delta(m)$ for every $u \in \text{Œ}$, $m \in \text{Æ}$.

Let us introduce the formal context by means of (5.16) and define γ and δ as follows: γ ascribes $\mathcal{M}_v$ to every $v \in \mathcal{E}(N)$ (see p. 64) and δ ascribes $\mathcal{M}^m$ to every $m \in \mathcal{K}_\ell^\diamond(N)$ (see p. 72). The condition (a) follows from (5.17), the condition (b) from (5.18) and the condition (c) follows directly from Observation 5.6. Thus, $\mathcal{U}(N)$ is a concept lattice.

The next observation is that $\mathcal{T}_\emptyset(N)$ is the null of $\mathcal{U}(N)$. By Observation 4.7 $\mathcal{T}_\emptyset(N) \in \mathcal{U}(N)$, by Lemma 4.6 it is a semi-graphoid and therefore $\mathcal{T}_\emptyset(N) \subseteq \mathcal{M}$.

To show that every $\mathcal{M}_v$, $v \in \mathcal{E}(N)$ is an atom of $\mathcal{U}(N)$ observe by (5.14) $\mathcal{M}_v \in \mathcal{U}(N)$ and assume $\mathcal{M} \in \mathcal{U}(N)$, $\mathcal{M} \subseteq \mathcal{M}_v$. If $v = u_{\langle i,j|K\rangle}$ then by Lemma 4.5 obtain $\mathcal{M}_v = \{\langle i,j|K\rangle, \langle j,i|K\rangle\} \cup \mathcal{T}_\emptyset(N)$. As $\mathcal{M}$ is a semi-graphoid $\mathcal{T}_\emptyset(N) \subseteq \mathcal{M}$. If $\mathcal{M} \neq \mathcal{T}_\emptyset(N)$ then either $\langle i,j|K\rangle$ or $\langle j,i|K\rangle$ belongs to $\mathcal{M}$ which implies by symmetry property $\mathcal{M}_v \subseteq \mathcal{M}$. The above mentioned fact implies with help of (5.17) that $\mathcal{U}(N)$ is an atomistic lattice.

The fact that $\mathcal{T}(N)$ is the unit of $\mathcal{U}(N)$ is evident: $\mathcal{T}(N) \in \mathcal{U}(N)$ by Lemma 5.2. To show that every $\mathcal{M}^s$, $s \in \mathcal{K}_\ell^\diamond(N)$ is a coatom of $\mathcal{U}(N)$ observe $\mathcal{M}^s \in \mathcal{U}(N)$ by (5.15) and assume $\mathcal{M} \in \mathcal{U}(N)$, $\mathcal{M}^s \subseteq \mathcal{M}$. By (5.14) and Lemma 5.7 write $\mathcal{M} = \bigcap_{r\in\mathcal{R}} \mathcal{M}^r$ where $\mathcal{R} = \{r \in \mathcal{K}_\ell^\diamond(N); \mathcal{M} \subseteq \mathcal{M}^r\}$. If $\mathcal{R} \setminus \{s\} \neq \emptyset$ then $\mathcal{M}^s \subseteq \mathcal{M} \subseteq \mathcal{M}^r$ for some $r \in \mathcal{K}_\ell^\diamond(N)$, $r \neq s$ which contradicts the fact $\mathcal{M}^s \setminus \mathcal{M}^r \neq \emptyset$ implied by Lemma 5.5. Therefore $\mathcal{R} \subseteq \{s\}$: if $\mathcal{R} = \emptyset$ then $\mathcal{M} = \mathcal{T}(N)$, if $\mathcal{R} = \{s\}$ then $\mathcal{M} = \mathcal{M}^s$. The above fact implies together with (5.18) that $\mathcal{U}(N)$ is a coatomistic lattice.

To evidence that $\{\mathcal{M}_v; v \in \mathcal{E}(N)\}$ are all atoms of $\mathcal{U}(N)$ realize that every atom is join-irreducible and use the well-known fact that every supremum-dense set must contain all join-irreducible elements (see Section 2.4.2 in Chapter 1 of [28]). Indeed, $\{\mathcal{M}_v, v \in \mathcal{E}(N)\}$ is supremum-dense in $\mathcal{U}(N)$ by (5.17). Analogously, the fact that $\{\mathcal{M}^m; m \in \mathcal{K}_\ell^\diamond(N)\}$ are all coatoms of $\mathcal{U}(N)$ follows from (5.18) and the fact that every infimum-dense subset must contain all meet-irreducible elements, in particular coatoms. □.

Remark 5.9 However, the formal context (5.16) is not the only option. For example, one can alternatively take combinatorial imsets in place of Œ and combinations of ℓ-skeletal imsets with non-negative integer coefficients in place of Æ (but the incidence relation is defined in the same way like in (5.16)). The second option is to put Œ $= \mathcal{S}(N)$ and Æ $= \mathcal{K}_\ell(N) \cap \mathbb{Z}^{\mathcal{P}(N)}$ (see Figure 7.11 for illustration). The third option is Œ $= con(\mathcal{E}(N))$ and Æ $= \mathcal{K}_\ell(N)$. Moreover, one can consider alternative standardization instead of ℓ-standardization (see Remark 5.3). Specific combined option is Œ $= \mathcal{E}(N)$ and Æ $= \mathcal{K}(N)$.

On the other hand, the formal context (5.16) is distinguished in a certain sense. Theorem 5.3 implies that every object of (5.16) defines a join-irreducible concept and every attribute of (5.16) defines a meet-irreducible concept. Thus, the context (5.16) is reduced in sense of Definition 24, Chapter 1 in [28]. Formal context of this type is unique for a given finite concept lattice up to respective isomorphism of formal contexts (see Proposition 12 in Chapter 1 of [28]). △

Another point of view on the lattice $(\mathcal{U}(N), \subseteq)$ is the following. It order-isomorphic to the class of structural imsets $\mathcal{S}(N)$ factorized with respect to corresponding facial equivalence (see Chapter 6). This understanding is suitable from computational point of view since the operation of supremum in the lattice corresponds to summing of structural imsets; see Consequence 6.1 in Section 6.2.1. A dual point of view is also possible: elements of $\mathcal{K}_\ell(N) \cap \mathbb{Z}^{\mathcal{P}(N)}$ factorized with respect to the corresponding equivalence can be taken into consideration. The following observation says that infimum is realizable by means of summing supermodular functions (imsets).

Observation 5.7 Let $\mathcal{R}$ be a finite set of supermodular functions over N. Then

$$\inf \{\mathcal{M}^m; m \in \mathcal{R}\} = \bigcap_{m\in\mathcal{R}} \mathcal{M}^m = \mathcal{M}^r \quad \text{where} \quad r = \sum_{m\in\mathcal{R}} m. \tag{5.19}$$

Proof: To show $\mathcal{M}^r \subseteq \mathcal{M}^m$ for $m \in \mathcal{R}$ take $\langle A,B|C\rangle \in \mathcal{M}^r$, write $0 = \langle r, u_{\langle A,B|C\rangle}\rangle = \sum_{m\in\mathcal{R}} \langle m, u_{\langle A,B|C\rangle}\rangle$ and use Observation 5.1(ii). The same equality can be used to show that, for every supermodular function s over N, the requirement $\mathcal{M}^s \subseteq \mathcal{M}^m$ for $m \in \mathcal{R}$ implies $\mathcal{M}^s \subseteq \mathcal{M}^r$. □

REMARK 5.10 The lattice $\mathcal{U}(N)$ is also order-isomorphic to a *face lattice* (see Section 0.3 in [28]) namely the lattice of faces of a certain polyhedral cone. For example, one can take the cone $con(\mathcal{E}(N)) \subseteq \mathbb{R}^{\mathcal{P}(N)}$. Another option is the cone $\mathcal{K}_\ell^\diamond(N)$ endowed with reverse inclusion $\supseteq$, respectively the cones $\mathcal{K}_\ell^\diamond(N)$, $\mathcal{K}_o^\diamond(N)$. In fact, original terminology from [108] was motivated by this point of view (see Remark 6.2 on p. 93). △

EXAMPLE 5.1 The lattice $\mathcal{U}(N)$ has only 1 element in case $|N| = 1$, namely $\mathcal{T}(N) = \mathcal{T}_\emptyset(N)$ and only 2 elements in case $|N| = 2$, namely $\mathcal{T}_\emptyset(N)$ and $\mathcal{T}(N)$. However, it has 22 elements in case $N = \{a,b,c\}$: the respective Hasse diagram is shown in Figure 5.6. Every node of the diagram contains a schematic description of the respective independence model with help of elementary independence statements. Note that Figure 5.6 also shows the lattice of semi-graphoids over $\{a,b,c\}$ as they coincide with structural independence models in this case. In fact, every structural model over $\{a,b,c\}$ is even a CI model. The number of structural models over 4 variables is 22108 [106]. ♢

OBSERVATION 5.8 Let u be a structural imset over N with $|N| = 3$. Then there exists a discrete probability measure P over N such that $\mathcal{M}_u = \mathcal{M}_P$.

Proof: This is an easy consequence of the fact that $(\mathcal{U}(N), \subseteq)$ is coatomistic (see Theorem 5.3) and Lemma 2.9 from Section 2.3.7. Six respective constructions of perfectly Markovian measures for the unit and coatoms of $\mathcal{U}(N)$ (see Figure 5.6) were already given: see Observation 2.2, Observation 2.3, Example 2.1 and Example 2.2. □

REMARK 5.11 This is to explain the relation of this theory to polymatroidal description of CI model used in [66]. Polymatroid is defined as a non-decreasing submodular function $h : \mathcal{P}(N) \to \mathbb{R}$ such that $h(\emptyset) = 0$. Some polymatroids can be obtained as multiples of entropy functions of discrete measures relative to the counting measure mentioned in Remark 4.4 - see [61]. The formal independence model induced by a polymatroid h consists of $\langle A,B|C\rangle \in \mathcal{T}(N)$ such that $\langle h, u_{\langle A,B|C\rangle}\rangle = 0$. Since $-h$ is a supermodular function, there is no difference between the model induced by a polymatroid h and the model produced by the supermodular function $-h$. Thus, models induced by polymatroids are just the structural independence models. There is an one-to-one correspondence between certain polymatroids and u-standardized supermodular function - see §5.2 in [116]. △

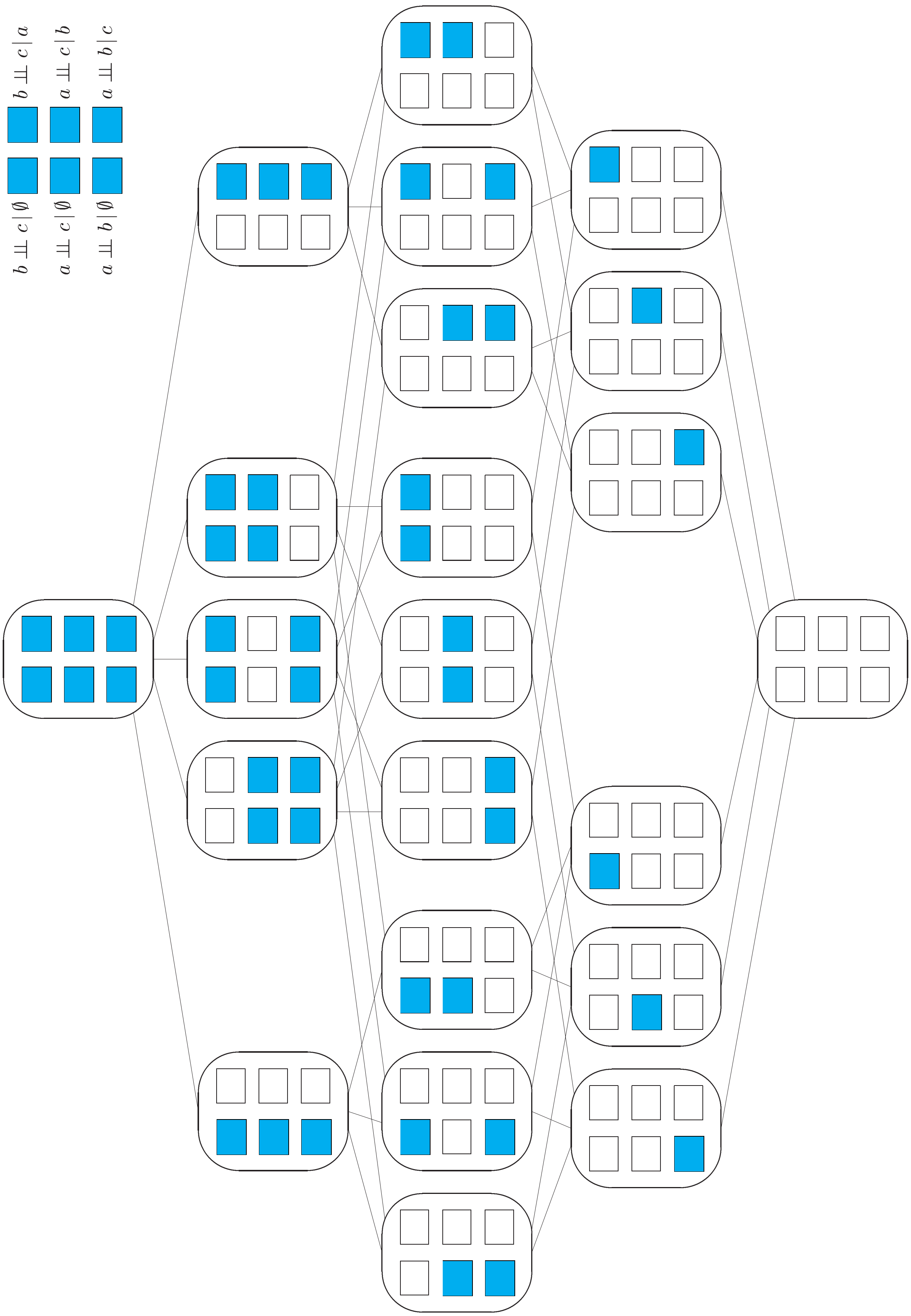


Figure 5.6: Concept lattice of CI models over $N = \{a, b, c\}$ (rotated).

Chapter 6

Markov equivalence

This chapter deals with implication and equivalence problem for structural imsets. First, the question how to understand the concept of Markov equivalence (and implication) is discussed and two types of equivalence are compared. The rest of the chapter is devoted to the stronger type of equivalence, called the *facial equivalence* and to the respective implication between structural imsets. Two characterization of facial implication, which are analogous to graphical characterization of Markov equivalence mentioned in Chapter 3, are given and related implementation tasks are discussed.

6.1 Two concepts of equivalence

Basically, there are two different ways of defining the concept of Markov equivalence for graphs which appear to be equivalent in case of classic graphical models, e.g. UG models, DAG models and CG models - see Sections 3.1 - 3.3. The first option is *distribution equivalence* which is the requirement that the classes of Markovian measures over N within a certain distribution framework coincide. By a distribution framework is understood a class Ψ of probability measures over N. Thus, distribution equivalence is always understood relative to a distribution framework.

The second option is *model equivalence* which is the requirement that the induced formal independence models coincide. This type of equivalence is not related to a distribution framework. Clearly, because of the definition of Markovian measure, model equivalence implies distribution equivalence. The converse is true in case of faithfulness (see Section 1.1 p. 8). That is, if a perfectly Markovian measure within the considered distribution framework Ψ exists for every graph (from the respective class of graphs) then distribution equivalence relative to Ψ implies model equivalence. This is the case of classic chain graphs relative to the class of discrete measures (see Section 3.3) and the case of alternative chain graphs relative to the class of non-degenerate Gaussian measures (see Section 3.5.5). Nevertheless, distribution and model equivalence coincide even under weaker assumption that the considered class of measures is perfect for every graph (see Remark 3.2 on p. 38). On the other hand, if the distribution framework is somehow limited then it may happen that model and distribution equivalence differ. For example, it happens in case that the class Ψ of measures with prescribed one-dimensional marginals P_i on fixed measurable spaces $(\mathsf{X}_i, \mathcal{X}_i)$, $i \in N$ is considered and P_k is a *degenerated measure* for every $k \in M$, $M \subseteq N$ which means that $P_k(\mathsf{A}) \in \{0,1\}$ for every $\mathsf{A} \in \mathcal{X}_k$. Then $i \perp\!\!\!\perp j \,|\, K \; [P]$ for every $P \in \Psi$ and $\langle i,j|K\rangle \in \mathcal{E}(N)$ with $i \in M$ (use Lemma 10.1) and one can show that all undirected graphs over N which have the same induced subgraph for $N \setminus M$ are distribution equivalent.

Remark 6.1 Note that one may consider even the third type of equivalence of graphs, namely the parametrization equivalence. This approach is based on the following interpretation of some types of graphs, e.g. ancestral graphs [85] and joint-response chain graphs [18]. Every edge of a graph of this type represents a real parameter, a specific distribution framework Ψ (usually the class of non-degenerate Gaussian measures over N) is considered and every collection of edge-parameters determines uniquely a probability measure from Ψ factorized in a particular way. Every graph of this type is then identified with the class of *parametrized distributions* which often coincides with the class of Markovian distributions within Ψ (e.g. in case of maximal ancestral graphs [85]). Two graphs can be called *parametrization equivalent* if their classes of parametrized distributions coincide. Of course, parametrization equivalence substantially depends on the considered distribution framework and may not coincide with distribution (Markov) equivalence - for example in case of general ancestral graphs [85].

The mentioned point of view motivates a general question whether (some) structural imsets may lead to a specific way of parametrization of the corresponding class of Markovian distribution (see Direction 4 in Chapter 8). △

However, in usual situations distribution and model equivalence coincide which means that the concept of Markov equivalence of graphs is unambiguously defined. Then the task to characterize Markov equivalence in graphical terms is correctly set. Several solutions of this general equivalence question (see Section 1.1, p. 8) were exemplified in Chapter 3. The aim of this chapter is to examine the same equivalence question for structural imsets. The problem is that in case of structural imsets one has to distinguish two above mentioned types of Markov equivalence and choose one of them as a basis of further study.

6.1.1 Facial and Markov equivalence

Two structural imsets u, v over N are *facially equivalent* if they induce the same CI model, i.e. $\mathcal{M}_u = \mathcal{M}_v$. Then one writes $u \rightleftharpoons v$. Let Ψ be a class of probability measures over N and $\Psi(u)$ denotes the class of Markovian measures with respect to u relative to Ψ:

$$\Psi(u) = \{P \in \Psi\, ;\; A \perp\!\!\!\perp B \,|\, C \;[P] \;\text{ whenever }\; \langle A, B|C\rangle \in \mathcal{M}_u \,\}\,. \tag{6.1}$$

Two structural imsets u and v over N are *Markov equivalent relative to* Ψ if $\Psi(u) = \Psi(v)$. The following observation is evident.

Observation 6.1 Facially equivalent structural imsets are Markov equivalent relative to any class Ψ of probability measures over N.

Clearly, Markov equivalence relative to Ψ implies Markov equivalence relative to any subclass $\widetilde{\Psi} \subseteq \Psi$. Natural question is whether the converse of Observation 6.1 holds for a reasonable class Ψ. The answer is negative even for the class of marginally continuous measures which involves the widest class of measures for which the method of structural imsets is safely applicable, namely the class of measures with finite multiinformation (see Section 4.1). This is illustrated by the following example.

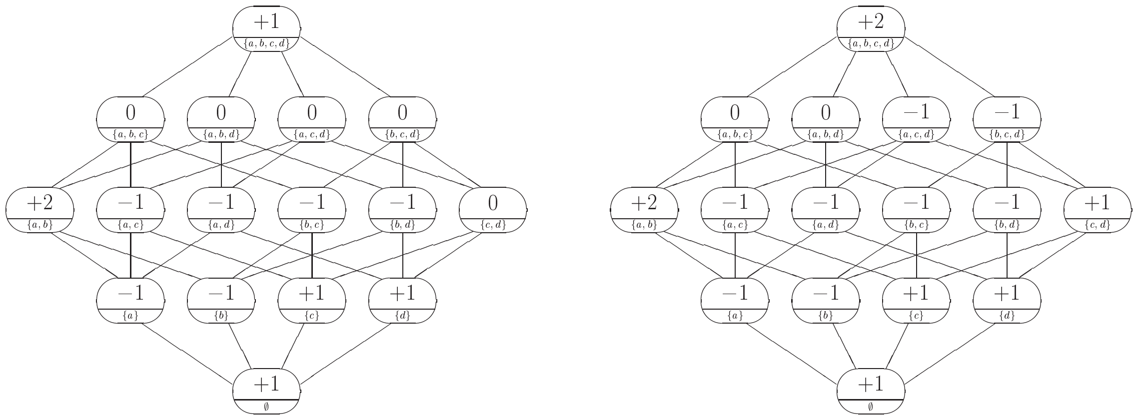


Figure 6.1: Two Markov equivalent structural imsets which are not facially equivalent.

EXAMPLE 6.1 There exist two structural imsets over $N = \{a, b, c, d\}$ which are Markov equivalent relative to the class of marginally continuous probability measures over N but which are not facially equivalent. Consider the imsets (see Figure 6.1)

$$u = u_{\langle c,d|\{a,b\}\rangle} + u_{\langle a,b|\emptyset\rangle} + u_{\langle a,b|\{c\}\rangle} + u_{\langle a,b|\{d\}\rangle} \quad \text{and} \quad v = u + u_{\langle a,b|\{c,d\}\rangle} \,.$$

Clearly, $\mathcal{M}_u \subseteq \mathcal{M}_v$ but $\langle a, b|\{c, d\}\rangle \in \mathcal{M}_v \setminus \mathcal{M}_u$ as shown in Example 4.1. On the other hand, by Consequence 2.1 every marginally continuous measure P over N which is Markovian with respect to u satisfies $a \perp\!\!\!\perp b \,|\, \{c, d\}\ [P]$. Hence, one can show that P is Markovian with respect to u iff it is Markovian with respect to v. $\diamondsuit$

In fact, the above mentioned phenomenon is a consequence of the fact that structural imsets do not satisfy the faithfulness requirement from Section 1.1 - see Remark 5.7 on p. 83. However, in case $|N| \leq 3$ every structural imset has a discrete perfectly Markovian measure over N - see Observation 5.8. In particular, if $|N| \leq 3$ then facial equivalence coincides both with Markov equivalence relative to the class of discrete measures and with Markov equivalence relative to the class of measures with finite multiinformation (use Observation 6.1).

In the rest of this chapter attention is restricted to facial equivalence and related facial implication. One reason is that facial implication is not adulterated by considering a specific class of distributions Ψ. Therefore, one has a better chance that the respective deductive mechanism can be implemented on a computer. Moreover, in my opinion, facial equivalence represents pure theoretical basis of Markov equivalence. Indeed, it will be shown later (see Lemma 6.3) that for a reasonable distribution framework Ψ every Markov equivalence class relative to Ψ decomposes into facial equivalence classes and just one of these classes consists of 'Ψ-representable' structural imsets, that is imsets having perfectly Markovian measures in Ψ. Thus, to describe CI structures arising within Ψ one can limit oneself to structural imsets of this type and facial equivalence on the considered subclass of stuctural imsets coincides with Markov equivalence relative to Ψ.

6.2 Facial implication

Let u, v are structural imsets over N. One says that u *facially implies* v and writes $u \rightharpoonup v$ if $\mathcal{M}_v \subseteq \mathcal{M}_u$. Observe that u is facially equivalent to v iff $u \rightharpoonup v$ and $v \rightharpoonup u$.

REMARK 6.2 This is to explain the motivation of above terminology. The adjective 'facial' was used already in [108] to name the respective deductive mechanism for structural imsets. This was motivated by an analogy with the theory of convex polytopes where the concept of face has a central role [12]. Indeed, one can consider the collection of all faces of the cone $con(\mathcal{E}(N))$ and introduce the following implication of structural imsets: u implies v if every face of $con(\mathcal{E}(N))$ which contains u contains also v. The original definition of facial implication of structural imsets used in [108] was nothing but modification of this requirement. It appeared to be equivalent to the condition $\mathcal{M}_v \subseteq \mathcal{M}_u$: one can show this using the results from [108], although it is not explicitly stated there. $\triangle$

6.2.1 Direct characterization of facial implication

LEMMA 6.1 Let u, v are structural imsets over N. Then $u \rightharpoonup v$ iff

$$\exists\, l \in \mathbb{N} \quad l \cdot u - v \ \text{ is a structural imset}\,, \tag{6.2}$$

which is under assumption that v is a combinatorial imset equivalent to the requirement

$$\exists\, k \in \mathbb{N} \quad k \cdot u - v \ \text{ is a combinatorial imset}\,. \tag{6.3}$$

Proof: Suppose $u \rightharpoonup v$ and write $n \cdot v = \sum_{w \in \mathcal{E}(N)} k_w \cdot w$ where $n \in \mathbb{N}, k_w \in \mathbb{Z}^+$. If $k_w > 0$ and $w = u_{\langle i,j|K\rangle}$ then $\langle i,j|K\rangle \in \mathcal{M}_v \subseteq \mathcal{M}_u$. Thus, there exists $l_w \in \mathbb{N}$ such that $l_w \cdot u - w \in \mathcal{S}(N)$. Put $l = \sum_{w \in \mathcal{E}(N), k_w > 0} k_w \cdot l_w$ and observe that

$$l \cdot u - v = (l \cdot u - n \cdot v) + (n-1) \cdot v = \sum_{w \in \mathcal{E}(N), k_w > 0} k_w \cdot (l_w \cdot u - w) + (n-1) \cdot v \in \mathcal{S}(N)$$

since $\mathcal{S}(N)$ is closed under summing. Thus, (6.2) was verified. Conversely, suppose (6.2) and consider $\langle A, B|C\rangle \in \mathcal{M}_v$. Find $k \in \mathbb{N}$ such that $k \cdot v - u_{\langle A,B|C\rangle} \in \mathcal{S}(N)$. As $\mathcal{S}(N)$ is closed under summing conclude

$$(k \cdot l) \cdot u - u_{\langle A,B|C\rangle} = k \cdot (l \cdot u - v) + (k \cdot v - u_{\langle A,B|C\rangle}) \in \mathcal{S}(N)\,,$$

which implies $\langle A, B|C\rangle \in \mathcal{M}_u$.

Evidently (6.3) implies (6.2). On contrary, suppose (6.2) and that v is a combinatorial imset. Take $n \in \mathbb{N}$ such that $n \cdot (l \cdot u - v)$ is combinatorial and put $k = n \cdot l$. As $v \in \mathcal{C}(N)$ and $\mathcal{C}(N)$ is closed under summing $k \cdot u - v = n \cdot (l \cdot u - v) + (n-1) \cdot v$ is a combinatorial imset. $\square$

REMARK 6.3 Basic difference between (6.2) and (6.3) is that testing whether an o-standardized imset is combinatorial is decidable in finitely many steps and the number of these steps is known! Indeed, if an o-standardized imset $w = k \cdot u - v$ is combinatorial, then the degree $deg\,(w)$ can be directly computed by Observation 4.3. It is the number of elementary imsets which have to be summed to obtain w. The only combinatorial imset of

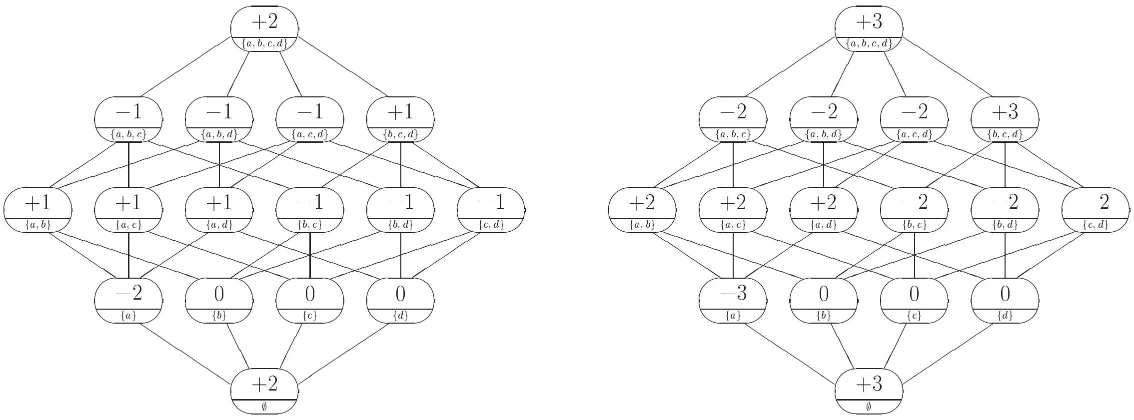


Figure 6.2: Structural imsets u and $2 \cdot u - v$ from Example 6.2.

degree 0 is the zero imset and an imset w with $deg(w) = n \in \mathbb{N}$ is combinatorial iff there exists an elementary imset $u_{\langle i,j|K \rangle}$ such that $w - u_{\langle i,j|K \rangle}$ is a combinatorial imset of degree $n-1$. Since the class of elementary imsets is known testing can be done recursively.

Note that one can modify the proof of Lemma 6.1 to show that (6.3) is equivalent to $u \rightharpoonup v$ even in case that v is a structural imset such that

$$\exists\, k, n \in \mathbb{N} \quad n \cdot v \ \text{ and } \ (n \cdot k - 1) \cdot v \ \text{ are combinatorial imsets}. \tag{6.4}$$

The condition (6.4) is formally weaker than the requirement that v is a combinatorial imset. However, their difference may appear to be illusory. So far, I do not know an example of a structural imset which is not a combinatorial imset - see Question 7. △

A natural question is how big the number $l \in \mathbb{N}$ from (6.2) could be. The following example shows that it may happen that $l > 1$.

Example 6.2 There exists a combinatorial imset u over $N = \{a, b, c, d\}$ and a semi-elementary imset v such that $2 \cdot u - v$ is a structural imset (and therefore $u \rightharpoonup v$) but $u - v$ is not a structural imset. Put

$$u = u_{\langle a,b|\emptyset \rangle} + u_{\langle a,c|\emptyset \rangle} + u_{\langle c,d|b \rangle} + u_{\langle b,d|c \rangle} + u_{\langle a,d|bc \rangle} + u_{\langle b,c|ad \rangle}\,, \qquad v = u_{\langle a,bcd|\emptyset \rangle} \tag{6.5}$$

and observe that

$$\begin{aligned} 2 \cdot u - v = 2 \cdot u - u_{\langle a,bcd|\emptyset \rangle} \quad = \quad & u_{\langle a,b|\emptyset \rangle} + u_{\langle a,c|\emptyset \rangle} + u_{\langle a,d|\emptyset \rangle} + \\ & u_{\langle c,d|b \rangle} + u_{\langle b,d|c \rangle} + u_{\langle b,c|d \rangle} + \\ & u_{\langle b,c|ad \rangle} + u_{\langle b,d|ac \rangle} + u_{\langle c,d|ab \rangle} \end{aligned} \tag{6.6}$$

is a combinatorial imset (see Figure 6.2 for illustration) and therefore $u \rightharpoonup v$. To see that $u - v$ is not a structural imset (see the left-hand picture of Figure 6.3 for illustration) consider the multiset $m_\circ$ shown in the right-hand picture of Figure 6.3. It is a supermodular multiset by Observation 5.1(iii). As $\langle m_\circ, u - v \rangle = -1$ the imset $u - v$ is not a structural imset by Observation 5.1(i).

On the other hand, one has $u \rightharpoonup w$ for an elementary imset $w = u_{\langle a,d|\emptyset \rangle}$. Indeed, one has

$$u - w = u_{\langle a,bc|\emptyset \rangle} + u_{\langle b,c|d \rangle} + u_{\langle b,d|ac \rangle} + u_{\langle c,d|ab \rangle}$$

which means that the constant l from (6.2) can be lower in this case. ♢

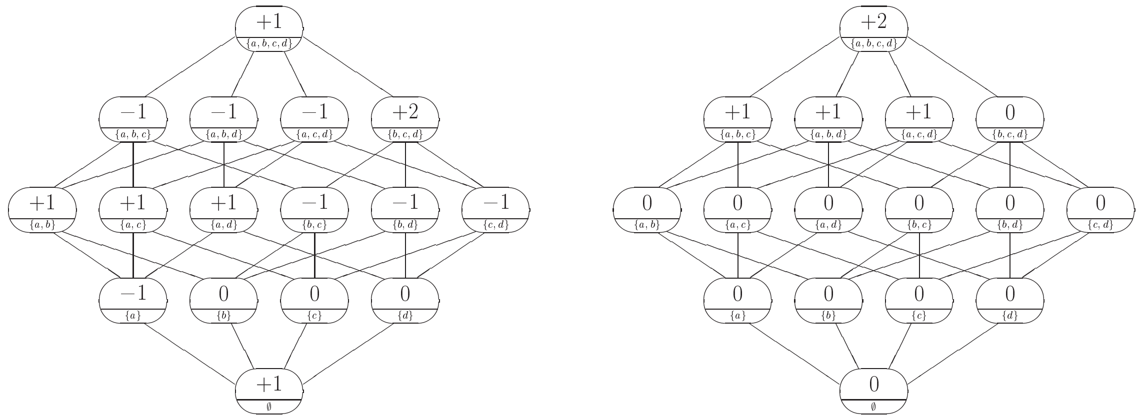


Figure 6.3: The imsets $u - v$ and the supermodular multiset $m_\circ$ from Example 6.2.

Remark 6.4 Example 6.2 shows that verification whether a semi-elementary imset is facially implied by a structural imset requires multiplication of the structural imset by 2 at least. Note that later Consequence 6.4 in Section 6.3.2 can be modified by replacing the class of elementary imsets by the class of semi-elementary imsets to get an upper estimate of the constant in (6.2) in this case. One can show using the results of [101] that in case $|N| = 4$

$$\max\{\,\langle r, w\rangle\,;\, r \in \mathcal{K}_\ell^\diamond(N)\,,\ w \text{ is a semi-elementary imset over } N\,\} = 2$$

which implies then that 2 is the minimal integer l_* satisfying

$$\forall\, u \in \mathcal{S}(N) \ \ v \text{ semi-elementary imset over } N \quad u \rightharpoonup v \ \text{ iff } \ l_* \cdot u - v \in \mathcal{S}(N)\,.$$

Note that one has $l_* = 1$ in case $|N| \leq 3$ for the same reason. △

The following consequence of Lemma 6.1 was already announced in Section 5.4.2.

Consequence 6.1 Let $\mathcal{Q}$ be a finite set of structural imsets over N. Then

$$\sup\{\mathcal{M}_u; u \in \mathcal{Q}\} = \mathcal{M}_v \quad \text{for } v = \sum_{u\in\mathcal{Q}} u \tag{6.7}$$

where the supremum is understood in the lattice $(\mathcal{U}(N), \subseteq)$.

Proof: To show $\mathcal{M}_u \subseteq \mathcal{M}_v$ for $u \in \mathcal{Q}$ take $\langle A, B|C\rangle \in \mathcal{M}_u$, find $k \in \mathbb{N}$ such that $k \cdot u - u_{\langle A,B|C\rangle} \in \mathcal{S}(N)$ and write

$$k \cdot v - u_{\langle A,B|C\rangle} = k \cdot \sum_{w\in\mathcal{Q}\setminus\{u\}} w + (k \cdot u - u_{\langle A,B|C\rangle}) \in \mathcal{S}(N)\,.$$

To show, for every structural imset w over N, that the assumption $\mathcal{M}_u \subseteq \mathcal{M}_w$ for $u \in \mathcal{Q}$ implies $\mathcal{M}_v \subseteq \mathcal{M}_w$ use Lemma 6.1. Indeed, the assumption means that $l_u \in \mathbb{N}$ with $l_u \cdot w - u \in \mathcal{S}(N)$ exists for every $u \in \mathcal{Q}$. Put $l = \sum_{u\in\mathcal{Q}} l_u$ and observe $l \cdot w - v = \sum_{u\in\mathcal{Q}}(l_u \cdot w - u) \in \mathcal{S}(N)$. □

REMARK 6.5 The definition of facial implication can be extended as follows. A (finite) set of structural imsets $\mathcal{Q}$ facially implies a structural imset w (write $\mathcal{Q} \rightharpoonup w$) if $\mathcal{M}_w \subseteq \mathcal{M}$ for every structural independence model $\mathcal{M}$ such that $\bigcup_{u\in\mathcal{Q}} \mathcal{M}_u \subseteq \mathcal{M}$. However, by Consequence 6.1 this condition is equivalent to the requirement $\mathcal{M}_w \subseteq \sup_{u\in\mathcal{Q}} \mathcal{M}_u \equiv \mathcal{M}_v$ where $v = \sum_{u\in\mathcal{Q}} u$. Thus, the extension of facial implication of this type is not needed because it is covered by the current definition of facial implication. △

6.2.2 Skeletal characterization of facial implication

LEMMA 6.2 Let u, v are structural imsets over N. Then $u \rightharpoonup v$ iff

$$\forall\, m \in \mathcal{K}_\ell^\diamond(N) \quad \langle m, v\rangle > 0 \;\Rightarrow\; \langle m, u\rangle > 0, \tag{6.8}$$

which is equivalent to the condition

$$\langle m, v\rangle > 0 \;\Rightarrow\; \langle m, u\rangle > 0 \quad \text{for every supermodular function } m \text{ over } N. \tag{6.9}$$

Moreover, the condition (6.8) is also equivalent to the requirement

$$l_* \cdot u - v \in \mathcal{S}(N) \quad \text{whenever } l_* \in \mathbb{N} \text{ such that } l_* \geq \langle r, v\rangle \text{ for every } r \in \mathcal{K}_\ell^\diamond(N). \tag{6.10}$$

Proof: Evidently (6.9) ⇒ (6.8). Conversely, if (6.8) then observe by Lemma 5.3 and Observation 5.1(i) that $\langle m, v\rangle > 0$ implies $\langle m, u\rangle > 0$ for every ℓ-standardized supermodular function $m \in \mathcal{K}_\ell(N)$. However, every supermodular function is strongly equivalent to a function of this type by Lemma 5.2 which means that (6.9) holds.

By Lemma 6.1 $u \rightharpoonup v$ iff the condition (6.2) holds. However, by Lemma 5.4 it is equivalent to the condition

$$\exists\, l \in \mathbb{N} \quad \forall\, m \in \mathcal{K}_\ell^\diamond(N) \quad l \cdot \langle m, u\rangle \geq \langle m, v\rangle\,, \tag{6.11}$$

which implies (6.8). The next step is to show that (6.8) implies

$$\forall\, m \in \mathcal{K}_\ell^\diamond(N) \quad l_* \cdot \langle m, u\rangle \geq \langle m, v\rangle \quad \text{for } \; l_* \in \mathbb{N} \; \text{ from (6.10)}\,. \tag{6.12}$$

Indeed, if $m \in \mathcal{K}_\ell^\diamond(N)$ such that $\langle m, v\rangle \leq 0$ then $l_* \cdot \langle m, u\rangle \geq 0 \geq \langle m, v\rangle$ by Observation 5.1(iii). If $m \in \mathcal{K}_\ell^\diamond(N)$ such that $\langle m, v\rangle > 0$ then (6.8) implies $\langle m, u\rangle > 0$. However, as both m and u are imsets $\langle m, u\rangle \in \mathbb{Z}$ and therefore $\langle m, u\rangle \geq 1$ and the assumption about l_* implies $l_* \cdot \langle m, u\rangle \geq l_* \geq \langle m, v\rangle$. The condition (6.12) then implies $l_* \cdot u - v \in \mathcal{S}(N)$ by Lemma 5.4. Thus, (6.8)⇒(6.10). Since $\mathcal{K}_\ell^\diamond(N)$ is finite $l_* \in \mathbb{N}$ satisfying the requirement from (6.10) exists which means that (6.10) implies $u \rightharpoonup v$ by Lemma 6.1. □

The role of the way of standardization of supermodular functions is not substantial in the above result. One can easily derive an analogous result with the u-skeleton respectively with the o-skeleton in place of the ℓ-skeleton by a similar procedure (see Remark 5.6).

It follows from Lemma 6.1 that one has $u \rightharpoonup u_{\langle A,B|C\rangle}$ for a structural imset u over N and $\langle A, B|C\rangle \in \mathcal{T}(N)$ iff $\langle A, B|C\rangle \in \mathcal{M}_u$. Therefore Lemma 6.2 can be viewed as an alternative criterion of testing whether a disjoint triplet over N is represented in a structural imset over N. Note that Lemma 6.1 is suitable in the situation one wants to confirm the hypothesis that $u \rightharpoonup v$ while Lemma 6.2, namely the conditions (6.8) and (6.9), is suitable in the situation one wants to disprove $u \rightharpoonup v$. This is illustrated by Example 6.3 below. Well, the relation of these two criteria of facial implication is analogous to the relation of moralization and d-separation criteria in case of DAG models (see Section 3.2).

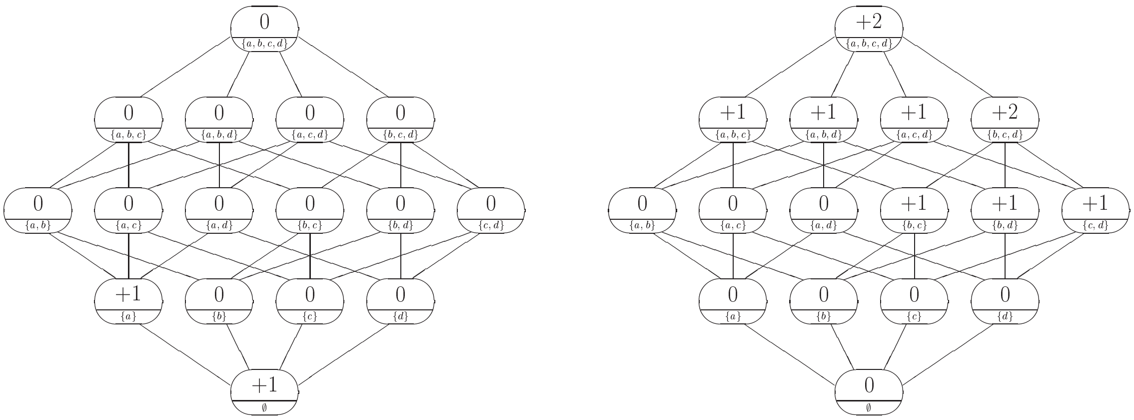


Figure 6.4: Strongly equivalent elements of the u-skeleton and the ℓ-skeleton.

EXAMPLE 6.3 Suppose $N = \{a, b, c, d\}$, consider the combinatorial imsets u and v from (6.5) in Example 6.2 and a semi-elementary imsets $w = u_{\langle b,acd|\emptyset\rangle}$. The fact $u \rightharpoonup v$ was verifed in Example 6.2 using direct characterization of facial implication, namely by means of the condition (6.3) of Lemma 6.1 with $k = 2$. To disprove $u \rightharpoonup w$ consider supermodular imset $m^{\{a\}\downarrow}$ (see p. 34) which is shown in the left-hand picture of Figure 6.4. Observe $\langle m^{\{a\}\downarrow}, w\rangle = 1$, $\langle m^{\{a\}\downarrow}, u\rangle = 0$ and apply Lemma 6.2, the condition (6.9), to see that $\neg(u \rightharpoonup w)$. Note that $m^{\{a\}\downarrow}$ belongs to the u-skeleton $\mathcal{K}_u^\diamond(N)$ and the corresponding element of the ℓ-skeleton (see Remark 5.6) is in the right-hand picture of Figure 6.4. $\diamondsuit$

An easy consequence of Lemma 6.2 is the following criterion of facial equivalence of structural imsets.

CONSEQUENCE 6.2 Let u, v be structural imsets over N. Then $u \rightleftharpoons v$ iff

$$\forall\, m \in \mathcal{K}_\ell^\diamond(N) \quad \langle m, u\rangle > 0 \quad \text{iff} \quad \langle m, v\rangle > 0\,, \tag{6.13}$$

which is equivalent to the condition that $\langle m, u\rangle > 0 \;\Leftrightarrow\; \langle m, v\rangle > 0$ for every supermodular function m over N.

Note that the skeletal criteria of testing facial implication and equivalence are effective in particular in case $|N| \leq 4$ since $|\mathcal{K}_\ell^\diamond(N)|$ is small in this case - see Remark 5.4. They are still implementable in case $|N| = 5$; a computer program which realizes facial implication of elementary imsets over five-element set can be found at

http://www.utia.cas.cz/user_data/studeny/fivevar.htm .

As the ℓ-skeleton is not at disposal in case $|N| \geq 6$ the only available criterion in that case is the criterion from Lemma 6.1.

6.2.3 Adaptation to a distribution framework

Let us consider a class Ψ of probability measures over N (a distribution framework) which satisfies the following two conditions:

$$\text{for every } P \in \Psi \text{ there exists a structural imset } u \text{ over } N \text{ such that } \mathcal{M}_u = \mathcal{M}_P, \tag{6.14}$$

$$\text{for every pair } P, Q \in \Psi \text{ there exists } R \in \Psi \text{ such that } \mathcal{M}_R = \mathcal{M}_P \cap \mathcal{M}_Q. \qquad (6.15)$$

There are at least three examples of distribution frameworks satisfying these two natural conditions: the class of measures with finite multiinformation, the class of discrete measures and the class of positive discrete measures (see Theorem 5.2 and Lemma 2.9). The goal of this section is to show that after suitable restriction of the class of structural imsets facial equivalence and Markov equivalence relative to Ψ coincide.

A structural imset u over N is representable in Ψ, shortly *Ψ-representable*, if there exists $P \in \Psi$ which is perfectly Markovian with respect to u, i.e. $\mathcal{M}_u = \mathcal{M}_P$. Evidently, every structural imset which is facially equivalent to a Ψ-representable structural imset is Ψ-representable as well. The class of Ψ-representable structural imsets over N will be denoted by $\mathcal{S}_\Psi(N)$.

Lemma 6.3 Let Ψ be a class of probability measures over N satisfying (6.14) and (6.15) and $u \in \mathcal{S}(N)$. Then the class of structural imsets Markov equivalent to u relative to Ψ is the union of a finite collection $\mho$ of facial equivalence classes ordered by relation $\rightharpoonup$. Moreover, the poset $(\mho, \rightharpoonup)$ has the greatest element which is the only class of facial equivalence $\wp \in \mho$ consisting of Ψ-representable imsets.

Proof: The first claim of the lemma follows easily from Observation 6.1. Let us put $\mathcal{M} = \bigcap_{P \in \Psi(u)} \mathcal{M}_P$ where $\Psi(u)$ is defined by (6.1) and

$$\Phi = \{ P \in \Psi\, ; \quad A \perp\!\!\!\perp B \,|\, C \;[P] \quad \text{whenever } \langle A, B|C \rangle \in \mathcal{M} \,\}\,.$$

The inclusion $\Psi(u) \subseteq \Phi$ follows directly from the definition of $\mathcal{M}$. The fact $\mathcal{M}_u \subseteq \mathcal{M}$ implies $\Phi \subseteq \Psi(u)$ and therefore $\Phi = \Psi(u)$. As $\mathcal{T}(N)$ is finite the set $\{\mathcal{M}_P; P \in \Psi(u)\}$ is also finite and one can show by repetitive application of the assumption (6.15) that $R \in \Psi(u)$ such that $\mathcal{M}_R = \mathcal{M}$ exists. By (6.14) a structural imset v with $\mathcal{M}_v = \mathcal{M}_R = \mathcal{M}$ exists, which means $\Phi = \Psi(v)$. Thus, u and v are Markov equivalent relative to Ψ. As $R \in \Psi$ is perfectly Markovian with respect to v the imset v is Ψ-representable.

Suppose that $w \in \mathcal{S}(N)$ such that $\Psi(w) = \Psi(u)$ and observe that

$$\mathcal{M}_w \subseteq \bigcap_{P \in \Psi(w)} \mathcal{M}_P = \bigcap_{P \in \Psi(u)} \mathcal{M}_P = \mathcal{M} = \mathcal{M}_v\,.$$

Thus $v \rightharpoonup w$ and the class $\wp$ of imsets facially equivalent to v is the greatest element of $(\mho, \rightharpoonup)$. If w is Ψ-representable then $Q \in \Psi$ with $\mathcal{M}_w = \mathcal{M}_Q$ exists and $Q \in \Psi(w) = \Psi(u)$ which implies $\mathcal{M} \subseteq \mathcal{M}_Q = \mathcal{M}_w \subseteq \mathcal{M}$. Hence, $\mathcal{M}_w = \mathcal{M} = \mathcal{M}_v$ which says $w \rightleftharpoons v$. □

Remark 6.6 Note that $(\mho, \rightharpoonup)$ is even a join semi-lattice. Indeed, $v \rightharpoonup \tilde{w} \rightharpoonup w$ and $\Psi(w) = \Psi(v)$ implies $\Psi(\tilde{w}) = \Psi(v)$ for $v, \tilde{w}, w \in \mathcal{S}(N)$. Hence, $\Psi(u+w) = \Psi(v)$ for structural imsets u, w with $\Psi(u) = \Psi(w) = \Psi(v)$ where $v \in \mathcal{S}(N)$ belongs to the greatest element $\wp$ of $\mho$ mentioned in Lemma 6.3. By Consequence 6.1 $\mathcal{M}_{u+w} = \mathcal{M}_u \vee \mathcal{M}_w$ which means that $u+w$ represents the join of u and w in $(\mho, \rightharpoonup)$. On the other hand, $(\mho, \rightharpoonup)$ need not be closed under the operation of meet in the lattice of structural imsets.

It may happen that $\mho$ consists of one class of facial equivalence only. This means that the class of imsets Markov equivalent to u coincides with the class of imsets facially equivalent to u. For example, this phenomenon is quite common in case $|N| = 4$ for the class of discrete measures over N: one has 18300 Markov equivalence classes and 22108 facial equivalence classes then [106, 107]. △

The following fact immediately follows from Lemma 6.3.

Consequence 6.3 Let Ψ be a class of probability measures over N satisfying (6.14) and (6.15). Consider the collection of Ψ-representable structural imsets $\mathcal{S}_\Psi(N)$ over N. Facial and Markov equivalence relative to Ψ coincide for imsets from $\mathcal{S}_\Psi(N)$.

In the considered case the class $\mathcal{S}_\Psi(N)$ satisfies both the requirements of faithfulness and the requirement of completeness relative to the class of CI structures arising within Ψ which were mentioned in Section 1.1. Thus, a theoretical solution of those problems is at disposal but a practical question how to recognize imsets from $\mathcal{S}_\Psi(N)$ remains to be solved then.

Remark 6.7 The idea of implementation of respective deductive mechanism on a computer is as follows. Except usual algebraic operations with structural imsets one needs to implement an additional operation which ascribes the respective Ψ-representable structural imset $v \in \mathcal{S}_\Psi(N)$ to every structural imset $u \in \mathcal{S}(N)$. Suppose that $u_1, \ldots, u_n$, $n \geq 1$ are structural imsets which represent input pieces of information about CI structure induced by an unknown distribution P which is known to belong to a given distribution framework Ψ (a subclass of the class of measures with finite multiinformation which satisfies (6.15)). The sum $u = \sum_{i=1}^{n} u_i$ then represents aggregated information about CI structure of P. But within the considered distribution framework Ψ over more can be deduced: one should find the respective $v \in \mathcal{S}_\Psi(N)$ which represents necessary conclusions of input pieces of information about CI structure of any $P \in \Psi$.

Nevertheless, possible inherent complexity of the problem of description of the lattice of CI structures arising within Ψ cannot be avoided. Indeed, implementation of the operation ascribing respective $v \in \mathcal{S}_\Psi(N)$ to every $u \in \mathcal{S}(N)$ may appear to be complicated (see Remark 5.7 for analogous consideration). Hopefully, the presented approach helps to decompose the original problem properly. $\triangle$

6.3 Testing facial implication

This section deals with implementation tasks connected with direct characterization of facial implication.

6.3.1 Testing structural imsets

The first natural question is how to recognize a structural imset. One possible method is given by Theorem 5.1 but, as explained in Remark 5.4, that method is not feasible in case $|N| \geq 6$. Thus, only the direct definition of structural imset is available in general. Therefore one needs to know whether the corresponding procedure is decidable. As explained in Remark 6.3 testing of combinatorial imsets is quite clear. One needs to know whether the natural number by which a structural imset must be multiplied to get a combinatorial imset is somehow limited.

Lemma 6.4 There exists $n \in \mathbb{N}$ such that

$$\forall \text{ imset } u \text{ over } N \quad u \in \mathcal{S}(N) \text{ iff } n \cdot u \in \mathcal{C}(N)\,. \tag{6.16}$$

Proof: One can apply Theorem 16.4 from [90] with says that every pointed rational polyhedral cone $C \subseteq \mathbb{R}^n$, $n \geq 1$ has a (unique) minimal integral *Hilbert basis* generating C that is (minimal) finite set $B \subseteq \mathbb{Z}^n$ such that

$$\forall\, \boldsymbol{x} \in C \cap \mathbb{Z}^n \quad \boldsymbol{x} = \sum_{\boldsymbol{y} \in B} k_{\boldsymbol{y}} \cdot \boldsymbol{y} \ \text{ for some } \ k_{\boldsymbol{y}} \in \mathbb{Z}^+$$

and $con(B) = C$ (which implies $B \subseteq C$). One can apply this result to the rational polyhedral cone $con(\mathcal{E}(N)) \subseteq \mathbb{R}^{\mathcal{P}(N)}$ which is pointed by Observation 4.1 as $\langle m_*, t\rangle > 0$ for every non-zero $t \in con(\mathcal{E}(N))$. Moreover, by Fact from Section 10.8.2 an imset u over N belongs to $con(\mathcal{E}(N))$ iff it is structural. Thus, a finite set of structural imsets $\mathcal{H}(N)$ exists such that

$$\forall\, u \in \mathcal{S}(N) \quad u = \sum_{v \in \mathcal{H}(N)} k_v \cdot v \ \text{ for some } \ k_v \in \mathbb{Z}^+ . \tag{6.17}$$

One can find $n(v) \in \mathbb{N}$ for every $v \in \mathcal{H}(N)$ such that $n(v) \cdot v$ is a combinatorial imset and put $n = \prod_{v \in \mathcal{H}(N)} n(v)$. Clearly, $n \cdot v \in \mathcal{C}(N)$ for every $v \in \mathcal{H}(N)$ and (6.17) implies that $n \cdot u \in \mathcal{C}(N)$ for every $u \in \mathcal{S}(N)$. □

Natural question is what is the minimal $n \in \mathbb{N}$ satisfying (6.16). I do not know the answer in case $|N| \geq 5$ (see Theme 12 on p. 143). But if $|N| \leq 4$ then $n = 1$: let me formulate as a separate observation the main result of [101].

Observation 6.2 If $|N| \leq 4$ then the class of structural imsets over N coincides with the class of combinatorial imsets over N.

Remark 6.8 The least $n \in \mathbb{N}$ satisfying (6.16) may appear to be too high. Alternative approach to direct testing structural imsets could be based on the concept of minimal integral Hilbert basis $\mathcal{H}(N)$ mentioned in the proof of Lemma 6.4 (see Theme 11). It follows from the proof of Theorem 16.4 of [90] that $\mathcal{H}(N)$ has the form

$$\mathcal{H}(N) = \{v \in \mathcal{S}(N);\ v \neq 0 \ \&\ \neg[v = v_1 + v_2 \ \text{ where } \ v_1, v_2 \in \mathcal{S}(N),\ v_1 \neq 0 \neq v_2]\,\}.$$

The fact that every elementary imset generates an extreme ray of $con(\mathcal{E}(N))$ allows to derive $\mathcal{E}(N) \subseteq \mathcal{H}(N)$. Of course, $\mathcal{H}(N) = \mathcal{E}(N)$ if $|N| \leq 4$ by Observation 6.2. The idea is to characterize $\mathcal{H}(N)$ in general. Then every imset u over N can be effectively tested whether it can be written as a combination of imsets from $\mathcal{H}(N)$ with non-negative integral coefficients. Indeed, one can modify the procedure described in Remark 6.3. △

6.3.2 Grade

Another natural question arising in connection with Lemma 6.1 is whether there exists $l \in \mathbb{N}$ such that

$$\forall\, u \in \mathcal{S}(N) \ \ v \in \mathcal{E}(N) \quad u \rightharpoonup v \ \text{ iff } \ l \cdot u - v \in \mathcal{S}(N). \tag{6.18}$$

The answer is yes. Evidently, if $l \in \mathbb{N}$ satisfies (6.18) then every $l' \in \mathbb{N}$, $l' \geq l$ satisfies it as well. Therefore one is interested in minimal $l \in \mathbb{N}$ satisfying (6.18) which appears to depend on N. Actually, it depends on $|N|$ only because of inherent one-to-one correspondence between $\mathcal{E}(N)$ and $\mathcal{E}(M)$, respectively between $\mathcal{S}(N)$ and $\mathcal{S}(M)$ for sets of variables N and M of the same cardinality. The following number is a good candidate for the minimal $l \in \mathbb{N}$ satisfying (6.18).

Supposing $|N| \geq 2$ let us call the *grade*, denoted by $gra(N)$, the natural number

$$gra(N) = \max\left\{ \langle r, w\rangle;\ r \in \mathcal{K}_\ell^\diamond(N) \quad w \in \mathcal{E}(N) \right\}. \tag{6.19}$$

Evidently, $gra(N)$ depends on $|N|$ only. Lemma 6.2, the condition (6.10), implies this:

Consequence 6.4 If $|N| \geq 2$ then $l = gra(N)$ satisfies (6.18).

Consequence 6.4 leads to an effective criterion of testing facial implication of elementary imsets in case $|N| \leq 4$ which utilizes the fact that structural and combinatorial imsets coincide in this case.

Consequence 6.5 Suppose that $2 \leq |N| \leq 4$, u is a structural imset over N and v an elementary imset over N. Then $u \rightharpoonup v$ iff $u - v$ is a combinatorial imset.

Proof: The first observation is that if $|N| \leq 4$ then $\langle m, v\rangle \in \{0, 1\}$ for every $m \in \mathcal{K}_\ell^\diamond(N)$ and $v \in \mathcal{E}(N)$ - see [101]. Thus, $gra(N) = 1$ and by Consequence 6.4 one has $u \rightharpoonup v$ iff $u - v \in \mathcal{S}(N)$ which is equivalent to $u - v \in \mathcal{C}(N)$ by Observation 6.2. □

However as shown in [116] $gra(N) = 7$ in case $|N| = 5$. In fact, an example from Section 4.3 of [116] shows that the minimal natural number l for which (6.18) holds is just 7 in case $|N| = 5$. The question what is the minimal $l \in \mathbb{N}$ satisfying (6.18) (c.f. p. 143) is partially answered by the following lemma.

Lemma 6.5 Suppose that $|N| \geq 2$. Then the minimal $l_* \in \mathbb{N}$ satisfying

$$\forall\, u \in \mathcal{C}(N)\ \forall v \in \mathcal{E}(N) \quad u \rightharpoonup v \ \text{ iff } \ l_* \cdot u - v \in \mathcal{S}(N) \tag{6.20}$$

is the upper integer part of

$$gra_*(N) = \max_{m \in \mathcal{K}_\ell^\diamond(N)} \frac{\max\left\{\langle m, w\rangle\,;\, w \in \mathcal{E}(N)\right\}}{\min\left\{\langle m, w\rangle\,;\, w \in \mathcal{E}(N)\ \langle m, w\rangle \neq 0\right\}}. \tag{6.21}$$

Proof: To show that every $l_* \in \mathbb{N}$ with $l_* \geq gra_*(N)$ satisfies (6.20) the procedure from the proof of Lemma 6.2 can be used. The only modification is that in case $m \in \mathcal{K}_\ell^\diamond(N)$ with $\langle m, u\rangle > 0$ the fact $u \in \mathcal{C}(N)$ implies $\langle m, u\rangle \geq \min\left\{\langle m, w\rangle\,;\, w \in \mathcal{E}(N)\ \langle m, w\rangle \neq 0\right\}$ which allows to write

$$l_* \cdot \langle m, u\rangle \geq gra_*(N) \cdot \min_{w \in \mathcal{E}(N), \langle m, w\rangle \neq 0} \langle m, w\rangle \geq \max_{w \in \mathcal{E}(N)} \langle m, w\rangle \geq \langle m, v\rangle.$$

To show that for every $l \in \mathbb{N}$ with $l < gra_*(N)$ there exists a combinatorial imset u and an elementary imset v such that $u \rightharpoonup v$ and $l \cdot u - v \notin \mathcal{S}(N)$ choose and fix $m \in \mathcal{K}_\ell^\diamond(N)$ for which the maximum in (6.21) is achieved. Then choose $\tilde{w} = u_{\langle i,j|K\rangle} \in \mathcal{E}(N)$ minimizing non-zero value $\langle m, w\rangle$, $w \in \mathcal{E}(N)$ and $v \in \mathcal{E}(N)$ maximizing $\langle m, w\rangle$ for $w \in \mathcal{E}(N)$. By Consequence 5.4 a $\tilde{u} \in \mathcal{C}(N)$ with $\mathcal{M}^m = \mathcal{M}_{\tilde{u}}$ exists. Put $u = \tilde{u} + \tilde{w}$. By Lemma 6.1 $\mathcal{M}^m = \mathcal{M}_{\tilde{u}} \subseteq \mathcal{M}_u$. As $\langle i, j|K\rangle \in \mathcal{M}_u \setminus \mathcal{M}^m$ the fact that m is a skeletal imset (see

Section 5.2) implies $\mathcal{M}_u = \mathcal{T}(N) \supseteq \mathcal{M}_v$ which means $u \rightharpoonup v$. On the other hand, by Observation 5.6 $\langle m, \tilde{u} \rangle = 0$ and therefore

$$l < gra_*(N) = \frac{\langle m, v \rangle}{\langle m, \tilde{w} \rangle} = \frac{\langle m, v \rangle}{\langle m, u \rangle} \quad \text{implies} \quad \langle m, l \cdot u - v \rangle < 0,$$

which means $l \cdot u - v \notin \mathcal{S}(N)$ by Lemma 5.4. □

Remark 6.9 Note that the type of the skeleton is not material in the above result. In fact, the ℓ-skeleton can be replaced either by the u-skeleton or by the o-skeleton and the respective constant $gra_*(N)$ has the same value. Indeed, it follows from Consequence 5.3 that for every skeletal supermodular function $\widetilde{m}$ there exists $\alpha > 0$ such that $\langle \widetilde{m}, u \rangle = \alpha \cdot \langle m, u \rangle$ for every $u \in \mathcal{E}(N)$ where $m \in \mathcal{K}_\ell^\diamond(N)$ is the unique model equivalent element of the ℓ-skeleton. Thus, the ratios maximalized in (6.21) are invariants of classes of model equivalence of skeletal imsets. Note that if $|N| \leq 5$ then

$$\forall\, m \in \mathcal{K}_\ell^\diamond(N) \quad \min\, \{\langle m, u \rangle;\ u \in \mathcal{E}(N)\ \langle m, u \rangle \neq 0\} = 1\,, \tag{6.22}$$

which implies that $gra(N) = gra_*(N)$ in this case. Thus, if the hypothesis (6.22) holds in general (see Question 8) then $gra(N)$ is the least $l \in \mathbb{N}$ satisfying (6.18) by Consequence 6.4 and Lemma 6.5. Note that an analogue of (6.22) holds for $|N| \leq 5$ and the u-skeleton (because of the operation of reflection mentioned on p. 131 or in Section 5.1.3 of [116]) but not for the o-skeleton. This is maybe the main difference between o-standardization and ℓ-standardization. △

6.4 Invariants of facial equivalence

This section deals with some of those attributes of structural imsets which are either invariable with respect to facial equivalence or characterize classes of facial equivalence.

Let u be a structural imset over N. By *effective domain* of u denoted by $\mathcal{D}_u^*$ is understood the class of sets $T \subseteq N$ such that $S' \subseteq T \subseteq S$ for some $S', S \subseteq N$ with $u(S'), u(S) > 0$, that is $\mathcal{D}_u^* = (\mathcal{D}_u^+)^\downarrow \cap (\mathcal{D}_u^+)^\uparrow$. Recall that $\mathcal{U}_u = (\mathcal{D}_u^+)^\downarrow$ is nothing but the *upper class* of u from Section 4.2.3. The *region* of u, denoted by $\mathcal{R}_u$, is the class of subsets of N obtained as follows:

$$\mathcal{R}_u = \bigcup_{\langle i,j|K \rangle \in \mathcal{M}_u \cap \mathcal{T}_\epsilon(N)} \{K, iK, jK, ijK\} = \bigcup_{\langle A,B|C \rangle \in \mathcal{M}_u \setminus \mathcal{T}_\emptyset(N)} \{C, AC, BC, ABC\}\,. \tag{6.23}$$

Note that the equality of the unions in (6.23) over the class of elementary triplets and over the class of all non-trivial triplets can be easily derived from the fact that $\mathcal{M}_u$ is a semi-graphoid (Lemma 4.6) by means of Lemma 2.2 on p. 16.

Lemma 6.6 Given a structural imset u over N one has $\mathcal{R}_u \subseteq \mathcal{D}_u^*$. If $u, v \in \mathcal{S}(N)$ such that $u \rightleftharpoons v$ then $\mathcal{U}_u = \mathcal{U}_v$, $\mathcal{D}_u^* = \mathcal{D}_v^*$ and $\mathcal{R}_u = \mathcal{R}_v$. Moreover, $S \notin \mathcal{R}_u$ for $S \subseteq N$ iff $w(S) = 0$ for every $w \in \mathcal{S}(N)$ which is facially equivalent to u.

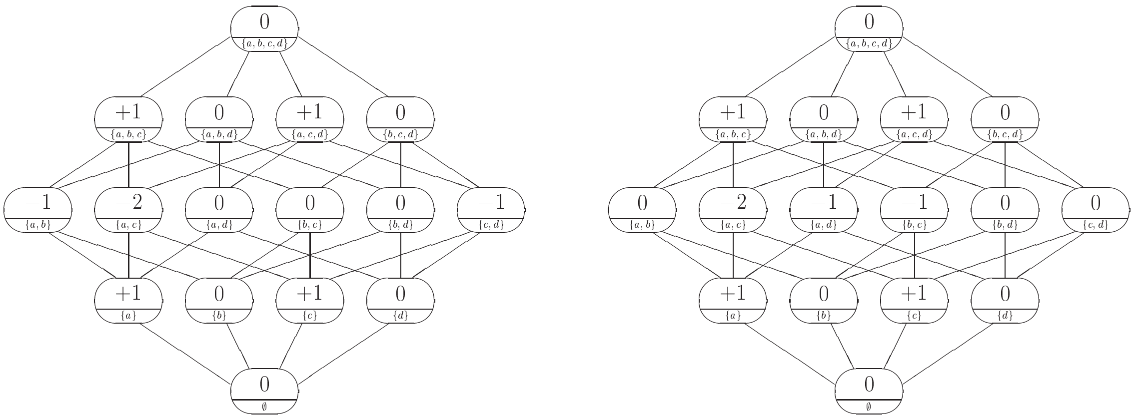


Figure 6.5: Two structural imsets with the same effective domain but different regions.

Proof: To show $\mathcal{R}_u \subseteq \mathcal{D}_u^*$ consider elementary triplet $\langle i,j|K\rangle \in \mathcal{M}_u$ and find $k \in \mathbb{N}$ with $k \cdot u - u_{\langle i,j|K\rangle} \in \mathcal{S}(N)$. As $\langle m^{ijK\uparrow}, u_{\langle i,j|K\rangle}\rangle > 0$ by Observation 5.1 $k \cdot \langle m^{ijK\uparrow}, u\rangle > 0$. Hence, $u(T) > 0$ for some $ijK \subseteq T \subseteq N$ which means $K, iK, jK, ijK \in (\mathcal{D}_u^+)^{\downarrow}$. Analogously, $\langle m^{K\downarrow}, u_{\langle i,j|K\rangle}\rangle > 0$ implies $\langle m^{K\downarrow}, u\rangle > 0$ and $u(S') > 0$ for some $S' \subseteq K$ which means $K, iK, jK, ijK \in (\mathcal{D}_u^+)^{\uparrow}$. The next observation is that $S \in ((\mathcal{D}_u^+)^{\downarrow})^{\max}$ iff $\langle m^{S\uparrow}, u\rangle > 0$ and $\langle m^{T\uparrow}, u\rangle = 0$ for every $T \supset S$. Indeed, $S \in ((\mathcal{D}_u^+)^{\downarrow})^{\max}$ means $u(S) > 0$ and $u(T) = 0$ for $T \supset S$ by Observation 4.5 ($\mathcal{L}_u \subseteq \mathcal{U}_u$) and one can show by reverse induction on $|T|$ that $u(T) = 0$ for every $S \subset T \subseteq N$ iff $\langle m^{T\uparrow}, u\rangle = 0$ for every $S \subset T \subseteq N$. Analogous arguments allow to show that $S \in ((\mathcal{D}_u^+)^{\uparrow})^{\min}$ iff $\langle m^{S\downarrow}, u\rangle > 0$ and $\langle m^{T\downarrow}, u\rangle = 0$ for every $T \subset S$ (replace $\subseteq$ by $\supseteq$). However, by Consequence 6.2 and Observation 5.1 the conditions $\langle m^{A\uparrow}, u\rangle > 0$, $\langle m^{A\uparrow}, u\rangle = 0$, $\langle m^{A\downarrow}, u\rangle > 0$, $\langle m^{A\downarrow}, u\rangle = 0$ for $A \subseteq N$ are invariable with respect to facial equivalence. Therefore, $\mathcal{U}_u$ and $\mathcal{D}_u^*$ are invariable as well. Analogous claim about $\mathcal{R}_u$ is evident because of its definition in terms of $\mathcal{M}_u$.

To show that $S \notin \mathcal{R}_u$ implies $u(S) = 0$ write $n \cdot u = \sum_{v \in \mathcal{E}(N)} k_v \cdot v$ where $n \in \mathbb{N}$, $k_v \in \mathbb{Z}^+$ and observe $\langle i,j|K\rangle \in \mathcal{M}_u$ whenever $k_v > 0$ for $v = u_{\langle i,j|K\rangle}$. As $S \notin \mathcal{R}_u$ one has $v(S) = 0$ for every $v \in \mathcal{E}(N)$ of this kind which implies $n \cdot u(S) = 0$. The consideration holds for any $w \in \mathcal{S}(N)$ with $u \rightleftharpoons w$ in place of u. Conversely, suppose $S \in \mathcal{R}_u$, take elementary triplet $\langle i,j|K\rangle \in \mathcal{M}_u$ with $S \in \{K, iK, jK, ijK\}$ and observe that $w = u + k \cdot u_{\langle i,j|K\rangle}$ is facially equivalent to u for every $k \in \mathbb{N}$ by Lemma 6.1. One can find $k \in \mathbb{N}$ such that $w(S) \neq 0$. $\square$

However, the effective domain and the region of a structural imset may differ as the following example shows.

Example 6.4 There exist two structural imsets u, v over $N = \{a,b,c,d\}$ with the same effective domain but different regions. Consider the imset $u = u_{\langle b,c|a\rangle} + u_{\langle a,d|c\rangle}$ shown in the left-hand picture of Figure 6.5 and the imset $v = u_{\langle c,d|a\rangle} + u_{\langle a,b|c\rangle}$ shown in the right-hand picture of Figure 6.5. The set $\{a,d\}$ belongs to the effective domain $\mathcal{D}_u^* = \mathcal{D}_v^*$ and to the region $\mathcal{R}_v$ but it does not belong to the region $\mathcal{R}_u$.

On the other hand, regions and effective domains of structural imsets over N coincide in case $|N| \leq 3$ (c.f. Section 7.5.1 and Figure 7.8). $\diamondsuit$

Remark 6.10 The significance of the concept of effective domain is that it allows to restrict the considered class of elementary imsets when one tests whether an o-standardized imset is combinatorial - see Remark 6.3. Indeed, if $u = \sum_{v \in \mathcal{E}(N)} k_v \cdot v$ with $k_v \in \mathbb{Z}^+$ then for every $v = u_{\langle i,j|K \rangle}$ with $k_v > 0$ one has $\langle i,j|K \rangle \in \mathcal{M}_u$ and therefore by Lemma 6.6 $K, iK, jK, ijK \in \mathcal{R}_u \subseteq \mathcal{D}_u^*$. Thus, a reduced class of elementary imsets $v = u_{\langle i,j|K \rangle}$ satisfying $K, iK, jK, ijK \in \mathcal{D}_u^*$ can be considered. Observe that the effective domain $\mathcal{D}_u^*$ can be identified directly on basis of u. This is the main difference from the region $\mathcal{R}_u$ which gives even stronger restriction of the class of considered elementary imsets but the region cannot be immediately recognized only on basis of u. It is a characteristics of the respective class of facially equivalent structural imsets and can be identified on basis of the imset only partially as mentioned in Lemma 6.6.

However, by Observation 4.3 one can compute the corresponding level-degrees of u for $l = 0, \ldots, |N| - 2$ which may result in further restriction of the class of considered imsets - in particular if some of the level-degrees vanish. $\triangle$

Effective domains are attributes of structural imsets which allow to distinguish immediately imsets which are not facially equivalent. A natural question whether there exists a complete collection of invariant properties of similar type in sense that for every pair of structural imsets u and v over N at least one property of this type exists in which they differ. Consequence 6.2 gives a positive answer to this question. Indeed, every skeletal imset $m \in \mathcal{K}_\ell^\diamond(N)$ is associated with an invariant attribute of a structural imset u over N, namely the fact whether the scalar product $\langle m, u \rangle$ vanishes or not. The collection of these attributes is complete in the above mentioned sense.

However, as explained in Remark 5.4 this criterion does not seem feasible in case $|N| \geq 6$. Therefore, one is interested in invariants analogous to the effective domain or in relatively simple characteristics of facial equivalence like the region - see Direction 3 in Chapter 8. For example, if $|N| \leq 3$ then a completely distinguishing class of attributes is the effective domain together with the minimal lower classes (see Section 7.4.2) or the pattern (see Section 7.5.1).

Chapter 7

The problem of representative choice

This chapter deals with the problem of choice of suitable representative within a class of facially equivalent structural imsets. It is an advanced subtask of general equivalence question mentioned in Section 1.1 studied in the framework of structural imsets - an analogous question has already been treated in graphical frameworks - see Chapter 3 (the concept of essential graph and the concept of the largest chain graph). A few principles of representative choice are introduced and discussed in this chapter. Special attention is devoted to the representation of graphical models by structural imsets. The last section describes some other ideas whose aim is unique description of structural models.

7.1 Baricentral imsets

An imset u over N is called *baricentral* if it has the form

$$u = \sum_{w \in \mathcal{E}(N),\, u \rightharpoonup w} w \quad \text{or equivalently} \quad u = \frac{1}{2} \cdot \sum_{\langle a,b|C\rangle \in \mathcal{M}_u \cap \mathcal{T}_\epsilon(N)} u_{\langle a,b|C\rangle}. \tag{7.1}$$

Evidently, every elementary imset is baricentral and every baricentral imset u is a combinatorial imset with the degree $|\{w \in \mathcal{E}(N); u \rightharpoonup w\}|$. Moreover, the definition implies that every class of facial equivalence of structural imsets contains exactly one baricentral imset. Nevertheless, semi-elementary imset need not be baricentral. Given a semi-elementary imset $u_{\langle A,B|C\rangle}$ for $\langle A,B|C\rangle \in \mathcal{T}(N)$ the respective facially equivalent baricentral imset even need not be its multiple of despite the fact that the formulas

$$deg(u_{\langle A,B|C\rangle}) = |A| \cdot |B| \qquad |\{w \in \mathcal{E}(N)\,;\, u_{\langle A,B|C\rangle} \rightharpoonup w\}| = |A| \cdot |B| \cdot 2^{|A|-1} \cdot 2^{|B|-1} \tag{7.2}$$

suggest that it may be the case.

Example 7.1 There exists a semi-elementary imset v over $N = \{a,b,c,d\}$ such that no multiple $k \cdot v$, $k \in \mathbb{N}$ is a baricentral imset. Put $v = u_{\langle a,bcd|\emptyset\rangle}$ - see the left-hand picture of Figure 7.1. Then $u \rightharpoonup w \in \mathcal{E}(N)$ iff $w = u_{\langle a,e|C\rangle}$ where $e \in \{b,c,d\}$, $C \subseteq \{b,c,d\} \setminus \{e\}$ (c.f. Lemma 2.2). The respective baricentral imset u is shown in the right-hand picture of Figure 7.1. Observe that $12 = deg(u) = 4 \cdot deg(v)$ but $u \neq 4 \cdot v$ since the level-degrees of u and v are not proportional: $deg(v,l) = 1$ for $l = 0,1,2$ while $deg(u,0) = deg(u,2) = 3$ and $deg(u,1) = 6$. On the other hand, $u = 3 \cdot v + u_{\langle a,b|c\rangle} + u_{\langle a,c|d\rangle} + u_{\langle a,d|b\rangle}$. ♢

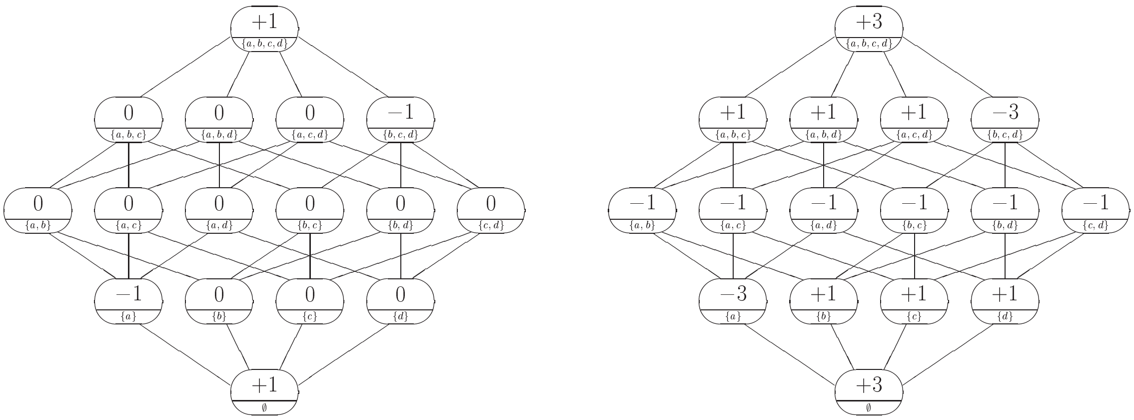


Figure 7.1: Respective non-proportional semi-elementary and baricentral imsets.

The significance of baricentral imsets consists in the fact that testing facial implication between them is very simple.

OBSERVATION 7.1 Let u, v are baricentral imsets over N. Then $u \rightharpoonup v$ iff $u - v$ is a combinatorial imset.

Proof: If $u \rightharpoonup v$ then $v \rightharpoonup w$ implies $u \rightharpoonup w$ for every $w \in \mathcal{E}(N)$ and $u - v \in \mathcal{C}(N)$ follows from (7.1). The converse follows from Lemma 6.1. $\square$

Note that testing combinatorial imsets is clear (see Remark 6.3). Analogous result holds if v is replaced by a semi-elementary imset (see Consequence 7.4 below). In particular, the whole induced model $\mathcal{M}_u$ can be easily identified on basis of a baricentral imset u over N (i.e. without 'multiplication').

REMARK 7.1 The terminology 'baricentral imset' was inspired by geometric idea that the class of structural imsets which are facially implied by $v \in \mathcal{S}(N)$ is nothing but the class of imsets belonging to the cone $con(\{w \in \mathcal{E}(N); v \rightharpoonup w\})$ (c.f. Remark 6.2 on p. 93). Thus, a (minimal) balanced combination of all extreme imsets of this cone forms the 'baricentre' of the cone.

Natural question is what is the number of baricentral imsets over N. A rough upper estimate can be obtained as follows. Suppose $n = |N| \geq 2$ and $S \subseteq N$, $|S| = k$. Then a limited number of $w \in \mathcal{E}(N)$ takes a non-zero value $w(S) \in \{-1, +1\}$. In particular, by (7.1) every baricentral imset u over N takes the value $u(S)$ in a finite set $\text{Ł}(S) \subseteq \mathbb{Z}$ which depends on $k = |S|$ only. Concretely, $\text{Ł}(S) = \{-k(n-k), \ldots, \binom{k}{2} + \binom{n-k}{2}\}$ for $2 \leq k \leq n-2$, $\text{Ł}(S) = \{0, \ldots, \binom{n}{2}\}$ for $k \in \{0, n\}$ and $\text{Ł}(S) = \{-n+1, \ldots, \frac{(n-1)(n-2)}{2}\}$ for $k \in \{1, n-1\}$. Thus, $|\text{Ł}(S)| = \binom{n}{2} + 1$ for every $S \subseteq N$. Since baricentral imsets are o-standardized it suffices to know their values for $2^n - n - 1$ sets only. Therefore, every baricentral imset can be represented as a function on $\{S \subseteq N; |S| \geq 2\}$ taking values in $\text{Ł}(S)$ for every S. The number of these functions is

$$\beta_n = \{\binom{n}{2} + 1\}^{2^n - n - 1}$$

which serves as an upper estimate of the number of baricentral imsets over N and therefore of the number of structural models over N.

The following gives an indirect comparison of memory demands when a structural model over N is represented either in the form of a baricentral imset or 'directly'. By Lemma 2.2 every semi-graphoid $\mathcal{M}$ over N is determined by $\mathcal{M} \cap \mathcal{T}_\epsilon(N)$. Thus, owing to symmetry property (see p. 15) it can be represented as a function on $\mathcal{E}(N)$ taking value in a two-element set. As $|\mathcal{E}(N)| = \binom{n}{2} \cdot 2^{n-2}$ the number of these functions is

$$\gamma_n = 2^{n \cdot (n-1) \cdot 2^{n-3}}.$$

One has $\beta_2 = \gamma_2$, $\beta_3 = 2^8 > 2^6 = \gamma_3$, $\beta_4 = 7^{11} > 2^{24} = \gamma_4$ and $\beta_5 = 11^{26} > 2^{80} = \gamma_5$. On the other hand $\beta_n \leq 2^{(n-2)\cdot(2^n-n-1)} < 2^{\binom{n}{2}\cdot 2^{n-2}} = \gamma_n$ for $n \geq 6$ so that 'asymptotically' the number of considered integral functions on $\{S \subseteq N; |S| \geq 2\}$ is lower than the number of binary functions on $\mathcal{E}(N)$! Only $(n-2)$ bits suffices to represent elements of $\mathrm{L}(S)$ for $S \subseteq N$ in case $n \geq 6$ which means that memory demands are slightly lower in case of representation by baricentral imsets. $\triangle$

On the other hand, the actual number of baricentral imsets (i.e. structural models) for $n = 2, 3, 4$ are much lower than the estimates from Remark 7.1 - see Example 5.1. The lattice of baricentral imsets over $\{a, b, c\}$ (ordered by $\rightharpoonup$) is shown in Figure 7.2.

Baricentral imsets provide quite good solution of the problem of representative choice from computational point of view. However, the question of getting respective baricentral imset from any given structural imset remains to be solved satisfactorily. For example, formulas ascribing respective baricentral imsets to graphical models are needed (see Theme 3 in Chapter 8). Relative disadvantage of baricentral imsets is that they do not seem to offer easy interpretation in comparison with 'standard' imsets for DAG models mentioned below.

7.2 Standard imsets

Some classic graphical models can be represented by certain 'standard' structural imsets which seem to exhibit important characteristics of the models. These standard representatives of graphical models may differ from baricentral representatives and seem to be more suitable from the point of view of interpretation. They are introduced in this section together with relevant basic facts. Note that the motive of later Sections 7.3 and 7.4 is to find out whether these exceptional representatives reflect some deeper principles so that the concept of standard imset could be extended even beyond the framework of graphical models.

7.2.1 Translation of DAG models

Let G be an acyclic directed graph over N. By a *standard imset for G* will be understood the imset u_G over N given by

$$u_G = \delta_N - \delta_\emptyset + \sum_{c \in N} \delta_{pa_G(c)} - \delta_{c \cup pa_G(c)}. \tag{7.3}$$

Lemma 7.1 Let G be an acyclic directed graph over N. Then the imset $u = u_G$ is a combinatorial imset and $\mathcal{M}_u = \mathcal{M}_G$. Moreover, $deg(u_G) = \frac{1}{2} \cdot |N| \cdot (|N|-1) - |\mathcal{A}(G)|$ where $|\mathcal{A}(G)|$ is the number of arrows in G.

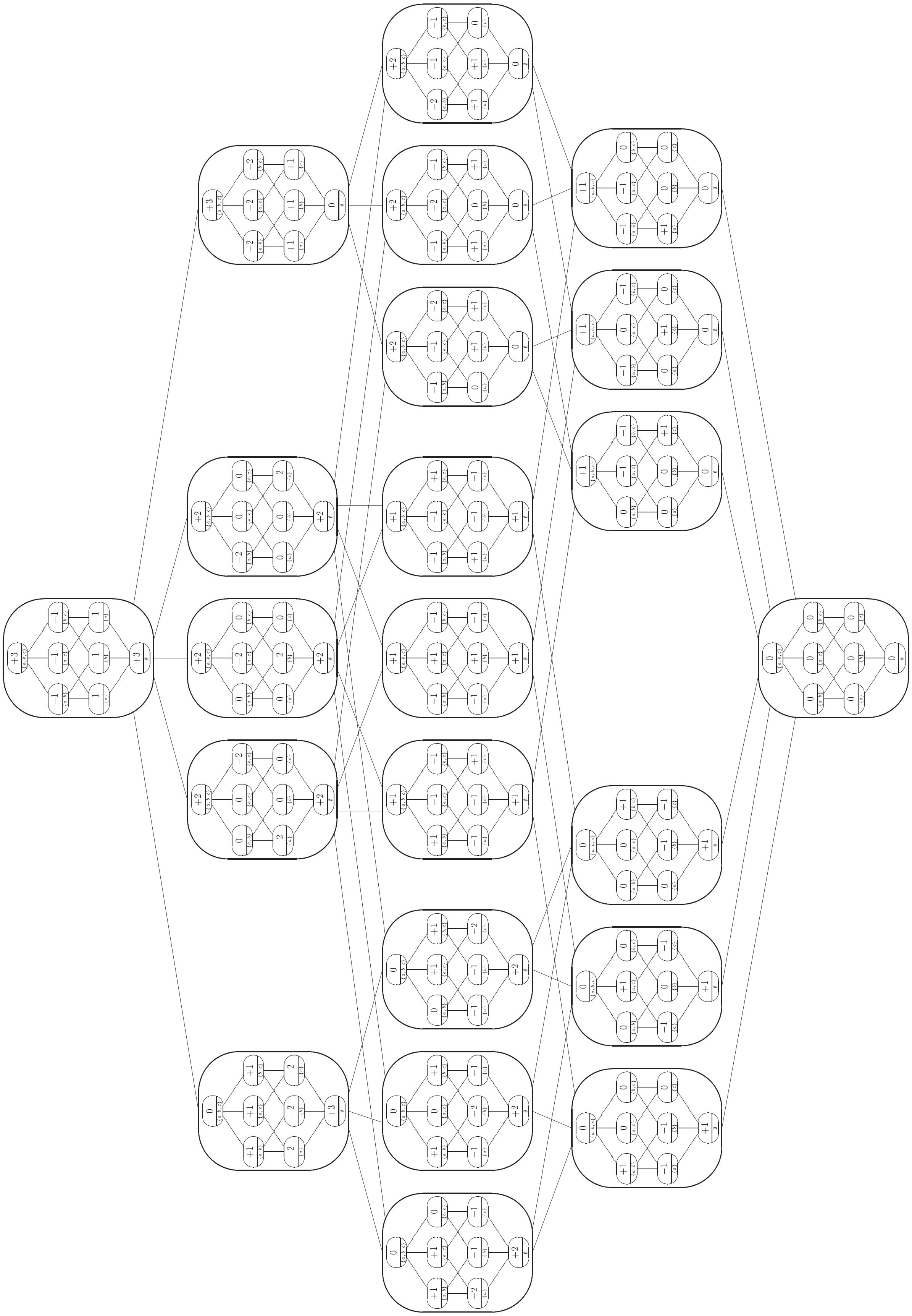


Figure 7.2: Baricentral imsets over $N = \{a, b, c\}$ (rotated).

Proof: Consider a fixed ordering $a_1, \ldots, a_n$, $n \geq 1$ of nodes of G consonant with direction of arrows and the corresponding causal input list (see Remark 3.4)

$$\langle\, a_j, a_1 \ldots a_{j-1} \setminus pa_G(a_j) \,|\, pa_G(a_j) \,\rangle \quad \text{for } j = 1, \ldots, n\,. \tag{7.4}$$

Introduce u_j as the semi-elementary imset corresponding to the j-th triplet from (7.4) for $j = 1, \ldots, n$ and write

$$u = \sum_{j=1}^{n} u_j = \sum_{j=1}^{n} \delta_{\{a_1,\ldots,a_j\}} - \delta_{\{a_1,\ldots,a_{j-1}\}} - \delta_{a_j pa(a_j)} + \delta_{pa(a_j)} = u_G$$

since almost all terms $\delta_{\{a_1,\ldots,a_j\}}$ are cancelled. Thus, u_G is a combinatorial imset and the substitution of $deg(u_j) = j - 1 - |pa_G(a_j)|$ into $deg(u) = \sum_{j=1}^{n} deg(u_j)$ gives the desired formula for $deg(u_G)$. Note that the formula above implies that $\sum_{j=1}^{n} u_j$ actually does not depend on the choice of causal input list.

Since $\mathcal{M}_u$ is a semi-graphoid containing (7.4) the result from [119] saying that $\mathcal{M}_G$ is the least semi-graphoid containing (7.4) implies $\mathcal{M}_G \subseteq \mathcal{M}_u$. For converse inclusion use the result from [31] implying that a discrete probability measure P over N with $\mathcal{M}_G = \mathcal{M}_P$ exists and Theorem 5.2 saying that $v \in \mathcal{S}(N)$ with $\mathcal{M}_P = \mathcal{M}_v$ exists. Since the list (7.4) belongs to $\mathcal{M}_G$ one has $v \rightharpoonup u_j$ for $j = 1, \ldots, n$ and therefore by Lemma 6.1 $v \rightharpoonup \sum_{j=1}^{n} u_j = u$ which means $\mathcal{M}_u \subseteq \mathcal{M}_v = \mathcal{M}_G$. □

REMARK 7.2 In fact, it was shown in the proof of Lemma 7.1 that $u_G \in \mathcal{S}_\Psi(N)$ where Ψ is the class of discrete measures over N (c.f. Section 6.2.3). △

Standard imsets appear to a suitable tool for testing Markov equivalence of acyclic directed graphs.

CONSEQUENCE 7.1 Let G, H be acyclic directed graphs over N. Then $\mathcal{M}_G = \mathcal{M}_H$ if and only if $u_G = u_H$.

Proof: By Lemma 7.1 $u_G = u_H \Rightarrow \mathcal{M}_G = \mathcal{M}_H$. The converse implication is shown in [42] as Consequence 3.1 concluding Remark 2 there. □

REMARK 7.3 Every semi-elementary imset over N is a standard imset for an acyclic directed graph over N. Indeed, given $\langle A, B|C\rangle \in \mathcal{T}(N)$ consider a total ordering of nodes of N in which the nodes of C precede the nodes of A which precede the nodes of B and these precede the nodes of $N \setminus ABC$. Take an undirected graph over N in which every pair of distinct nodes is a line except pairs $[a, b]$ where $a \in A$, $b \in B$. Consider a directed graph G which has this undirected graph as the underlying graph and has the direction of arrows consonant with the total ordering above. Then it makes no problem to see by the procedure in the proof of Lemma 7.1 that $u_G = u_{\langle A,B|C\rangle}$. △

7.2.2 Translation of decomposable models

Decomposable models, that is independence models induced by triangulated undirected graphs form an important class of graphical models - see Section 3.4.1. Let H be a triangulated undirected graph over N and $\mathcal{C}$ is the class of all its cliques. By *standard imset for* H will be understood the imset u_H over N given by

$$u_H = \delta_N + \sum_{\emptyset \neq \mathcal{B} \subseteq \mathcal{C}} (-1)^{|\mathcal{B}|} \cdot \delta_{\bigcap \mathcal{B}} \,. \tag{7.5}$$

It is shown below that u_H is a combinatorial imset (Consequence 7.2), the next lemma helps to compute u_H efficiently.

Lemma 7.2 Let H be a triangulated undirected graph over N and $\varrho : C_1, \ldots, C_m$, $m \geq 1$ is a sequence of (all) its cliques satisfying the running intersection property (see (3.1) on p. 45). Then

$$u_H = \delta_N - \sum_{i=1}^{m} \delta_{C_i} + \sum_{i=2}^{m} \delta_{S_i} \tag{7.6}$$

where $S_i = C_i \cap (\bigcup_{j<i} C_j)$ for $i = 2, \ldots, m$ are respective separators. In particular, the right-hand side of (7.6) does not depend on the choice of ϱ and can also be written as follows:

$$u_H = \delta_N - \sum_{C \in \mathcal{C}} \delta_C + \sum_{S \in \mathcal{S}} w(S) \cdot \delta_S \tag{7.7}$$

where $\mathcal{C}$ is the class of cliques of H, $\mathcal{S}$ is the class of separators and $w(S)$ denotes the multiplicity of a separator $S \in \mathcal{S}$.

Proof: The idea is to verify (7.6) by induction on $m = |\mathcal{C}|$. It is evident in case $m \leq 2$. If $m \geq 3$ then put $\mathcal{C}' = \mathcal{C} \setminus \{C_m\}$, $T = \bigcup \mathcal{C}'$ and $H' = H_T$. Observe that $C_1, \ldots, C_{m-1}$ is a sequence of all cliques of H' satisfying the running intersection property. Write by (7.5)

$$u_H = \delta_N - \delta_T + u_{H'} + \sum_{C_m \in \mathcal{B} \subseteq \mathcal{C}} (-1)^{|\mathcal{B}|} \cdot \delta_{\bigcap \mathcal{B}}. \tag{7.8}$$

Running intersection property says $S_m = C_m \cap (\bigcup_{j<m} C_j) \subseteq C_k$ for some $k < m$. This allows to write

$$\sum_{C_m \in \mathcal{B} \subseteq \mathcal{C}} (-1)^{|\mathcal{B}|} \cdot \delta_{\bigcap \mathcal{B}} = \sum_{C_m \in \mathcal{A} \subseteq \mathcal{C} \setminus \{C_k\}} \{(-1)^{|\mathcal{A}|} \cdot \delta_{\bigcap \mathcal{A}} - (-1)^{|\mathcal{A}|} \cdot \delta_{\bigcap \mathcal{A} \cap C_k}\} = -\delta_{C_m} + \delta_{C_m \cap C_k}$$

where the last equality holds because every term in braces vanishes whenever $|\mathcal{A}| \geq 2$: the inclusion $\bigcap \mathcal{A} \subseteq C_m \cap (\bigcup_{j<m} C_j) \subseteq C_k$ says $\bigcap \mathcal{A} \cap C_k = \bigcap \mathcal{A}$. Hence, by (7.8) and the induction hypotheses applied to H' (over T) get

$$u_H = \delta_N - \delta_T + (\delta_T - \sum_{i=1}^{m-1} \delta_{C_i} + \sum_{i=2}^{m-1} \delta_{S_i}) - \delta_{C_m} + \delta_{S_m},$$

which gives (7.6). □

REMARK 7.4 Note that (7.7) implies that the product formula induced by u_H (see Section 4.3) is nothing but well-known formula (3.2) characterizing Markovian measures with respect to triangulated graphs mentioned in Section 3.4.1. Thus, this classic result can be viewed as a special case of Theorem 4.1 on structural imsets. △

Decomposable models can be viewed as DAG models (see Figure 3.6). The reader may ask whether 'standard' translation of DAG models and decomposable models leads to the same imset. Positive answer is given by the following lemma.

LEMMA 7.3 Let H be a triangulated graph over N and $C_1, \ldots, C_m$, $m \geq 1$ is a sequence of its cliques satisfying the running intersection property. Put $R_i = C_i \setminus \bigcup_{j<i} C_j$ for $i = 1, \ldots, m$ and consider a total ordering of nodes of N in which nodes of R_i precede nodes of R_{i+1} for $i = 1, \ldots, m-1$. Let G is an acyclic directed graph over N having H as the underlying graph such that the direction of arrows in G is consonant with the constructed total ordering of nodes. Then $\mathcal{M}_G = \mathcal{M}_H$ and $u_H = u_G$.

Proof: The first observation is this:

$$\forall\, c \in N \quad \forall\, a, b \in pa_G(c) \quad a \neq b \;\Rightarrow\; [a, b] \text{ is an edge in } G. \tag{7.9}$$

Indeed, $c \in R_l$ for uniquely determined $l \leq m$. If $a \in pa_G(c)$ then $a \in \bigcup_{j \leq l} R_j = \bigcup_{j \leq l} C_j$ and $\{a, c\}$ belongs to a clique of H. Let C_i be the first clique in the sequence $C_1, \ldots, C_m$ containing $\{a, c\}$. Necessarily $i \leq l$ as otherwise $a, c \in C_i \cap (\bigcup_{j \leq l} C_j) \subseteq C_i \cap (\bigcup_{j<i} C_j) = S_i$ and by the running intersection property $a, c \in S_i \subseteq C_k$ for $k < i$ which contradicts the definition of C_i. However, as $c \notin C_j$ for $j < i$ necessarily $i = l$. Hence, $pa_G(c) \subseteq C_l$ which implies (7.9).

Now, both G and H can be viewed as (classic) chain graphs over N with the same underlying graph (see Section 3.3). To show $\mathcal{M}_G = \mathcal{M}_H$ by well-known graphical characterization [25] (see p. 44) it suffices to show that they have the same complexes. But H has no complexes and G as well because of (7.9).

The equality $u_G = u_H$ can be derived using (7.6) in Lemma 7.2. Indeed, if $d^i_*, \ldots, d^i_\dagger$ is the chosen ordering within R_i, $i = 1, \ldots, m$ then (7.3) gives

$$u_G = \delta_N - \delta_\emptyset + \sum_{i=1}^{m} \sum_{d=d^i_*}^{d^i_\dagger} \{\delta_{pa(d)} - \delta_{d \cup pa(d)}\} = \delta_N - \delta_\emptyset + \sum_{i=1}^{m} \{\, \delta_{S_i} - \delta_{C_i} \,\}$$

(where $S_1 = \emptyset$) as $pa_G(d^i_*) = S_i$ and $d^i_\dagger \cup pa_G(d^i_\dagger) = C_i$ for $i = 1, \ldots, m$ and all remaining terms within the inside sum are cancelled. □

CONSEQUENCE 7.2 Let H be a triangulated undirected graph over N. Then $u = u_H$ is a combinatorial imset, $\mathcal{M}_H = \mathcal{M}_u$ and u coincides with the standard imset for any acyclic directed graph G for which $\mathcal{M}_G = \mathcal{M}_H$.

Proof: This follows directly from Lemma 7.3, Lemma 7.1 and Consequence 7.1. □

REMARK 7.5 Because of Remark 7.2 the preceding consequence implies that $u_H \in \mathcal{S}_\Psi(N)$ where Ψ is the class of discrete measures over N. △

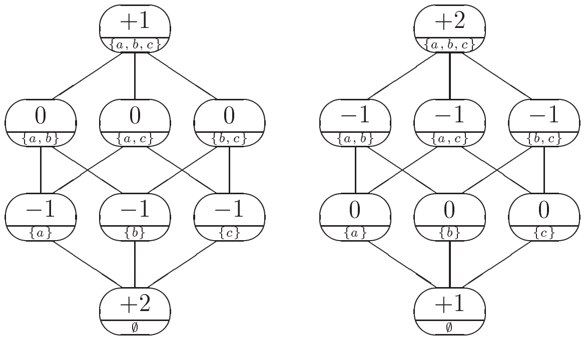


Figure 7.3: Two distinct equivalent imsets of the least degree.

7.3 Imsets of the least degree

One of possible approaches to the choice of representative from a class of facially equivalent structural imsets $\wp$ is to choose a combinatorial imset of the least degree (see Section 4.2.2 p. 59 for this concept). Note that $\wp$ contains combinatorial imsets by Consequence 5.4. The definition of degree implies that only finitely many combinatorial imsets over a fixed set N with prescribed degree exists. In particular, the set of combinatorial imsets of the least degree in $\wp$ is finite. By an *imset of the least degree* will be understood a combinatorial imset u which has the least degree within the class of combinatorial imsets v with $\mathcal{M}_u = \mathcal{M}_v$. Nevertheless, the class $\wp$ may contain more than one imset of the least degree.

EXAMPLE 7.2 There exists a class of facially equivalent structural imsets over $N = \{a, b, c\}$ which has two different imsets of the least degree. Consider the class $\wp$ of $w \in \mathcal{S}(N)$ with $\mathcal{M}_w = \mathcal{T}(N)$. Then both $u = u_{\langle b,c|a\rangle} + u_{\langle a,b|\emptyset\rangle} + u_{\langle a,c|\emptyset\rangle}$ and $v = u_{\langle a,b|c\rangle} + u_{\langle a,c|b\rangle} + u_{\langle b,c|\emptyset\rangle}$ (they are shown in Figure 7.3) have the least degree 3 within the class of combinatorial imsets from $\wp$. Observe that $\mathcal{L}_u = \{a, b, c\}^{\downarrow}$ while $\mathcal{L}_v = \{ab, ac, bc\}^{\downarrow}$. Note that the fact that u and v are all imsets of this kind can be verified using the procedure described later in Section 7.3.2. $\diamondsuit$

LEMMA 7.4 Standard imset for an acyclic directed graph G over N is an imset of the least degree.

Proof: Let $v \in \mathcal{C}(N)$ with $v \rightleftharpoons u$ where $u = u_G$. To verify

$$deg(v) \geq deg(u) = |\,\{(a, b) \in N \times N;\ a \neq b\,,\ [a, b] \text{ is not an edge in } G\,\}\,|$$

(see Lemma 7.1) write $v = \sum_{w \in \mathcal{E}(N)} k_w \cdot w$ for $k_w \in \mathbb{Z}^+$ and show that for every $a, b \in N$, $a \neq b$ such that $[a, b]$ is not an edge in G there exists $w = u_{\langle a,b|K\rangle} \in \mathcal{E}(N)$ with $k_w > 0$ for some $K \subseteq N \setminus ab$. Indeed, otherwise $\langle m, v\rangle = 0$ for $m = m^{ab\uparrow}$ as $\langle m, w\rangle > 0$ for $w \in \mathcal{E}(N)$ iff $w = u_{\langle a,b|K\rangle}$ for some $K \subseteq N \setminus ab$. Hence, by Observation 5.6 $\mathcal{M}_v \subseteq \mathcal{M}^m$. But the moralization criterion (see Section 3.2) says $\langle a, b\,|\,pa_G(a)pa_G(b)\rangle \in \mathcal{M}_G \setminus \mathcal{M}^m = \mathcal{M}_u \setminus \mathcal{M}^m$ which implies a contradictory conclusion $\mathcal{M}_v \neq \mathcal{M}_u$. $\square$

The previous lemma implies by Remark 7.3 this fact.

Consequence 7.3 Every semi-elementary imset is an imset of the least degree.

The method of finding of all imsets of the least degree within a given equivalence class mentioned in Example 7.2 is based on the fact that every imset of this type determines a certain minimal generator of the respective induced independence model. The method uses a computer program and its theoretical justification is given in the rest of this section.

7.3.1 Strong facial implication

Let u, v be combinatorial imsets over N. Let us say that u *strongly facially implies* v and write $u \rightsquigarrow v$ if $u - v$ is a combinatorial imset. Clearly, $u \rightsquigarrow v$ implies $u \rightharpoonup v$ by Lemma 6.1. The relation $\rightsquigarrow$ is a partial ordering on $\mathcal{C}(N)$ (for antisymmetry use Observation 4.4). Its advantage in comparison with $\rightharpoonup$ is that it can be easily tested (see Remark 6.3).

Observation 7.2 Every imset u of the least degree is minimal with respect to $\rightsquigarrow$ within the class $\{v \in \mathcal{C}(N)\,;\, v \rightleftharpoons u\}$.

Proof: If $u \neq v$ are combinatorial imsets and $u \rightsquigarrow v$ then $deg(u) - deg(v) = deg(u - v) > 0$ as the only combinatorial imset of degree 0 is the zero imset. □

However, the question whether the converse implication holds remains open (see Question 4 on p. 129).

7.3.2 Minimal generators

The point is that imsets satisfying the condition from Observation 7.2 correspond to specific minimal generators with respect to a closure operation on subsets of $X = \mathcal{T}(N)$ (see Section 10.2 p. 153 for related definitions). Indeed, the class $\mathcal{U}(N)$ is a closure system of subsets of $X = \mathcal{T}(N)$ by Observation 5.7 and one can introduce the respective closure operation $cl_{\mathcal{U}(N)}$ on subsets of $\mathcal{T}(N)$. Thus, by the *structural closure* of $\mathcal{G} \subseteq \mathcal{T}(N)$ is understood the least structural model containing $\mathcal{G}$ defined by

$$cl_{\mathcal{U}(N)}(\mathcal{G}) = \bigcap_{\mathcal{G} \subseteq \mathcal{M} \in \mathcal{U}(N)} \mathcal{M} \qquad \text{for } \mathcal{G} \subseteq \mathcal{T}(N).$$

A set $\mathcal{G} \subseteq \mathcal{T}(N)$ is called a *structural generator* of $\mathcal{M} \in \mathcal{U}(N)$ if $\mathcal{M} = cl_{\mathcal{U}(N)}(\mathcal{G})$; if moreover $\mathcal{G}$ consists of elementary triplets $\mathcal{G} \subseteq \mathcal{T}_\epsilon(N)$ then it is called an *elementary generator* of $\mathcal{M}$. A structural (elementary) generator of $\mathcal{M}$ is called *minimal* if no its proper subset is a structural generator of $\mathcal{M}$. As every structural model $\mathcal{M}$ over N is a semi-graphoid by Lemma 2.2 $\mathcal{M} \cap \mathcal{T}_\epsilon(N)$ is an elementary generator of $\mathcal{M}$. This implies the following observation.

Observation 7.3 Every structural model $\mathcal{M}$ over N has minimal elementary generator.

Remark 7.6 Note that the concept of (minimal) generator can be understood with respect to arbitrary closure operation on subsets of $\mathcal{T}(N)$, for example with respect to semi-graphoid closure operation or any closure operation introduced by means of syntactic inference rules of semi-graphoid type. The concept of *complexity* of a model (with respect to a closure operation) which can be introduced as the least cardinality of a generator appears to be an interesting characteristic of the model [114]. △

The following lemma provides a method of finding all imsets of the least degree.

LEMMA 7.5 Let $\mathcal{M}$ be a structural model over set N endowed with a total ordering $\preceq$ and $\bar{\wp} = \{v \in \mathcal{C}(N); \mathcal{M}_v = \mathcal{M}\}$. Then every minimal element u of $\bar{\wp}$ with respect to $\rightsquigarrow$ has the form

$$u = \sum_{w \in \mathcal{E}(N)} k_w \cdot w \qquad \text{where} \;\; k_w \in \{0,1\} \tag{7.10}$$

and $\mathcal{G} = \{\langle i,j|K\rangle \in \mathcal{T}_\epsilon(N)\,;\; k_{u_{\langle i,j|K\rangle}} = 1 \;\; i \prec j\}$ is a minimal elementary generator of $\mathcal{M}$.

Proof: Write $u = \sum_{w \in \mathcal{E}(N)} k_w \cdot w$ where $k_w \in \mathbb{Z}^+$. If $k_w \geq 2$ for some $w \in \mathcal{E}(N)$ then $u - w \in \mathcal{C}(N)$ is facially equivalent to u and $u \rightsquigarrow u - w$. Therefore necessarily (7.10) holds and $\mathcal{G}$ is in one-to-one correspondence with elements of $\mathcal{E}(N)$ having non-zero coefficients there. To show that $\mathcal{G}$ is a structural generator of $\mathcal{M}$ consider $\mathcal{M}' \in \mathcal{U}(N)$ with $\mathcal{G} \subseteq \mathcal{M}'$. As it is a semi-graphoid $\mathcal{M}_w \subseteq \mathcal{M}'$ for $w = u_{\langle i,j|K\rangle}$, $\langle i,j|K\rangle \in \mathcal{G}$ and by Consequence 6.1 $\mathcal{M}_u \subseteq \mathcal{M}'$. Thus $\mathcal{M}_u \subseteq cl_{\mathcal{U}(N)}(\mathcal{G})$ and the converse inclusion follows from $\mathcal{M}_u \in \mathcal{U}(N)$ and $\mathcal{G} \subseteq \mathcal{M}_u$. To show that no proper subset $\mathcal{F} \subset \mathcal{G}$ is a generator introduce $v = \sum_{\langle i,j|K\rangle \in \mathcal{F}} u_{\langle i,j|K\rangle} \in \mathcal{C}(N)$. Observe that $cl_{\mathcal{U}(N)}(\mathcal{F}) = \mathcal{M}_v$ (by an analogous procedure). If $cl_{\mathcal{U}(N)}(\mathcal{F}) = \mathcal{M}_u$ then $u \rightleftharpoons v$ and $u \rightsquigarrow v \neq u$ which contradicts the assumption. $\square$

In case $|N| \leq 4$ all minimal elementary generators of a structural model $\mathcal{M}$ over N can be found by a computer program written by my colleague P. Boček [9]. Thus, owing to Observation 7.2 given $\mathcal{M} \in \mathcal{U}(N)$ the list of imsets of the least degree inducing $\mathcal{M}$ can be obtained by reducing the list of imsets u satisfying (7.10). Reduction is sometimes necessary as the following example shows.

EXAMPLE 7.3 There exists a structural model $\mathcal{M}$ over $\{a,b,c,d\}$ and an elementary generator $\mathcal{G} \subseteq \mathcal{M} \cap \mathcal{T}_\epsilon(N)$ such that $v = \sum_{\langle i,j|K\rangle \in \mathcal{G}} u_{\langle i,j|K\rangle}$ is not an imset of the least degree. Let us consider the independence model from Example 3.1 on p. 41 (restriction of a DAG model). Then both imsets

$$u = u_{\langle a,d|\emptyset\rangle} + u_{\langle a,c|d\rangle} + u_{\langle b,d|a\rangle}, \qquad v = u_{\langle a,c|\emptyset\rangle} + u_{\langle b,d|\emptyset\rangle} + u_{\langle a,d|b\rangle} + u_{\langle a,d|c\rangle}$$

are defined by means of minimal elementary generators of $\mathcal{M}$ but $4 = deg(v) > deg(u) = 3$ (the imsets are shown in Figure 7.4). Note that $v = u + u_{\langle a,d|\emptyset\rangle}$ and $u + v$ is a baricentral imset. Moreover, u is the unique imset of the least degree among facially equivalent imsets while v is an imset with the least lower class (see Section 7.4.2) among facially equivalent imsets. $\diamondsuit$

The following consequence easily follows from Lemma 7.5, the definition of baricentral imset and Consequence 7.3.

CONSEQUENCE 7.4 If u is a baricentral imset over N and v is an imset of the least degree with $u \rightharpoonup v$ then $u - v$ is a combinatorial imset. In particular, for every $\langle A,B|C\rangle \in \mathcal{T}(N)$, $\langle A,B|C\rangle \in \mathcal{M}_u$ iff $u - u_{\langle A,B|C\rangle}$ is a combinatorial imset.

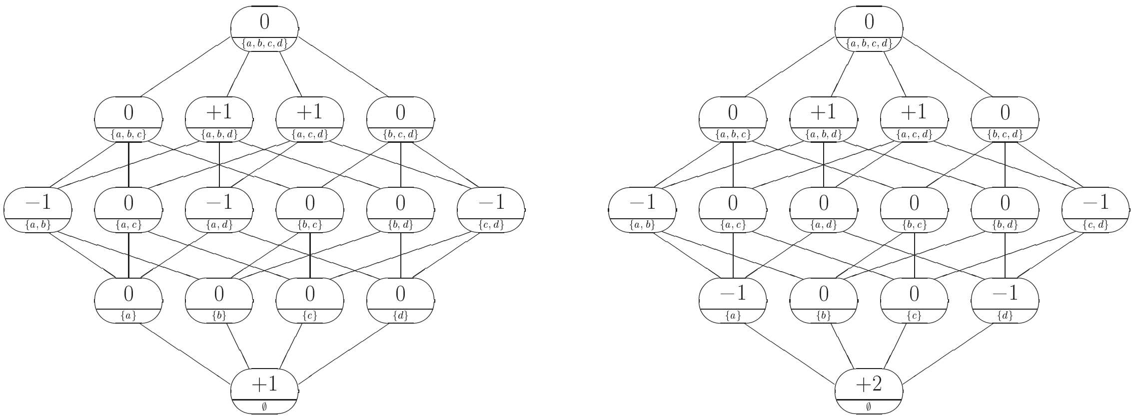


Figure 7.4: Imset of the least degree respectively with the least lower class.

7.4 Width

Recall that the lower class $\mathcal{L}_u$ of a structural imset u is contained in the upper class $\mathcal{U}_u$ (see Section 4.2.3, p. 60) but they differ for non-zero u. Moreover, by Consequence 4.3 marginals of Markovian measures for sets in $\mathcal{L}_u$ determine the marginals for sets in $\mathcal{U}_u$. The upper class is an invariant of facial equivalence (see Section 6.4) but the lower class is not as demonstrated by Example 7.2. In the considered example, the imset shown in the left-hand picture of Figure 7.3 tells more about which marginals determine the whole Markovian measure in comparison with the imset shown in the right-hand picture of Figure 7.3. Thus, facially equivalent imsets need not be equiinformative from this point of view. This consideration motivates an informal concept of *width of a structural imset* u which is the class $\mathcal{U}_u \setminus \mathcal{L}_u$.

7.4.1 Determining and unimarginal classes

Let us take more general view on some results of Section 4.4. Suppose that $\mathcal{M}$ is a structural model over N. The upper class $\mathcal{U} \equiv \mathcal{U}_u$ of subsets of N and the class of probability measures over N which are Markovian with respect to u do not depend on the choice of $u \in \mathcal{S}(N)$ with $\mathcal{M}_u = \mathcal{M}$; they are determined by $\mathcal{M}$ only.

A descending class $\mathcal{D} \subseteq \mathcal{U}$ of subsets of N will be called *determining for* $\mathcal{M}$ if the only descending class $\mathcal{E}$ with $\mathcal{D} \subseteq \mathcal{E} \subseteq \mathcal{U}$ such that $AC, BC \in \mathcal{E} \Rightarrow ABC \in \mathcal{E}$ for every $\langle A, B|C\rangle \in \mathcal{M}$ is the class $\mathcal{E} = \mathcal{U}$. A descending class $\mathcal{D} \subseteq \mathcal{U}$ will be called *unimarginal for* $\mathcal{M}$ if every pair of Markovian measures over N (with respect to $u \in \mathcal{S}(N)$ satisfying $\mathcal{M} = \mathcal{M}_u$) whose marginals for sets from $\mathcal{D}$ coincide has the same marginals for sets from $\mathcal{U}$. Evidently, whenever $\mathcal{D} \subseteq \mathcal{U}$ is determining resp. unimarginal then every descending class $\mathcal{D}'$ with $\mathcal{D} \subseteq \mathcal{D}' \subseteq \mathcal{U}$ is determining resp. unimarginal as well. Therefore, one is interested in *minimal determining classes* for $\mathcal{M}$, that is determining classes $\mathcal{D} \subseteq \mathcal{U}$ for $\mathcal{M}$ such that no proper descending subclass $\mathcal{D}' \subset \mathcal{D}$ is a determining class for $\mathcal{M}$. In particular, I am interested in the question for which $\mathcal{M}$ the *least determining class* (resp. the *least unimarginal class*) for $\mathcal{M}$ exists which is the (unique) determining (resp. unimarginal) class $\mathcal{D} \subseteq \mathcal{U}$, i.e. one has $\mathcal{D} \subseteq \mathcal{D}'$ for every descending determining (resp. unimarginal) class $\mathcal{D}' \subseteq \mathcal{U}$. DAG models appear to be examples of structural models of this type (see Section 7.4.3).

OBSERVATION 7.4 Every determining system is unimarginal.

Proof: If $\mathcal{D} \subseteq \mathcal{U}$ is determining and P, Q Markovian measures then put $\mathcal{E} = \{S \in \mathcal{U}; P^S = Q^S\}$ and observe that $\mathcal{D} \subseteq \mathcal{E}$ and $AC, BC \in \mathcal{E} \Rightarrow ABC \in \mathcal{E}$ for every $\langle A, B|C\rangle \in \mathcal{M}$ (use the 'uniqueness principle' mentioned in the proof of Consequence 4.3). □

Recall that Consequence 4.2 says that the lower class $\mathcal{L}_u$ is a determining class for $\mathcal{M}_u$ whenever $u \in \mathcal{S}(N)$. Thus, one can summarize implications as follows:

$$\text{lower class} \quad \Rightarrow \quad \text{determining class} \quad \Rightarrow \quad \text{unimarginal class.} \tag{7.11}$$

Note that a determining class need not be a lower class (see Example 7.4) and the question whether every unimarginal class is determining remains open - see Question 10 on p. 146. However, it is known that these concepts essentially coincide for DAG models (see Consequence 7.5 in Section 7.4.3).

REMARK 7.7 The concept of unimarginal class can be alternatively introduced as a concept relative to a distribution framework Ψ. Then every unimarginal class relative to Ψ is unimarginal relative to any subframework $\Psi' \subseteq \Psi$. But unimarginal classes may differ for different frameworks. Given a distribution framework Ψ and a structural model $\mathcal{M}$ one can ask what are minimal unimarginal classes for $\mathcal{M}$ relative to Ψ (see Theme 20 in Chapter 8). △

7.4.2 Imsets with the least lower class

A structural imset $u \in \mathcal{S}(N)$ is called *an imset with the least lower class* if $\mathcal{L}_u \subseteq \mathcal{L}_v$ for every $v \in \mathcal{S}(N)$ with $u \rightleftharpoons v$. Some classes of facial equivalence contain imsets of this type, for example the imset u from Example 7.2 on p. 112. On the other hand, as subsequent example shows there are classes of facial equivalence which do not have these imsets but several *imsets with minimal lower class*, i.e. imsets $u \in \mathcal{S}(N)$ such that no facially equivalent $v \in \mathcal{S}(N)$ with $\mathcal{L}_v \subset \mathcal{L}_u$ exists.

EXAMPLE 7.4 There exists a structural model $\mathcal{M}$ over $N = \{a, b, c, d\}$ such that

- the collection $\wp = \{u \in \mathcal{S}(N); \mathcal{M}_u = \mathcal{M}\}$ has three imsets with distinct minimal lower classes,
- 16 distinct minimal determining classes for $\mathcal{M}$ exist and none of them is a lower class for any $u \in \wp$.

Introduce $\mathcal{M} \subseteq \mathcal{T}(N)$ and a class of elementary imsets $\mathcal{K}$ as follows

$$\mathcal{M} = \{\, \langle A, B|C\rangle \in \mathcal{T}(N);\ |C| \geq 1 \,\}, \qquad \mathcal{K} = \{\, u_{\langle i,j|K\rangle} \in \mathcal{E}(N);\ |K| \geq 1 \,\}.$$

Observe that $\mathcal{M} = \mathcal{M}^m$ for the supermodular imset m shown in the left-hand picture of Figure 7.5 and $\mathcal{K} = \{w \in \mathcal{E}(N); \mathcal{M}_w \subseteq \mathcal{M}\}$. Introduce a combinatorial imset $u = \sum_{w \in \mathcal{K}} k_w \cdot w$ where

$$k_w = \begin{cases} 4 & \text{iff } w = u_{\langle a,b|K\rangle} \in \mathcal{K} \text{ or } w = u_{\langle c,d|K\rangle} \in \mathcal{K}\,, \\ 1 & \text{for remaining } w \in \mathcal{K}\,, \end{cases}$$

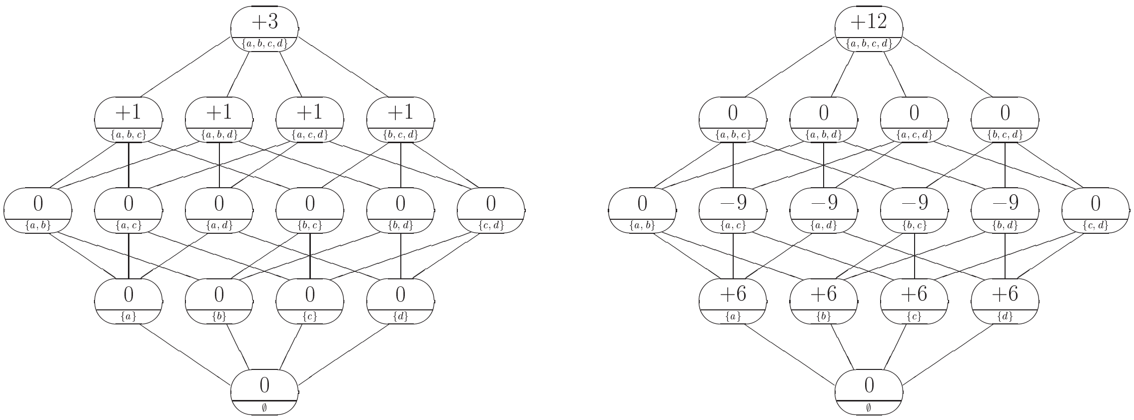


Figure 7.5: Multiset producing $\mathcal{M}$ and respective imset with minimal lower class.

(u is shown in the right-hand picture of Figure 7.5). Since $\mathcal{M}_u \cap \mathcal{T}_\epsilon(N) = \mathcal{M}^m \cap \mathcal{T}_\epsilon(N)$ by Lemma 2.2 $\mathcal{M}_u = \mathcal{M}$. The first step to show that u is an imset with minimal lower class is an observation that for every $v \in \mathcal{S}(N)$ with $\mathcal{M}_v = \mathcal{M}$ one has $v(S) < 0$ for some $S \in \mathcal{S} \equiv \{abc, ab, ac\}$. Indeed, put $\mathcal{K}' = \{u_{\langle a,d|K\rangle} \in \mathcal{E}(N)\,;\, |K| \geq 1\}$ write $n \cdot v = \sum_{w\in\mathcal{K}} l_w \cdot w$ where $n \in \mathbb{N}$, $l_w \in \mathbb{Z}^+$ and observe that $l_w > 0$ for some $w \in \mathcal{K}'$ because $\langle m^{ad\uparrow}, w\rangle = 0$ for $w \in \mathcal{K} \setminus \mathcal{K}'$ while $\langle m^{ad\uparrow}, w\rangle > 0$ for $w \in \mathcal{K}'$ (use Observation 5.6 to derive contradiction with the assumption $\mathcal{M}_v = \mathcal{M}$ in case $l_w = 0$ for $w \in \mathcal{K}'$). Put $s = \sum_{S\in\mathcal{S}} \delta_S$ and observe $\langle s, w\rangle \leq 0$ for $w \in \mathcal{K}$ and $\langle s, w\rangle = -1$ for $w \in \mathcal{K}'$. This implies $n \cdot \langle s, v\rangle < 0$ and the desired conclusion.

Analogous observation can be made for any class $\mathcal{S}'$ consisting of a three-element subset of N and a pair of its two-element subsets. If $v \in \mathcal{S}(N)$ satisfies $\mathcal{M}_v = \mathcal{M}$ and $\mathcal{L}_v \subseteq \mathcal{L}_u$ then the observation above necessitates $v(ac) < 0$ and one can derive analogously $v(ad), v(bc), v(bd) < 0$ which implies $\mathcal{L}_v = \mathcal{L}_u$. Therefore u is an imset with minimal lower class and permutation of variables gives two other examples of facially equivalent imsets with distinct minimal lower classes.

On the other hand, the class $\mathcal{D}_1 = \{ab, bc, cd\}^{\downarrow}$ is a determining class for $\mathcal{M}$ as $\langle a, c|b\rangle$, $\langle b, d|c\rangle$, $\langle a, d|bc\rangle \in \mathcal{M}$. One can show that $\mathcal{D}_1$ is minimal and an analogous conclusion can be made for any class obtained by permutation of variables either from $\mathcal{D}_1$ or $\mathcal{D}_2 = \{ab, ac, ad\}^{\downarrow}$. This list of minimal determining classes for $\mathcal{M}$ can be shown to be complete. $\diamondsuit$

7.4.3 Exclusivity of standard imsets

The standard imset u_G for an acyclic directed graph G appears to be an exclusive imset within the class of structural imsets u with $\mathcal{M}_u = \mathcal{M}_G$. The first step to show that it is an imset with the least lower class is the following lemma.

Lemma 7.6 Let $\mathcal{M}$ be a structural model over N, $\wp = \{u \in \mathcal{S}(N)\,;\, \mathcal{M}_u = \mathcal{M}\}$ and $\mathcal{U} = \mathcal{U}_u$ for (any) $u \in \wp$. If $S \in \mathcal{U}$ has the form $S = cT$ where $c \in N$ and

$$\mathcal{M} \cap \{\langle a, c|K\rangle \in \mathcal{T}_\epsilon(N); a \in T\} = \emptyset = \mathcal{M} \cap \{\langle a, b|K\rangle \in \mathcal{T}_\epsilon(N); a, b \in T\ c \in K\} \quad (7.12)$$

then every (descending) unimarginal class for $\mathcal{M}$ contains S.

Proof: The first observation is that any probability measure $P = Q \times \prod_{i\in N\setminus S} P_i$ where Q is a probability measure over S with $Q^T = \prod_{i\in T} Q_i$ and P_i, Q_i are arbitrary one-dimensional probability measures is Markovian with respect to $u \in \wp$. Indeed, by Lemma 2.2 it suffices to verify $\mathcal{M} \cap \mathcal{T}_\epsilon(N) \subseteq \mathcal{M}_P$. Suppose $\langle a, b|K\rangle \in \mathcal{M} \cap \mathcal{T}_\epsilon(N)$. If $a \notin S$ then $a \perp\!\!\!\perp N \setminus a \,|\, \emptyset \; [P]$ implies $a \perp\!\!\!\perp b \,|\, K \; [P]$, analogously in case $b \notin S$. If $a, b \in S$ then (7.12) implies $a, b \in T$ and $c \notin K$ so that $a \perp\!\!\!\perp N \setminus ac \,|\, \emptyset \; [P]$ implies $a \perp\!\!\!\perp b \,|\, K \; [P]$ as well. The second step is a construction: put $\mathsf{X}_i = \{0, 1\}$ and define a pair of probability measures Q_1, Q_2 on $\mathsf{X}_S = \prod_{i\in S} \mathsf{X}_i$:

$$Q_1([x_i]_{i\in S}) = 2^{-|S|} \qquad Q_2([x_i]_{i\in S}) = \begin{cases} 2^{-|S|} + \varepsilon & \text{if } \sum_{i\in S} x_i \text{ is even,} \\ 2^{-|S|} - \varepsilon & \text{if } \sum_{i\in S} x_i \text{ is odd,} \end{cases}$$

where $0 < \varepsilon < 2^{-|S|}$. Then put $P_j = Q_j \times \prod_{i\in N\setminus S} P_i$ for $j = 1, 2$ where P_i are some fixed probability measures on X_i, $i \in N \setminus S$. Observe that both P_1 and P_2 is Markovian with respect to $u \in \wp$, $P_1^S \neq P_2^S$ and $P_1^L = P_2^L$ whenever $L \subseteq N$, $S \setminus L \neq \emptyset$.

Finally, suppose for contradiction that $S \notin \mathcal{D}$ where $\mathcal{D} \subseteq \mathcal{U}$ is an unimarginal class for $\mathcal{M}$. This implies $P_1^L = P_2^L$ for $L \in \mathcal{D}$ and therefore $P_1^L = P_2^L$ for $L \in \mathcal{U}$ which contradicts the fact $S \in \mathcal{U}$. □

CONSEQUENCE 7.5 Given an acyclic directed graph G over N the lower class $\mathcal{L}_u$ for $u = u_G$ is the least unimarginal and the least determining class for $\mathcal{M}_G$. In particular, u_G is an imset with the least lower class.

Proof: By (7.11) and Lemma 7.1 is $\mathcal{L}_u$ a determining resp. unimarginal class for $\mathcal{M}_G$. If $\mathcal{D}$ is a lower class for $v \in \mathcal{S}(N)$ with $\mathcal{M}_v = \mathcal{M}_G$ resp. a determining class for $\mathcal{M}_G$ then it is an unimarginal class for $\mathcal{M}_G$ by (7.11). Lemma 7.6 can be then used to verify $\mathcal{L}_u \subseteq \mathcal{D}$. Indeed, if $S \in \mathcal{L}_u^{\max}$ then $S \in \mathcal{U}_u$ by Observation 4.5 and $S = c\, pa_G(c)$ for some $c \in N$ by (7.3). The moralization criterion (see Section 3.2) allows to verify that the condition (7.12) for $T = pa_G(c)$ is fulfilled for $\mathcal{M} = \mathcal{M}_G$. □

REMARK 7.8 Thus, the standard imset u_G for an acyclic directed graph G over N is both an imset of the least degree (see Lemma 7.4) and an imset with the least lower class. Note that a computer program [9] helped to show that in case $|N| \leq 4$ the converse holds, i.e. the only imset satisfying these two conditions is the standard imset u_G (for a given graph G). The question whether these two requirements determine standard imsets for acyclic directed graphs in general remains open - see Question 9 on p. 145. △

An interesting feature of the standard imset u for a DAG models is that it vanishes outside $\mathcal{L}_u \cup \mathcal{U}_u^{\max}$. On the other hand, a lot of other equivalent structural imsets with the same lower class exist, sometimes even imsets which take only strictly positive values in $\mathcal{R}_u \setminus \mathcal{L}_u$. The following example shows that imsets of this kind need not exist.

EXAMPLE 7.5 There exists a DAG model $\mathcal{M}$ over $N = \{a, b, c, d\}$ such that no $u \in \mathcal{S}(N)$ with the least lower class $\mathcal{L}_u$ among imsets with $\mathcal{M}_u = \mathcal{M}$ is strictly positive on $\mathcal{R}_u \setminus \mathcal{L}_u$ (for the notion of range $\mathcal{R}_u$ see Section 6.4). Consider a directed graph G shown in the left-hand picture of Figure 7.6, the corresponding standard imset is in the right-hand picture. Put $s = \delta_{ac} + \delta_{acd}$ and observe that one has $\langle s, u_{\langle i,j|K\rangle}\rangle = 0$ for every $\langle i, j|K\rangle \in \mathcal{M}_G \cap \mathcal{T}_\epsilon(N)$. This implies $\langle s, v\rangle = 0$ for every $v \in \mathcal{S}(N)$ with $\mathcal{M}_v = \mathcal{M}_G$. Since $ac, acd \in \mathcal{R}_u \setminus \mathcal{L}_u$ it implies $v(ac) = v(acd) = 0$ for every v of this kind which moreover satisfies $\mathcal{L}_v = \mathcal{L}_u$. Note that an analogous consideration can be made for $\tilde{s} = \delta_{bd} + \delta_{abd}$. ◇

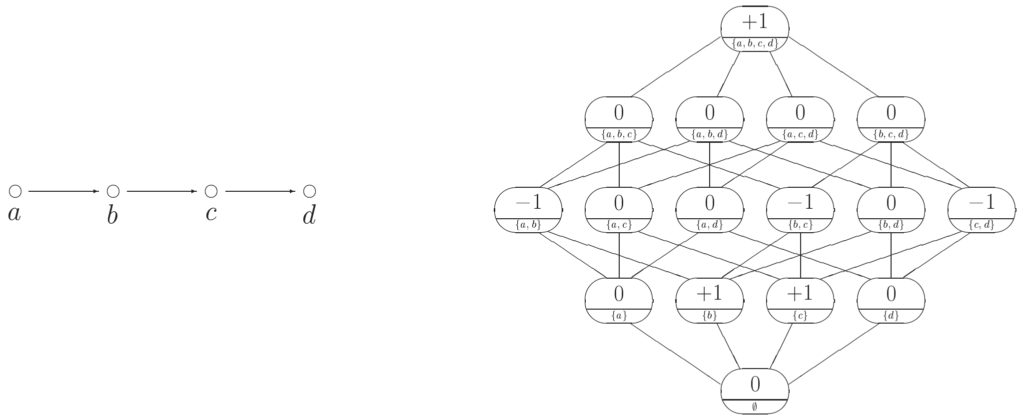


Figure 7.6: An acyclic directed graph and the respective standard imset.

7.5 Other ways of representation

This section describes further ideas how structural models can be possibly represented. It is only a rough outline of those approaches which look promising and which have to be examined in more details.

7.5.1 Pattern

Recall that one of feasible methods of representing a class of Markov equivalent acyclic directed graphs is to use a graph which is not an acyclic directed graph but which somehow exhibits common features of the graphs within the equivalence class - see Section 3.2, p. 41. This motivated an analogous idea in the framework of structural imsets.

Observation 7.5 Let $\mathcal{M}$ be a structural model over N, $\wp = \{u \in \mathcal{S}(N)\,;\, \mathcal{M}_u = \mathcal{M}\}$. Then the set $\{u(S)\,;\, u \in \wp\}$ for $S \subseteq N$ has one of the following four forms: $\{0\}, \mathbb{Z}, \{m, m+1, \ldots\}$ for some $m \in \mathbb{Z}^+$ and $\{\ldots, -l-1, -l\}$ for some $l \in \mathbb{Z}^+$.

Proof: If $S \notin \mathcal{R}_u$ for $u \in \wp$ then $\{v(S); v \in \wp\} = \{0\}$ by Lemma 6.6. If $S \in \mathcal{R}_u$ then $w = u_{\langle i,j|K\rangle}$ with $\langle i,j|K\rangle \in \mathcal{M} \cap \mathcal{T}_\epsilon(N)$ and $w(S) \neq 0$ exists by (6.23). Observe that for every $u \in \wp$ and $k \in \mathbb{N}$ one has $u + k \cdot w \in \wp$ which leads to the remaining cases. □

Pattern of a structural imset u over N can be introduced as an undirected graph which has the region $\mathcal{R}_u$ as the set of nodes and the set of lines is the collection of pairs of the form $\{K, iK\}, \{iK, ijK\}$ for some $\langle i,j|K\rangle \in \mathcal{M}_u \cap \mathcal{T}_\epsilon(N)$. Moreover, the nodes of the pattern have assigned symbolic values depending on the class $\wp$ of $v \in \mathcal{S}(N)$ with $u \rightleftharpoons v$:

$$\begin{array}{lll}\nabla(S) = + & \text{if} & \{v(S)\,;\, v \in \wp\} \subseteq \mathbb{Z}^+,\\ \nabla(S) = - & \text{if} & \{v(S)\,;\, v \in \wp\} \subseteq \mathbb{Z}^- \equiv \{-l\,;\, l \in \mathbb{Z}^+\},\\ \nabla(S) = \pm & \text{if} & \{v(S)\,;\, v \in \wp\} = \mathbb{Z}.\end{array}$$

The *symbolic function* ∇ can be formally extended to $\mathcal{P}(N)$ by putting

$$\nabla(S) = 0 \quad \text{if} \quad \{v(S)\,;\, v \in \wp\} = \{0\}.$$

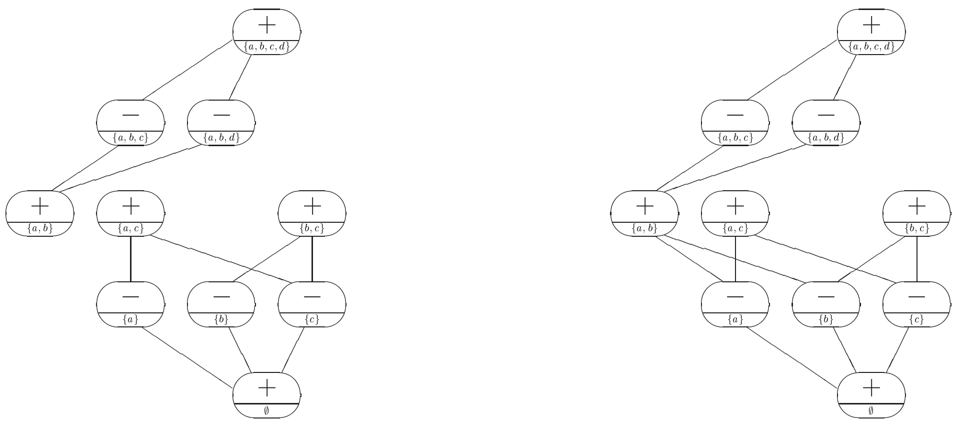


Figure 7.7: Patterns of imsets from Example 7.6.

An *evaluated pattern* is obtained by modification in values of ∇: one writes $+m$ where $m = \min\{v(S)\,;\, v \in \wp\}$ instead of $+$ and $-l$ where $l = \min\{-v(S)\,;\, v \in \wp\}$ instead of $-$. Note that the signs $+$ and $-$ are kept to distinguish the cases $\nabla(S) = +0$ (i.e. $m = 0$) and $\nabla(S) = -0$ (i.e. $l = 0$) from the case $\nabla(S) = 0$ (i.e. $\{v(S)\,;\, v \in \wp\} = \{0\}$). Observe that (evaluated) patterns characterize classes of facial equivalence. Note that symbolic functions ∇ distinguish all classes of facial equivalence if $|N| \leq 3$ - see Figure 7.8 for illustration. On the other hand, this is not true in general as the following example shows.

Example 7.6 There exist structural imsets over $N = \{a, b, c, d\}$ which are not facially equivalent but which have the same symbolic function. Let us put

$$u = u_{\langle a,c|\emptyset\rangle} + u_{\langle b,c|\emptyset\rangle} + u_{\langle c,d|ab\rangle}\,, \qquad v = u + u_{\langle a,b|\emptyset\rangle}\,.$$

The pattern of u is in the left-hand picture and the pattern of v in the right-hand picture of Figure 7.7. $\diamondsuit$

Remark 7.9 The question whether patterns distinguish all classes of facial equivalence remains open (see Direction 3). The following modification of the concept of (evaluated) pattern can possibly cure the problem if the answer is negative. More general values of the symbolic function can be considered. One can distinguish between 'upper plus' (denoted by $\hat{+}$) and the 'lower plus' (denoted by $\check{+}$):

$$\begin{array}{ll} \nabla(S) = \hat{+} & \text{if } \exists\,\langle i,j|K\rangle \in \mathcal{M}_u \cap \mathcal{T}_\epsilon(N) \text{ with } ijK = S, \\ \nabla(S) = \check{+} & \text{if } \exists\,\langle i,j|K\rangle \in \mathcal{M}_u \cap \mathcal{T}_\epsilon(N) \text{ with } K = S. \end{array}$$

Alternatively, one can turn the pattern into a graph with directed and bidirected edges. Indeed, every $\langle i,j|K\rangle \in \mathcal{M}_u \cap \mathcal{T}_\epsilon(N)$ generates four arrows $iK \to ijK$, $jK \to ijK$, $iK \to K$, $jK \to K$ and every pair of arrows $S \to T$ and $T \to S$ can be replaced by a bidirected edge $S \leftrightarrow T$. $\triangle$

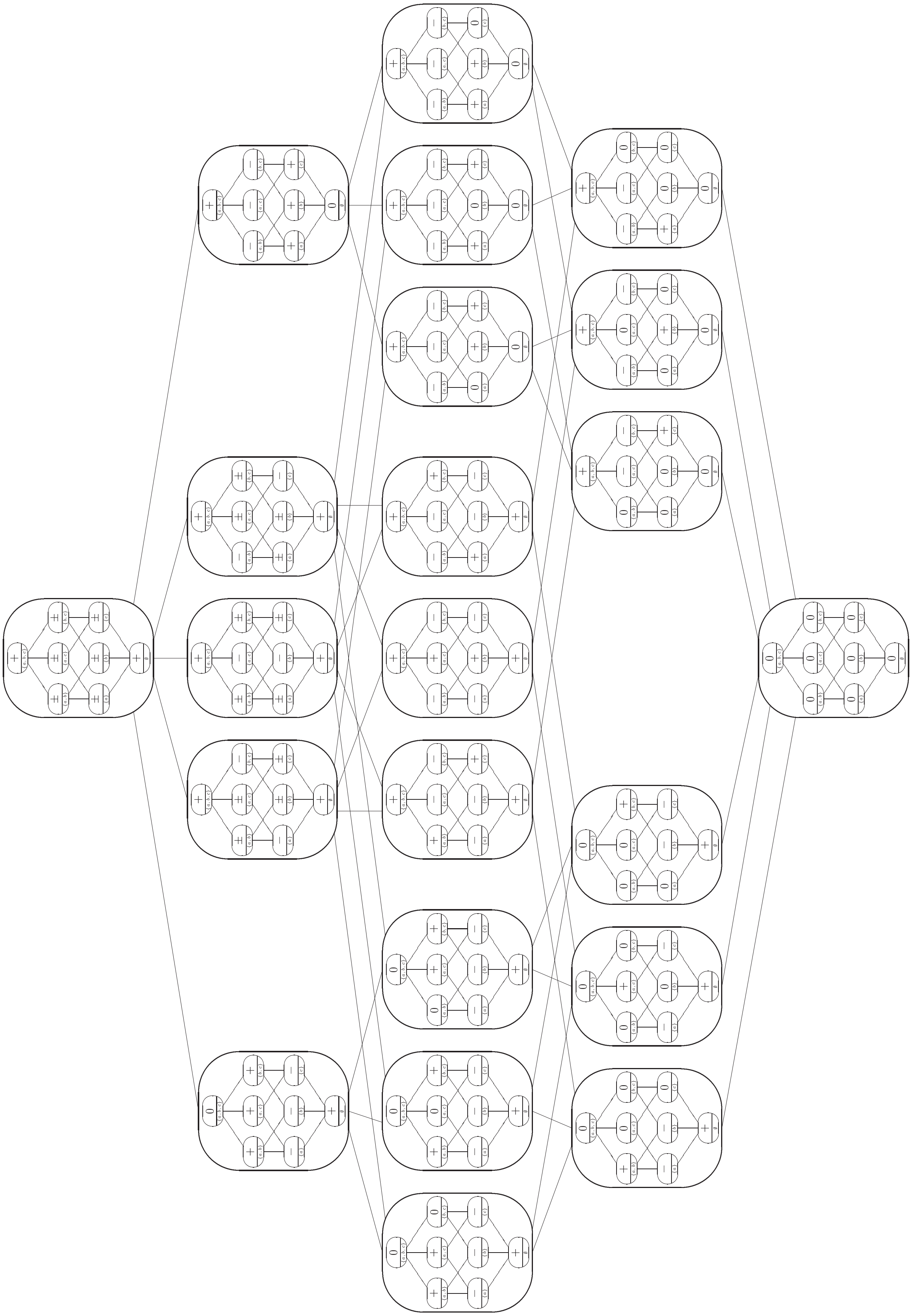


Figure 7.8: Symbolic functions over $N = \{a, b, c\}$ (rotated).

7.5.2 Dual description

Two approaches to the description of independence models by imsets were distinguished in Chapter 5. Every structural model is induced by a structural imset and produced by a supermodular imset (see Consequence 5.4) and both methods can be viewed as mutually dual approaches. As mentioned before Observation 5.7 (p. 87) one can take a dual point of view and describe structural models as independence models produced by ℓ-standardized supermodular imsets.

Dual baricentral imsets

One can introduce an analogue of facial equivalence and implication for ℓ-standardized supermodular imsets: $m \in \mathcal{K}_\ell(N) \cap \mathbb{Z}^{\mathcal{P}(N)}$ implies $r \in \mathcal{K}_\ell(N) \cap \mathbb{Z}^{\mathcal{P}(N)}$ if $\mathcal{M}^m \subseteq \mathcal{M}^r$ and they are equivalent if they produce the same model. Moreover, every imset of this kind is a non-negative rational combination of ℓ-skeletal imsets (see Lemma 5.3) so that these play the role which is analogous to the role of elementary imsets within the class of structural imsets (c.f. Theorem 5.3). Following this analogy an ℓ-standardized supermodular imset m over N will be called a *dual baricentral imset* if it has the form.

$$m = \sum_{r \in \mathcal{K}_\ell^\diamond(N),\, \mathcal{M}^m \subseteq \mathcal{M}^r} r\,. \tag{7.13}$$

The corresponding poset of dual baricentral imsets is shown in Figure 7.9.

Coportraits

Let me explain dual perspective in more details with help of the concept of Galois connection from Section 5.4. It was explained there (p. 85) that the poset of structural models $(\mathcal{U}(N), \subseteq)$ can be viewed as a concept lattice given by the formal context (5.16). More specifically, it follows from Lemma 2.2 that every structural model $\mathcal{M}$ over N is in one-to-one correspondence with a set of elementary imsets over N, namely with

$$\{v \in \mathcal{E}(N)\,;\, v = u_{\langle i,j|K\rangle} \text{ where } \langle i,j|K\rangle \in \mathcal{M} \cap \mathcal{T}_\epsilon(N)\,\}. \tag{7.14}$$

In particular, every $u \in \mathcal{S}(N)$, respectively $m \in \mathcal{K}_\ell(N) \cap \mathbb{Z}^{\mathcal{P}(N)}$ corresponds through $\mathcal{M}_u$ respectively through $\mathcal{M}^m$ to a subset of Œ $= \mathcal{E}(N)$:

$$\begin{aligned} \mathcal{E}_u &\equiv \{v \in \mathcal{E}(N)\,;\, v = u_{\langle i,j|K\rangle},\, \langle i,j|K\rangle \in \mathcal{M}_u\} = \{v \in \mathcal{E}(N)\,;\, u \rightharpoonup v\}, \\ \mathcal{E}^m &\equiv \{v \in \mathcal{E}(N)\,;\, v = u_{\langle i,j|K\rangle},\, \langle i,j|K\rangle \in \mathcal{M}^m\} = \{v \in \mathcal{E}(N)\,;\, \langle m, v\rangle = 0\}. \end{aligned} \tag{7.15}$$

Thus, every structural model can be identified with a set of objects of the formal context (5.16). In fact, it is an extent of a formal concept so that structural models correspond more or less to the description in terms of objects. However, as explained in Remark 5.8 every formal concept can be also described by means of its intent, i.e. in terms of attributes. In this case the set of attributes is the ℓ-skeleton Æ $= \mathcal{K}_\ell^\diamond(N)$ which motivates the following definition.

By *coportrait* of a structural imset u over N will be understood the set of skeletal imsets $\mathcal{H}_u$ given by

$$\mathcal{H}_u = \{r \in \mathcal{K}_\ell^\diamond(N)\,;\, \langle r, u\rangle = 0\}. \tag{7.16}$$

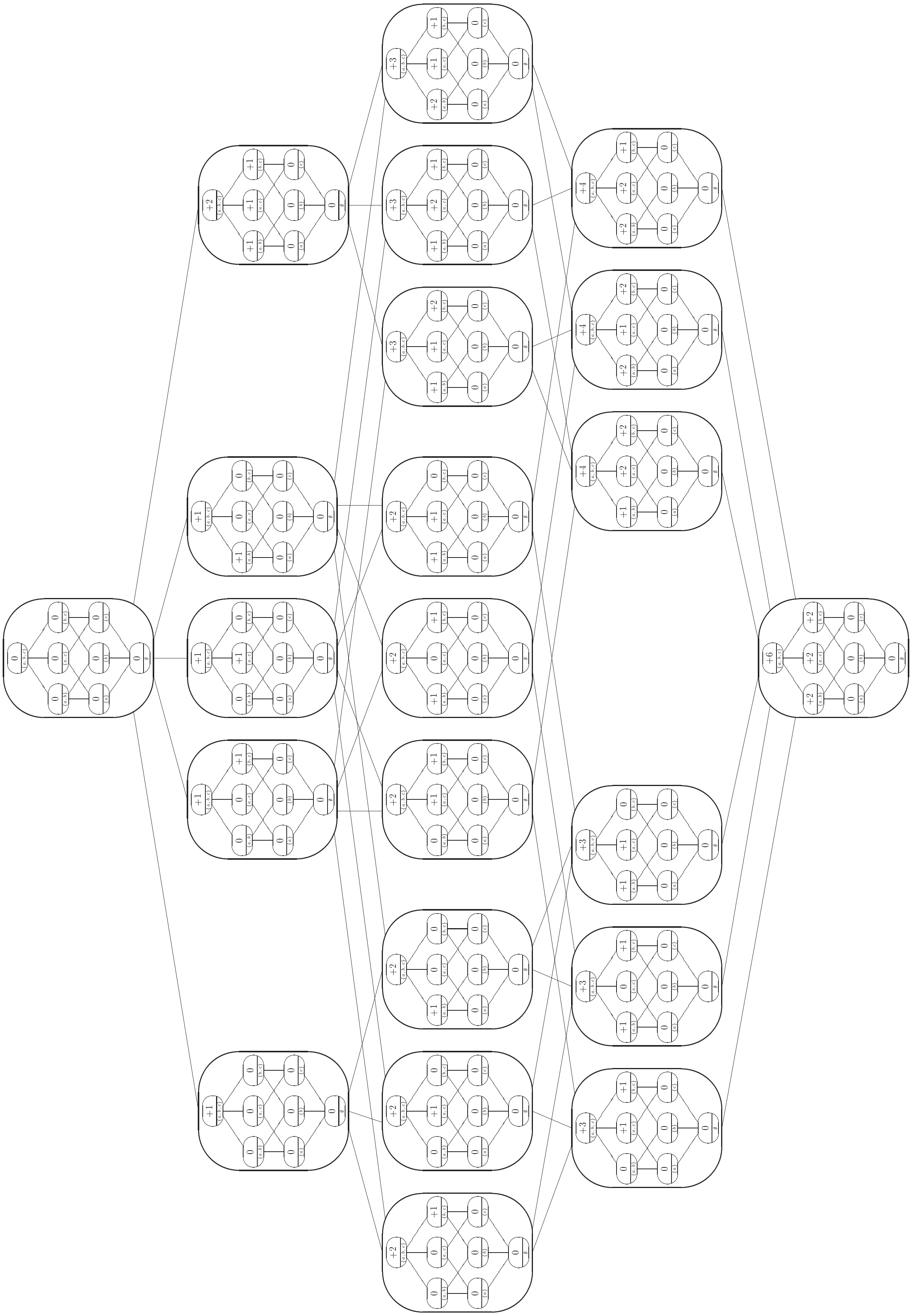


Figure 7.9: Dual baricentral imsets over $N = \{a, b, c\}$ (rotated).

Indeed $\mathcal{H}_u = \{r \in \mathcal{K}_\ell^\diamond(N)\,;\, \langle r, v\rangle = 0 \text{ for every } v \in \mathcal{E}_u\}$ which means that $\mathcal{H}_u$ is nothing but $\mathcal{E}_u^{\triangleright}$. As $\mathcal{E}_u^{\triangleright\triangleleft} = \mathcal{E}_u$ the pair $(\mathcal{E}_u, \mathcal{H}_u)$ is a formal concept in sense of Section 5.4.1. By Consequence 6.2 two structural imsets are facially equivalent iff they have the same coportrait. Thus, every class of facial equivalence is uniquely represented by the respective coportrait. The lattice of all coportrait over 3 variables is shown in Figure 7.10.

REMARK 7.10 This is to explain terminology. The idea of dual description of a structural model was presented already in [108] where the concept of *portrait of* $u \in \mathcal{S}(N)$ was introduced as the set of skeletal imsets

$$\{r \in \mathcal{K}_\ell^\diamond(N)\,;\, \langle r, u\rangle > 0\}. \tag{7.17}$$

Thus, coportrait $\mathcal{H}_u$ is nothing but the relative complement of (7.17) in $\mathcal{K}_\ell^\diamond(N)$ which motivated my terminology here. Provided the ℓ-skeleton is known (7.17) and (7.16) are equiinformative but the concept of coportrait seems more natural from theoretical point of view (in light of Galois connection). Despite this fact I decided to keep former terminology and do not rename things. The reason why I preferred in [108] (7.17) to (7.16) was my anticipation that for $|N| \geq 2$ the relative occurence of zeros in $\{\langle m, u\rangle; m \in \mathcal{K}_\ell^\diamond(N), u \in \mathcal{E}(N)\}$ exceeds the relative occurence of non-zero values (which seem to be true in explored cases). Practical consequence should be that portraits have less cardinality than coportraits for structural imsets inducing 'a lot of' independence statements. $\triangle$

Nevertheless, the method of dual description of structural models is limited to the situation when the skeleton is known. Of course, as explained in Remark 5.6 the type of the skeleton is not substantial since the use of the u-skeleton resp. the o-skeleton instead of the ℓ-skeleton leads to an 'isomorphic' concept of portrait and coportrait.

Global view

Of course, owing to Consequence 5.4 and (7.15) every coportrait can also be written in the form $\mathcal{H}^m = (\mathcal{E}^m)^{\triangleright}$ where $m \in \mathcal{K}_\ell(N) \cap \mathbb{Z}^{\mathcal{P}(N)}$. Note that one can show analogously to the proof of Lemma 6.1 that

$$\mathcal{H}^m = \{r \in \mathcal{K}_\ell^\diamond(N)\,;\; k \cdot m - r \in \mathcal{K}_\ell(N) \text{ for some } k \in \mathbb{N}\}\,.$$

Therefore, the mutual relation of a structural imset u and the corresponding set of elementary imsets $\mathcal{E}_u$ given by (7.15) is completely analogous to the mutual relation of an ℓ-standardized supermodular imset m and the corresponding set of ℓ-skeletal imsets $\mathcal{H}^m$. The global view on all four above mentioned approaches to the description of a structural model is indicated by Figure 7.11. One can use

1. a set of elementary imsets,
2. a structural imset,
3. a set of ℓ-skeletal imsets,
4. an ℓ-standardized supermodular imset.

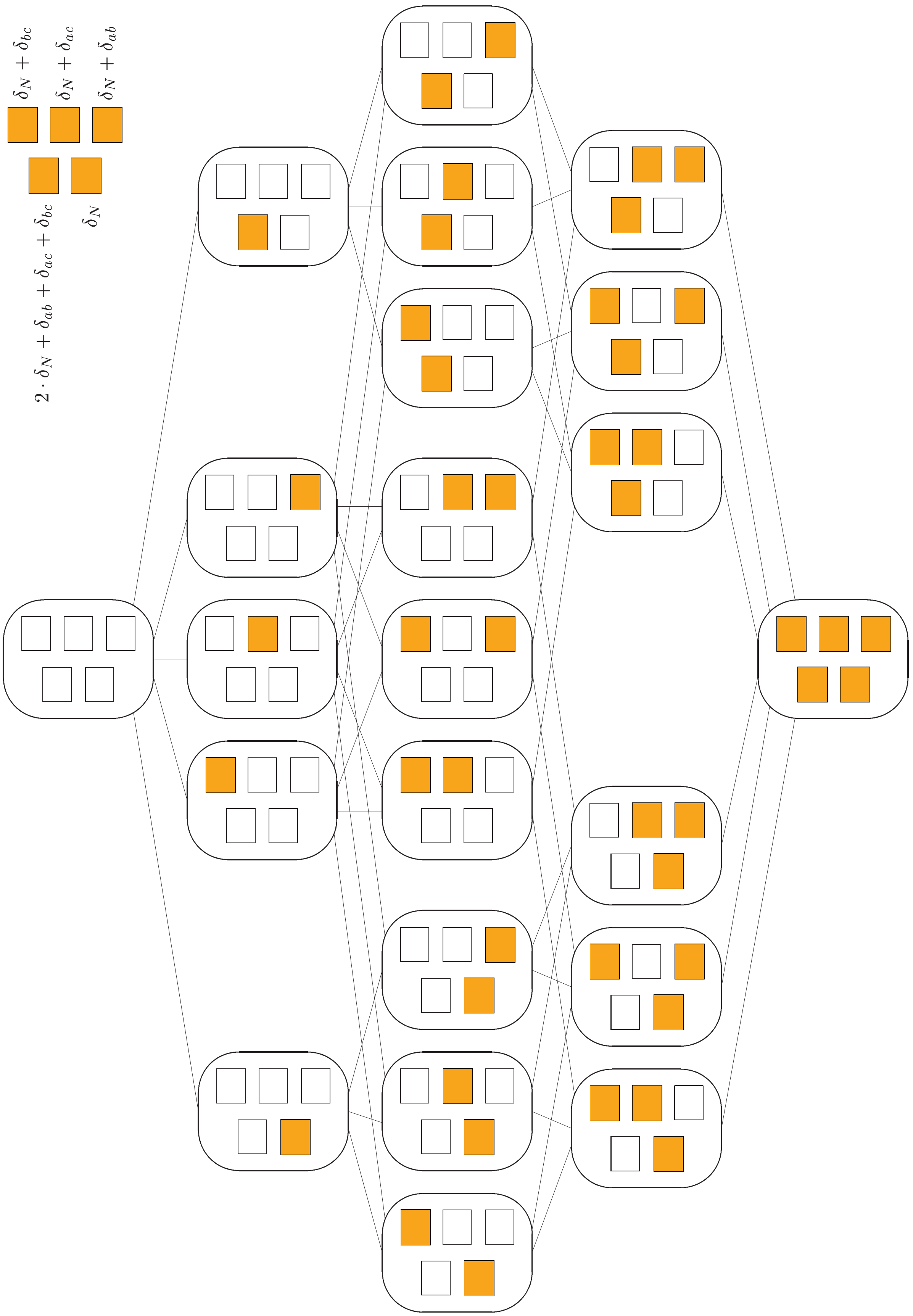


Figure 7.10: Coportraits of structural imsets over $N = \{a, b, c\}$ (rotated).

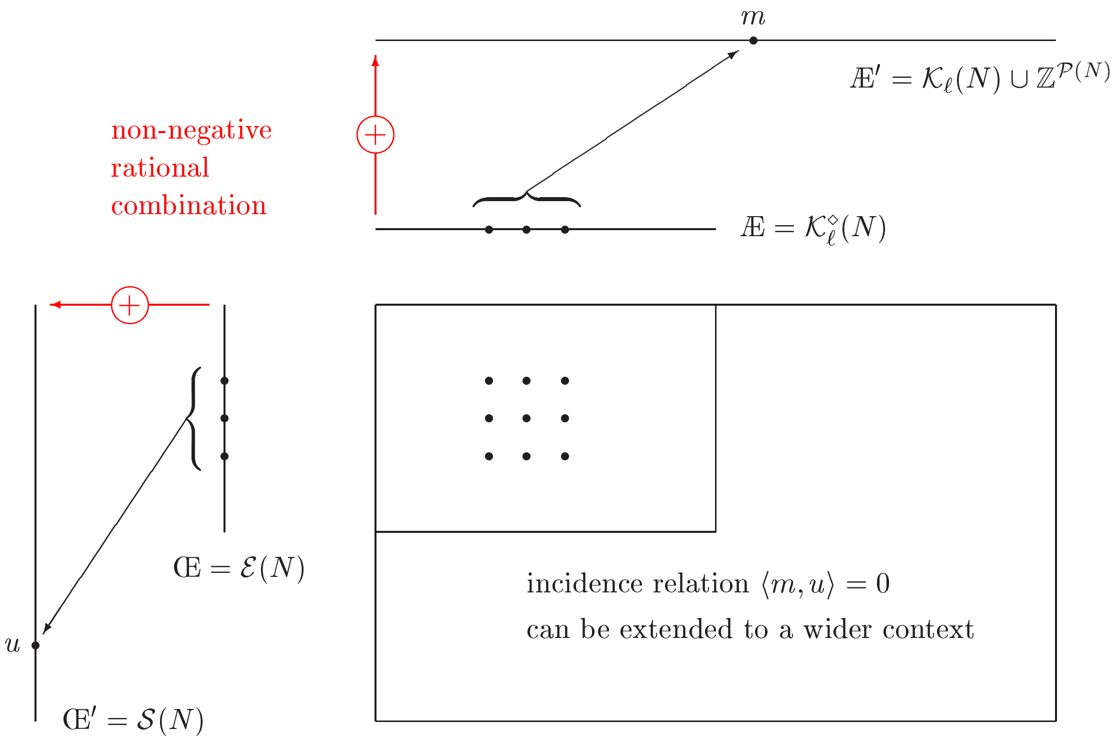


Figure 7.11: Extension of Galois connection for structural models - illustration.

Recall that the set of elementary imsets can be viewed as direct translation of the considered structural model, the structural imset is obtained by non-negative rational combination of elementary imsets, the set of skeletal imsets is obtained by Galois connection and the supermodular imset by non-negative rational combination of skeletal imsets.

Let me emphasize that unlike the case of general Galois connection described in Section 5.4.1 additional superstructure of summing elementary respectively skeletal imsets is at disposal. This fact allows to describe (and later implement) respective relation among formal concepts (namely the relation 'be a subconcept') with help of algebraic operations, more precisely by means of arithmetic of integers! This is the main asset of the described approach.

Dual minimal generators

However the dual approach exhibits some different mathematical properties. One can introduce an analogue of the concept of combinatorial imset, i.e. an imset which is a non-negative integral combination of ℓ-skeletal imsets. But there is no analogue of the concept of degree for imsets of this type: the sum of two ℓ-skeletal imsets from the first line of Figure 5.1 on p. 79 equals to the sum of three ℓ-skeletal imsets from the second line of Figure 5.1.

Nevertheless, one can introduce an analogue of the concept of strong facial implication (see Section 7.3.1) and prove an analogue of Lemma 7.5. Indeed, following the idea indicated in Remark 5.8 one can introduce the *lattice of coportraits*

$$\mathcal{W}(N) = \{\mathcal{H} \subseteq \mathcal{K}_\ell^\diamond(N)\,;\ \mathcal{H} = \mathcal{H}_u \ \text{ for } u \in \mathcal{S}(N)\}$$

and observe that it coincides with the collection of closed sets with respect to a closure operation on subsets of Æ $= \mathcal{K}_\ell^\diamond(N)$

$$\mathcal{H} \longmapsto \mathcal{H}^{\lhd\rhd} = \{m \in \mathcal{K}_\ell^\diamond(N)\,;\, \langle m,v\rangle = 0 \text{ for } v \in \mathcal{E}(N) \text{ with } \langle r,v\rangle = 0 \text{ for every } r \in \mathcal{H}\,\}\,.$$

One can show then that imsets minimal with respect to the 'dual strong facial implication' correspond to minimal generators with respect to this closure operation on subsets of $\mathcal{K}_\ell^\diamond(N)$. An interesting fact is that in case $|N| = 3$ every class of respective equivalence of imsets of above type has unique minimal imset in the described sense. In other words, no analogue of Example 7.2 is valid for dual representation if $|N| \leq 3$. Moreover, the corresponding dual baricentral imset is a multiple of the 'minimal' imset in that case. Well, perhaps the dual approach indeed exhibits better mathematical properties in this sense than the approach based on structural imsets. Nevertheless, I tend to believe that the phenomenon mentioned above is a coincidence and not a general feature of the dual approach (see Theme 6).

Chapter 8

Open problems

The goal of this chapter is to gather open problems and present a few topics omitted in the previous chapters. Open problem are classified according to the degree of vagueness in three categories. *Questions* are clear inquiries formulated as mathematical problems. Formal definitions of related concepts were given and expected answer is yes or no. *Themes* (of research) are wider areas of mutually related problems. Their formulation is slightly less specific (but still in mathematical terms) and they may deserve some clarification of involved concepts. *Directions* (of research) are very wide groups of problems with recognized common motivation source. They are formulated quite vaguely and may become a topic of research in forthcoming years. The secondary criterion of classification of open problems is their topic: the division of this chapter into sections was inspired by the motivation account from Section 1.1.

8.1 Unsolved theoretical problems

In this section open problems concerning theoretical groundings are gathered. Some of them were already mentioned earlier. They are classified by their topics.

8.1.1 Miscellaneous topics

Distributions

There are several open problems related to Sections 4.1 and 6.2.3.

Question 1 Let P and Q are probability measures over N defined on the a product of measurable spaces $(\mathsf{X}_N, \mathcal{X}_N) = \prod_{i\in N}(\mathsf{X}_i, \mathcal{X}_i)$ which have finite multiinformation (p. 24). Has their convex combination $\alpha \cdot P + (1-\alpha)\cdot Q$, $\alpha \in [0,1]$ finite multiinformation as well?

Theme 1 Is there any (direct) formula for the multiinformation function of a non-degenerate CG measure (see p. 54) in terms of their canonical or moment characteristics? Alternatively, is there any (iterative) method of its computing?

Note that owing to Lemma 2.7 equivalent formulation of Theme 1 is to find a formula for entropy of a CG measure P with respect to $\prod_{i\in N}\mu_i$ where $\{\mu_i\,;\, i\in N\}$ is the standard reference system for P (see p. 62). I am more likely sceptical about the existence of a direct formula of this kind. The following natural question, motivated by the results of Section 6.2.3, concerns Gaussian distribution framework.

Question 2 Let P, Q be non-degenerate Gaussian measures on $\mathbb{R}^N$ (see p. 165). Is there a non-degenerate Gaussian measure R on $\mathbb{R}^N$ such that $\mathcal{M}_R = \mathcal{M}_P \cap \mathcal{M}_Q$?

Graphs

Further open problems are related to Chapter 3. The following problem, named the "inclusion problem" in [42] can be viewed as an advanced subtask of the equivalence task (see Section 1.1).

Theme 2 Let G, H be acyclic directed graphs over N (see p. 154). Is there any graphical characterization of the inclusion $\mathcal{M}_G \subseteq \mathcal{M}_H$ (see Section 3.2)? Is it possible to characterize $\mathcal{M}_G \subseteq \mathcal{M}_H$ in terms of a simple algebraic relation of standard imsets u_G and u_H (p. 107)?

Note that suitable graphical characterization of Markov equivalence (i.e. of $\mathcal{M}_G = \mathcal{M}_H$) was found (p. 40) and I would appreciate an analogous solution of Theme 2 which is in terms of invariants of Markov equivalence. The following question concerning factorizably equivalent chain graphs was already mentioned in Section 3.3.

Question 3 Let $(\mathsf{X}_N, \mathcal{X}_N) = \prod_{i\in N}(\mathsf{X}_i, \mathcal{X}_i)$ be a fixed sample space with non-trivial $\mathcal{X}_i$ for every $i \in N$. Let G, H be (classic) chain graphs over N (p. 154) such that $\mathcal{M}_G = \mathcal{M}_H$ (see p. 43). Does the class of factorizable measures on $(\mathsf{X}_N, \mathcal{X}_N)$ with respect to G (see Remark 3.7 on p. 45) coincide with the class of factorizable measures with respect to H?

Structural imsets

There are some unsolved problems related to Chapter 7.

Theme 3 Let G be a chain graph over N (see p. 154). Is there any direct formula for the baricentral imset u over N (see p. 105) with $\mathcal{M}_u = \mathcal{M}_G$? Can every supermodular function m over N (see p. 71 and 72) be effectively 'translated' into a baricentral imset u over N with $\mathcal{M}_u = \mathcal{M}^m$?

Direction 1 Develop an effective criterion which decides whether a given structural imset is a baricentral imset.

Question 4 Let $\wp$ be a class of facially equivalent structural imsets over N (see p. 91) and $u \in \wp$ be a combinatorial imset minimal with respect to strong facial implication $\rightsquigarrow$ (see p. 113). Is u an imset of minimal degree in $\wp$?

8.1.2 Classification of skeletal imsets

Basic problem related to skeletal imsets (see Section 5.2) is the following one.

Theme 4 Is there any suitable characterization of skeletal imsets which allows to find the skeleton $\mathcal{K}_\ell^\diamond(N)$ for any finite non-empty set of variables N? How does $|\mathcal{K}_\ell^\diamond(N)|$ depend on $|N|$?

Note that [86] offers a characterization of extreme supermodular functions but the result is more likely a criterion whether a given ℓ-standardized supermodular function is skeletal (more precisely, it can be used for this purpose). However, the criterion does not seem suitable for the purpose of computer implementation. Therefore, the result of [86] does not solve the problem of finding the skeleton for every N. A promising idea how to tackle the problem is indicated in the rest of Section 8.1.2. A related task is the task to classify submaximal structural models. One can fix a way of standardization of skeletal imsets (see Remark 5.6) as submaximal structural models are in one-to-one correspondence with the elements of the respective skeleton. Every permutation $\pi : N \to N$ of variables can be extended to a permutation of the power set $\pi : \mathcal{P}(N) \to \mathcal{P}(N)$ and this step allows one to introduce permutable equivalence on the class of skeletal imsets: any skeletal imset m is equivalent in this sense to the composition $m\pi$ (see also [116]). Of course, every permutation of skeletal imsets defines a permutation of respective produced independence models. Basic way of classification is division of the class of (standardized) skeletal imsets into classes of permutable equivalence. Every equivalence class then represents a *type of a skeletal imset.* For example, 5 standardized skeletal imsets decompose into 3 types in case $|N| = 3$, 37 imsets decompose into 10 types in case $|N| = 4$ and 117978 imsets decompose into 1319 types in case $|N| = 5$.

Level equivalence

Nevertheless, perhaps even more precise way of classification of skeletal imsets exists. Suppose that $m \in \mathcal{K}(N)$ is a skeletal imset over N; let respective symbols m_ℓ, m_u and m_o denote the respective model equivalent element of the ℓ-skeleton, the u-skeleton and the o-skeleton obtained by formulas from Remark 5.6 (p. 80-81). Thus, m (more precisely, the respective class of model equivalent skeletal imsets) defines a certain equivalence on the class of subsets of N:

$$\forall\, S, T \subseteq N \quad S \sim_m T \;\Leftrightarrow\; [\, m_o(S) = m_o(T),\; m_\ell(S) = m_\ell(T) \text{ and } m_u(S) = m_u(T)\,]\,. \tag{8.1}$$

The equivalence classes of $\sim_m$ could be interpreted as areas in which these standardized skeletal imsets have the same values; in other words, they correspond to *levels of values.* In fact, I conjecture that the following hypothesis is true.

Question 5 Let $m \in \mathcal{K}_\ell^\diamond(N)$, m' is model equivalent skeletal imset over N (see p. 72) and $S, T \subseteq N$ such that $S \sim_m T$. Is then necessarily $m'(S) = m'(T)$?

Remark 8.1 As recognized in case $|N| = 4$ the o-standardized representative m_o cannot be omitted in (8.1) and $m_o(S) = m_o(T)$ is often equivalent to $S \sim_m T$ for $S, T \subseteq N$. It seems that m_ℓ resp. m_u can be omitted in (8.1) but not both. Therefore I think that o-standardization is the best standardization for the purpose of level equivalence. $\triangle$

Two skeletal imsets over N will be called *level equivalent* if they induce the same equivalence on subsets of $\mathcal{P}(N)$.

Observation 8.1 Let m^1, m^2 are level equivalent skeletal imsets over N and π is a permutation of N (extended to $\mathcal{P}(N)$). Then $m^1\pi$ and $m^2\pi$ are level equivalent.

Proof: This is a hint only. Given a skeletal imset m over N, put $r = m\pi$ and observe with help of formulas from Remark 5.6 that $r_\ell = m_\ell\pi$, $r_u = m_u\pi$ and $r_o = m_o\pi$. Hence, for every $S,T \subseteq N$ one has $S \sim_r T$ iff $\pi(S) \sim_m \pi(T)$ which implies the desired fact immediately. $\square$

REMARK 8.2 Further interesting operation with supermodular functions can be introduced with help of specific self-transformation ι of $\mathcal{P}(N)$:

$$\iota(S) = N \setminus S \quad \text{for every } S \subseteq N\,.$$

Given a supermodular function m over N one can introduce $z = m\iota$ and observe (see Section 5.1.3 in [116]) that z is also a supermodular function over N called the *reflection* of m. Indeed, the reflection of z is again m. Moreover, one can show using the formulas from Remark 5.6 that $z_\ell = m_u\iota$, $z_u = m_\ell\iota$ and $z_o = m_o\iota$. Consequently, for every $S,T \subseteq N$ one has $S \sim_z T$ iff $\iota(S) \sim_m \iota(T)$. An interesting fact is that in case $|N| \leq 4$ one has $S \sim_m T$ iff $N \setminus S \sim_m N \setminus T$ for every $m \in \mathcal{K}_o^\diamond(N)$ (see Example 8.1 below). In particular, m and z are level equivalent in this case. Nevertheless, the question whether the above hypotheses holds in general is open. $\triangle$

QUESTION 6 Let $m \in \mathcal{K}_o^\diamond(N)$ and $S,T \subseteq N$ such that $S \sim_m T$. Is then necessarily $N \setminus S \sim_m N \setminus T$?

Supertypes

Natural consequence of Observation 8.1 is that the concept of permutable equivalence can be extended to classes of level equivalence. Then every class of this extended permutable equivalence decomposes into several classes of level equivalence which decompose into individual (standardized) skeletal imsets. Thus, every class of permutable equivalence of this kind represents a *supertype*. For example, two supertypes exists in case $|N| = 3$ and five supertypes in case $|N| = 4$. An interesting fact is that every equivalence (8.1) on $\mathcal{P}(N)$, $|N| = 4$ defined by a skeletal imset m can be described by means of at most two 'cardinal' criteria which distribute sets $S \subseteq N$ to their equivalence classes (= levels) on basis of the cardinality of the intersection of S with one or two given disjoint subsets of N. Every equivalence on $\mathcal{P}(N)$ of this kind is therefore determined by a certain system of disjoint subsets of N having at most two components. This is illustrated by the following example.

EXAMPLE 8.1 One can distinguish five types of 'cardinal' criteria distributing subsets S of $N = \{a,b,c,d\}$ to levels which correspond to five supertypes of skeletal imsets.

1. The criterion $|S \cap \{a,b\}|$ divides $\mathcal{P}(N)$ into 3 levels - see the upper picture of Figure 8.1. The corresponding class of level equivalence has 1 standardized imset but the class of permutable equivalence has 6 classes of level equivalence. Therefore, the respective supertype involves 6 standardized skeletal imsets.

2. The criterion $|S \cap \{a,b,c\}|$ divides $\mathcal{P}(N)$ into 4 levels - see the lower picture of Figure 8.1. The corresponding class of level equivalence has 2 imsets, the class of permutable equivalence has 4 classes of level equivalence. Hence, the supertype involves 8 imsets. An example of a skeletal imset of this type is in Figure 6.4 where both ℓ-standardized and u-standardized versions are given.

3. The criterion $|S \cap \{a,b,c,d\}|$ divides $\mathcal{P}(N)$ into 5 levels - see the upper picture of Figure 8.2. The corresponding class of level equivalence has 3 imsets while the corresponding class of permutable equivalence has just one level equivalence class. Thus, the supertype involves 3 imsets.

4. Composed criterion $[\,|S \cap \{a,b,c\}|\,,\,|S \cap \{d\}|\,]$ divides $\mathcal{P}(N)$ into 8 levels - see the lower picture of Figure 8.2. The corresponding class of level equivalence has 2 imsets, the class of permutable equivalence has 4 classes of level equivalence and the supertype involves 8 imsets. An example of an imset of this kind is $m_\circ$ in the right-hand picture of Figure 6.3.

5. Composed criterion $[\,|S \cap \{a,b\}|\,,\,|S \cap \{c,d\}|\,]$ divides $\mathcal{P}(N)$ into 9 levels - see Figure 8.3. The corresponding class of level equivalence has 4 imsets while the corresponding class of permutable equivalence has 3 classes of level equivalence. The supertype involves 12 imsets; an example is the imset $m_\dagger$ from Figure 4.3. $\diamondsuit$

Endeavour described in Section 8.1.2 can be summarized as follows.

Theme 5 Can classification of supertypes of skeletal imsets by cardinal criteria be extended to a general case?

Moreover, in case of succesful solving of Themes 4 and 5 the following open problem may appear to be interesting.

Theme 6 Find out whether an ℓ-standardized supermodular imset producing a structural model $\mathcal{M}$ over N which is minimal with respect to dual strong facial implication (see p. 127) is uniquely determined. If yes, is the respective dual baricentral imset (p. 122) its multiple?

8.2 Operations with structural models

This section is an overview of basic operations with structural models. It is shown how they can be realized with help of operations with imsets (either with supermodular or with structural ones).

8.2.1 Reductive operations

These operations assign a model over T, $\emptyset \neq T \subseteq N$ to a structural model over N.

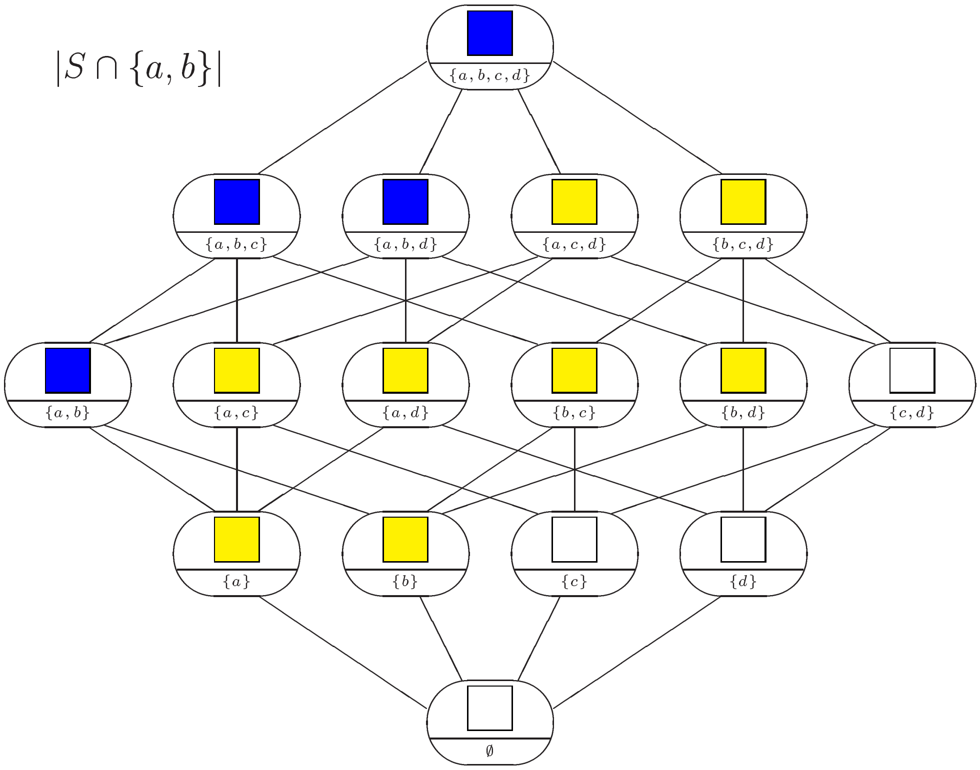
Contraction

Suppose $\mathcal{M} \subseteq \mathcal{T}(N)$, $\emptyset \neq T \subseteq N$ and $X \subseteq N \setminus T$. The model

$$\mathcal{M}_{T|X} = \{\langle A,B|C\rangle \in \mathcal{T}(T)\,;\ \langle A,B|CX\rangle \in \mathcal{M}\,\} \tag{8.2}$$

will be called the *contraction of $\mathcal{M}$ to T conditioned by X.*

Observation 8.2 If $\mathcal{M} \in \mathcal{U}(N)$ then $\mathcal{M}_{T|X} \in \mathcal{U}(T)$.



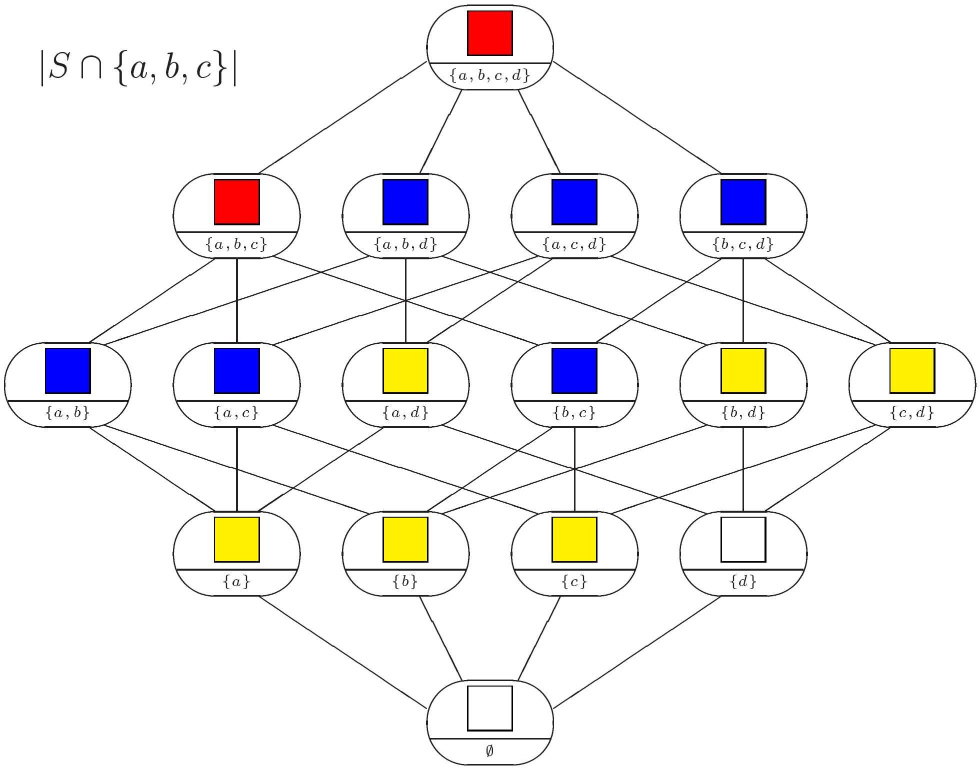
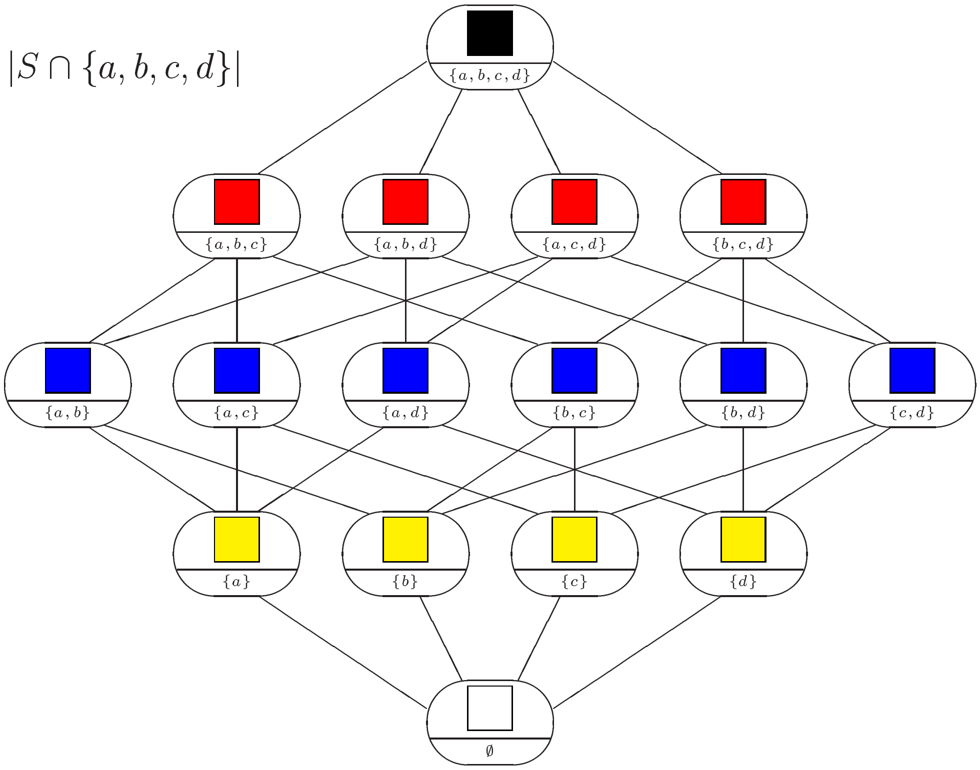


Figure 8.1: Cardinality criteria and respective levels for $N = \{a, b, c, d\}$.



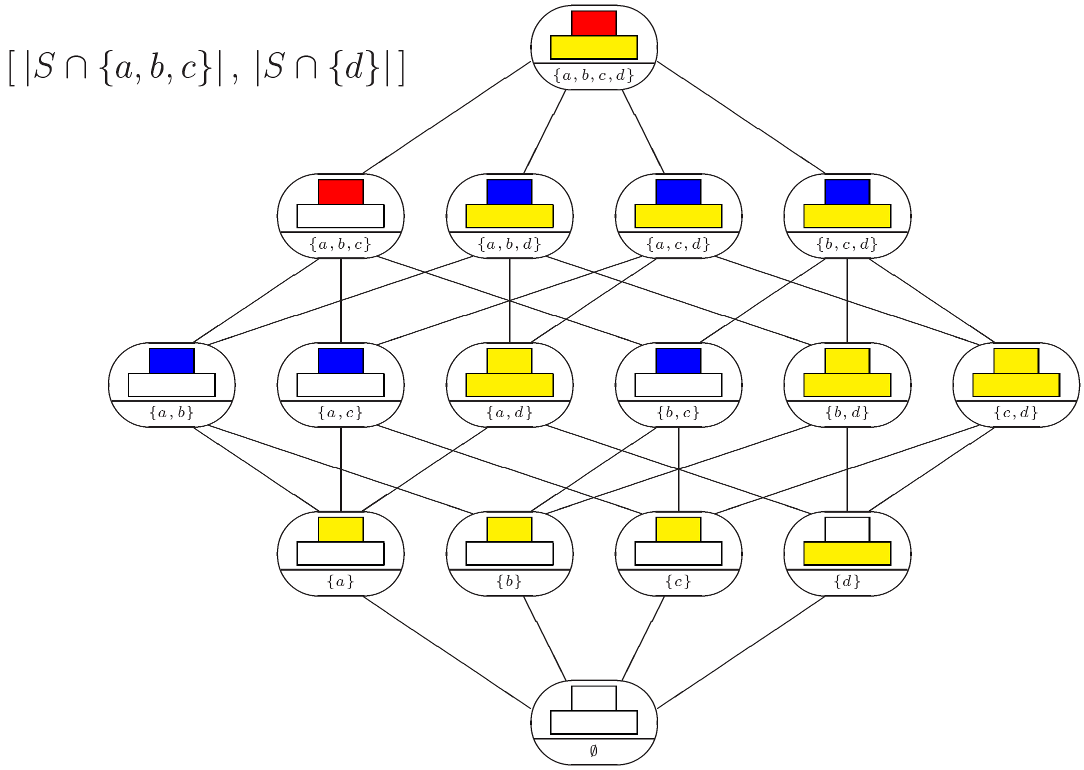


Figure 8.2: Further cardinality criteria and respective levels for $N = \{a, b, c, d\}$.

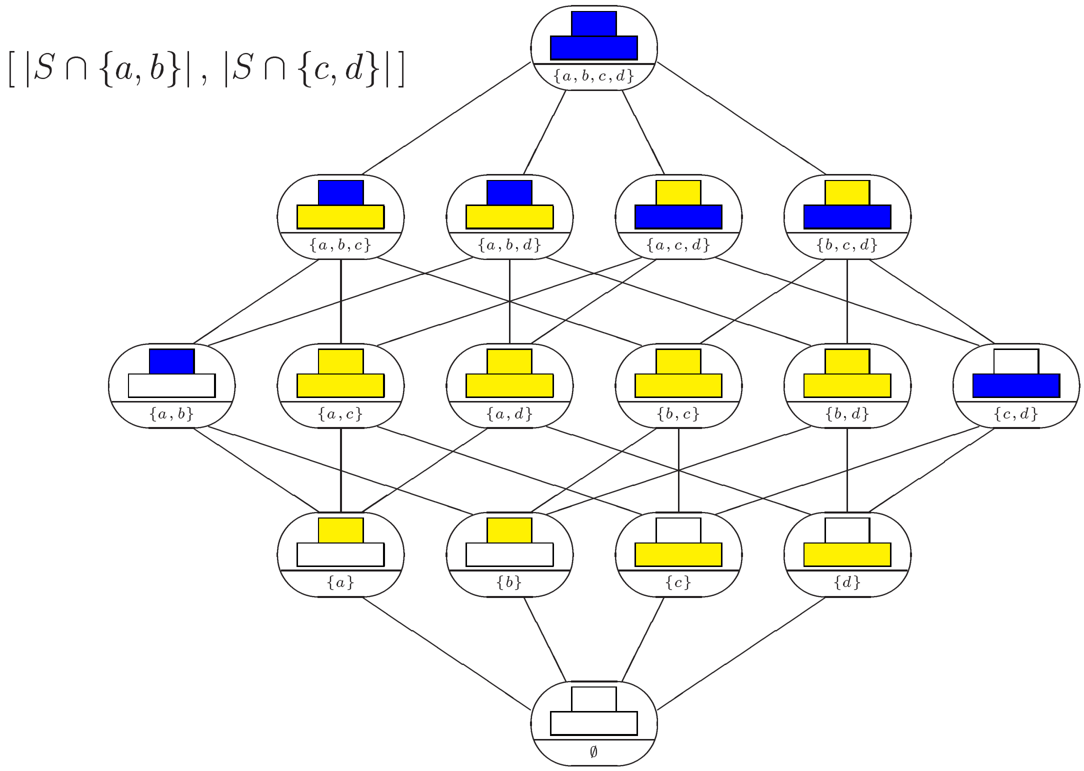


Figure 8.3: The last cardinality criterion and respective levels for $N = \{a,b,c,d\}$.

Proof: Given $m \in \mathbb{R}^{\mathcal{P}(N)}$ define $m_{T|X} \in \mathbb{R}^{\mathcal{P}(T)}$ by the formula

$$m_{T|X}(S) = m(S \cup X) \quad \text{for } S \subseteq T,$$

and observe $\langle m_{T|X}, u_{\langle A,B|C\rangle}\rangle = \langle m, u_{\langle A,B|CX\rangle}\rangle$ for every $\langle A,B|C\rangle \in \mathcal{T}(T)$. By Observation 5.1 derive that $m \in \mathcal{K}(N)$ implies $m_{T|X} \in \mathcal{K}(T)$. The equality above also implies $\langle A,B|C\rangle \in \mathcal{M}^{m_{T|X}}$ iff $\langle A,B|CX\rangle \in \mathcal{M}^{m}$. $\square$

Thus, conditioned contraction corresponds to linear operation $m \mapsto m_{T|X}$ with producing supermodular functions (imsets).

Remark 8.3 Note that the model $\mathcal{M}_{T|X}$ given by (8.2) was named *minor* of a semigraphoid $\mathcal{M}$ over N in [66] while the term 'contraction' was confined to the case $X = N\setminus T$ there. Moreover, the operation (8.2) applied to various graphical models was systematically treated in [85] and [45] under name 'marginalizing and conditioning'. My terminology is a compromise which reflects the idea that the operation (8.2) is simultaneously contraction and conditioning and fits best other names of operations with structural imsets mentioned below. $\triangle$

Restriction

Recall that the restriction $\mathcal{M}_T$ of $\mathcal{M} \subseteq \mathcal{T}(N)$ to $\emptyset \neq T \subseteq N$ was already introduced on p. 15 as $\mathcal{M} \cap \mathcal{T}(T)$. Of course, $\mathcal{M}_T$ is nothing but contraction $\mathcal{M}_{T|\emptyset}$ conditioned by the empty set. Hence, Observation 8.2 implies this.

Consequence 8.1 If $\mathcal{M} \in \mathcal{U}(N)$ and $\emptyset \neq T \subseteq N$ then $\mathcal{M}_T \in \mathcal{U}(T)$.

Note that it was shown in the proof of Observation 8.2 that the restriction of $\mathcal{M}$ corresponds to the restriction of producing supermodular function, i.e. $(\mathcal{M}^m)_T = \mathcal{M}^{m_T}$ where

$$m_T(S) = m(S) \quad \text{for every } S \subseteq T \text{ and } m \in \mathcal{K}(N).$$

On the other hand, an analogous statement for inducing structural imsets does not hold as the following example shows.

Example 8.2 There is no linear operation with structural imsets which corresponds to the restriction of induced structural models. Put $N = \{a, b, c\}$ and $T = \{a, b\}$. Take $v = u_{\langle a,c|\emptyset\rangle}$ and $w = u_{\langle a,b|c\rangle}$. By Lemma 4.5 $(\mathcal{M}_v)_T = (\mathcal{M}_w)_T = \mathcal{T}_\emptyset(T)$. On the other hand, $v + w = u_{\langle a,bc|\emptyset\rangle}$ which means $(\mathcal{M}_{u+v})_T = \mathcal{T}(T)$. Supposing there exists a linear mapping

$$u \in \mathcal{S}(N) \;\longmapsto\; u_T \in \mathcal{S}(T)$$

such that $(\mathcal{M}_u)_T = \mathcal{M}_{u_T}$ observe that that $v_T = 0 = w_T$ (use Lemma 6.1 and Observation 4.4 to conclude that the only structural imset over T inducing $\mathcal{T}_\emptyset(T)$ is the zero imset). By linearity of the mapping derive $(v + w)_T = 0$ which contradicts $(\mathcal{M}_{v+w})_T = \mathcal{T}(T)$. $\diamondsuit$

The preceding example motivates the next open problem.

Theme 7 Let u be a structural imset over N and $\emptyset \neq T \subseteq N$. Is there any direct formula for a structural imset inducing $(\mathcal{M}_u)_T$ is terms of u?

One can distinguish two versions of the problem. First, one can be interested in an algebraic formula which provides at least one structural imset over T inducing $(\mathcal{M}_u)_T$ for any $u \in \mathcal{S}(N)$. Second, one may wish to have an expression for the baricentral imset of $\mathcal{M}_T$ on basis of the baricentral imset of $\mathcal{M} \in \mathcal{U}(N)$. Nevertheless, both desired formulas must be non-linear as demonstrated by Example 8.2.

Remark 8.4 Restriction of every CI model induced by a probability measure P over N to $\emptyset \neq T \subseteq N$ is a CI model over T induced by the respective marginal P^T. In fact, conditioned contraction $\mathcal{M}_{T|X}$ of a CI model $\mathcal{M} \subseteq \mathcal{T}(N)$ induced by a discrete measure over N (where $X \subseteq N \setminus T$) is a CI model induced by a discrete measure over T [61]. $\triangle$

Other special operations

Given $\mathcal{M} \subseteq \mathcal{T}(N)$ and $\emptyset \neq T \subseteq N$ by the *full-conditioned contraction* is understood the model

$$\mathcal{M}_{T|*} = \{\, \langle A, B|C\rangle \in \mathcal{T}(T)\,;\ \forall\, X \subseteq N \setminus T \ \langle A, B|CX\rangle \in \mathcal{M} \,\}.$$

Clearly, $\mathcal{M}_{T|*} = \bigcap_{X \subseteq N \setminus T} \mathcal{M}_{T|X}$ and Observation 8.2 implies with help of Observation 5.7 the following fact.

Consequence 8.2 If $\mathcal{M} \in \mathcal{U}(N)$ and $\emptyset \neq T \subseteq N$ then $\mathcal{M}_{T|*} \in \mathcal{U}(T)$.

The consideration above also implies that full-conditioned contraction corresponds to the following linear operation with producing supermodular functions:

$$m \in \mathbb{R}^{\mathcal{P}(N)} \quad \longmapsto \quad m_{T|*}(S) = \sum_{K, K\cap T = S} m(K) \quad \text{for every } S \subseteq T\,.$$

REMARK 8.5 The *restriction of a structural imset* u over N to $\mathcal{P}(T)$ where $\emptyset \neq T \subseteq N$ need not be a structural imset. For example, consider $N = \{a,b,c\}$, $T = \{a,b\}$ and $u = u_{\langle b,c|a\rangle}$. Nevertheless, it is a structural imset if u vanishes outside $\mathcal{P}(T)$. This fact can be verified using Lemma 5.4 and Observation 5.1 with help of an observation that every supermodular function over T can be extended to a supermodular function over N - see the mapping (8.5) defined below. △

Nevertheless, the same linear mapping can be interpreted as a mapping assigning a structural imset over T to a structural imset over N named *contraction* :

$$u \in \mathbb{R}^{\mathcal{P}(N)} \quad \longmapsto \quad u_{[T]}(S) = \sum_{K, K\cap T = S} u(K) \quad \text{for } S \subseteq T. \tag{8.3}$$

OBSERVATION 8.3 If $u \in \mathcal{S}(N)$, $\emptyset \neq T \subseteq N$ then $u_{[T]} \in \mathcal{S}(T)$. Moreover,

$$\mathcal{M}'_{[T]} \equiv \{\, \langle A\cap T, B\cap T | C \cap T\rangle \,;\, \langle A,B|C\rangle \in \mathcal{M}_u \,\} \subseteq \mathcal{M}_{[T]} \equiv \mathcal{M}_{u_{[T]}}\,.$$

Proof: The first fact follows from linearity of the mapping (8.3) and the formula $\{u_{\langle A,B|C\rangle}\}_{[T]} = u_{\langle A\cap T, B\cap T|C\cap T\rangle}$ for any $\langle A,B|C\rangle \in \mathcal{T}(N)$. If $\langle A,B|C\rangle \in \mathcal{M}_u$ then $k\cdot u - u_{\langle A,B|C\rangle} \in \mathcal{S}(N)$ for some $k \in \mathbb{N}$ and $k \cdot u_{[T]} - \{u_{\langle A,B|C\rangle}\}_{[T]} \in \mathcal{S}(T)$ says $\langle A\cap T, B\cap T | C\cap T\rangle \in \mathcal{M}_{u_{[T]}}$. □

On the other hand, the inclusion can be strict as the following example shows.

EXAMPLE 8.3 There exists a structural imset u over N and $\emptyset \neq T \subseteq N$ with $\mathcal{M}'_{[T]} \neq \mathcal{M}_{[T]}$. Put $N = \{a,b,c,d\}$, $T = \{a,b,c\}$ and $u = u_{\langle a,b|\emptyset\rangle} + u_{\langle a,b|cd\rangle}$. Then $u_{[T]} = u_{\langle a,bc|\emptyset\rangle}$ but $\langle a, bc|\emptyset\rangle \notin \mathcal{M}'_{[T]}$. In fact, $\mathcal{M}'_{[T]}$ is not a semi-graphoid as $\langle a,b|\emptyset\rangle, \langle a,b|c\rangle \in \mathcal{M}'_{[T]}$. ◇

This motivates the next hypothesis.

THEME 8 Let $\mathcal{M} = \mathcal{M}_u$ for $u \in \mathcal{S}(N)$ and $\emptyset \neq T \subseteq N$. Is it true that $\mathcal{M}_{[T]} \equiv \mathcal{M}_{u_{[T]}}$ where $u_{[T]}$ is given by (8.3) coincides with the structural closure (see p. 113) of $\mathcal{M}'_{[T]} \equiv \{\langle A\cap T, B\cap T|C\cap T\rangle\,;\, \langle A,B|C\rangle \in \mathcal{M}\}$? Find out whether $\mathcal{M}_{[T]}$ is a CI model induced by a discrete probability measure over T provided that $\mathcal{M}$ is a CI model induced by a discrete probability measure over N.

8.2.2 Expansive operations

These operations assign a model over N to a structural model over T, $\emptyset \neq T \subseteq N$. Main attention is devoted to *extensions*, that is operations which assign a model over N to $\mathcal{M} \in \mathcal{U}(T)$ whose restriction is again $\mathcal{M}$. These expansive operations are pinpointed. Another type of expansive operation is a *lift* which can be viewed as a counterpart of (conditioned) contraction.

Solid extension

Given $\emptyset \neq T \subseteq N$, $\mathcal{M} \subseteq \mathcal{T}(T)$ the model $so(\mathcal{M}, N)$ over N given by

$$so(\mathcal{M}, N) = \{\, \langle A, B|C\rangle \in \mathcal{T}(N)\,;\ \langle A\cap T, B\cap T|C\cap T\rangle \in \mathcal{M}\,\} \tag{8.4}$$

will be called the *solid extension* of $\mathcal{M}$ to N.

OBSERVATION 8.4 If $\mathcal{M} \in \mathcal{U}(T)$ then $so(\mathcal{M}, N) \in \mathcal{U}(N)$ and $so(\mathcal{M}, N)_T = \mathcal{M}$.

The proof is based on a special linear extensive operation which assigns an extension m to every supermodular r over T:

$$r \in \mathcal{K}(T) \;\longmapsto\; m(S) = r(S\cap T) \quad \text{for } S \subseteq N\,. \tag{8.5}$$

Proof: Observe that the mapping from (8.5) is linear and $\langle m, u_{\langle A,B|C\rangle}\rangle = \langle r, u_{\langle A\cap T, B\cap T|C\cap T\rangle}\rangle$ for every $\langle A, B|C\rangle \in \mathcal{T}(N)$. Hence, $m \in \mathcal{K}(N)$ by Observation 5.1. Supposing $\mathcal{M} \in \mathcal{U}(T)$ by (5.15) there exists $r \in \mathcal{K}(T)$ with $\mathcal{M} = \mathcal{M}^r$. Observe that $so(\mathcal{M}, N) = \mathcal{M}^m$. □

REMARK 8.6 Note that the solid extension of a CI model $\mathcal{M}$ induced by a probability measure over T is a CI model again, the respective probability measure P over N has the form $P = Q \times \prod_{i\in N\setminus T} P_i$ where Q induces $\mathcal{M}$ and P_i are probability measures on arbitrary measurable spaces $(\mathsf{X}_i, \mathcal{X}_i)$, $i \in N\setminus T$. Moreover, the solid extension is a maximal extension in sense that $so(\mathcal{M}, N) \subseteq \mathcal{M}' \in \mathcal{U}(N)$, $(\mathcal{M}')_T = \mathcal{M}$ implies $\mathcal{M}' = so(\mathcal{M}, N)$. Indeed, it suffices to verify that $\langle A, B|C\rangle \in \mathcal{M}' \;\Rightarrow\; \langle A\cap T, B\cap T|C\cap T\rangle \in \mathcal{M}'$: since $\langle A, C\setminus T\,|\,C\cap T\rangle \in so(\mathcal{M}, N) \subseteq \mathcal{M}'$ and $\mathcal{M}'$ is a semi-graphoid by contraction property derive $\langle A, B(C\setminus T)\,|\,C\cap T\rangle \in \mathcal{M}'$ and hence by weak union and symmetry $\langle A\cap T, B\cap T\,|\,C\cap T\rangle \in \mathcal{M}'$. On the other hand, the solid extension need not be unique maximal extension. For example, consider $N = \{a, b, c\}$, $T = \{a, b\}$ and $\mathcal{M} = \mathcal{T}_\emptyset(T)$. Then $\mathcal{M}' = \mathcal{M}_u$ with $u = u_{\langle a,b|c\rangle} + u_{\langle a,c|b\rangle} + u_{\langle b,c|a\rangle}$ is another maximal $\mathcal{M}' \in \mathcal{U}(N)$ with $(\mathcal{M}')_T = \mathcal{M}$. △

Lift

Given $\emptyset \neq T \subseteq N$, $X \subseteq N\setminus T$ and $\mathcal{M} \subseteq \mathcal{T}(T)$ the model

$$li(\mathcal{M}, N : X) = \mathcal{T}_\emptyset(N) \cup \{\, \langle A, B|CX\rangle; \langle A, B|C\rangle \in \mathcal{M}\,\} \tag{8.6}$$

will be called the *lift* of $\mathcal{M}$ to N conditioned by X. Basic observation is that the operation of lift corresponds to the following linear mapping from $\mathbb{R}^{\mathcal{P}(T)}$ to $\mathbb{R}^{\mathcal{P}(N)}$ which assigns a structural imset $v[N, X]$ over N to a structural imset v over T:

$$v \in \mathbb{R}^{\mathcal{P}(T)} \;\longmapsto\; v[N, X](S) = \begin{cases} v(S\cap T) & \text{if } S\setminus T = X, \\ 0 & \text{if } S\setminus T \neq X, \end{cases} \quad \text{for any } S \subseteq N\,. \tag{8.7}$$

LEMMA 8.1 Suppose that $\emptyset \neq T \subseteq N$, $X \subseteq N\setminus T$. The mapping given by (8.7) is a linear mapping such that $v[N, X] \in \mathcal{S}(N)$ whenever $v \in \mathcal{S}(T)$. Moreover, it holds $li(\mathcal{M}_v, N : X) = \mathcal{M}_{v[N,X]}$. In particular, $li(\mathcal{M}, N : X) \in \mathcal{U}(N)$ whenever $\mathcal{M} \in \mathcal{U}(T)$ and one has $\{li(\mathcal{M}, N : X)\}_{T|X} = \mathcal{M}$.

Proof: Observe that $\delta_D[N,X] = \delta_{DX}$ for $D \subseteq T$ which implies by linearity of (8.7) $u_{\langle A,B|C\rangle}[N,X] = u_{\langle A,B|CX\rangle}$ for every $\langle A,B|C\rangle \in \mathcal{T}(T)$. This gives the first two statements of the lemma. As $\mathcal{M}_{v[N,X]}$ is a semi-graphoid it contains $\mathcal{T}_\emptyset(N)$. If $\langle A,B|C\rangle \in \mathcal{M}_v$ then $k \cdot v - u_{\langle A,B|C\rangle} \in \mathcal{S}(T)$ for some $k \in \mathbb{N}$ and by linearity $k \cdot v[N,X] - u_{\langle A,B|CX\rangle} \in \mathcal{S}(N)$ which means $\langle A,B|CX\rangle \in \mathcal{M}_{v[N,X]}$. The converse inclusion $\mathcal{M}_{v[N,X]} \subseteq li(\mathcal{M}_v, N:X)$ can be shown in three steps.

1. $\langle A,B|D\rangle \in \mathcal{T}(N) \setminus \mathcal{T}_\emptyset(N), \quad \neg\{X \subseteq D\} \;\Rightarrow\; \langle A,B|D\rangle \notin \mathcal{M}_{v[N,X]}$.
   Indeed, use Lemma 6.2: $\langle m^{D\downarrow}, u_{\langle A,B|D\rangle}\rangle > 0$ while $\langle m^{D\downarrow}, v[N,X]\rangle = 0$ as $X \setminus S \neq \emptyset$ for every $S \subseteq D$.

2. $\langle A,B|D\rangle \in \mathcal{T}(N) \setminus \mathcal{T}_\emptyset(N), \quad \neg\{ABD \subseteq TX\} \;\Rightarrow\; \langle A,B|D\rangle \notin \mathcal{M}_{v[N,X]}$.
   Indeed, again use Lemma 6.2: $\langle m^{ABD\uparrow}, u_{\langle A,B|D\rangle}\rangle > 0$ while $\langle m^{ABD\uparrow}, v[N,X]\rangle = 0$ as $S \setminus T \neq X$ for every $S \supseteq ABD$.

3. $\langle A,B|C\rangle \in \mathcal{T}(T) \setminus \mathcal{M}_v \;\Rightarrow\; \langle A,B|CX\rangle \notin \mathcal{M}_{v[N,X]}$.
   By Lemma 6.2 a supermodular function r over T with $\langle r, u_{\langle A,B|C\rangle}\rangle > 0$ and $\langle r, v\rangle = 0$ exists. Let m be its extension given by (8.5). It is a supermodular function over N, $\langle m, u_{\langle A,B|CX\rangle}\rangle = \langle r, u_{\langle A,B|C\rangle}\rangle > 0$ and $\langle m, v[N,X]\rangle = \langle r, v\rangle = 0$.

Thus, if $\langle A,B|D\rangle \in \mathcal{M}_{v[N,X]} \setminus \mathcal{T}_\emptyset(N)$ then $X \subseteq D$ and $ABD \subseteq TX$ by 1. and 2. Therefore $\langle A,B|D\rangle = \langle A,B|CX\rangle$ where $\langle A,B|C\rangle \in \mathcal{T}(T)$ and $\langle A,B|C\rangle \in \mathcal{M}_v$ by 3. The next statement then follows from (5.14); the equality $\{li(\mathcal{M}, N:X)\}_{T|X} = \mathcal{M}$ is trivial. $\square$

Ascetic extension

Given $\emptyset \neq T \subseteq N$, $\mathcal{M} \subseteq \mathcal{T}(T)$ the model $as(\mathcal{M}, N)$ over N given by

$$as(\mathcal{M}, N) = \mathcal{T}_\emptyset(N) \cup \mathcal{M} \tag{8.8}$$

is called the *ascetic extension* of $\mathcal{M}$ to N. It is nothing but the lift $li(\mathcal{M}, N:X)$ with $X = \emptyset$. Lemma 8.1 therefore implies this.

Consequence 8.3 If $\mathcal{M} \in \mathcal{U}(T)$ then $as(\mathcal{M}, N) \in \mathcal{U}(N)$ and $as(\mathcal{M}, N)_T = \mathcal{M}$.

It follows directly from (8.8) that the ascetic extension is the least extension of $\mathcal{M} \in \mathcal{U}(T)$ in sense that $\mathcal{M}' \in \mathcal{U}(N)$, $(\mathcal{M}')_T = \mathcal{M}$ implies $as(\mathcal{M}, N) \subseteq \mathcal{M}'$. Let me remind that the proof of Lemma 8.1 implies that the ascetic extension is realized by means of a linear operation with inducing structural imsets, namely by means of the mapping

$$v \in \mathbb{R}^{\mathcal{P}(T)} \;\longmapsto\; u(S) = \begin{cases} v(S) & \text{if } S \subseteq T\,, \\ 0 & \text{otherwise}\,, \end{cases} \qquad \text{for every } S \subseteq N\,.$$

Note that one can show with help of Lemma 2.9 that the ascetic extension of a CI model induced by non-degenerate discrete measure over T is a CI model (over N).

Cellular extension

Given $\emptyset \neq T \subseteq N$, $\mathcal{M} \subseteq \mathcal{T}(T)$ the model $ce(\mathcal{M}, N)$ over N given by

$$ce(\mathcal{M}, N) = \mathcal{T}_{\emptyset}(N) \cup \{ \langle A, B|CX \rangle\, ;\, \langle A, B|C \rangle \in \mathcal{M} \quad X \subseteq N \setminus T \} \tag{8.9}$$

is called the *cellular extension* of $\mathcal{M}$ to N. In fact, $ce(\mathcal{M}, N) = \bigcup_{X \subseteq N \setminus T} li(\mathcal{M}, N : X)$. Basic observation is that the cellular extension corresponds to a linear mapping from $\mathbb{R}^{\mathcal{P}(N)}$ which assigns a structural imset u over N to a structural imset v over T:

$$v \in \mathbb{R}^{\mathcal{P}(T)} \;\longmapsto\; u(S) = v(S \cap T) \quad \text{for } S \subseteq N. \tag{8.10}$$

LEMMA 8.2 Suppose $\mathcal{M} \subseteq \mathcal{T}(T)$ and $\emptyset \neq T \subseteq N$. The mapping given by (8.10) is a linear mapping which assigns $u \in \mathcal{S}(N)$ to $v \in \mathcal{S}(T)$. Moreover, $\mathcal{M}_u = ce(\mathcal{M}_v, N)$. In particular, $ce(\mathcal{M}, N) \in \mathcal{U}(N)$ whenever $\mathcal{M} \in \mathcal{U}(T)$ and $ce(\mathcal{M}, N)_T = \mathcal{M}$.

Proof: It follows directly from (8.10) and (8.7) that $u = \sum_{X \subseteq N \setminus T} v[N, X]$. Thus, $u \in \mathcal{S}(N)$ by Lemma 8.1 and $\mathcal{M}_{v[N,X]} \subseteq \mathcal{M}_u$ by Lemma 6.1 for every $X \subseteq N \setminus T$. The converse inclusion $\mathcal{M}_u \subseteq ce(\mathcal{M}_v, N)$ can be shown in two steps.

1. $\langle U, V|W \rangle \in \mathcal{T}(N), \quad U \setminus T \neq \emptyset \neq V \;\Rightarrow\; \langle U, V|W \rangle \notin \mathcal{M}_u$.
   Indeed, choose $i \in U \setminus T$, $j \in V$ and use Lemma 6.2: $\langle m, u_{\langle U,V|W \rangle} \rangle > 0$ and $\langle m, u \rangle = \sum_{X \subseteq N \setminus T} \langle m, v[N, X] \rangle = 0$ for $m = m^{ij\uparrow}$. To verify the last equality use Observation 5.6 and show $\mathcal{M}_{v[N,X]} \subseteq \mathcal{M}^m$: one has $\langle m, u_{\langle A,B|CX \rangle} \rangle = 0$ for every $\langle A, B|C \rangle \in \mathcal{M}_v$ and $X \subseteq N \setminus T$ according to Lemma 8.1. This is clear whenever $i \in CX$, in case $i \notin CX$ the assumption $i \notin T \supseteq AB$ implies $i \notin ABCX$.

2. $\langle U, V|W \rangle \in \mathcal{T}(N), \quad U, V \subseteq T, \quad \langle U, V|W \cap T \rangle \notin \mathcal{M}_v \;\Rightarrow\; \langle U, V|W \rangle \notin \mathcal{M}_u$.
   Indeed, by Lemma 6.2 find a supermodular function r over T with $\langle r, u_{\langle U,V|W \cap T \rangle} \rangle > 0$ and $\langle r, v \rangle = 0$. Define a supermodular function m over N by (8.5) and observe $\langle m, u_{\langle U,V|W \rangle} \rangle = \langle r, u_{\langle U,V|W \cap T \rangle} \rangle > 0$ with $\langle m, u \rangle = \sum_{X \subseteq N \setminus T} \sum_{K \subseteq T} m(KX) \cdot u(KX) = \sum_{X \subseteq N \setminus T} \langle r, v \rangle = 0$. This implies $\langle U, V|W \rangle \notin \mathcal{M}_u$ by Lemma 6.2.

Thus, if $\langle U, V|W \rangle \in \mathcal{M}_u \setminus \mathcal{T}_{\emptyset}(N)$ then $U, V \subseteq T$ by 1. and $\langle U, V|W \cap T \rangle \in \mathcal{M}_v$ by 2. To derive further statement use (5.14); the last formula is trivial. $\square$

REMARK 8.7 This is to explain my reasons for the chosen terminology. The reader can observe on basis of (8.8), (8.9) and (8.4) that

$$as(\mathcal{M}, N) \subseteq ce(\mathcal{M}, N) \subseteq so(\mathcal{M}, N) \quad \text{for every } \mathcal{M} \subseteq \mathcal{T}(T)\,.$$

The inclusions may be strict as Example 8.4 below shows. The fact that $as(\mathcal{M}, N)$ is the least extension of $\mathcal{M}$ motivated the adjective 'ascetic' and the fact that $so(\mathcal{M}, N)$ is one of maximal extensions of $\mathcal{M}$ (see Remark 8.6) motivated the adjective 'solid'. The adjective 'cellular' was motivated by the fact that $ce(\mathcal{M}, N)$ is composed of several different 'cells', namely $li(\mathcal{M}, N : X)$ for $X \subseteq N \setminus T$. $\triangle$

EXAMPLE 8.4 There exists $\mathcal{M} \in \mathcal{U}(N)$, $\emptyset \neq T \subseteq N$ whose ascetic, cellular and solid extensions differ. Put $N = \{a, b, c\}$, $T = \{a, b\}$ and $\mathcal{M} = \mathcal{M}_u$ where $u = u_{\langle a,b|\emptyset \rangle}$. Then $\langle a, b|c \rangle \in ce(\mathcal{M}, N) \setminus as(\mathcal{M}, N)$ and $\langle c, ab|\emptyset \rangle \in so(\mathcal{M}, N) \setminus ce(\mathcal{M}, N)$. $\diamondsuit$

8.2.3 Accumulative operations

The aim to represent 'big' structural models effectively in memory of a computer motivates the need for suitable definition of decomposition of a structural model into 'less-dimensional' models.

Localization of UG models

The following intuitive consideration does not pretend preciseness, it serves as a motivation account only (some results cited below can be even misinterpreted). Final goal is to achieve an analogue of the result from [63] saying that every UG model has *canonical decomposition into prime* UG submodels. In fact, well-known concept of *decomposition of undirected graphs* from [53] is behind this approach. Recall that if G is an undirected graph over N and $\langle A,B|C\rangle \in \mathcal{T}(N)\setminus\mathcal{T}_{\emptyset}(N)$ such that C a complete set in G and $A \perp\!\!\!\perp B\,|\,C\;[G]$ (p. 37) then the pairs of undirected graphs (G_{AC}, G_{BC}) is called a *proper decomposition of* G. The graphs G_{AC} resp. G_{BC} can have possibly further proper decompositions which means that every UG model $\mathcal{M}_G$ can be gradually decomposed into UG submodels with no proper decomposition which can be named *prime* UG submodels of $\mathcal{M}_G$. In my view, the result of [63] can be paraphrased as follows: for every undirected graph G over N a unique triangulated supergraph H over N (see p. 45) exists whose cliques correspond to (maximal) subsets of N defining prime UG submodels of $\mathcal{M}_G$. Therefore, various computational operations with Markovian measures with respect to G can be done 'locally' - within prime UG submodels. I believe that well-known method of *local computation* applied mainly to DAG models [17] has analogous source of justification. Therefore I hope that these ideas can be extended to more general structural models. Suitable concept of decomposition of a structural model based on an accumulative operation with structural models is needed. One of possible proposals is mentioned below.

Composition

Given $\langle A,B|C\rangle \in \mathcal{T}(N)$ and $\mathcal{M}^1 \in \mathcal{U}(AC)$, $\mathcal{M}^2 \in \mathcal{U}(BC)$ such that $\mathcal{M}^1_C = \mathcal{M}^2_C$ by the *composition of* $\mathcal{M}^1$ *and* $\mathcal{M}^2$ will be understood the structural model $\mathcal{M}^1\otimes\mathcal{M}^2$ over ABC given by

$$\mathcal{M}^1 \otimes \mathcal{M}^2 = cl_{\mathcal{U}(ABC)}(as(\mathcal{M}^1, ABC) \cup as(\mathcal{M}^2, ABC) \cup \{\langle A,B|C\rangle\})\,. \tag{8.11}$$

In words, both $\mathcal{M}^1$ and $\mathcal{M}^2$ are embedded into $\mathcal{U}(ABC)$ by the ascetic extension, then $\langle A,B|C\rangle$ is added and the structural closure operation (see p. 113) is applied. It follows from the definition that $\mathcal{M}^1 \otimes \mathcal{M}^2 \in \mathcal{U}(ABC)$. Natural question related to the concept of *conditional product* from [21] is the following one.

THEME 9 Suppose $\langle A,B|C\rangle \in \mathcal{T}(N)$, $\mathcal{M}^1 \in \mathcal{U}(AC)$, $\mathcal{M}^2 \in \mathcal{U}(BC)$ with $\mathcal{M}^1_C = \mathcal{M}^2_C$. Is it true that $(\mathcal{M}^1 \otimes \mathcal{M}^2)_{AC} = \mathcal{M}^1$ and $(\mathcal{M}^1 \otimes \mathcal{M}^2)_{BC} = \mathcal{M}^2$? Can the domain of the operation $\otimes$ defined by (8.11) be restricted suitably so that the axioms of conditional product from [21] are fulfilled for it then?

Structural decomposition

Let $\mathcal{M}$ be a structural model over N and $U,V \subseteq N$ such that $UV = N$. One says that $\mathcal{M}$ decomposes into $\mathcal{M}_U$ and $\mathcal{M}_V$ or that $(\mathcal{M}_U, \mathcal{M}_V)$ forms a *structural decomposition of* $\mathcal{M}$ if

$\mathcal{M} = \mathcal{M}_U \otimes \mathcal{M}_V$. The decomposition is *proper* if $U \setminus V \neq \emptyset \neq V \setminus U$. A structural model will be called *indecomposable* if it has no proper structural decomposition. Note that $(\mathcal{M}_U)_{U\cap V} = (\mathcal{M}_V)_{U\cap V}$ which means that the composition $\mathcal{M}_U \otimes \mathcal{M}_V$ is always defined. Clearly, a necessary condition for the existence of structural decomposition $(\mathcal{M}_U, \mathcal{M}_V)$ is $U \setminus V \perp\!\!\!\perp V \setminus U \,|\, U \cap V \; [\mathcal{M}]$ (see p. 15). In particular, $\mathcal{M} = \mathcal{T}_\emptyset(N)$ is an indecomposable model. As $\mathcal{M}_U$ resp. $\mathcal{M}_V$ can be again decomposed one can obtain gradually a *full decomposition* of $\mathcal{M}$ into indecomposable models. Unfortunately, the hypothesis that every structural model has unique full decomposition of this type into 'prime' components is false. Indeed, consider the model from Example 7.4: one has $\mathcal{M}_{abc} \otimes \mathcal{M}_{abd} = \mathcal{M} = \mathcal{M}_{acd} \otimes \mathcal{M}_{bcd}$ and every $\mathcal{M}_S$ with $S \subseteq N$, $|S| = 3$ is indecomposable.

Theme 10 Let G be an undirected graph over N and $\langle A, B|C\rangle \in \mathcal{T}(N) \setminus \mathcal{T}_\emptyset(N)$ defines a proper decomposition of G. Is $((\mathcal{M}_G)_{AC}, (\mathcal{M}_G)_{BC})$ a structural decomposition of $\mathcal{M}_G$? Has any CG model (see p. 43) unique minimal 'canonical' decomposition into maximal indecomposable submodels?

It may be the case that uniqueness of 'canonical' decomposition cannot be achieved even under possible additional standardization requirements. This would confirm that the concept of structural decomposition is not suitable for the purpose mentioned earlier (p. 141). Then one should look for another type of decomposition (based on another accumulative operation with structural imsets) which generalizes decomposition of UG models.

Direction 2 Develop an analogue of the method of local computation for structural models based on conveniently defined concept of decomposition of structural models. Find sufficient conditions for decomposition of this type which can be verified by statistical tests or on basis of expert knowledge. Develop an analogy of Shenoy's pictorial method of valuation networks [92, 89] for local representation of structural imsets and Markovian measures in memory of a computer.

Well-known results about factorization of maximum-likelihood estimators [53, 17] should be generalized then as well.

8.3 Implementation tasks

These open problems are motivated by the task to implement facial implication on a computer. The most important question is probably the next one.

Question 7 Is every structural imset u over N already a combinatorial imset over N (see p. 59)?

If the answer to Question 7 is negative then the following two problems become topics of immediate interest.

Theme 11 Given a finite non-empty set of variables N, find the least finite class $\mathcal{H}(N)$ of structural imset such that

$$\forall\, u \in \mathcal{S}(N) \quad u = \sum_{v \in \mathcal{H}(N)} k_v \cdot v \qquad \text{for some } k_v \in \mathbb{Z}^+ .$$

Recall that the existence of the class $\mathcal{H}(N)$, named minimal integral Hilbert basis of $con(\mathcal{E}(N))$ follows from Theorem 16.4 in [90]. One has $\mathcal{E}(N) = \mathcal{H}(N)$ iff $\mathcal{S}(N) = \mathcal{C}(N)$.

THEME 12 Given a finite non-empty set of variables N, determine the least $n_* \in \mathbb{N}$ such that an imset over N is structural iff its multiple $n_* \cdot u$ is a combinatorial imset, i.e.

$$\forall\, u \in \mathbb{Z}^{\mathcal{P}(N)} \quad u \in \mathcal{S}(N) \;\Leftrightarrow\; n_* \cdot u \in \mathcal{C}(N).$$

Determine the least $n_{**} \in \mathbb{N}$ satisfying

$$\forall\, u \in \mathbb{Z}^{\mathcal{P}(N)} \quad u \in \mathcal{S}(N) \;\Leftrightarrow\; \exists\, n \in \mathbb{N}\; n \leq n_{**} \quad n \cdot u \in \mathcal{C}(N)\,.$$

Find out how the values n_* and n_{**} depend on $|N|$.

Note that $n_{**} \leq n_*$ and I am not able to decide whether the inequality is strict. Indeed, $n_* = 1 \Leftrightarrow n_{**} = 1 \Leftrightarrow \mathcal{S}(N) = \mathcal{C}(N)$. Further important question concerns the ℓ-skeleton.

QUESTION 8 Let $\mathcal{K}_\ell^\diamond(N)$ be the ℓ-skeleton over N (see p. 76) and $\mathcal{E}(N)$ the class of elementary imsets over N (p. 57). Is the equality

$$\min\,\{\, \langle m, u\rangle\,;\; u \in \mathcal{E}(N) \quad \langle m, u\rangle \neq 0 \,\} = 1$$

fulfilled for every $m \in \mathcal{K}_\ell^\diamond(N)$?

Note that the condition from Question 8 implies that $gra(N) = gra_*(N)$ (see p. 101). The following problem becomes relevant in case both Question 7 and Question 8 have negative answers.

THEME 13 How does depend the value of the least $l \in \mathbb{N}$ satisfying the condition

$$\forall\, u \in \mathcal{S}(N) \;\; \forall\, v \in \mathcal{E}(N) \quad u \rightharpoonup v \;\Leftrightarrow\; l \cdot u - v \in \mathcal{S}(N) \tag{8.12}$$

depend on $|N|$? Recall that facial implication $\rightharpoonup$ is defined on p. 93. Can one determine $gra(N)$ directly without finding the skeleton, i.e. without solving Theme 4?

Recall that if either Question 7 or Question 8 has positive answer then the least $l \in \mathbb{N}$ satisfying (8.12) is $gra(N)$.

THEME 14 Is there the least $l_* \in \mathbb{N}$ such that

$$\forall\, u \in \mathcal{S}(N) \;\; \forall\, v \in \mathcal{E}(N) \quad u \rightharpoonup v \;\Leftrightarrow\; l_* \cdot u - v \in \mathcal{C}(N)\,?$$

How does l_* depend on $|N|$ then? If there is no $l_* \in \mathbb{N}$ of this kind, find out, for a given $u \in \mathcal{S}(N)$, how the class of $k \in \mathbb{N}$ satisfying

$$\forall\, v \in \mathcal{E}(N) \quad u \rightharpoonup v \;\Leftrightarrow\; k \cdot u - v \in \mathcal{C}(N)$$

looks? Is there a structural imset $u \in \mathcal{S}(N)$ such that the condition (6.4) from Remark 6.3 is not fulfilled (that is $n \cdot u \notin \mathcal{C}(N)$ or $(k \cdot n - 1) \cdot u \notin \mathcal{C}(N)$ for every $k, n \in \mathbb{N}$)?

Theme 15 Given a finite non-empty set of variables N, is there $l_{\dagger} \in \mathbb{N}$ such that

$$\forall\, u \in \mathcal{S}(N)\ \forall\, v \text{ semi-elementary imset over } N \quad u \rightharpoonup v \;\Leftrightarrow\; l_{\dagger} \cdot u - v \in \mathcal{S}(N),$$

respectively $l_{\dagger\dagger} \in \mathbb{N}$ such that

$$\forall\, u \in \mathcal{S}(N)\ \forall\, v \text{ semi-elementary imset over } N \quad u \rightharpoonup v \;\Leftrightarrow\; l_{\dagger\dagger} \cdot u - v \in \mathcal{C}(N)?$$

How does $l_{\dagger}$ respectively $l_{\dagger\dagger}$ depend on $|N|$ then?

The following open problem also concerns facial implication.

Theme 16 Is there any method of testing facial implication which combines direct and skeletal criteria (see Lemma 6.1 on p. 93 and Lemma 6.2 on p. 96) and which is more suitable for efficient implementation on a computer?

The above formulation is partially vague, let me specify what I have in mind in more details. Direct criterion of facial implication $u \rightharpoonup v$ consists in testing whether $k \cdot u - v \in \mathcal{C}(N)$ for some $k \in \mathbb{N}$. This can be tested recursively as mentioned in Remark 6.3. However, plenty of 'transient' imset obtained during 'decomposition' procedure are not combinatorial imsets. This can be often recognized immediately by means of Theorem 5.1 (which is behind the skeletal criterion of $\rightharpoonup$) and save superfluous steps of the recursive 'decomposition' procedure. The observation that a 'transient' imset v is not combinatorial can be made on basis of the fact that $\langle m, v\rangle < 0$ for a 'standard' supermodular imset m over N, for example the imset $m^{A\uparrow}$ resp. $m^{A\downarrow}$ for $A \subseteq N$ (p. 34) or m_l for $l = 0, \ldots, |N|-2$ (p. 57). The point is that one need not to have the whole skeleton at disposal! In fact, Remark 6.10 is based just on observations of this type.

Memory demands

Another important problem is what are memory demands for representing a structural imset in memory of a computer. Informally, by *actual dimension* of the class of structural models $\mathcal{U}(N)$ is understood the 'minimal' number of binary attributes of elements of $\mathcal{U}(N)$ which can distinguish between every pair of distinct structural models.

Observation 8.5 The following inequality holds

$$\lceil \ln_2 |\mathcal{U}(N)|\, \rceil \;\leq\; \text{actual dimension of } \mathcal{U}(N) \;\leq \min\{\, |\mathcal{E}(N)|, |\mathcal{K}^{\diamond}_{\ell}(N)|\,\}.$$

Proof: If s binary attributes distinguish between elements of $\mathcal{U}(N)$ then s bites is enough to represent all elements of $\mathcal{U}(N)$. Hence $2^s \geq |\mathcal{U}(N)|$ gives the lower estimate. The fact that elementary, respectively skeletal imsets differentiate between structural models follows from Lemma 2.2 resp. Consequence 6.2 □

For example, the actual dimension of $\mathcal{U}(N)$ is 1 in case $|N| = 2$ as $|\mathcal{U}(N)| = 2$ and $|\mathcal{E}(N)| = |\mathcal{K}^{\diamond}_{\ell}(N)| = 1$ (while in case $|N| = 1$ one has $|\mathcal{U}(N)| = 1$ and $\mathcal{E}(N) = \emptyset = \mathcal{K}^{\diamond}_{\ell}(N)$). If $|N| = 3$ then $|\mathcal{U}(N)| = 22$ gives lower estimate $5 = \lceil\, ln_2\, |\mathcal{U}(N)|\,\rceil$ which is precise since $|\mathcal{K}^{\diamond}_{\ell}(N)| = 5 < 6 = |\mathcal{E}(N)|$. In case $|N| = 4$ one has $2^{14} < |\mathcal{U}(N)| = 22108 < 2^{15}$, $|\mathcal{E}(N)| = 24$ and $|\mathcal{K}^{\diamond}_{\ell}(N)| = 37$. Thus, Observation 8.5 implies:

CONSEQUENCE 8.4 If $|N| = 4$ then the actual dimension of $\mathcal{U}(N)$ is between 15 and 24.

Note that the inequality $|\mathcal{U}(N)| < 2^{15}$ in case $|N| = 4$ means that one can perhaps 'construct' 15 awkward artificial attributes which differentiate between the elements of $\mathcal{U}(N)$ (and perhaps they have even the form of 'functions' of 24 'elementary' characteristics). However, I am interested in those characteristics or attributes which have reasonable interpretation and can be generalized in sense that their generalization 'achieves' the actual dimension of $\mathcal{U}(N)$ for $|N| \geq 5$. In fact, I am interested in solution of the following vaguely defined problem.

THEME 17 Given a non-empty finite set of variables N, what is the least cardinality of a set of interpretable binary attributes which differentiate between structural models over N? How does it depend on $|N|$?

8.4 Interpretation and learning tasks

Open problems loosely motivated by 'practical' questions of interpretation and learning from Section 1.1 are gathered below.

8.4.1 Meaningful description of structural models

The following two open problems are motivated by the concept of standard imset for an acyclic directed graph from Section 7.2.1

QUESTION 9 Let G be an acyclic directed graph over N (p. 154). Is it true that the standard imset for G (see p. 107) is the only imset from the class of combinatorial imsets inducing $\mathcal{M}_G$ which is simultaneously an imset of the least degree (p. 112) and an imset with the least lower class (p. 116)?

Natural question is whether the concept of standard imset for a DAG model can be generalized.

THEME 18 Is there any consistent principle of unique choice of representatives of classes of facial equivalence (see p. 91) such that, for every acyclic directed graph G over N, the standard imset u_G is chosen from the class $\wp = \{u \in \mathcal{S}(N)\,;\, \mathcal{M}_u = \mathcal{M}_G\}$?

The above description is somewhat vague, let me specify more detailed hypotheses. Suppose $\mathcal{M} \in \mathcal{U}(N)$, $\wp = \{v \in \mathcal{S}(N)\,;\, \mathcal{M} = \mathcal{M}_v\}$ and $\mathcal{U} = \mathcal{U}_v$ for some $v \in \wp$ (it does not depend on $v \in \wp$ - see Lemma 6.6 on p. 102). Let us put

$$\mathcal{L}^1_* = \bigcup \{\mathcal{L} \subseteq \mathcal{U}\,;\; \mathcal{L} \text{ is a minimal lower class } \mathcal{L}_v \text{ for } v \in \wp\,\}$$

where minimality is understood with respect to inclusion ($\mathcal{L}_v$ is defined on p. 60). Respective hypotheses are that $u \in \wp$ with $\mathcal{L}_u = \mathcal{L}^1_*$ exists for every $\mathcal{M} \in \mathcal{U}(N)$ and that a combinatorial imset with the least degree among $u \in \wp \cap \mathcal{C}(N)$ satisfying $\mathcal{L}_u = \mathcal{L}^1_*$ is determined uniquely. If they were true then u_G is an imset of this kind for $\mathcal{M} = \mathcal{M}_G$

where G is an acyclic directed graph over N (by Lemma 7.4 on p. 112 and Consequence 7.5 on p. 118). The hypotheses can be modified by considering the class

$$\mathcal{L}_*^2 = \bigcup \{\mathcal{L} \subseteq \mathcal{U}\,;\ \mathcal{L} \text{ is a minimal determining class for } \mathcal{M}\}$$

or the class

$$\mathcal{L}_*^3 = \bigcup \{\mathcal{L} \subseteq \mathcal{U}\,;\ \mathcal{L} \text{ is a minimal unimarginal class for } \mathcal{M}\}$$

(see p. 115) in place of $\mathcal{L}_*^1$. Further open problem is motivated by Section 6.4.

Direction 3 Look for necessary conditions for facial equivalence of structural imsets formulated in terms of invariants of facial equivalence which are easy to verify and offer clear interpretation. The aim is to find a set of these conditions which is able to distinguish every pair of structural imsets which are not equivalent.

Desired complete set of interpretable invariants could then become a basis of alternative way of description of structural models which is suitable from the point of view of interpretation.

8.4.2 Distribution frameworks and learning

Below mentioned problems concern more or less the distribution framework. In my view, they are also related to general task of learning structural models (see Section 1.1, p. 9).

Theme 19 Let Ψ be a class of probability measures over N satisfying the conditions (6.14) and (6.15) from Section 6.2.3. Let $\Psi(u)$ denote the class of Markovian measures with respect to $u \in \mathcal{S}(N)$ given by (6.1) on p. 91 and $\mathcal{S}_\Psi(N)$ the class of Ψ-representable structural imsets over N (p. 98). Is the condition

$$\forall\, u, v \in \mathcal{S}_\Psi(N) \quad u \rightharpoonup v \;\Leftrightarrow\; \Psi(u) \subseteq \Psi(v)$$

fulfilled then? Does it hold under additional assumptions on Ψ?

Question 10 Let $\mathcal{M}$ be a structural model over N, $\mathcal{U} = \mathcal{U}_u$ is the upper class of $u \in \mathcal{S}(N)$ with $\mathcal{M}_u = \mathcal{M}$ and $\mathcal{D} \subseteq \mathcal{U}$ is an unimarginal class for $\mathcal{M}$ (see Section 7.4.1, p. 115). Is then $\mathcal{D}$ necessarily a determining class for $\mathcal{M}$?

The above question can also be formulated relative to a distribution framework Ψ (see Remark 7.7 on p. 116).

Theme 20 Let Ψ_1, Ψ_2 be classes of probability measures satisfying (6.14) and $\mathcal{M}$ be a structural model over N. May it happen that minimal unimarginal classes for $\mathcal{M}$ relative to Ψ_1 and Ψ_2 differ? More specifically, I am interested in the class of discrete measures (p. 13) in place of Ψ_1 and the class of non-degenerate Gaussian measures in place of Ψ_2 (p. 28).

The last two open problems are closely related to mathematical statistics. The first one is the 'parametrization problem'.

Direction 4 Find out for which structural imsets u over N and for which classes Ψ of probability measures with prescribed sample space $(\mathsf{X}_N, \mathcal{X}_N) = \prod_{i\in N}(\mathsf{X}_i, \mathcal{X}_i)$ a suitable parametrization of the class of Markovian measures $\Psi(u)$ with respect to u exists.

Note that I am interested in parametrization by means of 'independent' parameters i.e. situations in which elements of $\Psi(u)$ are in one-to-one correspondence with parameters belonging to a n-dimensional interval $[0,1]^n$ for some $n \in \mathbb{N}$. Preferable parametrizations are those in which parameters correspond to 'less-dimensional' marginal measures. Typical example is the parametrization of non-degenerate Gaussian measures which are Markovian with respect to an acyclic directed graph [7, 85].

Direction 5 Propose methods of learning structural models on basis of data (both statistical testing and estimation). Develop methods for statistical estimation of Markovian measures with respect to a given structural imset within a given distribution framework (i.e. fitting procedures).

I think that the most suitable methodological approach to statistical learning (of structural models) is to introduce suitable distance on the set of probability measures belonging to the considered distribution framework (with a fixed sample space). Then one can compute the distance of the empirical measure (computed from data) and the set of Markovian measures with respect to prospective structural imset. I hope that the equivalence result from Section 4.5 (see p. 68), namely the product formula (p. 61), may serve as a basis for some iterative fitting procedures.

Chapter 9

Conclusions

The aim of this chapter is to summarize the method(s) of description of probabilistic conditional independence (CI) structures mentioned in this work.

Chapter 3 is an overview of graphical methods of description of CI structures. These methods are suitable from the point of view of interpretation and some of them are good from the point of view of implementation (on a computer). However, they are not complete in sense that they are not able to describe all (discrete) probabilistic CI structures (see Section 3.6). Omission of this theoretical requirement may result in serious methodological errors in practical learning procedures (see Section 1.1). This fact motivated an effort to develop a non-graphical method of description of probabilistic CI structures by objects of discrete mathematics which complies with the requirement of completeness.

The method of description of probabilistic CI structures by means of *structural imsets* described in Chapter 4 and in subsequent chapters meets the above requirement of completeness for a quite wide class of distributions, namely for probability measures with finite multiinformation (Theorem 5.2). The class of measures with finite multiinformation involves three basic classes of distributions used in practice (in graphical modelling of CI structures), namely the class of discrete measures, the class of (non-degenerate) Gaussian measures and the class of (non-degenerate) CG measures (see Section 4.1).

Theorem 4.1 gives three equivalent definitions of a Markovian (probability) measure P with respect to a structural imset u. The *standard definition* requires that every CI statement 'represented' in u is indeed valid CI statement with respect to P (see Section 4.4.2). The second equivalent definition is the requirement that P satisfies the *product formula* induced by u (see Section 4.3). The product formula, which needs an auxiliary concept of a reference system of dominating measures, is perhaps important from the point of view of interpretation (of CI structures induced by structural imsets). First, it generalizes the well-known product formula for decomposable models (see (3.2) on p. 45); this happens when one takes in place of u the standard imset for the respective decomposable graph (see Section 7.2.2, Remark 7.4). Second, it can perhaps be viewed as a loose analogue of formulas defining log-linear models. Third, the product formula also illustrates how the *uniqueness principle* for Markovian measures with respect to a structural imset works. The principle, formulated in Consequence 4.3, says that the marginals of a Markovian measure P (with respect to a structural imset u) for the lower class $\mathcal{L}_u$ determine uniquely the marginals for the upper class $\mathcal{U}_u$. Indeed, in case of the standard imset induced by an acyclic directed graph (see Section 7.2) the upper class corresponds to the collection of all marginals (including the measure P itself) and the lower class describes the least possible collection of marginals determining (uniquelly) the measure P - see Consequence 7.5 in Section 7.4.3. The third equivalent definition of Markovian measure is the requirement that the scalar product of the *multiinformation function* m_P with u vanishes (see Section 4.5). This equivalent definition seems to be important from the point of view of computer implementation and maybe from the point of view of learning. Indeed, perhaps it can serve as a basis of a (future) learning method which can determine the most suitable structural imset on basis of a statistical estimate of the multiinformation function. However, the main significance is in bringing the point of view of algebra. This may facilitate computer implementation of the method on basis of arithmetic of integers.

The algebraic point of view is emphasized in Chapter 5. The multiinformation function is known to be an ℓ-standardized supermodular function (Consequence 2.2) and the cone $\mathcal{K}_\ell(N)$ of ℓ-standardized supermodular functions plays an important role in the presented approach. More precisely, $\mathcal{K}_\ell(N)$ is a pointed rational polyhedral cone which implies that it has finitely many extreme rays and every extreme ray contains just one non-zero normalized ℓ-standardized supermodular imset (Lemma 5.3) named *ℓ-skeletal imset*. Finite collection $\mathcal{K}_\ell^\diamond(N)$ of ℓ-skeletal imsets allows one to characterize dually structural imsets as o-standardized imsets with non-negative scalar products with ℓ-skeletal imsets (Theorem 5.1); in fact $\mathcal{K}_\ell^\diamond(N)$ is the least collection of normalized ℓ-standardized imsets of this sort.

Every supermodular function defines a certain formal CI structure (see Section 5.1.1) and an important fact is that the class of CI structures induced by structural imsets coincides with the class of CI structures which can be described by supermodular functions. The relation od these two different (but equivalent) methods of description of CI structures is characterized in Section 5.4 as a relation of duality. This is done with help of an algebraic concept of Galois connection interpreted in light of the theory of formal concept analysis [28]. The lattice of CI structures induced by structural imsets (or equivalently those which can be described by supemodular functions) is shown to be a finite concept lattice which is both atomistic and coatomistic (Theorem 5.3). Its atoms correspond to known elementary (structural) imsets while its coatoms correspond to ℓ-skeletal imsets.

Chapter 6 deals mainly with implementation of *facial implication* which corresponds to the inclusion of induced CI structures. Two characterization of facial implication $u \rightharpoonup v$ between structural imsets u, v are given. The first one (Lemma 6.1) characterizes it as a 'direct' arithmetical relation of u and v, namely that there exists $l \in \mathbb{N}$ such that $l \cdot u - v$ is a structural imset (resp. a combinatorial imset). The second one (Lemma 6.2) characterizes it with help of ℓ-skeletal imsets (respectively supermodular functions), namely by the requirement that u has non-zero scalar product with an ℓ-skeletal imset (a supermodular function) whenever v does so. The skeletal characterization then leads to a characterization of facially equivalent structural imsets (Consequence 6.2).

To transform the task of computer implementation of facial implication into a standard task of integer programming two observations are needed. The first observation (Consequence 6.4) is that in testing $u \rightharpoonup v$ for a structural imset u and an elementary imset v the number $l \in \mathbb{N}$ such that $l \cdot u - v$ is a structural imset is limited by a constant. The constant depends on the cardinality of the set of variables N only and it has the value 1 in case $|N| \leq 4$ and 7 in case $|N| = 5$. A good candidate for the the least constant of this type is the *grade* (see Section 6.3.2); the least constant of this kind for a combinatorial imset u and an elementary imset v is found in Lemma 6.5. The second observation is that in testing whether a given imset $\tilde{u}$ (e.g. $l \cdot u - v$ above) is a structural imset the coefficient $n \in \mathbb{N}$ such that $n \cdot \tilde{u}$ is a combinatorial imset (i.e. a sum of known elementary imsets) is also limited by a constant depending on the cardinality of N (Lemma 6.4). The constant is 1 in case $|N| \leq 4$; there is a hope that it is 1 in general (see Question 7).

Further results of Chapter 6 allow to adapt the described method of description of CI structures to a particular distribution framework, that is a class of probability measures over N satisfying certain basic conditions (see Section 6.2.3).

Chapter 7 concerns the problem of choice of a representative of a class of facially equivalent structural imsets. A good solution from the point of view of computer implementation seems to be the *baricentral imset* (Section 7.1). Indeed, facial implication between baricentral imsets is very simple: one has $u \rightharpoonup v$ for these imsets iff $u - v$ is a combinatorial imset (Observation 7.1). From the point of view of interpretation *standard imsets* for acyclic directed graphs (Section 7.2.1) and for triangulated undirected graphs (Section 7.2.2) seems to be suitable. First, they provide a simple translation of classic graphical models into the framework of structural imsets. Second, they offer an alternative (non-graphical) method of testing Markov equivalence for acyclic directed graphs (Consequence 7.1). Finally, they are exclusive for two theoretical reasons: standard imsets for acyclic directed graphs are combinatorial imsets of the least degree (Lemma 7.4) and imsets with the least lower class (Consequence 7.5).

Chapter 8 is an overview of open problems. The most important problems from the point of view of computer implementation of the method seem to be the question whether structural imsets indeed coincide with combinatorial imsets (Question 7) and the task to characterize ℓ-skeletal imsets (Theme 4). Note that ℓ-skeletal imsets are known in case $|N| \leq 5$ but not in general.

Well, the overall goal of the work was to present the method of description of probabilistic CI structures by means of structural imsets. This involves motivation, the description of present state and an outlook represented by the list of open problems. This non-graphical method removes inevitable limitation of graphical approaches and promises a chance of computer implementation by transforming the problem into standard tasks of integer programming. The work gives a theoretical solution; practical implemetation requires further research. However, in case of successful solving the task of computer implemetation the method can find wide application both in area of artificial intelligence and in area of multivariate statistics.

Chapter 10

Appendix

University graduates in mathematics should be familiar with the majority of the concepts and facts gathered in this chapter. However, certain misunderstanding can occur in their exact meaning and, moreover, graduates in other fields (e.g. computer science, statistics) may not be familiar with all basic facts. Thus, to avoid misunderstanding and to facilitate reading I decided to recall them here. Just to provide the reader with a reference source for well-known facts which can be easily utilized with help of the Index.

10.1 Classes of sets

By a *singleton* is understood a set containing one element only, the symbol $\emptyset$ is reserved for the *empty set*. The symbol $S \subseteq T$ (also $T \supseteq S$) denotes that S is a *subset* of T (alternatively T is a *superset* of S) which involves the situation $S = T$. However, *strict inclusion* is denoted as follows: $S \subset T$ or $T \supset S$ means that $S \subseteq T$ but $S \neq T$. The *power set* of a non-empty set X is the class of all its subsets $\{\,T;\, T \subseteq X\}$, denoted by $\mathcal{P}(X)$. The symbol $\bigcup \mathcal{D}$ denotes the *union* of a class $\mathcal{D} \subseteq \mathcal{P}(X)$; the symbol $\bigcap \mathcal{D}$ the *intersection* of a class $\mathcal{D} \subseteq \mathcal{P}(X)$. Supposing N is a non-empty finite set (of variables) a class $\mathcal{D} \subseteq \mathcal{P}(N)$ is called *ascending* if it is closed under supersets, i.e.

$$\forall\, S, T \subseteq N \quad S \in \mathcal{D},\ S \subseteq T \ \Rightarrow\ T \in \mathcal{D}\,.$$

Given $\mathcal{D} \subseteq \mathcal{P}(N)$, the *induced ascending class*, denoted by $\mathcal{D}^{\uparrow}$, is the least ascending class containing $\mathcal{D}$, i.e.

$$\mathcal{D}^{\uparrow} = \{\,T \subseteq N;\ \exists\, S \in \mathcal{D} \quad S \subseteq T\}.$$

Analogously, a class $\mathcal{D} \subseteq \mathcal{P}(N)$ is called *descending* if it is closed under subsets, i.e.

$$\forall\, S, T \subseteq N \quad S \in \mathcal{D},\ T \subseteq S \quad \Rightarrow\ T \in \mathcal{D}\,,$$

and given $\mathcal{D} \subseteq \mathcal{P}(N)$ the *induced descending class* $\mathcal{D}^{\downarrow}$ consists of subsets of sets in $\mathcal{D}$, i.e.

$$\mathcal{D}^{\downarrow} = \{\,T \subseteq N;\ \exists\, S \in \mathcal{D} \quad T \subseteq S\}.$$

A set $S \in \mathcal{D}$ where $\mathcal{D} \subseteq \mathcal{P}(N)$ is called a *maximal* set of $\mathcal{D}$ if $\forall\, T \in \mathcal{D} \quad S \subseteq T \ \Rightarrow\ S = T$; the class of maximal sets of $\mathcal{D}$ is denoted by $\mathcal{D}^{\max}$. Clearly, $\mathcal{D}^{\max} = (\mathcal{D}^{\downarrow})^{\max}$ and $\mathcal{D}^{\downarrow} = (\mathcal{D}^{\max})^{\downarrow}$. Dually, a set $S \in \mathcal{D}$ is called a *minimal* set of $\mathcal{D}$ if $\forall\, T \in \mathcal{D} \quad T \subseteq S \ \Rightarrow\ S = T$ and $\mathcal{D}^{\min}$ denotes the class of minimal sets of $\mathcal{D}$.

By a *permutation* of a finite non-empty set N will be understood a one-to-one mapping $\pi : N \to N$. It can be also viewed as a mapping on the power set $\mathcal{P}(N)$ which assigns $\pi(S) = \{\pi(x); x \in S\}$ to every $S \subseteq N$. Then given a real function $m : \mathcal{P}(N) \to \mathbb{R}$ the juxtaposition $m\pi$ will denote the *composition of m and π* defined by $S \mapsto m(\pi(S))$ for $S \subseteq N$.

10.2 Posets and lattices

Partially ordered set $(L,\preceq)$, shortly a *poset*, is a non-empty set L endowed with a *partial ordering* $\preceq$, that is, a binary relation on L which is

(i) reflexive: $\forall\, x \in L \quad x \preceq x\,,$
(ii) antisymmetric: $\forall\, x,y \in L \quad x \preceq y,\ y \preceq x \ \Rightarrow\ x = y\,,$
(iii) transitive: $\forall\, x,y,z \in L \quad x \preceq y,\ y \preceq z \ \Rightarrow\ x \preceq z\,.$

The phrase *total ordering* is used if moreover $\forall\, x,y \in L$ either $x \preceq y$ or $y \preceq x$. Given $x,y \in L$, one writes $x \prec y$ for $x \preceq y$ and $x \neq y$. If $x \prec y$ and there is no $z \in L$ such that $x \prec z$ and $z \prec y$ then x is called a *lower neighbour* of y and y is an *upper neighbour* of x. Given $M \subseteq L$ an element $x \in M$ is a *minimal element* of M with respect to $\preceq$ if there is no $z \in M$ with $z \prec x$, $y \in M$ is a *maximal element* of M with respect to $\preceq$ if there is no $z \in M$ with $z \succ x$.

Given $M \subseteq L$, the *supremum* of M in L, denoted by $\sup M$ and alternatively called the *least upper bound* of M is an element of $y \in L$ such that $z \preceq y$ for every $z \in M$ but $y \preceq y'$ for each $y' \in L$ with $z \preceq y'$ for every $z \in M$. Owing to antisymmetry of $\preceq$, the supremum of M is determined uniquely if it exists. Given $x,y \in L$, their *join* denoted by $x \vee y$ is the supremum of the set $\{x\} \cup \{y\}$. A poset in which every pair of elements has a join is called *join semi-lattice.*

Analogously, the *infimum* of $M \subseteq L$, denoted by $\inf M$, called also the *greatest lower bound* of M is an element of $x \in L$ such that $x \preceq z$ for every $z \in M$ but $x' \preceq x$ for each $x' \in L$ with $x' \preceq z$ for every $z \in M$. The *meet* of elements $x,y \in L$, denoted by $x \wedge y$ is the infimum of the set $\{x\} \cup \{y\}$. *Lattice* is a poset $(L,\preceq)$ such that for every $x,y \in L$ there exists both supremum $x \vee y$ and infimum $x \wedge y$ in L. A lattice is *distributive* if for every $x,y,z \in L$

$$x \wedge (y \vee z) = (x \wedge y) \vee (x \wedge z) \ \text{ and } \ x \vee (y \wedge z) = (x \vee y) \wedge (x \vee z)\,.$$

Typical example of a distributive lattice is a *ring of subsets* of a finite non-empty set N, that is a collection $\mathcal{R} \subseteq \mathcal{P}(N)$ which is closed under (finite) intersection and union. In particular, $\mathcal{P}(N)$ ordered by inclusion $\subseteq$ is a distributive lattice.

Complete lattice is a poset $(L,\preceq)$ such that every subset $M \subseteq L$ has the supremum and infimum in L. Note that it suffices to show that every $M \subseteq L$ has the infimum. Any finite lattice is an example of a complete lattice. By the *null element* of a complete lattice L is undestood the *least element* of L, that is $x_0 \in L$ such that $x_0 \preceq z$ for every $z \in L$; it is nothing but the supremum of the empty set in L. By the *unit element* is understood the *greatest elements* of L, that is $y_1 \in L$ such that $z \preceq y_1$ for every $z \in L$. An element x of a complete lattice is *join-irreducible* if $x \neq \sup\{z \in L\,;\, z \prec x\,\}$ and *meet-irreducible* if $x \neq \inf\{z \in L\,;\, x \prec z\,\}$. An element of a finite lattice is join-irreducible iff it has exactly one lower neighbour and it is meet-irreducible iff it has exactly one upper neighbour. The set of join-irreducible elements in a finite lattice $(L,\preceq)$ is the least set $M \subseteq L$ which is *supremum-dense* which means that for every $x \in L$ there exists $M' \subseteq M$ such that $x = \sup M'$. Analogously, the set of meet-irrreducible elements in L is the least set $M \subseteq L$ which is *infimum-dense*, i.e. for every $y \in L$ there exists $M' \subseteq M$ with $y = \inf M'$. Standard example of a join-irreducible element in a complete lattice is an *atom* of L which is an upper neighbour of the null element of L. By *coatom* of L is undestood a lower neighbour of the unit element of L. A complete lattice L is *atomistic* if the set of its atoms is supremum dense in L; equivalently if the only join-irreducible elements are atoms. It is *coatomistic* if the set of its coatoms is infimum dense in L, i.e. the only meet-irreducible elements are coatoms.

Two posets $(L_1,\preceq_1)$ and $(L_2,\preceq_2)$ are *order-isomorphic* if there exists a mapping $\phi : L_1 \mapsto L_2$ onto L_2 such that

$$x \preceq_1 y \ \Leftrightarrow\ \phi(x) \preceq_2 \phi(y) \quad \text{for every } \ x,y \in L_1\,.$$

The mapping ϕ is then a one-to-one mapping between L_1 and L_2 and it is called an *order-isomorphism*. If the poset $(L_1, \preceq_1)$ is a complete lattice then $(L_2, \preceq_2)$ is also a complete lattice and ϕ is even (complete) *lattice homomorphism* which means that

$$\phi(\sup M) = \sup\{\phi(z);\, z \in M\} \quad \phi(\inf M) = \inf\{\phi(z);\, z \in M\} \quad \text{for every } M \subseteq L_1\,.$$

General example of a complete lattice can be obtained by means of a *closure operation* on subsets of a set X, that is a mapping $cl : \mathcal{P}(X) \to \mathcal{P}(X)$ which is

(i) isotone: $\forall\, S, T \subseteq X \quad S \subseteq T \;\Rightarrow\; cl(S) \subseteq cl(T)$,
(ii) extensive: $\forall\, S \subseteq X \quad S \subseteq cl(S)$,
(iii) idempotent: $\forall\, S \subseteq X \quad cl(cl(S)) = cl(S)$.

A set $S \subseteq X$ is called *closed with respect to cl* if $S = cl(S)$. Given a closure operation cl on subsets of X the collection $\mathcal{K} \subseteq \mathcal{P}(X)$ of closed sets with respect to cl is closed under set intersection:

$$\mathcal{D} \subseteq \mathcal{K} \;\Rightarrow\; \bigcap \mathcal{D} \in \mathcal{K} \qquad (\text{by convention } \bigcap \mathcal{D} = X \text{ for } \mathcal{D} = \emptyset)\,.$$

Every collection $\mathcal{K} \subseteq \mathcal{P}(X)$ satisfying this requirement is called a *closure system* of subsets of X. The correspondence between cl and $\mathcal{K}$ is one-to-one since the formula

$$cl_{\mathcal{K}}(S) = \bigcap \{T \subseteq X\,;\, S \subseteq T \in \mathcal{K}\} \quad \text{for } S \subseteq X\,,$$

defines a closure operation on subsets of X having $\mathcal{K}$ as the collection of closed sets with respect to $cl_{\mathcal{K}}$ (see Theorem 1 in [28]). The poset $(\mathcal{K}, \subseteq)$ is then a complete lattice in which

$$\sup \mathcal{D} = cl\,(\bigcup \mathcal{D}\,) \quad \inf \mathcal{D} = \bigcap \mathcal{D} \qquad \text{for every } \mathcal{D} \subseteq \mathcal{K}\,.$$

Every complete lattice is order-isomorphic to a lattice of this type - see Proposition 3 in Chapter 1 of [28].

10.3 Graphs

A (classic) *graph* is specified by a non-empty finite set of *nodes* N and by a set of *edges* consisting of pairs of elements taken from N. Several types of edges are mentioned in this work, but classic graphs admit only two basic types of edges. An *undirected edge* or a *line* over N is an unordered pair $\{a, b\}$ where $a, b \in N$, $a \neq b$ (that is a two-element subset of N). A *directed edge* or an *arrow* over N is an ordered pair (a, b) where $a, b \in N$, $a \neq b$. Pictorial representation is clear: nodes are represented by small circles and edges by corresponding links beween them. Note that explicit requirement $a \neq b$ excludes any *loop*, that is an edge connecting a node with itself (loops are possible in some non-classic graphs).

A (classic) *graph with mixed edges* over (a set of nodes) N is given by a set of lines $\mathcal{L}$ over N and by a set of arrows $\mathcal{A}$ over N. Supposing $G = (N, \mathcal{L}, \mathcal{A})$ is a graph of this kind one writes '$a — b$ in G' in case $\{a, b\} \in \mathcal{L}$ and says that there exists a line between a and b in G. Similarly, in case $(a, b) \in \mathcal{A}$ one say that there exists an arrow from a to b in G and writes '$a \to b$ in G' or '$b \leftarrow a$ in G'. Pictorial representation naturally reflects notation in both cases.

If either $a — b$ in G, $a \to b$ in G or $a \leftarrow b$ in G, then one simply says that $[a, b]$ is an *edge* in G. Note explicitly that this definition allows (for a pair of distinct nodes $a, b \in N$) that each of $a — b$, $a \to b$ and $a \leftarrow b$ are simultaneously edges in G! If $\emptyset \neq T \subset N$, then the *induced subgraph* of G for T is the graph $G_T = (T, \mathcal{L}_T, \mathcal{A}_T)$ over T where $\mathcal{L}_T$ ($\mathcal{A}_T$) is the set of those lines (arrows) over T which are also in $\mathcal{L}$ (in $\mathcal{A}$). A *hybrid graph* over N is a graph with mixed edges

G without *multiple edges*. That means, for an ordered pair of distinct nodes (a,b), $a,b \in N$ at most one of three above mentioned options can occur.

A *route* from a node a to a node b (or between nodes a and b) in a graph G with mixed edges is a sequence of nodes $c_1,\ldots,c_n \in N$, $n \geq 1$ together with a sequence of edges $\epsilon_1,\ldots,\epsilon_{n-1} \in \mathcal{L} \cup \mathcal{A}$ (possibly empty in case $n=1$) such that $a = c_1$, $b = c_n$ and ϵ_i is either $c_i - c_{i+1}$, $c_i \to c_{i+1}$ or $c_i \leftarrow c_{i+1}$ for $i = 1,\ldots,n-1$. A route is called *undirected* if ϵ_i is $c_i - c_{i+1}$ for $i = 1,\ldots,n-1$, *descending* if ϵ_i is either $c_i - c_{i+1}$ or $c_i \to c_{i+1}$ for $i = 1,\ldots,n-1$ and *strictly descending* if $n \geq 2$ and ϵ_i is $c_i \to c_{i+1}$ for $i = 1,\ldots,n-1$. In particular, every undirected route is a descending route. A *path* is a route in which all nodes $c_1,\ldots,c_n$ are distinct, a *cycle* is a route where $n \geq 3$, $c_1 = c_n$ and $c_1,\ldots,c_{n-1}$ are distinct (and ϵ_2 is not a reverse copy of ϵ_1 in case $n = 3$). A *directed cycle* is a cycle which is a descending route and where ϵ_i is $c_i \to c_{i+1}$ at least once. The adjectives undirected and (strictly) directed are used for paths as well.

A node a is a *parent* of a node b in G or b is a *child* of a if $a \to b$ in G; a is an *ancestor* of b in G, dually b is a *descendant* of a if there exists a descending route (equivalently a descending path) from a to b in G. The set of parents of a node b in G will be denoted by $pa_G(b)$. Supposing $A \subseteq N$ the symbol $an_G(A)$ will denote the set of ancestors of the nodes of A in G. Analogously, a is a *strict ancestor* (b is a *strict descendant* of a) if there exist a strictly descending route from a to b. Similarly, a is *connected* to b in G if there exists an undirected route (equivalently an undirected path) between a and b. Clearly, the relation 'be connected' is an equivalence relation which decomposes N into equivalence classes, named *connectivity components*.

An *undirected graph* is a graph containing only lines (that is $\mathcal{A} = \emptyset$), a *directed graph* is a graph containing only arrows (that is $\mathcal{L} = \emptyset$). The *underlying graph* H of a graph with mixed edges $G = (N,\mathcal{L},\mathcal{A})$ is an undirected graph H over N such that $a - b$ in H iff $[a,b]$ is an edge in G. A set $A \subseteq N$ in an undirected graph H over N is *complete* if $a - b$ for every $a,b \in A$, $a \neq b$; a *clique* of H is a maximal complete subset of N.

A *acyclic directed graph* over N is a directed graph over N without directed cycles. It can be equivalently introduced as a directed graph G whose nodes can be ordered in a sequence $a_1,\ldots,a_k$, $k \geq 1$ such that if $[a_i,a_j]$ is an edge in G for $i < j$, then $a_i \to a_j$ in G. A *chain* for a hybrid graph G over N is a partition of N into ordered disjoint (non-empty) subsets $B_1,\ldots,B_n$, $n \geq 1$ called *blocks* such that, if $[a,b]$ is an edge in G with $a,b \in B_i$ then $a - b$, and if $[a,b]$ is an edge in G with $a \in B_i, b \in B_j, i < j$ then $a \to b$. A *chain graph* is a hybrid graph which admits a chain. It can be equivalently introduced as a hybrid graph without directed cycles (see [110] Lemma 2.1). Evidently, every undirected or acyclic directed graph is a chain graph.

Note that various other types of edges are used in advanced graphical approaches (see Section 3.5), e.g. bidirected edges, dashed lines, dashed arrows or even loops. From purely mathematical point of view these edges can be also introduced as either ordered or unordered pairs of nodes, but their meaning is different. Thus, because of different interpretation they have to be carefully distinguished from the above mentioned 'classic' edges. However, all the concepts introduced in Section 10.3 can be naturally extended to the graphs allowing edges of additional types.

10.4 Topological concepts

Metric space (X,ρ) is a non-empty set X endowed with a *distance* ρ which is a non-negative real function $\rho : \mathsf{X} \times \mathsf{X} \to [0,\infty)$, such that $\forall\, x,y,z \in \mathsf{X}$ one has

(i) $\rho(x,y) = 0$ iff $x = y$,

(ii) $\rho(x,y) = \rho(y,x)$,

(iii) $\rho(x,z) \leq \rho(x,y) + \rho(y,z)$.

A set $\mathsf{G} \subseteq \mathsf{X}$ is called *open* in (X, ρ) if for every $x \in \mathsf{G}$ there exists $\varepsilon > 0$ such that the *open ball* $U(x, \varepsilon) \equiv \{y \in \mathsf{X};\ \rho(x, y) < \varepsilon\}$ with center x and radius ε belongs to G. A set $\mathsf{F} \subseteq \mathsf{X}$ is *closed* if its complement $\mathsf{X} \setminus \mathsf{G}$ is open. A metric space is *separable* if it has a countable *dense set*, that is such a set $\mathsf{S} \subseteq \mathsf{X}$ that $\forall x \in \mathsf{X}\ \forall \varepsilon > 0$ there exists $y \in \mathsf{S} \cap U(x, \varepsilon)$. A metric space is *complete* if every Cauchy sequence $x_1, x_2, \ldots$ of elements of X, i.e. a sequence satisfying $\forall \varepsilon > 0\ \ \exists n \in \mathbb{N}$ such that $\forall k, l \geq n \quad \rho(x_k, x_l) < \varepsilon$, converges to an element $x \in \mathsf{X}$, i.e. $\forall \varepsilon > 0 \quad \exists n \in \mathbb{N}$ such that $\forall k \geq n \quad \rho(x_k, x) < \varepsilon$.

Classic example of a separable complete metric space is an arbitrary non-empty finite set X endowed with the *discrete distance* δ defined as follows:

$$\delta(x, y) = \begin{cases} 0 & \text{if } x = y\,, \\ 1 & \text{otherwise}\,. \end{cases}$$

Another common example is the set of n-dimensional real vectors $\mathbb{R}^n$, $n \geq 1$ endowed with the *Eucledian distance*

$$\varrho(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \qquad \text{for } \boldsymbol{x} = [x_1, \ldots, x_n],\ \boldsymbol{y} = [y_1, \ldots y_n]\,.$$

The set of real numbers $\mathbb{R}$ with $\varrho(x, y) = |x - y|$ is a special case.

Topological space (X, τ) is a non-empty set X endowed with a *topology* τ which is a class of subsets of X closed under finite intersection, arbitrary union, and involving both the empty set $\emptyset$ and X itself. Every metric space (X, ρ) is an example of a topological space because the class of open sets in (X, ρ) is a topology. A topological space of this kind is called *metrizable* and its topology is *induced by the distance* ρ. For instance, the set of real numbers $\mathbb{R}$ is often automatically understood as a topological space endowed with *Eucledian topology* induced by Eucledian distance. The *product of topological spaces* (X_1, τ_1) and (X_2, τ_2) is the Cartesian product $\mathsf{X}_1 \times \mathsf{X}_2$ endowed with the *product topology*, that is the class of sets $\mathsf{G} \subseteq \mathsf{X}_1 \times \mathsf{X}_2$ such that $\forall (x_1, x_2) \in \mathsf{G}$ there exist $\mathsf{G}_1 \in \tau_1$, $\mathsf{G}_2 \in \tau_2$ with $(x_1, x_2) \in \mathsf{G}_1 \times \mathsf{G}_2 \subseteq \mathsf{G}$. The product $\prod_{i \in N} (X_i, \tau_i)$ of any finite collection (X_i, τ_i), $i \in N$, $|N| \geq 2$ of topological spaces is defined analogously. For example, $\mathbb{R}^n$ $(n \geq 2)$ endowed with the topology induced by Eucledian distance can be viewed as the product of topological spaces $\mathsf{X}_i = \mathbb{R}$, $i \in \{1, \ldots, n\}$.

A real function $f : \mathsf{X} \to \mathbb{R}$ on a topological space (X, τ) is *continuous* if $\{x \in \mathsf{X}; f(x) < r\}$ belongs to τ for every $r \in \mathbb{R}$.

10.5 Measure-theoretical concepts

Measurable space $(\mathsf{X}, \mathcal{X})$ is a non-empty set X endowed with a *σ-algebra* $\mathcal{X}$ over X which is a class of subsets of X involving X itself and closed under countable union and complement. Given a class $\mathcal{A}$ of subsets of X, the least σ-algebra over X containing $\mathcal{A}$ (i.e. the intersection of all σ-algebras containing $\mathcal{A}$) is called the *σ-algebra generated by* $\mathcal{A}$ and denoted by $\sigma(\mathcal{A})$. In particular, if (X, τ) is a topological space, then the σ-algebra generated by its topology is the *Borel σ-algebra* or the σ-algebra of *Borel sets*. *Trivial* σ-algebra over X is the class $\{\emptyset, \mathsf{X}\}$, that is the σ-algebra generated by an empty class. Given a measurable space $(\mathsf{X}, \mathcal{X})$ the class of all σ-algebras $\mathcal{S} \subseteq \mathcal{X}$, ordered by inclusion, is a lattice. Indeed, for σ-algebras $\mathcal{S}, \mathcal{T} \subseteq \mathcal{X}$, their *supremum* $\mathcal{S} \vee \mathcal{T}$ is the σ-algebra generated by $\mathcal{S} \cup \mathcal{T}$, while their *infimum* $\mathcal{S} \wedge \mathcal{T}$ is simply the intersection $\mathcal{S} \cap \mathcal{T}$. The *product of measurable spaces* $(\mathsf{X}_1, \mathcal{X}_1)$ and $(\mathsf{X}_2, \mathcal{X}_2)$ is the Cartesian product $\mathsf{X}_1 \times \mathsf{X}_2$ endowed with the *product σ-algebra* $\mathcal{X}_1 \times \mathcal{X}_2$ which is generated by *measurable rectangles*, that is the sets of the form $\mathsf{A} \times \mathsf{B}$ where $\mathsf{A} \in \mathcal{X}_1$ and $\mathsf{B} \in \mathcal{X}_2$. The product $(\prod_{i \in N} \mathsf{X}_i, \prod_{i \in N} \mathcal{X}_i)$ of arbitrary finite collection of measurable spaces $(\mathsf{X}_i, \mathcal{X}_i)$, $i \in N$ where $|N| \geq 2$ is defined analogously.

A real function $f : \mathsf{X} \to \mathbb{R}$ on a measurable space $(\mathsf{X}, \mathcal{X})$ is *measurable* (sometimes one writes $\mathcal{X}$-measurable) if $\{x \in \mathsf{X}\,; f(x) < r\}$ belongs to $\mathcal{X}$ for every $r \in \mathbb{R}$. Typical example is the *indicator* χ_{A} of a set $\mathsf{A} \in \mathcal{X}$ defined as follows:

$$\chi_{\mathsf{A}}(x) = \begin{cases} 1 & \text{if} \quad x \in \mathsf{A}\,, \\ 0 & \text{if} \quad x \in \mathsf{X} \setminus \mathsf{A}\,. \end{cases}$$

Given a real measurable function $f : \mathsf{X} \to \mathbb{R}$, its *positive part* f^+ and *negative part* f^- are non-negative measurable functions defined by

$$f^+(x) = \max\{f(x), 0\}\,, \quad f^-(x) = \max\{-f(x), 0\} \quad \text{for } x \in X\,,$$

and one has $f = f^+ - f^-$ and $|f| = f^+ + f^-$.

Non-negative *measure* on a measurable space $(\mathsf{X}, \mathcal{X})$ is a function μ defined on $\mathcal{X}$, taking values in the interval $[0, \infty]$ (infinite values are allowed) which satisfies $\mu(\emptyset) = 0$ and is *countably additive*, that is the equality

$$\mu(\bigcup_{i=1}^{\infty} \mathsf{A}_i) = \sum_{i=1}^{\infty} \mu(\mathsf{A}_i)$$

holds for every countable collection of pairwise disjoint sets $\mathsf{A}_1, \mathsf{A}_2, \ldots$ in $\mathcal{X}$. It is a *finite measure* if $\mu(\mathsf{X}) < \infty$ and a *σ-finite measure* if there exists a sequence $\mathsf{B}_1, \mathsf{B}_2, \ldots$ of sets in $\mathcal{X}$ such that $\mathsf{X} = \bigcup_{i=1}^{\infty} \mathsf{B}_i$ and $\mu(\mathsf{B}_i) < \infty$ for every $i \in \mathbb{N}$. A trivial example of a finite measure is a non-empty finite set X endowed with the *counting measure* υ on $(\mathsf{X}, \mathcal{P}(\mathsf{X}))$ defined by $\upsilon(\mathsf{A}) = |\mathsf{A}|$ for every $\mathsf{A} \subseteq \mathsf{X}$. Classic example of a σ-finite measure is *Lebesgue measure* on $\mathbb{R}^n$, $n \geq 1$, endowed with the Borel σ-algebra $\mathcal{B}^n$. This measure can be introduced as the only non-negative measure λ on $(\mathbb{R}^n, \mathcal{B}^n)$ ascribing to every n-dimensional interval its volume, that is

$$\lambda\,(\prod_{i=1}^{n}[r_i, s_i)\,) = \prod_{i=1}^{n}(s_i - r_i) \qquad \text{whenever } r_i, s_i \in \mathbb{R},\ r_i < s_i,\ i = 1, \ldots, n\,.$$

Probability measure is a measure μ satisfying $\mu(\mathsf{X}) = 1$. It is *concentrated* on a set $\mathsf{B} \in \mathcal{X}$ if $\mu(\mathsf{B}) = 1$ or equivalently $\mu(\mathsf{X} \setminus \mathsf{B}) = 0$. Two real measurable functions f and g on $(\mathsf{X}, \mathcal{X})$ *equal μ-almost everywhere* if $\mu(\{x \in \mathsf{X},\ f(x) \neq g(x)\}) = 0$. Then one writes $f = g \ \ \mu$-a.e. Clearly, it is an equivalence relation.

The concept of integral is understood in sense of Lebesgue. Given a non-negative measure μ on $(\mathsf{X}, \mathcal{X})$ this construction (described for example in [87] Chapter 1) assigns a value $\int_{\mathsf{A}} f(x)\, d\mu(x)$ from $[0, \infty]$, called the *integral of f through A with respect to μ* to every non-negative measurable function f and arbitrary $\mathsf{A} \in \mathcal{X}$ (f can be defined on A only). A real measurable function f on $(\mathsf{X}, \mathcal{X})$ is called *μ-integrable* if the integral of its absolute value $\int_{\mathsf{X}} |f(x)|\, d\mu(x)$ is finite. Finite integral $\int_{\mathsf{A}} f(x)\, d\mu(x)$ (i.e. a real number) is then defined for every μ-integrable function f and $\mathsf{A} \in \mathcal{X}$. Note that supposing f is μ-integrable and g is $\mathcal{X}$-measurable function on X one has $f = g$ μ-a.e. iff $\int_{\mathsf{A}} f(x)\, d\mu(x) = \int_{\mathsf{A}} g(x)\, d\mu(x)$ for every $\mathsf{A} \in \mathcal{X}$ (and g is μ-integrable in both cases). Note that in case $(\mathsf{X}, \mathcal{X}) = (\prod_{i \in N} \mathsf{X}_i, \prod_{i \in N} \mathcal{X}_i)$ it is equivalent to apparently weaker requirement that the equality of integrals holds for every measurable rectangle A only. This follows from the fact that every two finite measures on $(\prod_{i \in N} \mathsf{X}_i, \prod_{i \in N} \mathcal{X}_i)$ which equal on measurable rectangles must coincide.

Sometimes, one needs to introduce (possibly infinite) integral even for a non-integrable real measurable function $f : \mathsf{X} \to \mathbb{R}$ by the formula

$$\int_{\mathsf{X}} f(x)\, d\mu(x) = \int_{\mathsf{X}} f^+(x)\, d\mu(x) - \int_{\mathsf{X}} f^-(x)\, d\mu(x)$$

provided that at least one of the integrals on the right-hand side is finite. Then one says that f is *μ-quasi-integrable* and the integral $\int_{\mathsf{X}} f(x)\, d\mu(x)$ is defined as a value in the interval $[-\infty,+\infty]$. Let us refer for elementary properties of Lebesgue integral to [87], Chapter 1.

Supposing μ and ν are measures on $(\mathsf{X},\mathcal{X})$ one says that ν is *absolutely continuous with respect to* μ and writes $\nu \ll \mu$ if $\mu(\mathsf{A}) = 0$ implies $\nu(\mathsf{A}) = 0$ for every $\mathsf{A} \in \mathcal{X}$. Basic measure-theoretical result is *Radon-Nikodym theorem* (see [87], Sections 6.9 and 6.10).

Theorem Supposing ν is a finite measure and μ a σ-finite measure on $(\mathsf{X},\mathcal{X})$ such that $\nu \ll \mu$ there exists a non-negative μ-integrable function f called *Radon-Nikodym derivative of ν with respect to μ* such that

$$\nu(\mathsf{A}) = \int_{\mathsf{A}} f(x)\; d\mu(x) \qquad \text{for every } \mathsf{A} \in \mathcal{X}\,.$$

Moreover, one can show (using Theorem 1.29 in [87]) that, for every $\mathcal{X}$-measurable function g on X, g is ν-integrable iff $g \cdot f$ is μ-integrable and

$$\int_{\mathsf{A}} g(x)\, d\nu(x) = \int_{\mathsf{A}} g(x) \cdot f(x)\; d\mu(x) \qquad \text{for every } \mathsf{A} \in \mathcal{X}\,.$$

According to the remark above, Radon-Nikodym derivative is determined uniquely only within equivalence μ-a.e. One writes $f = \frac{d\nu}{d\mu}$ to denote that a non-negative $\mathcal{X}$-measurable function f is (a version of) Radon-Nikodym derivative of ν with respect to μ.

Product of σ-finite measures μ_1 on $(\mathsf{X}_1,\mathcal{X}_1)$ and μ_2 on $(\mathsf{X}_2,\mathcal{X}_2)$ is the unique measure $\mu_1 \times \mu_2$ on $(\mathsf{X}_1 \times \mathsf{X}_2, \mathcal{X}_1 \times \mathcal{X}_2)$ defined on measurable rectangles as follows:

$$(\mu_1 \times \mu_2)\,(\mathsf{A} \times \mathsf{B}) = \mu_1(\mathsf{A}) \cdot \mu_2(\mathsf{B}) \quad \text{whenever } \mathsf{A} \in \mathcal{X}_1,\ \mathsf{B} \in \mathcal{X}_2\,.$$

Let us refer to [87] (Chapter 7, Example 7 and Sections 7.6, 7.7) for the proof of existence and uniqueness of (necessarily σ-finite) product measure $\mu_1 \times \mu_2$. Product of finitely many σ-finite measures $\prod_{i\in N} \mu_i$, $|N| \geq 2$ can be introduced analogously. Another basic measure-theoretical result is *Fubini theorem* (see [87], Section 7.8).

Theorem Let μ_1 be a σ-finite measure on $(\mathsf{X}_1,\mathcal{X}_1)$ and μ_2 a σ-finite measure on $(\mathsf{X}_2,\mathcal{X}_2)$. Suppose that f is a non-negative $\mathcal{X}_1 \times \mathcal{X}_2$-measurable function on $\mathsf{X}_1 \times \mathsf{X}_2$. Then the function $x_1 \mapsto \int_{\mathsf{X}_2} f(x_1,x_2)\, d\mu_2(x_2)$ is $\mathcal{X}_1$-measurable, the function $x_2 \mapsto \int_{\mathsf{X}_1} f(x_1,x_2)\, d\mu_1(x_1)$ is $\mathcal{X}_2$-measurable and one has

$$\int_{\mathsf{X}_1 \times \mathsf{X}_2} f(x_1,x_2)\; d(\mu_1 \times \mu_2)([x_1,x_2]) =$$

$$= \int_{\mathsf{X}_1} \int_{\mathsf{X}_2} f(x_1,x_2)\; d\mu_2(x_2)\, d\mu_1(x_1) = \int_{\mathsf{X}_2} \int_{\mathsf{X}_1} f(x_1,x_2)\; d\mu_1(x_1)\, d\mu_2(x_2)\,.$$

Whenever f is $\mu_1 \times \mu_2$-integrable real function on $\mathsf{X}_1 \times \mathsf{X}_2$ the same conclusion holds with the proviso that respective functions on X_i are defined μ_i-almost everywhere ($i = 1, 2$).

By a *measurable mapping* of a measurable space $(\mathsf{X},\mathcal{X})$ into a measurable space $(\mathsf{Y},\mathcal{Y})$ is understood a mapping $t : \mathsf{X} \to \mathsf{Y}$ such that for every $\mathsf{B} \in \mathcal{Y}$ the set $t_{-1}(\mathsf{B}) \equiv \{x \in \mathsf{X};\, t(x) \in \mathsf{B}\}$ belongs to $\mathcal{X}$. Note that a measurable function is a special case when Y is $\mathbb{R}$ endowed with the Borel σ-algebra. Every probability measure P on $(\mathsf{X},\mathcal{X})$ then *induces through* t a probability measure Q on $(\mathsf{Y},\mathcal{Y})$ defined by $Q(\mathsf{B}) = P(t_{-1}(\mathsf{B}))$ for every $\mathsf{B} \in \mathcal{Y}$.

Two measurable spaces $(\mathsf{X},\mathcal{X})$ and $(\mathsf{Y},\mathcal{Y})$ are *isomorphic* if there exists a one-to-one mapping $\varsigma:\mathcal{X}\to\mathcal{Y}$ which is onto $\mathcal{Y}$ and preserves countable union and complement. Then $\varsigma(\emptyset)=\emptyset$, $\varsigma(\mathsf{X})=\mathsf{Y}$ and countable intersection and inclusion are also preserved. The inverse mapping preserves these operations as well, and every measure μ on $(\mathsf{X},\mathcal{X})$ corresponds to a measure ν on $(\mathsf{Y},\mathcal{Y})$ defined by

$$\nu(\mathsf{B})=\mu(\varsigma_{-1}(\mathsf{B}))\quad\text{for }\ \mathsf{B}\in\mathcal{Y}\,,$$

and conversely. For example, given measurable spaces $(\mathsf{X}_1,\mathcal{X}_1)$ and $(\mathsf{X}_2,\mathcal{X}_2)$, the space $(\mathsf{X}_1,\mathcal{X}_1)$ is isomorphic to $(\mathsf{X}_1\times\mathsf{X}_2,\bar{\mathcal{X}}_1)$ endowed with the σ-algebra

$$\bar{\mathcal{X}}_1\equiv\{\mathsf{A}\times\mathsf{X}_2\,;\,\mathsf{A}\in\mathcal{X}_1\}\subseteq\mathcal{X}_1\times\mathcal{X}_2\,.$$

Supposing P is a probability measure on a measurable space $(\mathsf{X},\mathcal{X})$ and $\mathcal{A}\subseteq\mathcal{X}$ is a σ-algebra over X, the *restriction of P to $\mathcal{A}$* will be denoted by $P^{\mathcal{A}}$. Given $\mathsf{B}\in\mathcal{X}$, the *conditional probability of* B *given* $\mathcal{A}$ *with respect to* P is an $\mathcal{A}$-measurable function $h:\mathsf{X}\to[0,1]$ such that

$$P(\mathsf{A}\cap\mathsf{B})=\int_{\mathsf{A}} h(x)\;dP(x)\quad\text{for every }\ \mathsf{A}\in\mathcal{A}\,. \tag{10.1}$$

One can use Radon-Nikodym theorem with $(\mathsf{X},\mathcal{A})$, $\mu=P^{\mathcal{A}}$, and $\nu(\mathsf{A})=P(\mathsf{A}\cap\mathsf{B})$ for $\mathsf{A}\in\mathcal{A}$, to show that a function h satisfying (10.1) exists and is determined uniquely within equivalence $P^{\mathcal{A}}$-a.e. Let us write $h=P(\mathsf{B}|\mathcal{A})$ to denote that a $\mathcal{A}$-measurable function $h:\mathsf{X}\to[0,1]$ is (a version of) conditional probability of B given $\mathcal{A}$. Let us mention (without proof) two equivalent definitions of conditional probability. The first one, apparently weaker, says that $h=P(\mathsf{B}|\mathcal{A})$ iff the equality (10.1) holds for every $\mathsf{A}\in\mathcal{G}$, where $\mathcal{G}\subseteq\mathcal{A}$ is a class closed under finite intersection such that $\sigma(\mathcal{G})=\mathcal{A}$. The second one, apparently stronger, says that $h=P(\mathsf{B}|\mathcal{A})$ iff for every non-negative $\mathcal{A}$-measurable function $g:\mathsf{X}\to\mathbb{R}$ and $\mathsf{A}\in\mathcal{A}$ one has

$$\int_{\mathsf{A}\cap\mathsf{B}} g(x)\;dP(x)=\int_{\mathsf{A}} g(x)\cdot h(x)\;dP(x)\equiv\int_{\mathsf{A}} g(x)\cdot h(x)\;dP^{\mathcal{A}}(x)\,.$$

It follows from the definition of conditional probability that whenever $\mathcal{S}\subseteq\mathcal{T}\subseteq\mathcal{X}$ are σ-algebras, and $\mathsf{B}\in\mathcal{X}$ then every $\mathcal{S}$-measurable version of $P(\mathsf{B}|\mathcal{T})$ is a version of $P(\mathsf{B}|\mathcal{S})$. Sometimes, it happens that a certain fact or the value of an expression does not depend on the choice of a version of conditional probability. In this case the symbol $P(\mathsf{B}|\mathcal{A})$ is used in the corresponding formula to substitute arbitrary version of conditional probability of B given $\mathcal{A}$ (w.r.t. P).

Remark Having fixed just P on $(\mathsf{X},\mathcal{X})$ and a σ-algebra $\mathcal{A}\subseteq\mathcal{X}$ by a *regular version of conditional probability given* $\mathcal{A}$ is understood a function which ascribes to every $\mathsf{B}\in\mathcal{X}$ a version of $P(\mathsf{B}|\mathcal{A})$ such that, for every $x\in\mathsf{X}$, the mapping $\mathsf{B}\mapsto P(\mathsf{B}|\mathcal{A})(x)$ is a probability measure on $(\mathsf{X},\mathcal{X})$. Note that this concept is taken from [55], §26.1 and that a regular version of conditional probability may not exist in general (e.g. Example VI.1.35 in [98]). However, under certain topological assumptions, namely that X is a separable complete metric space and $\mathcal{X}$ is the class of Borel sets in X, its existence is guaranteed (see either [98], Theorem VI.1.21 or [74], Consequence of Theorem V.4.4). △

Supposing μ is a measure on the product of measurable spaces $(\mathsf{X}_1\times\mathsf{X}_2,\mathcal{X}_1\times\mathcal{X}_2)$ the *marginal measure* ν on $(\mathsf{X}_1,\mathcal{X}_1)$ is defined as follows:

$$\nu(\mathsf{A})=\mu(\mathsf{A}\times\mathsf{X}_2)\quad\text{for every }\ \mathsf{A}\in\mathcal{X}_1\,.$$

Every $\mathcal{X}_1$-measurable function h on X_1 can be viewed as $\mathcal{X}_1\times\mathcal{X}_2$-measurable function on $\mathsf{X}_1\times\mathsf{X}_2$. Then h is ν-integrable iff it is μ-integrable and

$$\int_{\mathsf{X}_1\times\mathsf{X}_2} h(x_1)\;d\mu([x_1,x_2])=\int_{\mathsf{X}_1} h(x_1)\;d\nu(x_1)\,.$$

A real function $\varphi : [0,\infty) \to \mathbb{R}$ is called *convex* if for all $r, s \in [0,\infty)$ and $\alpha \in [0,1]$

$$\varphi\left(\alpha \cdot r + (1-\alpha)\cdot s\right) \le \alpha \cdot \varphi(r) + (1-\alpha)\cdot \varphi(s)\,.$$

It is called *strictly convex* if this inequality is strict whenever $r \neq s$ and $\alpha \in (0,1)$. Further basic result is *Jensen's inequality* (one can modify the proof from [87], Section 3.3).

Theorem Let μ be a probability measure on $(\mathsf{X}, \mathcal{X})$, $f : \mathsf{X} \to [0,\infty)$ a μ-integrable function and $\varphi : [0,\infty) \to \mathbb{R}$ a convex function. Then

$$\varphi\Big(\int_{\mathsf{X}} f(x)\, d\mu(x)\Big) \le \int_{\mathsf{X}} \varphi\left(f(x)\right) d\mu(x).$$

In case φ is strictly convex the equality occurs if and only if f is constant μ-a.e., more exactly $f(x) = k$ for μ-a.e. $x \in \mathsf{X}$ where $k = \int_{\mathsf{X}} f(x)\, d\mu(x)$.

10.6 Conditional independence of σ-algebras

Let $\mathcal{A}, \mathcal{B}, \mathcal{C} \subseteq \mathcal{X}$ are σ-algebras in a measurable space $(\mathsf{X}, \mathcal{X})$ and P is a probability measure on it. One can say that $\mathcal{A}$ is *conditionally independent of* $\mathcal{B}$ *given* $\mathcal{C}$ *with respect to* P and write $\mathcal{A} \perp\!\!\!\perp \mathcal{B} \,|\, \mathcal{C}\ [P]$ if, for every $\mathsf{A} \in \mathcal{A}$ and $\mathsf{B} \in \mathcal{B}$, one has

$$P(\mathsf{A} \cap \mathsf{B}|\mathcal{C})(x) = P(\mathsf{A}|\mathcal{C})(x) \cdot P(\mathsf{B}|\mathcal{C})(x) \qquad \text{for } P^{\mathcal{C}}\text{-a.e. } x \in \mathsf{X}\,. \tag{10.2}$$

Note (without proof) that an apparently weaker equivalent formulation is as follows. It suffices to verify (10.2) only for $\mathsf{A} \in \tilde{\mathcal{A}}$ and $\mathsf{B} \in \tilde{\mathcal{B}}$ where $\tilde{\mathcal{A}} \subseteq \mathcal{A}$ respectively $\tilde{\mathcal{B}} \subseteq \mathcal{B}$ are classes closed under finite intersection such that $\sigma(\tilde{\mathcal{A}}) = \mathcal{A}$ respectively $\sigma(\tilde{\mathcal{B}}) = \mathcal{B}$.

LEMMA 10.1 Under the assumptions above $\mathcal{A} \perp\!\!\!\perp \mathcal{B} \,|\, \mathcal{C}\ [P]$ occurs iff for every $\mathsf{A} \in \mathcal{A}$ there exists a $\mathcal{C}$-measurable version of $P(\mathsf{A}|\mathcal{B} \vee \mathcal{C})$.

Proof: To show the necessity of the condition fix $\mathsf{A} \in \mathcal{A}$ and choose a version f of $P(\mathsf{A}|\mathcal{C})$. Write for every $\mathsf{B} \in \mathcal{B}, \mathsf{C} \in \mathcal{C}$ by definition of $P(\mathsf{A} \cap \mathsf{B}|\mathcal{C})$ and (10.2)

$$P(\mathsf{A} \cap \mathsf{B} \cap \mathsf{C}) = \int_{\mathsf{C}} P(\mathsf{A} \cap \mathsf{B}|\mathcal{C})(x)\, dP(x) = \int_{\mathsf{C}} P(\mathsf{A}|\mathcal{C})(x) \cdot P(\mathsf{B}|\mathcal{C})(x)\, dP(x),$$

and continue using the 'stronger' definition of $P(\mathsf{B}|\mathcal{C})$ and the fact $f = P(\mathsf{A}|\mathcal{C})$

$$\int_{\mathsf{C}} P(\mathsf{A}|\mathcal{C})(x) \cdot P(\mathsf{B}|\mathcal{C})(x)\, dP(x) = \int_{\mathsf{B}\cap\mathsf{C}} P(\mathsf{A}|\mathcal{C})(x)\, dP(x) = \int_{\mathsf{B}\cap\mathsf{C}} f(x)\, dP(x).$$

Since the class $\mathcal{G} = \{\mathsf{B} \cap \mathsf{C};\ \mathsf{B} \in \mathcal{B},\ \mathsf{C} \in \mathcal{C}\}$ is closed under finite intersection and $\mathcal{B} \vee \mathcal{C} = \sigma(\mathcal{G})$, by the 'weaker' definition of $P(\mathsf{A}|\mathcal{B} \vee \mathcal{C})$ conclude that $f = P(\mathsf{A}|\mathcal{B} \vee \mathcal{C})$.

To show the sufficiency fix $\mathsf{A} \in \mathcal{A}$ and $\mathsf{B} \in \mathcal{B}$. Take a $\mathcal{C}$-measurable version f of $P(\mathsf{A}|\mathcal{B} \vee \mathcal{C})$ and observe $f = P(\mathsf{A}|\mathcal{C})$. Then write by the definition of $P(\mathsf{A} \cap \mathsf{B}|\mathcal{C})$ and the fact $f = P(\mathsf{A}|\mathcal{B} \vee \mathcal{C})$ for every $\mathsf{C} \in \mathcal{C}$

$$\int_{\mathsf{C}} P(\mathsf{A} \cap \mathsf{B}|\mathcal{C})(x)\ dP^{\mathcal{C}}(x) = P(\mathsf{A} \cap \mathsf{B} \cap \mathsf{C}) = \int_{\mathsf{B}\cap\mathsf{C}} f(x)\ dP(x)\,,$$

and continue using the 'stronger' definition of $P(\mathsf{B}|\mathcal{C})$ and the fact $f = P(\mathsf{A}|\mathcal{C})$

$$\int_{\mathsf{B}\cap\mathsf{C}} f(x)\; dP(x) = \int_{\mathsf{C}} f(x)\cdot P(\mathsf{B}|\mathcal{C})(x)\; dP^{\mathcal{C}}(x) = \int_{\mathsf{C}} P(\mathsf{A}|\mathcal{C})(x)\cdot P(\mathsf{B}|\mathcal{C})(x)\; dP^{\mathcal{C}}(x).$$

Thus, the equality

$$\int_{\mathsf{C}} P(\mathsf{A}\cap\mathsf{B}|\mathcal{C})(x)\; dP^{\mathcal{C}}(x) = \int_{\mathsf{C}} P(\mathsf{A}|\mathcal{C})(x)\cdot P(\mathsf{B}|\mathcal{C})(x)\; dP^{\mathcal{C}}(x)$$

was verified for every $\mathsf{C}\in\mathcal{C}$ which implies (10.2). $\square$

The next lemma describes basic properties of conditional independence for σ-algebras.

LEMMA 10.2 Supposing P is a probability measure on $(\mathsf{X},\mathcal{X})$ and $\mathcal{A},\mathcal{B},\mathcal{C},\mathcal{D},\mathcal{E},\mathcal{F},\mathcal{G}\subseteq\mathcal{X}$ are σ-algebras, it holds

(i) $\mathcal{B}\subseteq\mathcal{C} \;\Rightarrow\; \mathcal{A}\perp\!\!\!\perp\mathcal{B}\,|\,\mathcal{C}\;[P]$,

(ii) $\mathcal{A}\perp\!\!\!\perp\mathcal{B}\,|\,\mathcal{C}\;[P] \;\Rightarrow\; \mathcal{B}\perp\!\!\!\perp\mathcal{A}\,|\,\mathcal{C}\;[P]$,

(iii) $\mathcal{A}\perp\!\!\!\perp\mathcal{E}\,|\,\mathcal{C}\;[P],\ \mathcal{F}\subseteq\mathcal{E},\ \mathcal{C}\subseteq\mathcal{G}\subseteq\mathcal{E}\vee\mathcal{C} \;\Rightarrow\; \mathcal{A}\perp\!\!\!\perp\mathcal{F}\,|\,\mathcal{G}\;[P]$,

(iv) $\mathcal{A}\perp\!\!\!\perp\mathcal{B}\,|\,\mathcal{D}\vee\mathcal{C}\;[P],\ \mathcal{A}\perp\!\!\!\perp\mathcal{D}\,|\,\mathcal{C}\;[P] \;\Rightarrow\; \mathcal{A}\perp\!\!\!\perp\mathcal{B}\vee\mathcal{D}\,|\,\mathcal{C}\;[P]$.

Proof: The condition (ii) follows immediately from symmetry in (10.2). For other properties use the equivalent definition from Lemma 10.1. For (i) realize that $\mathcal{B}\vee\mathcal{C}=\mathcal{C}$ and therefore every version of $P(\mathsf{A}|\mathcal{B}\vee\mathcal{C})$ is $\mathcal{C}$-measurable. In case (iii) observe $\mathcal{S}\equiv\mathcal{F}\vee\mathcal{G}\subseteq\mathcal{E}\vee\mathcal{C}\equiv\mathcal{T}$. The assumption $\mathcal{A}\perp\!\!\!\perp\mathcal{E}\,|\,\mathcal{C}\;[P]$ implies, for every $\mathsf{A}\in\mathcal{A}$, the existence of a $\mathcal{C}$-measurable version of $P(\mathsf{A}|\mathcal{T})$. Since $\mathcal{C}\subseteq\mathcal{G}\subseteq\mathcal{S}$ it is both $\mathcal{G}$-measurable and $\mathcal{S}$-measurable. Hence, it is a version of $P(\mathsf{A}|\mathcal{S})$. The existence of a $\mathcal{G}$-measurable version of $P(\mathsf{A}|\mathcal{S})$ means $\mathcal{A}\perp\!\!\!\perp\mathcal{F}\,|\,\mathcal{G}\;[P]$. To show (iv) fix $\mathsf{A}\in\mathcal{A}$ and by $\mathcal{A}\perp\!\!\!\perp\mathcal{B}\,|\,\mathcal{D}\vee\mathcal{C}\;[P]$ derive the existence of $(\mathcal{D}\vee\mathcal{C})$-measurable version f of $P(\mathsf{A}|\mathcal{B}\vee\mathcal{D}\vee\mathcal{C})$. Similarly, by $\mathcal{A}\perp\!\!\!\perp\mathcal{D}\,|\,\mathcal{C}\;[P]$ derive the existence of $\mathcal{C}$-measurable version g of $P(\mathsf{A}|\mathcal{D}\vee\mathcal{C})$. Observe that f is a version of $P(\mathsf{A}|\mathcal{D}\vee\mathcal{C})$ and by the 'uniqueness' of $P(\mathsf{A}|\mathcal{D}\vee\mathcal{C})$ derive that $f=g$ $P^{\mathcal{D}\vee\mathcal{C}}$-a.e. Hence, f and g are equal $P^{\mathcal{B}\vee\mathcal{D}\vee\mathcal{C}}$-a.e. and by 'uniqueness' of $P(\mathsf{A}|\mathcal{B}\vee\mathcal{D}\vee\mathcal{C})$ conclude that g is its version. This implies $\mathcal{A}\perp\!\!\!\perp\mathcal{B}\vee\mathcal{D}\,|\,\mathcal{C}\;[P]$. $\square$

CONSEQUENCE 10.1 Supposing P is a probability measure on $(\mathsf{X},\mathcal{X})$ *semi-graphoid properties* for σ-algebras hold, that is one has for σ-algebras $\mathcal{A},\mathcal{B},\mathcal{C},\mathcal{D}\subseteq\mathcal{X}$ (the symbol of P is omitted):

1. triviality: $\mathcal{A}\perp\!\!\!\perp\mathcal{B}\,|\,\mathcal{C}$ whenever $\mathcal{B}\vee\mathcal{C}=\mathcal{C}$,
2. symmetry: $\mathcal{A}\perp\!\!\!\perp\mathcal{B}\,|\,\mathcal{C} \;\Rightarrow\; \mathcal{B}\perp\!\!\!\perp\mathcal{A}\,|\,\mathcal{C}$,
3. decomposition: $\mathcal{A}\perp\!\!\!\perp\mathcal{B}\vee\mathcal{D}\,|\,\mathcal{C} \;\Rightarrow\; \mathcal{A}\perp\!\!\!\perp\mathcal{D}\,|\,\mathcal{C}$,
4. weak union: $\mathcal{A}\perp\!\!\!\perp\mathcal{B}\vee\mathcal{D}\,|\,\mathcal{C} \;\Rightarrow\; \mathcal{A}\perp\!\!\!\perp\mathcal{B}\,|\,\mathcal{D}\vee\mathcal{C}$,
5. contraction: $\mathcal{A}\perp\!\!\!\perp\mathcal{B}\,|\,\mathcal{D}\vee\mathcal{C}$ & $\mathcal{A}\perp\!\!\!\perp\mathcal{D}\,|\,\mathcal{C} \;\Rightarrow\; \mathcal{A}\perp\!\!\!\perp\mathcal{B}\vee\mathcal{D}\,|\,\mathcal{C}$.

Proof: Use Lemma 10.2; for decomposition use (iii) with $\mathcal{E}=\mathcal{B}\vee\mathcal{D}$, $\mathcal{F}=\mathcal{D}$, $\mathcal{G}=\mathcal{C}$; for weak union put $\mathcal{E}=\mathcal{B}\vee\mathcal{D}$, $\mathcal{F}=\mathcal{B}$, $\mathcal{G}=\mathcal{D}\vee\mathcal{C}$ instead. $\square$

10.7 Relative entropy

Suppose that P is a finite measure and μ a σ-finite measure on a measurable space $(\mathsf{X},\mathcal{X})$. In case $P \ll \mu$ choose a version of Radon-Nikodym derivative $\frac{dP}{d\mu}$, accept the convention $0\cdot\ln 0 \equiv 0$ and introduce

$$H(P|\mu) = \int_{\mathsf{X}} \frac{dP}{d\mu}(x)\cdot \ln\frac{dP}{d\mu}(x)\; d\mu(x). \tag{10.3}$$

Provided that the function $\frac{dP}{d\mu}\cdot\ln\frac{dP}{d\mu}$ is μ-quasi-integrable we call the integral the *relative entropy of P with respect to μ*. Of course, quasi-integrability and the value of $H(P|\mu)$ does not depend on the choice of a version of $\frac{dP}{d\mu}$. It follows from the definition of Radon-Nikodym derivative that $H(P|\mu)$ can be equivalently introduced as the integral

$$H(P|\mu) = \int_{\mathsf{X}} \ln\frac{dP}{d\mu}(x)\; dP(x)\,, \tag{10.4}$$

provided that $\ln\frac{dP}{d\mu}$ is P-quasi-integrable. Hence, the relative entropy of P with respect to μ is finite iff $\ln\frac{dP}{d\mu}$ is P-integrable. Let us note that, in general $P \ll \mu$ does not imply the existence of the integral in (10.3) and in case $H(P|\mu)$ is defined, it can take any value in the interval $[-\infty,+\infty]$. However, when both P and μ are probability measures (and $P \ll \mu$) the existence of $H(P|\mu)$ is guaranteed and it can serve as a measure of similarity of P to μ.

Lemma 10.3 Supposing P and μ are probability measures on $(\mathsf{X},\mathcal{X})$ such that $P \ll \mu$ the relative entropy of P with respect to μ is defined and $H(P|\mu)\geq 0$. Moreover $H(P|\mu)=0$ iff $P=\mu$.

Proof: Apply Jensen's inequality to the case $f=\frac{dP}{d\mu}$ and $\varphi(r)=r\cdot\ln r$ for $r>0$, $\varphi(0)=0$. Since P is a probability measure $\int_{\mathsf{X}} f(x)\; d\mu(x) = 1$ and $\varphi(1)=0$ gives the lower estimate. Moreover, since φ is strictly convex $H(P|\mu)=0$ iff $f(x)=1$ for μ-a.e. $x\in\mathsf{X}$ which is equivalent to the requirement $P=\mu$. $\square$

Let us emphasize that the assumption that $H(P|\mu)$ is finite involves the requirement $P\ll\mu$.

10.8 Finite-dimensional subspaces and cones

Throughout Section 10.8 the set of n-dimensional real vectors $\mathbb{R}^n, n\geq 1$ is fixed. It is a topological space endowed with Eucledian topology. Given $\boldsymbol{x},\boldsymbol{y}\in\mathbb{R}^n$ and $\alpha\in\mathbb{R}$ the *sum of vectors* $\boldsymbol{x}+\boldsymbol{y}\in\mathbb{R}^n$ and the *scalar multiple* $\alpha\cdot\boldsymbol{x}\in\mathbb{R}^n$ are defined componentwisely. The symbol $\mathbf{0}$ denotes the *zero vector* which has 0 as all its components. Given $\mathsf{A}\subseteq\mathbb{R}^n$ the symbol $-\mathsf{A}$ denotes the set $\{-\boldsymbol{x}\,;\boldsymbol{x}\in\mathsf{A}\}$ where $-\boldsymbol{x}$ denotes the scalar multiple $(-1)\cdot\boldsymbol{x}$. The *scalar product* of two vectors $\boldsymbol{x}=[x_i]_{i=1}^n$ and $\boldsymbol{y}=[y_i]_{i=1}^n$ is the number $\langle\boldsymbol{x},\boldsymbol{y}\rangle=\sum_{i=1}^n x_i\cdot y_i$.

10.8.1 Linear subspaces

A set $\mathsf{L}\subseteq\mathbb{R}^n$ is a *linear subspace* if $\mathbf{0}\in\mathsf{L}$ and L is closed under linear combinations, i.e.

$$\forall\boldsymbol{x},\boldsymbol{y}\in\mathsf{L}\quad\forall\alpha,\beta\in\mathbb{R}\qquad \alpha\cdot\boldsymbol{x}+\beta\cdot\boldsymbol{y}\in\mathsf{L}\,.$$

Every linear subspace of $\mathbb{R}^n$ is a closed set with respect to Eucledian topology. A set $\mathsf{A}\subseteq\mathbb{R}^n$ *linearly generates* a subspace $\mathsf{L}\subseteq\mathbb{R}^n$ if every element of L is a linear combination of elements of A, i.e.

$$\forall\,\boldsymbol{x}\in\mathsf{L}\quad\exists\,\mathsf{B}\subseteq\mathsf{A}\text{ finite such that } \boldsymbol{x}=\sum_{\boldsymbol{y}\in\mathsf{B}}\alpha_{\boldsymbol{y}}\cdot\boldsymbol{y}\;\text{ for some }\alpha_{\boldsymbol{y}}\in\mathbb{R},\;\boldsymbol{y}\in\mathsf{B}.$$

By convention $\mathbf{0}$ is the empty linear combination which means that $\emptyset$ linearly generates the subspace $\{\mathbf{0}\}$. A finite set $\mathsf{A} \subseteq \mathbb{R}^n$ is *linearly independent* if

$$\forall\, \alpha_{\boldsymbol{y}} \in \mathbb{R},\ \boldsymbol{y} \in \mathsf{A} \quad \sum_{\boldsymbol{y} \in \mathsf{A}} \alpha_{\boldsymbol{y}} \cdot \boldsymbol{y} = 0 \ \Rightarrow\ [\alpha_{\boldsymbol{y}} = 0 \text{ for every } \ \boldsymbol{y} \in \mathsf{A}]\,.$$

In particular, a set containing $\mathbf{0}$ is never linearly independent. *Linear basis* of a subspace $\mathsf{L} \subseteq \mathbb{R}^n$ is any finite linearly independent set $\mathsf{A} \subseteq \mathsf{L}$ which (linearly) generates L. Every linear subspace $\mathsf{L} \subseteq \mathbb{R}^n$ has a basis which is possibly empty in case $\mathsf{L} = \{\mathbf{0}\}$. Different bases of L have the same number of elements called the *dimension* of L. The dimension is the number between 0 (for $\mathsf{L} = \{\mathbf{0}\}$) and n (for $\mathsf{L} = \mathbb{R}^n$).

One says that a subspace $\mathsf{L} \subseteq \mathbb{R}^n$ is a *direct sum* of subspaces $\mathsf{L}_1, \mathsf{L}_2 \subseteq \mathbb{R}^n$ and writes $\mathsf{L} = \mathsf{L}_1 \oplus \mathsf{L}_2$ if $\mathsf{L}_1 \cap \mathsf{L}_2 = \{\mathbf{0}\}$, $\mathsf{L}_1 \subseteq \mathsf{L}$, $\mathsf{L}_2 \subseteq \mathsf{L}$ and $\mathsf{L}_1 \cup \mathsf{L}_2$ generates L. Then every $\boldsymbol{x} \in \mathsf{L}$ can be written in the form $\boldsymbol{x} = \boldsymbol{y} + \boldsymbol{z}$ where $\boldsymbol{y} \in \mathsf{L}_1$, $\boldsymbol{z} \in \mathsf{L}_2$ and this decomposition of $\boldsymbol{x}$ is unique. Moreover, the dimension of L is the sum of dimensions of L_1 and L_2. The *orthogonal complement* of a set $\mathsf{A} \subseteq \mathbb{R}^n$ is the set

$$\mathsf{A}^{\perp} = \{\boldsymbol{x} \in \mathbb{R}^n;\ \langle \boldsymbol{x}, \boldsymbol{y} \rangle = 0 \text{ for every } \boldsymbol{y} \in \mathsf{A}\}\,.$$

It is always a linear subspace. Moreover, for every linear subspace $\mathsf{L} \subseteq \mathbb{R}^n$ one has $\mathbb{R}^n = \mathsf{L} \oplus \mathsf{L}^{\perp}$ and $\mathsf{L} = (\mathsf{L}^{\perp})^{\perp}$.

10.8.2 Convex cones

A set $\mathsf{K} \subseteq \mathbb{R}^n$ is a *convex cone* if $\mathbf{0} \in \mathsf{K}$ and K is closed under conical combinations, i.e.

$$\forall\, \boldsymbol{x}, \boldsymbol{y} \in \mathsf{K} \quad \forall\, \alpha, \beta \geq 0 \quad \alpha \cdot \boldsymbol{x} + \beta \cdot \boldsymbol{y} \in \mathsf{K}\,.$$

By *closed convex cone* is understood a convex cone which is a closed set with respect to Eucledian topology on $\mathbb{R}^n$. An example of a closed convex cone is a linear subspace. Another example is the *dual cone* A^* to a set $\mathsf{A} \subseteq \mathbb{R}^n$ defined by

$$\mathsf{A}^* = \{\boldsymbol{y} \in \mathbb{R}^n;\ \langle \boldsymbol{x}, \boldsymbol{y} \rangle \geq 0 \text{ for every } \boldsymbol{x} \in \mathsf{A}\}\,.$$

This is a general example as $\mathsf{K} \subseteq \mathbb{R}^n$ is a closed convex cone iff $\mathsf{K} = \mathsf{A}^*$ for some $\mathsf{A} \subseteq \mathbb{R}^n$ (see Consequence 1 in [105]). Further example of a closed cone is the *conical closure* $con(\mathsf{B})$ of a non-empty set $\emptyset \neq \mathsf{B} \subseteq \mathbb{R}^n$ ($con(\emptyset) = \{\mathbf{0}\}$ by convention):

$$con(\mathsf{B}) = \{\boldsymbol{x} \in \mathsf{R}^n\,;\ \boldsymbol{x} = \sum_{\boldsymbol{z} \in \mathsf{C}} \alpha_{\boldsymbol{z}} \cdot \boldsymbol{z} \text{ for some } \alpha_{\boldsymbol{z}} \geq 0 \text{ and finite } \emptyset \neq \mathsf{C} \subseteq \mathsf{B}\}.$$

Note that, $con(\mathsf{B}) = \mathsf{B}^{**}$ for every finite $\mathsf{B} \subseteq \mathbb{R}^n$ (see Fact 6 and Proposition 1 in [105]). A cone $con(\mathsf{B})$ with finite B called a *polyhedral cone*; by a *rational polyhedral cone* is understood the conical closure of a finite set of rational vectors $\mathsf{B} \subseteq \mathbb{Q}^n$. Basic fact is that a set $\mathsf{K} \subseteq \mathbb{R}^n$ is a polyhedral cone iff $\mathsf{K} = \mathsf{A}^*$ for a finite $\mathsf{A} \subseteq \mathbb{R}^n$. Analogously, K is a rational polyhedral cone iff $\mathsf{K} = \mathsf{A}^*$ for a finite set of rational vectors $\mathsf{A} \subseteq \mathbb{Q}^n$ (see Proposition 5 in [105]). Note that these facts can be viewed as consequences (or analogues) of a well-known result from convex analysis saying that polytopes coincide with bounded polyhedrons. Let us call by a *face* of a polyhedral cone K a convex cone $\mathsf{F} \subseteq \mathsf{K}$ such that $\forall\, \boldsymbol{x}, \boldsymbol{y} \in \mathsf{K} \quad \boldsymbol{x} + \boldsymbol{y} \in \mathsf{F}$ implies $\boldsymbol{x}, \boldsymbol{y} \in \mathsf{F}$. Note that this is a modification of the usual definition of a face of a closed convex set from [12] and the definitions coincide for non-empty subsets F of a polyhedral cone K. One can show that a face of a polyhedral cone is a polyhedral cone (c.f. Consequence 8.4 in [12]).

A closed cone $\mathsf{K} \subseteq \mathbb{R}^n$ is *pointed* if $\mathsf{K} \cap (-\mathsf{K}) = \{\mathbf{0}\}$. Apparently stronger equivalent definition says that a closed cone K is pointed iff there exists $\boldsymbol{y} \in \mathbb{R}^n$ such that $\langle \boldsymbol{x}, \boldsymbol{y} \rangle > 0$ for every $\boldsymbol{x} \in \mathsf{K} \setminus \{\mathbf{0}\}$ (see Proposition 2 in [105]). By a *ray* generated by non-zero vector $\mathbf{0} \neq \boldsymbol{x} \in \mathbb{R}^n$ is understood the set $\mathsf{R}_{\boldsymbol{x}} = \{\alpha \cdot \boldsymbol{x} \, ; \, \alpha \geq 0\}$. Clearly, every cone contains whole ray R with any of its non-zero vectors $\mathbf{0} \neq \boldsymbol{x} \in \mathsf{R}$ which then necessarily generates R. Given a closed convex cone $\mathsf{K} \subseteq \mathbb{R}^n$ a ray $\mathsf{R} \subseteq \mathsf{K}$ is called *extreme* (in K) if

$$\forall \boldsymbol{x}, \boldsymbol{y} \in \mathsf{K} \quad \boldsymbol{x} + \boldsymbol{y} \in \mathsf{R} \ \text{ implies } \boldsymbol{x}, \boldsymbol{y} \in \mathsf{R}.$$

A closed cone K has extreme rays iff it is pointed and contains a non-zero vector $\mathbf{0} \neq \boldsymbol{x} \in \mathsf{K}$. Moreover, every pointed closed convex cone $\mathsf{K} \subseteq \mathbb{R}^n$ is a conical combination of its extreme rays, more exactly $\mathsf{K} = con(\mathsf{B})$ for every $\mathsf{B} \subseteq \mathsf{K}$ such that $\mathsf{M} \cap (\mathsf{R} \setminus \{\mathbf{0}\}) \neq \emptyset$ for each extreme ray $\mathsf{R} \subseteq \mathsf{K}$. Note that this fact can be viewed as a consequence of well-known Krein-Millman theorem for bounded closed convex sets (see Proposition 4 in [105]). A pointed closed cone is a polyhedral cone iff it has finitely many extreme rays. Moreover, it is a rational polyhedral cone iff it has finitely many extreme rays and every its extreme ray is generated by a rational vector (see Consequence 5 in [105]). Basic result of Section 5.2 are based on the following specific property of pointed rational polyhedral cones.

Lemma Let $\mathsf{K} \subseteq \mathbb{R}^n$ be a pointed rational polyhedral cone and R is an extreme ray of K. Then there exists $\boldsymbol{q} \in \mathbb{Q}^n$ such that $\langle \boldsymbol{q}, \boldsymbol{x} \rangle = 0$ for any $\boldsymbol{x} \in \mathsf{R}$ and $\langle \boldsymbol{q}, \boldsymbol{y} \rangle > 0$ whenever $\mathbf{0} \neq \boldsymbol{y}$ belongs to other (extreme) ray of K.

Another useful fact is that every conical combination of integral vectors is necessarily a rational conical combination (see Lemma 10 in [105]).

Fact Supposing $\mathsf{B} \subseteq \mathbb{Z}^n$ every $\boldsymbol{x} \in con(\mathsf{B}) \cap \mathbb{Z}^n$ has the form $\boldsymbol{x} = \sum_{\boldsymbol{y} \in \mathsf{B}} \alpha_{\boldsymbol{y}} \cdot \boldsymbol{y}$ where $\alpha_{\boldsymbol{y}} \in \mathbb{Q}$, $\alpha_{\boldsymbol{y}} \geq 0$ for every $\boldsymbol{y} \in \mathsf{B}$.

10.9 Concepts from multivariate analysis

The concepts and facts mentioned in this section are commonly used in mathematical statistics, in particular in its special area known as *multivariate analysis*. The proofs of the facts from Section 10.9.1 can be found in textbooks of matrix calculus, for example [23], Chapters 1 and 2. The proofs of basic facts from Section 10.9.3 are in any reasonable textbook of statistics, see e.g. [3], section V.1.

10.9.1 Matrices

Given non-empty finite sets N, M by a real $N \times M$-matrix will be understood a real function on $N \times M$, that is an element of $\mathbb{R}^{N \times M}$. The corresponding values are indicated by subscripts so that one writes $\mathbf{\Sigma} = (\sigma_{ij})_{i \in N, j \in M}$ to explicate the components of a matrix $\mathbf{\Sigma}$ of this type. Note that this approach slightly differs from classic understanding of the concept of matrix where the index sets have settled pre-orderings, e.g. $N = \{1, 2, \ldots, n\}$ and $M = \{1, \ldots, m\}$. This enables one to write certain formulas involving matrices in much more elegant way.

The result of *matrix multiplication* of an $N \times M$-matrix $\mathbf{\Sigma}$ and an $M \times K$-matrix $\mathbf{\Gamma}$ (where N, M, K are non-empty finite sets) is an $N \times K$-matrix denoted by $\mathbf{\Sigma} \cdot \mathbf{\Gamma}$. A real *vector $\boldsymbol{v}$ over N*, that is an element of $\mathbb{R}^N$ will be here understood as a column vector so that it should appear in matrix multiplication with an $N \times N$-matrix $\mathbf{\Sigma}$ from left: $\mathbf{\Sigma} \cdot \boldsymbol{v}$. *Null matrix* or vector having all components zero is denoted by $\mathbf{0}$; *unit matrix* by $\boldsymbol{I}$. An $N \times N$-matrix $\mathbf{\Sigma} = (\sigma_{ij})_{i,j \in N}$ is *symmetric* if $\sigma_{ij} = \sigma_{ji}$ for every $i, j \in N$; *regular* if there exists (uniquely determined) *inverse*

$N \times N$-matrix $\mathbf{\Sigma}^{-1}$ such that $\mathbf{\Sigma} \cdot \mathbf{\Sigma}^{-1} = \boldsymbol{I} = \mathbf{\Sigma}^{-1} \cdot \mathbf{\Sigma}$. The *transpose* of $\mathbf{\Sigma}$ will be denoted by $\mathbf{\Sigma}^{\top}$, the *determinant* by $\det(\mathbf{\Sigma})$.

By a *generalized inverse* of a real $N \times N$-matrix $\mathbf{\Sigma}$ will be understood any $N \times N$ matrix $\mathbf{\Sigma}^{-}$ such that $\mathbf{\Sigma} \cdot \mathbf{\Sigma}^{-} \cdot \mathbf{\Sigma} = \mathbf{\Sigma}$. A matrix of this sort always exists, but it is not determined uniquely unless $\mathbf{\Sigma}$ is regular when it coincides with $\mathbf{\Sigma}^{-1}$ (see [82], Section 1b.5). However, the expressions in which generalized inverses are commonly used usually do not depend on their choice.

A real symmetric $N \times N$-matrix $\mathbf{\Sigma}$ is called *positive semi-definite* if $\boldsymbol{v}^{\top} \cdot \mathbf{\Sigma} \cdot \boldsymbol{v} \geq 0$ for every $\boldsymbol{v} \in \mathbb{R}^N$, and *positive definite* if $\boldsymbol{v}^{\top} \cdot \mathbf{\Sigma} \cdot \boldsymbol{v} > 0$ for every $\boldsymbol{v} \in \mathbb{R}^N$, $\boldsymbol{v} \neq \mathbf{0}$. Note that $\mathbf{\Sigma}$ is positive definite iff it is regular and positive semi-definite. In that case is $\mathbf{\Sigma}^{-1}$ positive definite as well. Given an $N \times N$-matrix $\mathbf{\Sigma} = (\sigma_{ij})_{i,j \in N}$ and $A, B \subseteq N$ the symbol $\mathbf{\Sigma}_{A \cdot B}$ will be used to denote $A \times B$*-submatrix*, that is $\mathbf{\Sigma}_{A \cdot B} = (\sigma_{ij})_{i \in A, j \in B}$. Supposing $\mathbf{\Sigma}$ is positive definite (semi-definite) and $A \subseteq N$ its *main submatrix* $\mathbf{\Sigma}_{A \cdot A}$ is positive definite (semi-definite) as well. Note that the operation $\mathbf{\Sigma} \mapsto \mathbf{\Sigma}_{A \cdot A}$ plays sometimes the role of 'marginalizing' (but only for positive semi-definite matrices). On the other hand, supposing $\mathbf{\Sigma}$ is only regular, $\mathbf{\Sigma}_{A \cdot A}$ need not be regular.

Suppose that $\mathbf{\Sigma}$ is a real $N \times N$-matrix, non-empty sets $A, C \subseteq N$ are disjoint and $\mathbf{\Sigma}_{C \cdot C}$ is regular. Then one can introduce *Schur complement* $\mathbf{\Sigma}_{A|C}$ as an $A \times A$ matrix as follows:

$$\mathbf{\Sigma}_{A|C} = \mathbf{\Sigma}_{A \cdot A} - \mathbf{\Sigma}_{A \cdot C} \cdot (\mathbf{\Sigma}_{C \cdot C})^{-1} \cdot \mathbf{\Sigma}_{C \cdot A}$$

with the convention $\mathbf{\Sigma}_{A|\emptyset} \equiv \mathbf{\Sigma}_{A \cdot A}$. Note that $\mathbf{\Sigma}_{AC \cdot AC}$ is regular iff $\mathbf{\Sigma}_{A|C}$ (and $\mathbf{\Sigma}_{C \cdot C}$) is regular and $(\mathbf{\Sigma}_{A|C})^{-1} = ((\mathbf{\Sigma}_{AC \cdot AC})^{-1})_{A \cdot A}$ then. Moreover, the following 'transitivity principle' holds: supposing $A, B, C \subseteq N$ are pairwise disjoint and $\mathbf{\Sigma}$ is an $N \times N$-matrix such that both $\mathbf{\Sigma}_{C \cdot C}$ and $\mathbf{\Sigma}_{BC \cdot BC}$ is regular one has $\mathbf{\Sigma}_{A|BC} = (\mathbf{\Sigma}_{AB|C})_{A|B}$. An important fact is that whenever $\mathbf{\Sigma}$ is positive definite then $\mathbf{\Sigma}_{A|C}$ is positive definite as well. Thus, the operation $\mathbf{\Sigma}_{AC \cdot AC} \mapsto \mathbf{\Sigma}_{A|C}$ often plays the role of 'conditioning' (for positive definite matrices only).

However, one sometimes needs to define 'conditional' matrix $\mathbf{\Sigma}_{A|C}$ even in case that $\mathbf{\Sigma}_{C \cdot C}$ is not regular. Thus, supposing $\mathbf{\Sigma}$ is a positive semi-definite matrix one can introduce $\mathbf{\Sigma}_{A|C}$ by means of generalized inverse $(\mathbf{\Sigma}_{C \cdot C})^{-}$ as follows

$$\mathbf{\Sigma}_{A|C} = \mathbf{\Sigma}_{A \cdot A} \; - \; \mathbf{\Sigma}_{A \cdot C} \cdot (\mathbf{\Sigma}_{C \cdot C})^{-} \cdot \mathbf{\Sigma}_{C \cdot A}.$$

Note that this matrix does not depend on the choice of generalized inverse (use [82], Section 8a.2(v)) and that in case of positive definite matrix is coincides with the above mentioned Schur complement. Therefore, the concept of 'conditioning' is extended to positive semi-definite matrices.

10.9.2 Statistical characteristics of probability measures

Remark Elementary concept of mathematical statistics is a *random variable* which is a real measurable function ξ on a certain (intentionally unspecified) measurable space $(\Omega, \mathcal{A})$ where Ω is interpreted as the 'universum' of elementary events and $\mathcal{A}$ as the collection of 'observable' random events. Moreover, it is assumed that $(\Omega, \mathcal{A})$ admits a probability measure $\boldsymbol{P}$. Then every *random vector*, that is a finite collection of random variables $\boldsymbol{\xi} = [\xi_i]_{i \in N}$ where $|N| \geq 2$ induces on $\mathbb{R}^N = \prod_{i \in N} \mathsf{X}_i$ with $\mathsf{X}_i = \mathbb{R}$, endowed with the Borel σ-algebra $\mathcal{B}^N$ (= the product of Borel σ-algebras on $\mathbb{R}$ in this case) a probability measure P called the *distribution* of $\boldsymbol{\xi}$

$$P(\mathsf{A}) = \boldsymbol{P}\left(\{w \in \Omega;\; \boldsymbol{\xi}(w) \in \mathsf{A}\}\right) \qquad \text{for every Borel set } \mathsf{A} \subseteq \mathbb{R}^N.$$

The measurable space $(\mathbb{R}^N, \mathcal{B}^N)$ is then called the (joint) *sample space*. Note that 'generalized' random variables taking values in alternative sample spaces (e.g. finite sets instead of $\mathbb{R}$) are sometimes considered as well.

The area of interest of mathematical statistics is not the 'underlying' theoretical probability $\boldsymbol{P}$ but the induced probability measure P on the sample space. Indeed, despite the fact that textbooks of statistics introduce various numerical characteristics of random vectors, these numbers actually do not characterize random vectors themselves but their distributions, that is induced Borel probability measures on $\mathbb{R}^N$. The purpose of many statistical methods is then simply to estimate these numerical characteristics from data. Definitions of basic ones are recalled this section. $\triangle$

Let P be probability measure on $(\prod_{i\in N} \mathsf{X}_i, \prod_{i\in N} \mathcal{X}_i) = (\mathbb{R}^N, \mathcal{B}^N)$ where $|N| \geq 2$. Let x_i denote the i-th component $(i \in N)$ of a vector $\boldsymbol{x} \in \mathbb{R}^N$. In case that, for every $i \in N$, the function $\boldsymbol{x} \mapsto x_i$, $\boldsymbol{x} \in \mathbb{R}^N$ (which is $\mathcal{B}^N$-measurable) is P-integrable one can define the *expectation* as a real vector $\boldsymbol{e} = (e_i)_{i\in N} \in \mathbb{R}^N$ having components

$$e_i = \int_{\mathbb{R}^N} x_i \; dP(\boldsymbol{x}) = \int_{\mathsf{X}_i} y \; dP^{\{i\}}(y) \quad \text{for } i \in N.$$

If moreover the function $\boldsymbol{x} \mapsto (x_i - e_i) \cdot (x_j - e_j)$ is P-integrable for every $i, j \in N$ one defines the *covariance matrix* of P as an $N \times N$-matrix $\boldsymbol{\Sigma} = (\sigma_{ij})_{i,j\in N}$ with elements

$$\sigma_{ij} = \int_{\mathbb{R}^N} (x_i - e_i) \cdot (x_j - e_j) \; dP(\boldsymbol{x}) = \int_{\mathsf{X}_i \times \mathsf{X}_j} (y - e_i) \cdot (z - e_j) \; dP^{\{i,j\}}(y, z) \quad \text{for } i, j \in N.$$

Alternative names are 'variance matrix', 'dispersion matrix' [82], or even 'variance-covariance matrix' [124]. Elementary fact is that covariance matrix is always positive semi-definite; the converse is also valid (see the next section).
Supposing P has a covariance matrix $\boldsymbol{\Sigma} = (\sigma_{ij})_{i,j\in N}$ such that $\sigma_{ii} > 0$ for every $i \in N$ one can introduce the *correlation matrix* $\boldsymbol{\Gamma} = (\rho_{ij})_{i,j\in N}$ by the formula

$$\rho_{ij} = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii} \cdot \sigma_{jj}}} \qquad \text{for } i, j \in N.$$

Note that the situation above occurs whenever $\boldsymbol{\Sigma}$ is regular (= positive definite) and $\boldsymbol{\Gamma}$ is then a positive definite matrix with $\gamma_{ii} = 1$ for every $i \in N$.

10.9.3 Multivariate Gaussian distributions

Definition of a general Gaussian measure on $\mathbb{R}^N$ is not straightforward. First, one has to introduce one-dimensional Gaussian measure $\mathcal{N}(r, s)$ on $\mathbb{R}$ with parameters $r, s \in \mathbb{R}$, $s \geq 0$. In case $s > 0$ one can do so by defining Radon-Nikodym derivative with respect to Lebesgue measure on $\mathbb{R}$

$$f(x) = \frac{1}{\sqrt{2\pi s}} \cdot \exp^{-\frac{(x-r)^2}{2s}} \qquad \text{for } x \in \mathbb{R}.$$

In case $s = 0$ is $\mathcal{N}(r, 0)$ defined as the Borel measure on $\mathbb{R}$ concentrated on $\{r\}$.

Then supposing $\boldsymbol{e} \in \mathbb{R}^N$ and $\boldsymbol{\Sigma}$ is a positive semi-definite $N \times N$-matrix ($|N| \geq 1$) one can introduce the *Gaussian measure* $\mathcal{N}(\boldsymbol{e}, \boldsymbol{\Sigma})$ as a Borel measure P on $\mathbb{R}^N$ such that, for every $\boldsymbol{v} \in \mathbb{R}^N$, P induces through measurable mapping $\boldsymbol{x} \mapsto \boldsymbol{x}^\top \cdot \boldsymbol{v}$, $\boldsymbol{x} \in \mathbb{R}^N$ one-dimensional Gaussian measure $\mathcal{N}(\boldsymbol{v}^\top \cdot \boldsymbol{e}, \boldsymbol{v}^\top \cdot \boldsymbol{\Sigma} \cdot \boldsymbol{v})$ on $\mathbb{R}$. Let us note that a measure of this kind always exists and is determined uniquely by the requirement above. Moreover, P has then the expectation $\boldsymbol{e}$ and the covariance matrix $\boldsymbol{\Sigma}$. This indicates why parameters were designed in this way and shows that every positive semi-definite matrix is the covariance matrix of a Gaussian measure.

Linear transformation of a Gaussian measure $\mathcal{N}(\boldsymbol{e},\boldsymbol{\Sigma})$ by a mapping $\boldsymbol{x}\mapsto\boldsymbol{y}+\boldsymbol{\Lambda}\cdot\boldsymbol{x}$, $\boldsymbol{x}\in\mathbb{R}^N$ where $\boldsymbol{y}\in\mathbb{R}^M$, $\boldsymbol{\Lambda}\in\mathbb{R}^{M\times N}$, $|M|\geq 1$ is again a Gaussian measure $\mathcal{N}(\boldsymbol{y}+\boldsymbol{\Lambda}\cdot\boldsymbol{e},\,\boldsymbol{\Lambda}\cdot\boldsymbol{\Sigma}\cdot\boldsymbol{\Lambda}^\top)$. In particular, the marginal of a Gaussian measure is again a Gaussian measure

$$P=\mathcal{N}(\boldsymbol{e},\boldsymbol{\Sigma}),\ \emptyset\neq A\subseteq N\ \Rightarrow\ P^A=\mathcal{N}(\boldsymbol{e}_A,\boldsymbol{\Sigma}_{A\cdot A})\,. \tag{10.5}$$

Note that this explains the interpretation of $\boldsymbol{\Sigma}_{A\cdot A}$ as a 'marginal' matrix. Very important fact is that independence is characterized by means of the covariance matrix

$$P=\mathcal{N}(\boldsymbol{e},\boldsymbol{\Sigma})\ \ A,B\subseteq N\ \ A\cap B=\emptyset\ \ \Rightarrow\ \ [\,P^{AB}=P^A\times P^B\quad\text{iff}\quad\boldsymbol{\Sigma}_{A\cdot B}=\boldsymbol{0}\,]\,. \tag{10.6}$$

In general, Gaussian measure $\mathcal{N}(\boldsymbol{e},\boldsymbol{\Sigma})$ is concentrated on a certain shifted linear subspace which can be described as follows

$$\{\,x\in\mathbb{R}^N;\ \forall\boldsymbol{v}\in\mathbb{R}^N\quad\boldsymbol{v}^\top\cdot\boldsymbol{\Sigma}=\boldsymbol{0}\ \Rightarrow\ \boldsymbol{v}^\top\cdot(x-\boldsymbol{e})=0\,\}\,. \tag{10.7}$$

In case $\boldsymbol{\Sigma}$ is regular, the subspace is whole $\mathbb{R}^N$ and $P=\mathcal{N}(\boldsymbol{e},\boldsymbol{\Sigma})$ can be introduced directly by its Radon-Nikodym derivative with respect to Lebesgue measure on $\mathbb{R}^N$

$$f(x)=\frac{1}{\sqrt{(2\pi)^{|N|}\cdot\det(\boldsymbol{\Sigma})}}\cdot\exp^{-\frac{(x-\boldsymbol{e})^\top\cdot\boldsymbol{\Sigma}^{-1}\cdot(x-\boldsymbol{e})}{2}}\qquad\text{for } x\in\mathbb{R}^N\,.$$

This version of Radon-Nikodym derivative is strictly positive and continuous with respect to Eucledian topology on $\mathbb{R}^N$. Moreover, it is the unique continuous version within the class of all possible versions of Radon-Nikodym derivative of P with respect to Lebesgue measure λ. This simple fact motivates an implicit convention used commonly in statistical literature; only continuous versions, called *densities* are taken into consideration. The convention is in concordance with usual way of 'marginalizing' since, for $\emptyset\neq A\subset N$, by integrating continuous density

$$f_A(x)=\int\limits_{\mathsf{X}_{N\setminus A}} f(x,y)\ d\lambda(y)\qquad\text{for }\ x\in\mathbb{R}^A\,,$$

one gets a continuous density again. This also motivates a natural way of definition of (continuous) *conditional density* for disjoint $A,C\subseteq N$ by the formula

$$f_{A|C}(x|z)=\frac{f_{AC}(xz)}{f_C(z)}\qquad\text{for }\ x\in\mathbb{R}^A,\ z\in\mathbb{R}^C,$$

where the ratio is zero whenever $f_C(z)=0$ by convention, and the definition of conditional measure for every $z\in\mathbb{R}^C$

$$P_{A|C}(\mathsf{A}|z)=\int\limits_{\mathsf{A}} f_{A|C}(x|z)\ d\lambda_A(x)\qquad\text{for every Borel set } \mathsf{A}\subseteq\mathbb{R}^A,$$

which appears to be a regular version of conditional probability on $\mathbb{R}^A$ given C. Let us emphasize that just the acceptance of the convention above leads to its 'uniqueness' for every $z\in\mathbb{R}^C$. It is again a Gaussian measure

$$\begin{aligned}&P=\mathcal{N}(\boldsymbol{e},\boldsymbol{\Sigma}),\quad A,C\subseteq N,\quad A\cap C=\emptyset\neq A\ \Rightarrow\\ &\qquad P_{A|C}(*|z)=\mathcal{N}(\boldsymbol{e}_A+\boldsymbol{\Sigma}_{A\cdot C}\cdot(\boldsymbol{\Sigma}_{C\cdot C})^{-1}\cdot(z-\boldsymbol{e}_C)\,,\,\boldsymbol{\Sigma}_{A|C})\end{aligned} \tag{10.8}$$

(see [82], Section 8a.2(v)), called sometimes *conditional Gaussian measure*. Important feature is that its covariance matrix does not depend on z. This maybe explains the meaning of Schur complement $\boldsymbol{\Sigma}_{A|C}$, sometimes called the *conditional covariance matrix*.

However, conditioning can be introduced 'uniquely' even in case of degenerate Gaussian measure for $z \in \mathbb{R}^C$ belonging to respective shifted linear subspace mentioned in (10.7). It is again a Gaussian measure, given by (10.8) but $(\mathbf{\Sigma}_{C\cdot C})^{-1}$ is replaced by a generalized inverse $(\mathbf{\Sigma}_{C\cdot C})^{-}$. As shown in [82], section 8a.2(v), this conditional measure does not depend on the choice of generalized inverse.

The last important fact is that in case $\mathbf{\Sigma}$ is positive definite the measure $P = \mathcal{N}(\boldsymbol{e}, \mathbf{\Sigma})$ has finite relative entropy with respect to Lebesgue measure λ on $\mathbb{R}^N$, namely

$$H(P \,|\, \lambda) \;=\; \frac{-|N|}{2} \cdot \ln(2\pi) - \frac{|N|}{2} - \frac{1}{2}\ln(\det(\mathbf{\Sigma})), \tag{10.9}$$

see [82], section 8a.6 (note that Rao's entropy is nothing but minus relative entropy).

Index

The first part of the Index is the list of notions in alphabetic order. The reference usually indicates the page containing the definition. Several entries are in italics. These entries indicate either concepts from literature not studied in details in this work (they are mentioned without definition only) or vaguely defined concepts. The second part of the Index is the list of elementary items like Figures, Lemmas etc. The third part of the Index is the list of special symbols.

SUBJECT INDEX

A

B

C

D

E

F

G

M

N

O

P

Q

R

S

T

U

V

W

Z

LIST OF ITEMS

Consequences

Conventions

Directions

Examples

Equations

Figures

Lemmas

Observations

Questions

Remarks

Themes

Theorems

THE LIST OF SYMBOLS

Simple conventional symbols

$\ll$	symbol for absolute continuity of measures 157
$\perp\!\!\!\perp$ $\top\!\!\!\top$	symbols for conditional independence and dependence 14
$\otimes$	symbol for conditional product 141
$\oplus$	symbol for direct sum (of linear spaces 162)
$\emptyset$	symbol for the empty set 151
$\rightleftharpoons$	symbol for facial equivalence 91
$\rightharpoondown$	symbol for facial implication 93
∞	symbol for infinity
$\int$	symbol for Lebegue integral 156
$\sim_m$	symbol for level equivalence induced by a skeletal imset m 130
$\wedge$ $\vee$	symbols for meet (infimum) and join (supremum) 152 (155)
$\cdot$	symbol for multiplication of matrices 163, (scalar) multiple 161
$\neg$	symbol for negation
$\preceq$ $\succeq$ $\prec$ $\succ$	symbols for partial ordering 152
$\prod$ $\times$	symbols for product (in general)
$\setminus$	symbol for set difference, e.g. $A \setminus B$
$\subseteq$ $\supseteq$ $\subset$ $\supset$	symbols for set inclusion 151
$\bigcup$ $\cup$ $\bigcap$ $\cap$	symbols for set union and intersection 151
$\leadsto$	symbol for strong facial implication 113
0	zero, zero imset 34
$\mathbf{0}$	zero vector 161, null matrix 163

Composite conventional symbols

$\mid * \mid$	absolute value, cardinality 12
$\lfloor * \rfloor$	lower integer part: $\lfloor a \rfloor = \max\{z \in \mathbb{Z}\,;\, z \leq a\}$ for $a \in \mathbb{R}$
$\lceil * \rceil$	upper integer part: $\lceil a \rceil = \min\{z \in \mathbb{Z}\,;\, a \leq z\}$ for $a \in \mathbb{R}$
$\binom{*}{*}$	combination number: $\binom{n}{k} = \frac{1 \cdot \ldots \cdot n}{1 \cdot \ldots \cdot k \cdot 1 \cdot \ldots \cdot (n-k)}$ for $n, k \in \mathbb{N}$, $k \leq n$
$* - *$	line (undirected edge): $a - b$ is a line between a and b 153
$* \to *$	arrow (directed edge): $a \to b$ is an arrow from a to b 153
$(*, *)$	open interval, ordered pair 153
$[*, *]$	closed interval, edge in a graph 153
$\langle *, * \rangle$	scalar product 36 161
$\{*, \ldots, *\}$	set containing elements of the list $*, \ldots, *$

Symbols from other languages

Æ	the set of attributes of a formal context 84
χ_A	indicator of a set A 156
δ_A	identificator of a set A 34
Δ	generic symbol for a set of discrete variables 54
Γ	generic symbol for a set of continuous variables 54
$\Im$	incidence relation of a formal context 84
λ	(occasionally) Lebesgue measure 156
∇	generic symbol for a symbolic function 119

Œ	the set of objects of a formal context 84
π	Ludolf's constant, generic symbol for a permutation 151
$\wp$	generic symbol for a single class of facial equivalence 98
Ψ	the class of measures over N (distribution framework) 38 90-91 97
$\Psi(u)$	the class of Markovian measures with respect to u in Ψ 91
σ-	sigma = a generic symbol for countable infinite operation
$\sigma(\mathcal{A})$	σ-algebra generated by a class of set $\mathcal{A}$ 155
$\boldsymbol{\Sigma}^{-1}$, $\boldsymbol{\Sigma}^{-}$	inverse and generalized inverse of a matrix $\boldsymbol{\Sigma}$ 163
$\boldsymbol{\Sigma}_{A\cdot B}$, $\boldsymbol{\Sigma}_{A\vert B}$	submatrix of a matrix $\boldsymbol{\Sigma}$, Schur complement 164
$\boldsymbol{\Sigma}^{\top}$, $\boldsymbol{v}^{\top}$	transpose of a matrix $\boldsymbol{\Sigma}$ and of a vector $\boldsymbol{v}$ 164
$\mho$	generic symbol for a set of classes of facial equivalence 98
υ	counting measure 156

Symbols in alphabetic order

$-\mathsf{A}$	the set $\{-x\,;\, x\in\mathsf{A}\}$ 161
$\mathsf{A}^{\perp}$, A^{*}	orthogonal complement of A, dual cone to A 162
$\langle A,B\vert C\rangle$	disjoint triplet over N 14
$A \perp\!\!\!\perp B\,\vert\, C\ [\boldsymbol{o}]$	(conditional) independence statement 14 13 37 40 43 64 72
$A \not\perp\!\!\!\perp B\,\vert\, C\ [\boldsymbol{o}]$	(conditional) dependence statement 14
$\mathcal{A} \perp\!\!\!\perp \mathcal{B}\,\vert\,\mathcal{C}\ [P]$	conditional independence for σ-algebras 159
$an_G(A)$	the set of ancestors of a set A in a graph G 154
$as\,(\mathcal{M},N)$	ascetic extension of a model $\mathcal{M}$ to N 139
$\mathcal{C}$	(occasionally) the set of cliques of an undirected graph 45
$\mathcal{C}(N)$	the class of combinatorial imsets over N 59
$ce\,(\mathcal{M},N)$	cellular extension of a model $\mathcal{M}$ to N 140
cl	generic symbol for a closure operation 153
$cl_{\mathcal{U}(N)}$	structural closure 113
$\mathcal{D}^{\downarrow}$, $\mathcal{D}^{\uparrow}$	induced descending and ascending class 151
$\mathcal{D}_u^{+}$, $\mathcal{D}_u^{-}$	positive and negative domain of an imset u 34
$\mathcal{D}_u^{*}$	the effective domain of a structural imset u 102
$\frac{d\nu}{d\mu}$	Radon-Nikodym derivative (density) of ν with respect to μ 157
$deg(u,l)$, $deg(u)$	(level-) degree of a combinatorial imset u 59
$\det(\boldsymbol{\Sigma})$	determinant of a matrix $\boldsymbol{\Sigma}$ 164
$\mathcal{E}(N)$, $\mathcal{E}_l(N)$	the class of elementary imsets over N (of degree l) 57
exp	the symbol for exponential function
f^{+}, f^{-}	positive and negative part of a function f 156
f_A	generic symbol for marginal density 20 166
$f^{\downarrow A}$	projection of a density f to A 20
$f_{A\vert C}$	generic symbol for conditional density 166
$f_{\boldsymbol{e},\boldsymbol{\Sigma}}$	density of non-degenerate Gaussian measure $\mathcal{N}(\boldsymbol{e},\boldsymbol{\Sigma})$ 28
G_T	induced subgraph of G for T 153
G^T	marginal undirected graph 39
$gra_*(N)$, $gra(N)$	(generalized) grade of the set of variables N 101
$\mathcal{H}(N)$	Hilbert basis of the cone $con(\mathcal{E}(N))$ 100
$\mathcal{H}_u$	coportrait of a structural imset 122
$H(P\vert\mu)$	relative entropy of P with respect to μ 161

$H(P\|\mu : Q)$	Q-perturbated relative entropy of P with respect to μ 54
$h_{P,\mu}$	entropy function 67
$\boldsymbol{I}$	unit matrix 163
$\langle i,j\|K\rangle$	elementary triplet over N 16
$i \perp\!\!\!\perp j \,\|\, K$	elementary independence statement over N 16
inf	infimum, the greatest lower bound 152
$\mathcal{K}(N)$	the class of supermodular functions over N 71
$\mathcal{K}_\ell(N)$	the class of ℓ-standardized supermodular functions over N 75
$\mathcal{K}_\ell^\diamond(N)$	the ℓ-skeleton
$\mathcal{K}_o^\diamond(N)$, $\mathcal{K}_u^\diamond(N)$	the o-skeleton, the u-skeleton
ℓ-	generic symbol for 'lower' standardization 35
$\mathcal{L}(N)$	the class of modular functions over N 73
$\mathcal{L}_u$	the lower class of a structural imset u 60
$li(\mathcal{M}, N : X)$	lift of a model $\mathcal{M}$ to N conditioned by X 138
ln	symbol for (natural) logarithm
$m^{A\uparrow}$, $m^{A\downarrow}$	identificators of supersets and subsets of a set A 34
m_l, m_*	(level-) degree identifiers 57 (59)
m_ℓ, m_u, m_o	elements of the ℓ-, u- and o-skeleton corresponding to m 80-81
m_P	multiinformation function induced by P 26
m_T	restriction of a supermodular function m to T 136
$m_\dagger$, $m_\circ$	special supermodular functions 67 (Figure 4.3) 95 (Figure 6.3)
$m_{T\|X}$	contraction of a function m to T conditioned by X 135
$\mathcal{M}^m$	independence model produced by a supermodular function m 72
$\mathcal{M}_{\boldsymbol{o}}$ ($\mathcal{M}_P$, $\mathcal{M}_G$, $\mathcal{M}_u$)	independence model induced by $\boldsymbol{o}$ (P, G, u) 14 37 40 43 64
$\mathcal{M}_T$	restriction of a model $\mathcal{M}$ to T 15 135
$\mathcal{M}_{[T]}$	model induced by the contraction $u_{[T]}$ of $u \in \mathcal{S}(N)$ to T 137
$\mathcal{M}_{T\|X}$	contraction of a model $\mathcal{M}$ to T conditioned by X 132
$\mathcal{M}_{T\|*}$	full-conditioned contraction of a model $\mathcal{M}$ to T 136
max, $\mathcal{D}^{\max}$	maximum, the class of maximal sets in $\mathcal{D}$ 151
min, $\mathcal{D}^{\min}$	minimum, the class of minimal sets in $\mathcal{D}$ 151
N	generic symbol for a non-empty finite set of variables 12
$\mathbb{N}$	the set of natural numbers 12
$\mathcal{N}(r,s)$	one-dimensional Gaussian measure 165
$\mathcal{N}(\boldsymbol{e}, \boldsymbol{\Sigma})$	Gaussian measure with expectation $\boldsymbol{e}$ and covariance matrix $\boldsymbol{\Sigma}$ 165 (28)
o-	generic symbol for 'orthogonal' standardization 35 75
P	generic symbol for a probability measure over N 12
P^A (μ^A)	marginal of a measure P (μ) for a set A 12
$P^{\mathcal{A}}$	restriction of P to a σ-algebra $\mathcal{A}$ 158
$\mathcal{P}(N)$ $\mathcal{P}(X)$	the power set 151 (34)
$P_{A\|C}$	conditional probability on X_A given C 12 (166)
$\bar{P}_{A\|C}$	its regular version (in Gaussian case 29)
$P(\mathsf{B}\|\mathcal{A})$	conditional probability of B given $\mathcal{A}$ with respect to P 158
P-a.e. (μ-a.e.)	almost everywhere with respect to P (μ) 156
$pa_G(b)$	the set of parents of a node b in a graph G 154
$\mathbb{Q}$, $\mathbb{Q}^n$	the set of rational numbers 12, rational vectors 162

$\mathbb{R}$, $\mathbb{R}^n$	the set of real numbers 12, real vectors 155
$\mathbb{R}^{\mathcal{P}(N)}$	the class of real functions on the power set $\mathcal{P}(N)$
$\mathcal{R}_u$	region of a structural imset u 102
R_u	range of a structural imset u 60
$R_{\boldsymbol{x}}$	ray generated by a vector $\boldsymbol{x}$ 163
$(\mathbb{R},\mathcal{B})$, $(\mathbb{R}^n,\mathcal{B}^n)$	the space of real numbers (vectors) with Borel σ-algebra 156 (28)
$\mathcal{S}$	(occasionally) the class of separators of a triangulated graph 45
$\mathcal{S}(N)$	the class of structural imsets over N 59
$\mathcal{S}_{\Psi}(N)$	the class of Ψ-representable structural imsets over N 98
$so\,(\mathcal{M},N)$	solid extension of a model $\mathcal{M}$ to N 138
sup	supremum, the least upper bound 152
$\mathcal{T}(N)$	the class of disjoint triplets over N 14
$\mathcal{T}_{\emptyset}(N)$, $\mathcal{T}_{\epsilon}(N)$	the classes of trivial and elementary triplets 16
u-	generic symbol for 'upper' standardization 36 74
u^+, u^-	positive and negative part of an imset u 34
u_G, u_H	standard imsets for graphs G and H 107 110
u_T	restriction of an imset u to T 136
$u_{[T]}$	contraction of a structural imset u to T 137
$u_{\langle A,B\|C\rangle}$ $u_{\langle i,j\|K\rangle}$	(semi-) elementary imset 58 57
$\mathcal{U}_u$	the upper class of a structural imset u 60
$\mathcal{U}(N)$	the class of structural models over N 85
$U(x,\varepsilon)$	ball with center x and radius ε 155
$\mathcal{W}(N)$	the class of coportraits over N 126
x_A	projection of a configuration x onto a set A 20
X_A	generic symbol for a sample space for A 12
$\mathcal{X}_A$, $\bar{\mathcal{X}}_A$	coordinate σ-algebra for A 12, its isomorphic representation 13
$(\mathsf{X},\mathcal{X})$ $(\mathsf{X}_A,\mathcal{X}_A)$	generic symbol for a measurable space 155 conventional notation 12
$\mathcal{X}_{\emptyset}$	trivial σ-algebra 12 (155)
$X^{\triangleright}$, $Y^{\triangleleft}$	Galois connection 84
$\mathbb{Z}$, $\mathbb{Z}^+$	the set of integers, the set of non-negative integers 12
$\mathbb{Z}^{\mathcal{P}(N)}$	the class of imsets over N 34

Bibliography

[1] M. Aigner: Combinatorial Theory, Springer-Verlag 1979.

[2] Z. An, D. A. Bell, J. G. Hughes: On the axiomatization of conditional independence, Kybernetes 21 (1992), n. 7, pp. 48-58.

[3] J. Anděl: Mathematical Statistics (in Czech), SNTL (Prague) 1985.

[4] S. A. Andersson, M. D. Perlman: Lattice models for conditional independence in multivariate normal distributions, Annals of Statistics 21 (1993), pp. 1318-1358.

[5] S. A. Andersson, D. Madigan, M. D. Perlman: A characterization of Markov equivalence classes for acyclic digraphs, Annals of Statistics 25 (1997), n. 2, pp. 505-541.

[6] S. A. Andersson, D. Madigan, M. D. Perlman, C. M. Triggs: A graphical characterization of lattice conditional independence models, Annals of Mathematics and Artificial Intelligence 21 (1997), pp. 27-50.

[7] S. A. Andersson, D. Madigan, M. D. Perlman: Alternative Markov properties for chain graphs, Scandinavian Journal of Statistics 28 (2001), n. 1, pp. 33-85.

[8] G. Birkhoff: Lattice Theory, AMS Colloquim Publications 1951.

[9] P. Boček: GENERATOR, a computer program, Institute of Information Theory and Automation, March 2001.

[10] R. R. Bouckaert: IDAGs: a perfect map for any distribution, in Symbolic and Quantitative Approaches to Reasoning and Uncertainty (M. Clarke, R. Kruse, S. Moral eds.), Lecture Notes in Computer Science 747, Springer-Verlag 1993, pp. 49-56.

[11] R. R. Bouckaert, M. Studený: Chain graphs: semantics and expressiveness, in Symbolic and Quantitative Approaches to Reasoning and Uncertainty (Ch. Froidevaux, J. Kohlas eds.), Lecture Notes in Artificial Intelligence 946, Springer-Verlag 1995, pp. 67-76.

[12] A. Brøndsted: An Introduction to Convex Polytopes, Springer-Verlag 1983 (Russian translation Mir 1988).

[13] L. M. de Campos, J. F. Huete, S. Moral: Probability intervals, a tool for uncertain reasoning, a technical report DECSAI-93206, July 1993, University of Granada.

[14] L. M. de Campos: Independency relationships in possibility theory and their application to learning belief networks, in Mathematical and Statistical Methods in Artificial Intelligence (G. Della Riccia, R. Kruse, R. Viertl eds.), Springer-Verlag 1995, pp. 119-130.

[15] L. M. de Campos: Characterization of decomposable dependency models, Journal of Artificial Intelligence Research 5 (1996), pp. 289-300.

[16] D. M. Chickering: A transformational characterization of equivalent Bayesian network structures, in Uncertainty in Artificial Intelligence 11 (P. Besnard, S. Hanks eds.), Morgan Kaufmann, 1995, pp. 87-98.

[17] R. G. Cowell, A. P. Dawid, S. L. Lauritzen, D. J. Spiegelhalter: Probabilistic Networks and Expert Systems, Springer-Verlag 1999.

[18] D. R. Cox, N. Wermuth: Multivariate Dependencies - Models, Analysis and Interpretation, Chapman and Hall 1996.

[19] A. P. Dawid: Conditional independence in statistical theory, Journal of the Royal Statistical Society series B 41 (1979), n. 1, pp. 1-31.

[20] A. P. Dawid: Conditional independence, in Encyclopedia of Statistical Science Update vol. 2 (S. Kotz, C. B. Read, D. L. Banks eds.), John Wiley 1999, pp. 146-155.

[21] A. P. Dawid, M. Studený: Conditional products, an alternative approach to conditional independence, in Artificial Intelligence and Statistics 99, Proceedings of the 7th Workshop (D. Heckerman, J. Whittaker eds.), Morgan Kaufmann 1999, pp. 32-40.

[22] A. P. Dawid: Separoids - a general framework for conditional independence and irrelevance, submitted to Annals of Mathematics and Artificial Intelligence.

[23] M. Fiedler: Special Matrices and Their Use in Numerical Mathematics (in Czech), SNTL (Prague) 1981.

[24] J.-P. Florens, M. Mouchart, J.-M. Rolin: Elements of Bayesian Statistics, Marcel Dekker 1990.

[25] M. Frydenberg: The chain graph Markov property, Scandinavian Journal of Statistics 17 (1990), n. 4, pp. 333-353.

[26] M. Frydenberg: Marginalization and collapsibility in graphical interaction models, Annals of Statistics 18 (1990), n. 2, pp. 790-805.

[27] L. C. van der Gaag, J.-J. Ch. Meyer: Informational independence, models and normal forms, International Journal of Intelligent Systems 13 (1998), n. 1, pp. 83-109.

[28] B. Ganter, R. Wille: Formal Concepts Analysis - Mathematical Foundations, Springer-Verlag 1999.

[29] D. Geiger: Towards the formalization of informational dependencie, technical report CSD-850053, R-102, UCLA Los Angeles, December 1987.

[30] D. Geiger, T. Verma, J. Pearl: Identifying independence in Bayesian networks, Networks 20 (1990), n. 5, pp. 507-534.

[31] D. Geiger, J. Pearl: On the logic of causal models, in Uncertainty in Artificial Intelligence 4 (R. D. Shachter, T. S. Lewitt, L. N. Kanal and J. F. Lemmer eds.), North-Holland 1990, pp. 3-14.

[32] D. Geiger, A. Paz, J. Pearl: Axioms and algorithms for inferences involving probabilistic independence, Information and Computation 91 (1991), n. 1, pp. 128-141.

[33] D. Geiger, J. Pearl: Logical and algorithmic properties of conditional independence and graphical models, Annals of Statistics 21 (1993), n. 4, pp. 2001-2021.

[34] D. Geiger, A. Paz, J. Pearl: On testing whether an embedded Bayesian network represents a probability model, in Uncertainty in Artificial Intelligence 10 (R. L. de Mantaras, D. Poole eds.), Morgan Kaufmann 1994, pp. 244-252.

[35] P. Hájek, T. Havránek, R. Jiroušek: Uncertain Information Processing in Expert Systems, CRC Press 1992.

[36] R. A. Horn, Ch. R. Johnson: Matrix Analysis, Cambridge University Press 1986 (Russian translation, Mir 1989).

[37] R. Jiroušek: Solution of the marginal problem and decomposable distributions, Kybernetika 27 (1991), pp. 403-412.

[38] K. G. Jöreskog, D. Sörbom: LISREL 7 - A Guide to the Program and Application, SPSS Inc. 1989.

[39] G. Kauermann: On a dualization of graphical Gaussian models, Scandinavian Journal of Statistics 23 (1996), n. 1, pp. 105-116.

[40] H. G. Kellerer: Verteilungsfunktionen mit gegebenem Marginalverteilungen (in German), Z. Wahrscheinlichkeitstheorie 3 (1964), pp. 247-270.

[41] H. Kiiveri, T. P. Speed, J. B. Carlin: Recursive causal models, Journal of Australian Mathematical Society series A 36 (1984), pp. 30-52.

[42] T. Kočka, R. R. Bouckaert, M. Studený: On the inclusion problem, research report n. 2010, Institute of Information Theory and Automation, Prague, February 2001.

[43] J. T. A. Koster: Gibbs factorization and the Markov property, unpublished manuscript.

[44] J. T. A. Koster: Markov properties of nonrecursive causal models, Annals of Statistics 24 (1996), n. 5, pp. 2148-2177.

[45] J. T. A. Koster: Marginalizing and conditioning in graphical models, submitted to Bernoulli.

[46] I. Kramosil: A note on non-axiomatizability of independence relations generated by certain probabilistic structures, Kybernetika 24 (1988), n. 2, pp. 439-446.

[47] S. L. Lauritzen, N. Wermuth: Mixed interaction models, research report R-84-8, Inst. Elec. Sys., University of Aalborg 1984.
Note that this report was later modified and became a basis of the paper [50].

[48] S. L. Lauritzen, T. P. Speed, K. Vijayan: Decomposable graphs and hypergraphs, Journal of Australian Mathematical Society A 36 (1984), n. 1, pp. 12-29.

[49] S. L. Lauritzen, D. J. Spiegelhalter: Local computation with probabilities on graphical structures and their application to expert systems, Journal of the Royal Statistical Society series B 50 (1988), n. 2, pp. 157-224.

[50] S. L. Lauritzen, N. Wermuth: Graphical models for associations between variables, some of which are qualitative and some quantitative, Annals of Statistics 17 (1989), n. 1, pp. 31-57.

[51] S. L. Lauritzen: Mixed graphical association models, Scandinavian Journal of Statistics 16 (1989), n. 4, pp. 273-306.

[52] S. L. Lauritzen, A. P. Dawid, B. N. Larsen, H.-G. Leimer: Independence properties of directed Markov fields, Networks 20 (1990), n. 5, pp. 491-505.

[53] S. L. Lauritzen: Graphical Models, Clarendon Press 1996.

[54] M. Levitz, M. D. Perlman, D. Madigan: Separation and completeness properties for AMP chain graph Markov models, submitted to Annals of Statistics.

[55] M. Loéve: Probability Theory, Foundations, Random Processes, D. van Nostrand 1955.

[56] F. M. Malvestuto: A unique formal system for binary decomposition of database relations, probability distributions and graphs, Information Sciences 59 (1992), pp. 21-52. + F. M. Malvestuto, M. Studený: Comment on "A unique formal ... graphs", Information Sciences 63 (1992), pp. 1-2.

[57] F. M. Malvestuto: Theory of random observables in relational data bases, Information Systems 8 (1983), n. 4, pp. 281-289.

[58] J. L. Massey: Causal interpretation of random variables (in Russian), Problemy Peredachi Informatsii 32 (1996), n. 1, pp. 112-116.

[59] F. Matúš: Ascending and descending conditional independence relations, in Information Theory, Statistical Decision Functions and Random Processes, Transactions of 11th Prague Conference, vol. B (S. Kubík, J. Á. Víšek eds.), Kluwer 1992, pp. 189-200.

[60] F. Matúš: On equivalence of Markov properties over undirected graphs, Journal of Applied Probability 29 (1992), n. 3, pp. 745-749.

[61] F. Matúš: Probabilistic conditional independence structures and matroid theory, backgrounds, International Journal of General Systems 22 (1994), n. 2, pp. 185-196.

[62] F. Matúš: Stochastic independence, algebraic independence and abstract connectedness, Theoretical Computer Science A 134 (1994), n. 2, pp. 445-471.

[63] F. Matúš: On the maximum-entropy extensions of probability measures over undirected graphs, in Proceedings of WUPES'94, September 11-15, 1994, Třešť, Czech Republic, pp. 181-198.

[64] F. Matúš, M. Studený: Conditional independences among four random variables I., Combinatorics, Probability and Computing 4 (1995), n. 4, pp. 269-278.

[65] F. Matúš: Conditional independences among four random variables II., Combinatorics, Probability and Computing 4 (1995), n. 4, pp. 407-417.

[66] F. Matúš: Conditional independence structures examined via minors, Annals of Mathematics and Artificial Intelligence 21 (1997), pp. 99-128.

[67] F. Matúš: Conditional independences among four random variables III., final conclusion, Combinatorics, Probability and Computing 8 (1999), n. 3, pp. 269-276.

[68] Ch. Meek: Causal inference and causal explanation with background knowlewdge, in Uncertainty in Artificial Intelligence 11 (P. Besnard, S. Hanks eds.), Morgan Kaufmann 1995, pp. 403-410.

[69] Ch. Meek: Strong completeness nad faithfulness in Bayesian networks, in Uncertainty in Artificial Intelligence 11 (P. Besnard, S. Hanks eds.), Morgan Kaufmann 1995, pp. 411-418.

[70] E. Mendelson: Introduction to Mathematical Logic, 2nd edition, D. van Nostrand 1979.

[71] M. Mouchart, J.-M. Rolin: A note on conditional independence with statistical applications, Statistica 44 (1984), n. 4, pp. 557-584.

[72] M. Mouchart, J.-M. Rolin: Letter to the editor, Statistica 45 (1985), n. 3, pp. 427-430.

[73] J. Moussouris: Gibbs and Markov properties over undirected graphs, Journal of Statistical Physics 10 (1974), n. 1, pp. 11-31.

[74] J. Neveu: Bases Mathématiques du Calcul des Probabilités (in French), Masson et Cie 1964.

[75] A. Paz: A fast version of Lauritzen's algorithm for checking representation of independencies, in Proceedings of 5th Workshop on Uncertainty Processing (WUPES'00) June 21-24, 2000, J. Hradec, Czech Republic, p. 181-190.

[76] A. Paz, R. Y. Geva, M. Studený: Representation of irrelevance relations by annotated graphs, Fundamenta Informaticae 42 (2000), pp. 149-199.

[77] J. Pearl, A. Paz: Graphoids, graph-based logic for reasoning about relevance relations, in Advances in Artificial Intelligence II (B. Du Boulay, D. Hogg, L. Steels eds.), North-Holland 1987, pp. 357-363.

[78] J. Pearl: Probabilistic Reasoning in Intelligent Systems - Networks of Plausible Inference, Morgan Kaufmann 1988.

[79] A. Perez: ε-admissible simplifications of the dependence structure of a set of random variables, Kybernetika 13 (1997), pp. 439-449.

[80] M. D. Perlman, L. Wu: Lattice conditional independence models for contingency tables with non-monotone missing data pattern, Journal of Statistical Planning 79 (1999), pp. 259-287.

[81] C. van Putten, J. H. van Shuppen: Invariance properties of conditional independence relation, Annals of Probability 13 (1985), n. 3, pp. 934-945.
Note that by [72] the majority of the results of this paper is almost identical with the results of [71].

[82] C. R. Rao: Linear Statistical Inference and Its Application, John Wiley 1965.

[83] T. S. Richardson: A polynomial-time algorithm for deciding Markov equivalence of directed cyclic graphical models, in Uncertainty in Artificial Intelligence 12 (E. Horvitz, F. Jensen eds.), Morgan Kaufmann 1996, pp. 462-469.

[84] T. S. Richardson: A discovery algorithm for directed cyclic graphs, in Uncertainty in Artificial Intelligence 12 (E. Horvitz, F. Jensen eds.), Morgan Kaufmann 1996, pp. 454-461.

[85] T. S. Richardson, P. Spirtes: Ancestral graph Markov models, report n. 374, Department of Statistics, University of Washington Seattle (2001), submitted to Annals of Statistics.

[86] J. Rosenmüller, H. G. Weidner: Extreme convex set functions with finite carrier - general theory, Discrete Mathematics 10 (1974), n. 3-4, pp. 343-382.

[87] W. Rudin: Real and Complex Analysis, McGraw-Hill 1974.

[88] Y. Sagiv, S. F. Walecka: Subset dependencies and completeness result for a subclass of embedded multivalued dependencies, Journal of Association for Computing Machinery 29 (1982), n. 1, pp. 103-117.

[89] G. Shafer: Probabilistic Expert Systems, CBMS-NSF Regional Conference Series in Applies Mathematics 67, SIAM 1996.

[90] A. Schrijver: Theory of Linear and Integer Programming, John Wiley 1986.

[91] L. S. Shapley: Cores of convex games, International Journal of Game Theory 1 (1971/1972), pp. 11-26.

[92] P. P. Shenoy: Conditional independence in valuation-based systems, International Journal of Approximate Reasoning 10 (1994), n. 3, pp. 203-234.

[93] P. P. Shenoy: Representing conditional independence relations by valuated networks, International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 2 (1994), pp. 143-165.

[94] P. Spirtes, C. Glymour, R. Scheines: Causation, Prediction and Search, Lecture Notes in Statistics 81, Springer-Verlag 1993.

[95] P. Spirtes: Directed cyclic graphical representations of feedback models, in Uncertainty in Artificial Intelligence 11 (P. Besnard, S. Hanks eds.), Morgan Kaufmann 1995, pp. 491-498.

[96] W. Spohn: Stochastic independence, causal independence and shieldability, Journal of Philosophical Logic 9 (1980), n. 1, pp. 73-99.

[97] W. Spohn: On the properties of conditional independence, in Patrick Suppes, Scientific Philosopher vol. 1, Probability and Probabilistic Causality (P. Humphreys ed.), Kluwer 1994, pp. 173-196.

[98] J. Štěpán: Probability Theory - Mathematical Foundations (in Czech), Academia (Prague) 1987.

[99] M. Studený: Asymptotic behaviour of empirical multiinformation, Kybernetika 23 (1987), n. 2, pp. 124-135.

[100] M. Studený: Multiinformation and the problem of characterization of conditional independence relations, Problems of Control and Information Theory 18 (1989), n. 1, pp. 3-16.

[101] M. Studený: Convex set functions I. and II., research reports n. 1733 and n. 1734, Institute of Information Theory and Automation, Prague, November 1991.

[102] M. Studený: Conditional independence relations have no finite complete characterization, in Information Theory, Statistical Decision Functions and Random Processes, Transactions of 11th Prague Conference, vol. B (S. Kubík, J. Á. Víšek eds.), Kluwer 1992, pp. 377-396.

[103] M. Studený: Multiinformation and conditional independence II., research report n. 1751, Institute of Information Theory and Automation, Prague, September 1992.

[104] M. Studený: Formal properties of conditional independence in different calculi of AI, in Symbolic and Quantitative Approaches to Reasoning and Uncertainty (M. Clarke, R. Kruse, S. Moral eds.), Lecture Notes in Computer Science 747, Springer-Verlag 1993, pp. 341-348.

[105] M. Studený: Convex cones in finite-dimensional real vector spaces, Kybernetika 29 (1993), n. 2, pp. 180-200.

[106] M. Studený: Structural semigraphoids, International Journal of General Systems 22 (1994), n. 2, pp. 207-217.

[107] M. Studený, P. Boček: CI-models arising among 4 random variables, in Proceedings of WUPES'94, September 11-15, 1994, Třešť, Czech Republic, pp. 268-282.

[108] M. Studený: Description of structures of conditional stochastic independence by means of faces and imsets (a series of 3 papers), International Journal of General Systems 23 (1994/1995), n. 2-4, pp. 123-137, 201-219, 323-341.

[109] M. Studený: Semigraphoids and structures of probabilistic conditional independence, Annals of Mathematics and Artificial Intelligence 21 (1997), n. 1, pp. 71-98.

[110] M. Studený: A recovery algorithm for chain graphs, International Journal of Approximate Reasoning 17 (1997), n. 2-3, pp. 265-293.

[111] M. Studený: On marginalization, collapsibility and precollapsibility, in Distributions with Given Marginals and Moment Problems (V. Beneš, J. Štěpán eds.), Kluwer 1977, pp. 191-198.

[112] M. Studený, R. R. Bouckaert: On chain graph models for description of conditional independence structures, Annals of Statistics 26 (1998), n. 4, pp. 1434-1495.

[113] M. Studený: Bayesian networks from the point of view of chain graphs, in Uncertainty in Artificial Intelligence 14 (G. F. Cooper, S. Moral eds.), Morgan Kaufmann 1998, pp. 496-503.

[114] M. Studený: Complexity of structural models, in Prague Stochastics'98, August 23-28, Prague 1998, pp. 521-528.

[115] M. Studený, J. Vejnarová: The multiinformation function as a tool for measuring stochastic dependence, in Learning in Graphical Models (M. I. Jordan ed.), Kluwer 1998, pp. 261-298.

[116] M. Studený, R. R. Bouckaert, T. Kočka: Extreme supermodular set functions over five variables, research report n. 1977, Institute of Information and Automation, Prague, January 2000.

[117] J. Vejnarová: Conditional independence in possibility theory, in Proceedings of ISIPTA'99 (1st International Symposium on Imprecise Probabilities and their Applications) (G. de Cooman, F. G. Cozman, S. Moral, P. Walley eds.), pp. 343-351.

[118] I. Vajda: Theory of Statistical Inference and Information, Kluwer 1989.

[119] T. Verma, J. Pearl: Causal networks, semantics and expressiveness, in Uncertainty in Artificial Intelligence 4 (R. D. Shachter, T. S. Lewitt, L. N. Kanal and J. F. Lemmer eds.), North-Holland 1990, pp. 69-76.

[120] T. Verma, J. Pearl: Equivalence and synthesis of causal models, in Uncertainty in Artificial Intelligence 6 (P. P. Bonissone, M. Henrion, L. N. Kanal, J. F. Lemmer eds.), Elsevier 1991, pp. 220-227.

[121] T. Verma, J. Pearl: An algorithm for deciding if a set of observed independencies has a causal explanation, in Uncertainty in Artificial Intelligence 8 (D. Dubois, M. P. Wellman, B. D'Ambrosio, P. Smets eds.), Morgan Kaufmann 1992, pp. 323-330.

[122] M. Volf, M. Studený: A graphical characterization of the largest chain graphs, International Journal of Approximate Reasoning 20 (1999), n. 3, pp. 209-236.

[123] S. Watanabe: Information theoretical analysis of multivariate correlation, IBM Journal of Research and Development 4 (1960), pp. 66-81.

[124] J. Whittaker: Graphical Models in Applied Multivariate Statistics, John Wiley 1990.

[125] Y. Xiang, S. K. M. Wong, N. Cercone: Critical remarks on single link search in learning belief networks, in Uncertainty in Artificial Intelligence 12 (E. Horvitz, F. Jensen eds.), Morgan Kaufmann 1996, pp. 564-571.

Acknowledgements

Děkuji svým rodičům za trpělivost a všestrannou pomoc a podporu. Moje neteř Petra bude určitě potěšena když si přečte, že oceňuji její morální podporu.

Many people deserve my thanks for a help with this work. In particular, I would like to thank Marie Kolářová for typing the text of this work in LaTeX. As concerns expert help I am indebted to my colleagues (and former co-authors) František Matúš and Phil Dawid for their remarks (even for some critical ones made by Fero), various advices, pointers to literature and discussion which helped me clarify my view on the topic. I have also profited from cooperation with other colleagues: some results presented in this work were achieved with help of computer programs written by Pavel Boček, Remco Bouckaert, Tomáš Kočka, Martin Volf and Jiří Vomlel. Moreover, I am indebted to my colleagues Radim Jiroušek, Otakar Kříž and Jiřina Vejnarová for encouragement to write this work which was quite important for me. The cooperation with my colleagues mentioned above involved joint theoretical research as well. As concerns technical help I would like to thank to Václav Kelar for making special LaTeX fonts for me and to Jarmila Pánková for helping me to print several pages with coloured pictures.

Finally, I am also indebted to other colleagues all over the world whose papers or books inspired me somehow in connection with this work. In particular, I would like to mention my PhD supervisor Albert Perez. However, except those who were mentioned above many others influences me. Let me name some of them: Steen Andersson, Luis de Campos, Robert Cowell, David Cox, Morten Frydenberg, Dan Geiger, Tomáš Havránek, Jan Koster, Ivan Kramosil, Steffen Lauritzen, Franco Malvestuto, Michel Mouchart, Chris Meek, Azaria Paz, Judea Pearl, Michael Perlman, Jean-Marie Rolin, Thomas Richardson, Jim Smith, Glenn Shafer, Prakash Shenoy, David Spiegelhalter, Peter Spirtes, Wolfgang Spohn, Nanny Wermuth, Joe Whittaker, Raymond Yeung and Zhen Zhang. The above list is not exhaustive; I apologize to those who were possibly omitted.

Let me emphasize that I profited from meeting a lot of colleagues who gave me inspiration during the seminar "Conditional Independence Structures" held from September 27 to October 17, 1999 in the Fields Institute for Research in Mathematical Sciences, Toronto, Canada, and during several events organized within the framework of ESF program "Highly Structured Stochastic Systems" for 1997-2000. In particular, let me thank to Helène Massam and Steffen Lauritzen who gave me a chance to participate actively in these wonderful events. For example, I remember stimulating atmosphere of HSSS research kitchen "Learning conditional independence models" held in Třešť, Czech Republic, in October 2000.