Institute of Information Theory and Automation
Academy of Sciences of the Czech Republic

On probabilistic conditional independence structures

Milan Studený

Prague, January 2003

The 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

The monograph was written in the Department of Decision-Making Theory of the Institute of Information Theory and Automation (Academy of Sciences of the Czech Republic) in Prague and 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).

Preface

The book is devoted to the mathematical description of probabilistic conditional independence structures. The topic of conditional independence, which belongs both within the scope of statistics and within the scope of artificial intelligence, has been in the center of my research activity for many years - since the late 1980's. I have been primarily influenced by researchers working in the area of graphical models but I gradually realized that the concept of conditional independence is not necessarily bound to the idea of graphical description and may have a broader impact. This observation led me to an attempt to develop a non-graphical method of description of probabilistic conditional independence structures which, in my view, ovecomes inherent limitation of graphical approaches. The method of structural imsets described in this book can be viewed as an algebraic approach to the description of conditional independence structures although it remains within the framework of discrete mathematics.

The basic idea of this approach was already presented in the middle of the 1990's - in a series of papers [129]. However, I was not satisfied with the original presentation of the approach for several reasons. First, the series of papers only dealt with the discrete case which is a kind of imperfection from the point of view of statistics. Second, the main message was dimmed by unnecessary mathematical peculiarities and important ideas were not perhaps pinpoited clearly. Third, the motivation was not explained in details. Another reason why I think that the original series of papers was difficult to read for researchers in the area of artificial intelligence is that 'practical' implementation aspects of the presented approach were suppressed there and the pictorial representation of considered mathematical objects, to which researchers interested in graphical models are accustomed to, was omitted.

Within the next five years further mathematical results were achieved which amended, supplemented and gave more precision of the original idea. I have also deliberated about suitable terminology and the way of presentation of the method of structural imsets which would be both acceptable to statisticians and researchers in the area of artificial intelligence and exact from the mathematical point of view as well. I wrote it up in my DrSc thesis [138] which became the basis of this monograph. Finally, after finishing the thesis I realized potential future practical application of the method to learning graphical models and decided to emphasize this by writing an additional chapter.

Thus, the aim of the monograph is to present the method of structural imsets in its full (present) extent which includes the motivation, the mathematical foundations, which I tried to present in a didactic form, indication of the area of application and open problems. More specifically, the motivation is explained in the first chapter. The second chapter recalls the basic concepts in the area of probabilistic conditional independence structures. The third chapter is an overview of classic graphical methods of description of conditional independence structures. The core of the method of structural imsets is presented in the next four chapters. The eighth chapter indicates the application of the method to learning graphical models. Open problems are gathered in the ninth chapter and needed elementary mathematical notions are recalled in the Appendix for readers convenience.

The book is intended for

- mathematicians who may be attracted by this particular application of mathematics in the area of artificial intelligence and statistics,
- researchers in statistics and informatics who may become interested in deeper mathematical groungings of the theory of (graphical) models of conditional independence structures,
- PhD students in the fields of mathematics, probability, statistics, informatics and computer science who may find inspiration in the book and perhaps make some progress either by solving open problems or by applying the presented theory in practice.

Many people deserve my thanks for a help with this piece of work. In particular, I would like to thank Marie Kolářová for typing the text of the monograph in LaTeX. As concerns expert help I am indebted to my colleagues (and former co-authors) Fero Matúš and Phil Dawid for their remarks (even for some critical ones made by Fero), various pieces of advice, pointers to the literature and discussion which helped me clarify the view on the topic of the book. I have also profited from the cooperation with other colleagues: some results presented here were achieved with the 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 their encouragement in writing my DrSc thesis which was quite important for me. The cooperation with all my colleagues mentioned above involved joint theoretical research as well. As concerns the 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 prepare several pages with colourful pictures. I am gratelful to Antonín Otáhal and Cheri Dohnal for correcting my (errors in) English.

Finally, I am also indebted to other colleagues all over the world whose papers, theses and books inspired me somehow in connection with this monograph. In particular, I would like to mention my PhD supervisor Albert Perez. However, except those who were already mentioned above many other colleagues influenced me. Let me name some of them: Steen Andersson, Luis de Campos, Max Chickering, 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 the HSSS research kitchen "Learning conditional independence models" held in Třešť, Czech Republic, in October 2000.

Prague, January 2003 M.S.

Special acknowledgements in Czech

Na tomto místě bych rád poděkoval zejména svým rodičům za trpělivost, lásku a všestrannou pomoc a podporu, kterou mi poskytli. Můj bratr a jeho rodina, zvláště pak moje neteř Petra, budou určitě potěšeni, když si přečtou, že oceňuji jejich morální podporu.

Contents

Chapter 1

Introduction

The central topic of this book 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 monograph 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

- the analysis of *contingency tables* which is an area of discrete statistics dealing with categorical data,
- the *multivariate analysis* which is a branch of statistics investigating mutual relationships among continuous real-valued variables,
- the *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 book 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 monograph 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 book, for example, to probability theory, mathematical statistics, topology and mathematical logic.

1.1 Motivation thoughts

The reader is asked to excuse the following 'methodological' considerations which perhaps explain 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.

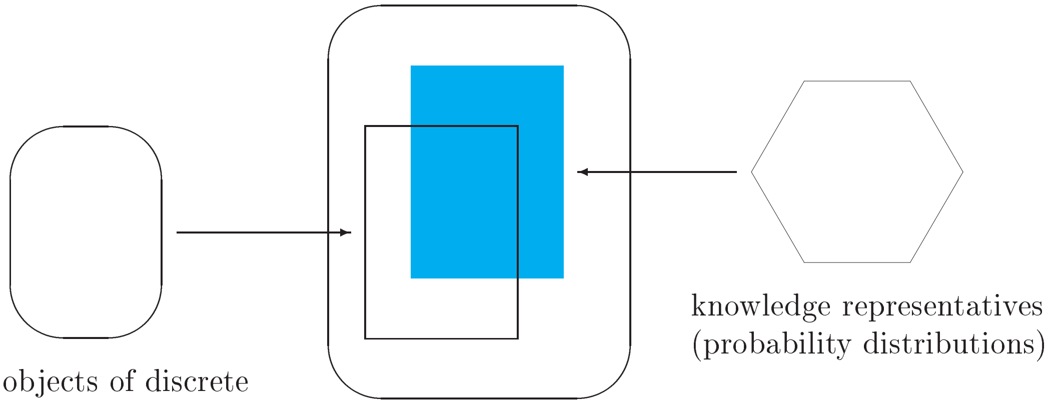


Figure 1.1: Theoretical groundings (an informal illustration).

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. 18), the class of non-degenerate Gaussian measures over N (see p. 34), the class of CG measures over N (see p. 61) or any specific parametrized class of distributions. In other words, certain *distribution framework* is specified (see Section 10.10.3). 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 p. 19 in Section 2.2.1). 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 and discrete functions, are used to describe CI structures. Thus, certain *universum of objects of discrete mathematics* is specified. 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' an induced independence model so that it can possibly describe one of the CI structures to be described.

The definition of the induced formal independnece model depends on the type of considered objects. Every particular universum of objects of discrete mathematics 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 the 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 universum 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 the 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 object from the considered universum of objects of discrete mathematics 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. If this is not the case an advanced subtask occurs, namely the task to characterize conveniently those formal independence models which can be described by the objects from the considered universum.
- *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 [115] 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 distribution framework (= relative to the class of discrete measures) and faithfulness in Gaussian distribution framework. Note that for classic graphical models the faithfulness is usually ensured while the 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, used also a traditional term "(strong) completeness of a separation graphical criterion" [42, 87, 133, 70]. However, according to the classification above, results of this type belong to the results gathered under the label 'faithfulness' (customary reasons for traditional terminology are explained in Remark 3.2 on p. 44). Thus, I distinguish between the 'completeness of a criterion' on one hand and the 'completeness of an universum of objects' (for the 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 problem of graphical characterization of inclusion quasiordering (see Section 8.3 and p. 164) 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

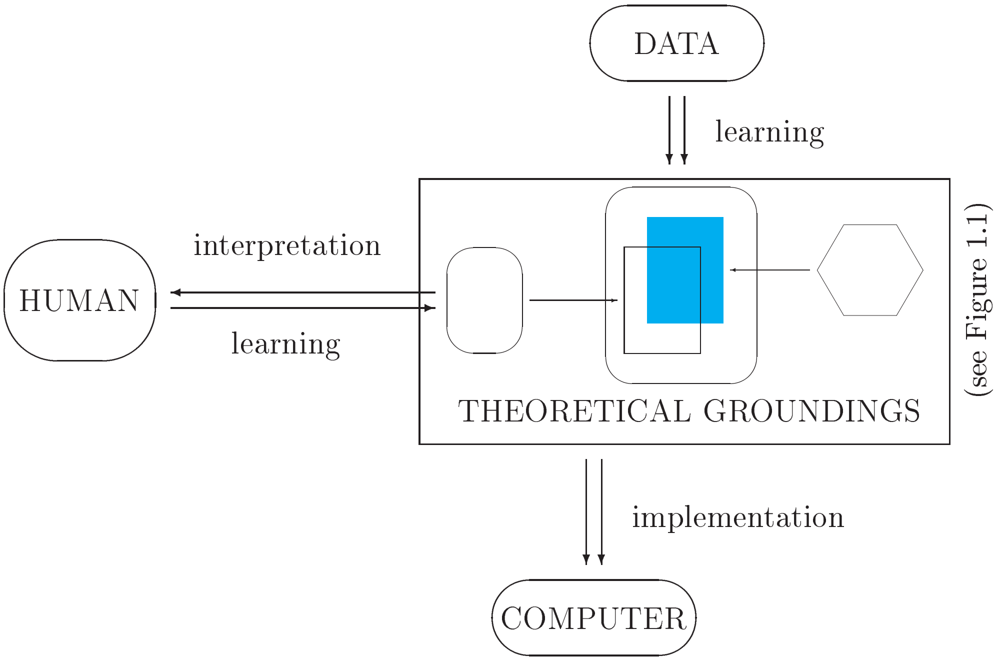


Figure 1.2: Practical questions (an informal illustration).

can be visualized in a way that they are understood easily and interpreted correctly as CI structures.

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

- *Implementation* is the question how to manage the corresponding computational tasks. An advanced subquestion is whether the acceptance of a particular CI structure allows one to do respective subsequent calculation with probability distributions effectively, namely whether the cosidered objects of discrete mathematics 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 [115] 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' [63] which is behind several working probabilistic expert systems [24, 46]. 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. 179) which has also a strong computational aspect.

The goal of these motivation thoughts is the idea that the practical questions are ultimately connected with the theoretical groundings. 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 the 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. 45) 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 Section 3.6). 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 the 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 the CI structure which can be described by a graph within the considered universum of graphs. However, in general, this edge removal represents the 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 again 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 universum of graphical models which is not complete. Note that an attitude like this was already criticized within the community of reseachers in artificial intelligence (see [147] and Remark 8.1). In my opinion, these repeated problems in solving practical question of learning are inevitable consequences of the 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 book, I propose a non-graphical method of description of probabilistic CI structures which primarily solves the completeness and has the potential to take care of practical questions.

1.2 Goals of the monograph

The aim of the book 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 author's earlier papers.

The second goal is to present the mathematical basis of an alternative method of description of probabilistic CI structures. The alternative method of structural imsets 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 text of the monograph is perhaps longer and more detailed than it could be. The reason is that not only experts in the fields and mathematicians are the expected audience. The 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 the author to solve the dilemma 'understandability' versus 'conciseness' in favour of preciseness and understandability.

1.3 Structure of the book

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 book.

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 the completeness (see p. 12) is needed.

Chapter 4 introduces a method of this type, namely the method of *structural imsets*. A class of distributions to which this method is applicable is specified - it is the class of distributions with finite multiinformation - and the concept of structural imset is defined. The main result of the chapter (Theorem 4.1) says that three possible ways of associating probability distributions and structural imsets are equivalent.

Chapter 5 compares two different, but equivalent, *ways of describing* CI structures by means of imsets. An algebraic point of view is emphasized in the chapter. It is shown there that every probabilistic CI structure induced by a distribution with finite multiinformation can be described by the method of structural imsets. Moreover, a duality relation between those two ways of description is established. An unifying point of view provided by the theory of formal concept analysis is offered.

Chapter 6 is devoted to an advanced question of Markov equivalence (see p. 12) within the framework of structural imsets. A characterization of equivalent imsets is given there and a lot of attention is devoted to implementation tasks. The respective *facial implication* of structural imsets is characterized in two different ways. One of them allows one to transform the task of computer implementation of facial implication into a standard task of integer programming. Moreover, the question of adaptation of the method of structural imsets to a particular distribution framework is discussed there (Section 6.5).

Chapter 7 deals with the problem of choice of a suitable representative of a class of equivalent structural imsets. Two approaches to this problem are offered. The concept of *baricentral imset* seems to be a good solution from the theoretical point of view in the general context while the concept of *standard imset* for an acyclic directed graph seems to be advantageous in the context of classic graphical models.

Chapter 8 concerns the question of learning. It is more likely an analytic review of methods of learning graphical models than a pure mathematical text. However, the goal is to show that the method of structural imsets can be applied in this area too. A solution to the problem of characterization of inclusion quasiordering is offered and the significance of standard imsets in the context of learning is explicated (Section 8.4).

Chapter 9 is an overview of open problems to be studied in order to tackle practical question (see p. 12-13).

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 the help of the Index.

Chapter 2

Basic concepts

Throughout the book 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 a graph. The set N will also serve as the basic set for non-graphical tools of discrete mathematics introduced in this monograph (semi-graphoids, imsets etc.).

The following convention will be used throughout the book: 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 the monograph 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 of these measurable spaces $(\prod_{i\in N} \mathsf{X}_i, \prod_{i\in N} \mathcal{X}_i)$. In this case I will use the symbol $(\mathsf{X}_A, \mathcal{X}_A)$ 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

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

Moreover, let us accept two natural 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 only admits 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

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

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. 191) is needed within this general approach. Let me emphasize that $P_{A|C}$ only depends on the marginal P^{AC} 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, if $C = \emptyset$ then 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

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

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 the uniqueness of $P_{A|C}$ mentioned above follows from basic measure-theoretical facts, for details see the Appendix, Section 10.5. △

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 $\mathsf{B} \in \mathcal{X}_B$

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

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 probability given C since these are determined uniquely 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

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

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. Note that the fact that the equality P^C-a.e. concides with the equality on a certain fixed set is a speciality of the discrete case. Other common equivalent definitions of conditional independence are mentioned in Section 2.3. △

However, the concept of conditional independence is not exclusively a probabilistic concept. This concept was introduced in several non-probabilistic frameworks, namely in various calculi for dealing with uncertainty in artificial intelligence - for details and overview see [125, 29, 112]. 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 [27], then mentioned by Mouchart and Rolin [90], and van Putten and van Shuppen [100]. Spohn [117] 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 [96]. Their terminology [97] 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 [17, 139], relational databases over N [107] and 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 the 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 this set of triplets 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 $\mathcal{T}(N)$. A subset $\mathcal{M} \subseteq \mathcal{T}(N)$ interpreted in this way will be called the *formal independence model*. 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

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

By the *restriction* of a formal independence model $\mathcal{M}$ over N to a set $\emptyset \neq T \subseteq N$ will be undestood the set $\mathcal{M} \cap \mathcal{T}(T)$ 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 [27] 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 [37, 78]. However, then the statement $A \perp\!\!\!\perp A \,|\, C \; [P]$ has a 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. § 2 in [78]) 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 monograph are pure conditional independence structures; therefore I limit myself to pure conditional independence statements. △

Remark 2.4 To avoid misunderstanding let me explain that the noun 'model' can have three different meanings in this monograph. First, it can be used in its general sense in which case it is usually used without an adjective. Second, it is a part of the phrase '(formal) independence model' in which case the word 'independence' indicates that one has in mind the concept introduced in this section. Note that this terminology comes from the area of artificial intelligence - see [97]. Third, it can be a part of the phrase 'statistical model' in which case the adjective 'statistical' indicates that one has in mind the concept mentioned in Section 10.10.2, that is, a class of probability measures. Note that this terminology is often used in statistics - see Remark on p. 202.

However, there is a simple reason why two different concepts are named by the same noun. The reason is that every formal independence model $\mathcal{M} \subseteq \mathcal{T}(N)$ can be understood as a statistical model $\mathbb{M}$, provided that a distribution framework Ψ (see Section 10.10.3) is fixed. Indeed, one can put

$$\mathbb{M} \; = \; \{ \, P \in \Psi \, ; \;\; A \perp\!\!\!\perp B \,|\, C \; [P] \quad \text{whenever } \langle A, B|C \rangle \in \mathcal{M} \, \} \, .$$

Every statistical model of this kind will be called the *statistical model of CI structure*. Note that this concept generalizes the classic concept of graphical model [146, 67]. Indeed, the reader can learn in Chapter 3 that a graph G having N as the set of nodes usually induces the class $\mathbb{M}_G$ of Markovian measures over N, that is, a statistical model. This graphical statistical model is, however, defined by means of the formal independence model $\mathcal{M}_G$. Note for explanation that the class $\mathbb{M}_G$ can also typically be introduced in another way - see p. 139 for an equivalent definition in case of acyclic directed graphs. △

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 $A \perp\!\!\!\perp \emptyset \,|\, C \; [\mathcal{M}]$,
2. symmetry $A \perp\!\!\!\perp B \,|\, C \; [\mathcal{M}]$ implies $B \perp\!\!\!\perp A \,|\, C \; [\mathcal{M}]$,
3. decomposition $A \perp\!\!\!\perp BD \,|\, C \; [\mathcal{M}]$ implies $A \perp\!\!\!\perp D \,|\, C \; [\mathcal{M}]$,
4. weak union $A \perp\!\!\!\perp BD \,|\, C \; [\mathcal{M}]$ implies $A \perp\!\!\!\perp B \,|\, DC \; [\mathcal{M}]$,
5. contraction $A \perp\!\!\!\perp B \,|\, DC \; [\mathcal{M}]$ 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 [97] 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. 193). 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 $\bar{\mathcal{X}}_{AB} = \bar{\mathcal{X}}_A \vee \bar{\mathcal{X}}_B$ for $A, B \subseteq N$ the rest follows from Consequence 10.1. □

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

REMARK 2.5 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 [31] which is a mathematical structure unifying variety of notions of 'irrelevance' arising in probability, statistics, artificial intelligence and other fields. △

2.2.3 Elementary independence statements

Well, to store a semi-graphoid over N in memory of a computer one need not to 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, the 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. [3, 76]). 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

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

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

Proof: (see also [76]) The necessity of the condition (2.2) is easily derivable using the decomposition and the weak union properties combined with the symmetry property.

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 or B is not a singleton. Owing to the symmetry property one can consider without the loss of generality $|B| \geq 2$, choose $b \in B$ and put $B' = B \setminus \{b\}$. By the induction assumption (2.2) implies both $A \perp\!\!\!\perp b \,|\, B'C\ [\mathcal{M}]$ 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 [76].

Remark 2.6 Another reduction of memory demands for semi-graphoid representation follows from the symmetry. Instead of keeping a pair of mutually symmetric statements $a \perp\!\!\!\perp b \,|\, K$ and $b \perp\!\!\!\perp a \,|\, K$ 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 $\langle a, b|K\rangle$ over N generates a semi-graphoid over N consisting of $\langle a, b|K\rangle$, its symmetric image $\langle b, a|K\rangle$ 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 [96, 97] formulated a conjecture that semi-graphoids coincide with conditional independence models induced by discrete probability measures. However, this conjecture was refuted in [121] by finding a further formal property of these models, which is 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 [3]. 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* because a result of this kind would have been a substantial step towards a syntactic description of these models in the sense of mathematical logic. Indeed, as explained in § 5 of [123], then it would have been possible to construct a deductive system which is an analogue of the notion "formal axiomatic theory" from [89]. 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 [123] (for a more didactic proof see [136]) 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 [81, 82, 84]; for an overview see [128] 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 2.7 On the other hand, several results on relative completeness of semi-graphoid properties were achieved. In [43] and independently in [79] models of "unconditional" stochastic independence (that is, submodels consisting of *unconditioned independence statements* of the form $A \perp\!\!\!\perp B \,|\, \emptyset$) were characterized by means of properties derivable from the 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 [44, 73]. As a specific relative completeness result can be interpreted the result from [130] saying that the semi-graphoid generated by a pair of conditional independence statements is always a conditional independence model induced by a discrete probability measure. Note that the problem of axiomatic characterization of CI models mentioned above differs from the problem of axiomatization (in the sense of mathematical logic) of a single CI structure over an infinite set of variables N which was treated in [59]. △

2.3 Classes of probability measures

There is no uniformly accepted conception of the notion of *probability distribution* in the 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 book; 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 more likely 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 also be 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 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 [121], 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 the induction on $|N|$ to get the desired conclusion. □

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

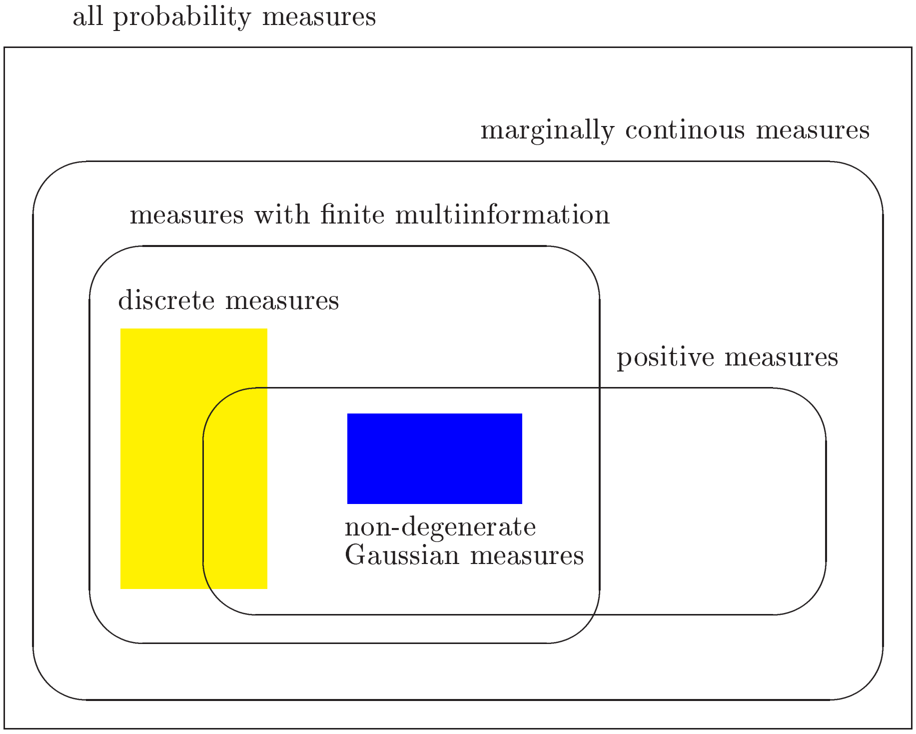


Figure 2.1: The relation of basic classes of probability measures over N.

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 2.8 Let us note without details (see Remark 1 in [121]) 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 [71]. 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 the 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 [33]. $\triangle$

The next step is 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 $\emptyset \neq A \subseteq B \subseteq N$ 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 2.9 This is to explain the way of definition of marginal densities in Convention 1. First, let me emphasize that the marginal density is not 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 the 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 the set $\{x \in \mathsf{X}_N;\, f^{\downarrow C}(x_C) = 0\}$ 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 can conclude that the condition (2.1) from p. 18 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, the factorization after $\mathcal{D}$ is equivalent to the 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 [67], § 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. 26), 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

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

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 the discrete case (see [3, 118] and Proposition 4.1 in [78]).

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

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

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. The *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 2.10 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' [145], 'dependence tightness' [98] or 'entaxy' [74]. 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 [136]). Asymptotic behaviour of 'empirical multiinformation' which can be used as a statistical estimate of multiinformation on basis of data was examined in [120]. △

To clarify the significance of multiinformation for the 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

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

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)

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

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 the 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. 26) 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

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

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

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

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. 193 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 $D \subseteq N$). Thus, (2.13) was derived. □

REMARK 2.11 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 if joint measures of this type exist then it may happen that none of them complies with the required conditional independence statement. For both examples see [30].

Thus, the assumption of marginal continuity implies the existence of the conditional product. Note that the regularity of conditional probabilities $P_{A|C}$ or $P_{B|C}$ is a more general sufficient condition for the existence of the conditional product (see Proposition 2 in [121]). 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 [136]. △

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,

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

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

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

Proof: The relation (2.15) is evident. Given $S \subseteq N$, put $\langle A, B|C\rangle = \langle S, N \setminus S \,|\, \emptyset\rangle$ 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

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

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

Thus, the class of probability measures over N having finite multiinformation is, by the 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 the relations (2.16) and (2.17) establish a basic method for handling conditional independence used in this monograph. Because these relations originate from information theory (the expresion in (2.16) is nothing but the conditional mutual information mentioned in Remark 2.11) 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 these tools. Consequence 2.2 also implies that the class of measures with finite multiinformation is closed under the operation of taking marginals. Note without further explanation 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 $-\infty < H(P^S \,|\, \prod_{i\in S} \mu_i) < \infty$ and $-\infty < H(P^{\{i\}} \,|\, \mu_i) < \infty$ for every $i \in S$. Then $0 \leq m_P(S) < \infty$ and

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

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 the positivity of density may depend on the choice of a 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 (the probabilistic approach to) artificial intelligence. Pearl [97] noticed that conditional independence models induced by these measures satisfy further special formal property (except the 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 [67] 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. 26). 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. □

Let us note that there are discrete probability measures whose induced conditional independence model is not a graphoid, that is, it does not satisfy the intersection property (see Example 2.3 on p. 38). 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 book *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 the 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 to be 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, if $\boldsymbol{\Sigma}$ is not regular then 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 2.3 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

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

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 $\boldsymbol{\Sigma}_{A|\emptyset} = \boldsymbol{\Sigma}_{A\cdot A}$ by convention. Elements of miscellaneous conditional covariance matrices can serve as convenient parameters of Gaussian measures - e.g. [8].

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 sole 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. 191) 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 $\forall\, \mathsf{A} \in \mathcal{X}_A$, $\forall\, \mathsf{B} \in \mathcal{X}_B$,

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

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 [119]). 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 <u>every</u> $\mathsf{A} \in \mathcal{Y}_A$ 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,

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

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 consist of zeros (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 [67] or Corollaries 6.3.3 and 6.3.4 in [146]).

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

$$a \perp\!\!\!\perp b \,|\, \emptyset \; [P] \;\Leftrightarrow\; \sigma_{ab} = 0 \;\Leftrightarrow\; \varrho_{ab} = 0\,,$$

and for distinct $a, b, c, \in N$

$$a \perp\!\!\!\perp b \,|\, \{c\} \; [P] \;\Leftrightarrow\; \sigma_{cc} \cdot \sigma_{ab} = \sigma_{ac} \cdot \sigma_{cb} \;\Leftrightarrow\; \varrho_{ab} = \varrho_{ac} \cdot \varrho_{cb}\,.$$

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

$$a \perp\!\!\!\perp b \,|\, N \setminus \{a, b\} \; [P] \;\Leftrightarrow\; \kappa_{ab} = 0\,.$$

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,

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

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 $\{a,b\}$ (see Section 10.9.1). Hence $((\mathbf{\Sigma}_{D|N\setminus D})^{-1})_{ab} = \kappa_{ab}$ for every such D. □

Remark 2.12 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

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

in the 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 the Gaussian case, specified by 'less' number of requirements than in the 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. △

Thus, one can expect many special 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 [97] 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 § 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 $(\mathbf{\Sigma}_{ABD|C})_{A\cdot BD} = \mathbf{0}$. □

However, the composition property 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

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

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: $(0,0,0)$, $(0,1,1)$, $(1,0,1)$, $(1,1,0)$. $\diamondsuit$

Another special property of Gaussian conditional independence models is the following one which was also mentioned by Pearl [97] in the context of graphical models.

8. weak transitivity
   $A \perp\!\!\!\perp B \,|\, C \; [\mathcal{M}]$ and $A \perp\!\!\!\perp B \,|\, Cd \; [\mathcal{M}]$ implies $A \perp\!\!\!\perp d \,|\, C \; [\mathcal{M}]$ or $d \perp\!\!\!\perp B \,|\, C \; [\mathcal{M}]$

for pairwise disjoint $A, B, C \subseteq N$, $d \in N \setminus ABC$.

Consequence 2.5 Let P be a Gaussian measure over N, sets $A, B, C \subseteq N$ are pairwise disjoint and $d \in N \setminus ABC$. Then

$$A \perp\!\!\!\perp B \,|\, C \; [P] \;\&\; A \perp\!\!\!\perp B \,|\, Cd \; [P] \;\; \Rightarrow \;\; \{\; A \perp\!\!\!\perp d \,|\, C \; [P] \text{ or } d \perp\!\!\!\perp B \,|\, C \; [P] \;\}.$$

Proof: It suffices to assume that A and B are singletons. Indeed, owing to Consequence 2.4 (and semi-graphoid properties) $A \perp\!\!\!\perp B \,|\, C$ is equivalent to the requirement condition $[a \perp\!\!\!\perp b \,|\, C\}$ for every $a \in A$, $b \in B]$ and similar observation is valid for all four involved CI statements. There is no pair $a \in A$, $b \in B$ with $\neg\{\, a \perp\!\!\!\perp d \,|\, C\,\}$ and $\neg\{\, d \perp\!\!\!\perp b \,|\, C\,\}$ because this contradicts the fact $\{\, a \perp\!\!\!\perp b \,|\, C \;\&\; a \perp\!\!\!\perp b \,|\, Cd\,\}$ implied by semi-graphoid properties. In other terms, either $[\forall\, a \in A \;\, a \perp\!\!\!\perp d \,|\, C]$ or $[\forall\, b \in B \;\, d \perp\!\!\!\perp b \,|\, C]$ and one can again use Consequence 2.4 to get the desired conclusion.

Lemma 2.8 allows one to reduce the general case to the case $C = \emptyset$. Indeed, one can consider $\mathbf{\Sigma}_{N \setminus C | C}$ in place of $\mathbf{\Sigma}$ which is also a positive semi-definite matrix (see Section 10.9.1) and therefore it is a covariance matrix of a Gaussian measure over $N \setminus C$ (see Section 10.9.3).

If $A = \{a\}$, $B = \{b\}$ and $C = \emptyset$ then apply Consequence 2.3 to the correlation matrix $\mathbf{\Gamma} = (\varrho_{ij})_{i,j \in N}$ of P: $0 = \varrho_{ab} = \varrho_{ad} \cdot \varrho_{db}$. Hence $\varrho_{ad} = 0$ or $\varrho_{db} = 0$ which yields the desired fact. $\square$

The above consequence makes it possible to construct the following example.

Example 2.2 There exists a pair P, Q of non-degenerate Gaussian measures over N with $|N| = 3$ such that $\mathcal{M} = \mathcal{M}_P \cap \mathcal{M}_Q$ is not CI model induced by any Gaussian measure over N. Indeed, put $N = \{a, b, c\}$ and define matrices $\mathbf{\Sigma} = (\sigma_{ij})_{i,j \in N}$ and $\mathbf{\Sigma}' = (\sigma'_{ij})_{i,j \in N}$ as follows: $\sigma_{ii} = \sigma'_{ii} = 1$ for $i \in N$, $\sigma_{bc} = \sigma_{cb} = \sigma'_{ac} = \sigma'_{ca} = \frac{1}{2}$ and $\sigma_{ij} = \sigma'_{ij} = 0$ for remaining $i, j \in N$. Put $P = \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$, $Q = \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}')$ and observe that $\mathcal{M}_P$ is the semi-graphoid closure of $\langle a, bc|\emptyset\rangle$ while $\mathcal{M}_Q$ is the semi-graphoid closure of $\langle b, ac|\emptyset\rangle$. Thus, $\langle a, b|c\rangle, \langle a, b|\emptyset\rangle \in \mathcal{M} \equiv \mathcal{M}_P \cap \mathcal{M}_Q$ while $\langle a, c|\emptyset\rangle \notin \mathcal{M}$ and $\langle c, b|\emptyset\rangle \notin \mathcal{M}$. By Consequence 2.5 $\mathcal{M}$ is not a Gaussian CI model. $\diamondsuit$

In fact, the counterexample above means that the poset of CI models induced by non-degenerate Gaussian measures over N (ordered by inclusion) is not a lattice. Note that in case $|N| = 3$ this poset coincides with the poset of DAG models (see Section 3.2) which is shown in Figure 7.3. However, if $|N| > 3$ then these posets differ - see Excercise 3.8 in [97].

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

CONSEQUENCE 2.6 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

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

□

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 valid universally.

EXAMPLE 2.3 There exists a Gaussian measure P over N with $|N| = 3$ such that

$$a \perp\!\!\!\perp b \,|\, \{c\}\ [P] \quad \text{and} \quad \neg(\, a \perp\!\!\!\perp b \,|\, \emptyset\ [P]\,) \quad \text{for arbitrarily chosen distinct } a, b, c \in N.$$

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 the 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 constructing 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 such that $\mathcal{M}_R = \mathcal{M}_P \cap \mathcal{M}_Q$ 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 the 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, by usual definition of conditional indepencence for σ-algebras, to the requirement : $\forall\, \mathsf{A}^x \in \bar{\mathcal{X}}_A,\ \mathsf{B}^x \in \bar{\mathcal{X}}_B$

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

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. 190) and by the definition of R verify that

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

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

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). $\square$

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

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

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

$$\forall \text{ elementary triplet } \langle a,b|K\rangle \text{ with } |abK| = l \qquad a \perp\!\!\!\perp b\,|\,K\ [P_{[A]}] \Leftrightarrow abK \neq A.$$

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

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

CONSEQUENCE 2.7 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. $\square$

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 the 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 2.13 The word 'multiset' is taken from combinatorial theory [1] while the word 'imset' is an abbreviation for **i**nteger-valued **multiset**. Later in this monograph 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:

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

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 $A \subseteq N$:

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

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 [9]). 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 monograph) is to write it as a combination of simpler 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 book, 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

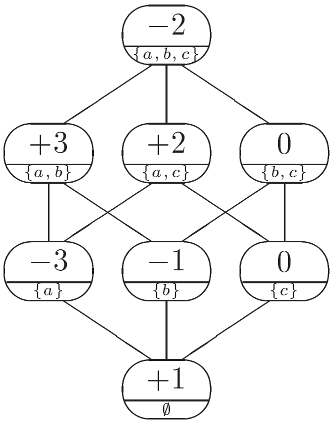


Figure 2.2: Hasse diagram of an imset over $N = \{a, b, c\}$.

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 $m : \mathcal{P}(N) \to \mathbb{R}$ 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* [97]. 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 [96] in terms of a finite number of formal properties:

1. triviality $\quad A \perp\!\!\!\perp \emptyset \,|\, C\ [G]$,
2. symmetry $\quad A \perp\!\!\!\perp B \,|\, C\ [G]$ implies $B \perp\!\!\!\perp A \,|\, C\ [G]$,
3. decomposition $\quad A \perp\!\!\!\perp BD \,|\, C\ [G]$ implies $A \perp\!\!\!\perp D \,|\, C\ [G]$,
4. strong union $\quad A \perp\!\!\!\perp B \,|\, C\ [G]$ implies $A \perp\!\!\!\perp B \,|\, DC\ [G]$,
5. intersection $\quad A \perp\!\!\!\perp B \,|\, DC\ [G]$ and $A \perp\!\!\!\perp D \,|\, BC\ [G]$ implies $A \perp\!\!\!\perp BD \,|\, C\ [G]$,
6. transitivity $\quad A \perp\!\!\!\perp B \,|\, C\ [G]$ 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 [92] where undirected graphs were used to model geometric arrangement in space. Several types of

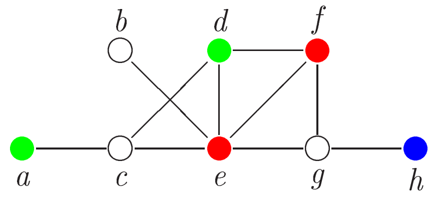


Figure 3.1: The set $C = \{e, f\}$ separates between $A = \{a, d\}$ and $B = \{h\}$.

Markov conditions were later introduced [66] 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. [77]) but coincide in case of positive measures [66].

Note that similar story was observed in case of acyclic directed graphs and chain graphs (for an overview see § 3.2 of [67]) and has recently been repeated in case of advanced graphical models ' (see Section 3.5). However, in this book, attention is only paid to the result of this development, that is to graphical criteria which correspond to respective global Markov conditions. $\triangle$

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

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

and *perfectly Markovian* if the converse implication holds as well. It was shown in [44] (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 the faithfulness (in the sense of Section 1.1) is ensured for the universum of undirected graphs and discrete distribution framework.

Remark 3.2 This is to explain certain habitual terminology used sometimes in the literature. What is claimed in this rhe remark also holds 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 probability measures Ψ implies the following weaker result. Whenever a disjoint triplet $t \in \mathcal{T}(N)$ 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 not valid conditional independence statement with respect to P. Some authors [35, 57, 51] say then that *the class of measures Ψ is perfect* with respect to G. Thus, Theorem 2.3 from [36] says that the class of CG measures (see Section 4.1.3) 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 the literature [42, 133, 70] 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 the *strong completeness* is then meant the existence of a perfectly Markovian measure over N with prescribed non-trivial sample space X_N [87, 70]. $\triangle$

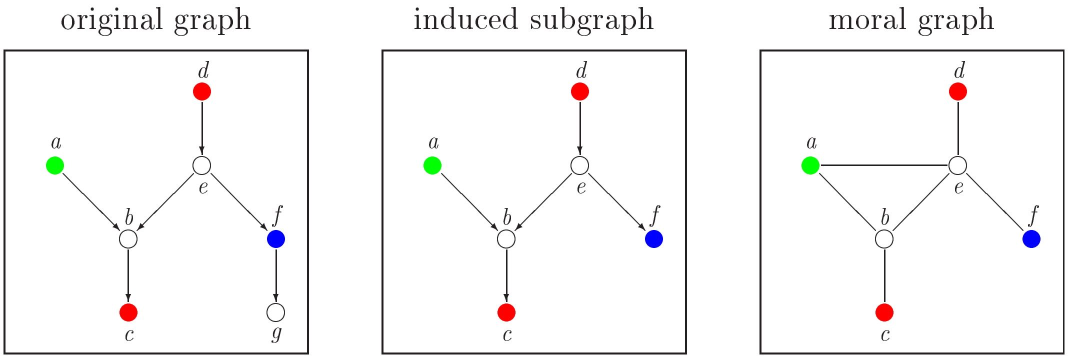


Figure 3.2: Testing $\langle a, f \,|\, \{c, d\}\rangle$ according to the moralization criterion.

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 G and H are Markov equivalent iff $\mathcal{M}_G = \mathcal{M}_H$. Moreover, the observation that $a - b$ in G iff $\neg(a \perp\!\!\!\perp b \,|\, N \setminus \{a, b\}\ [G]\,)$ implies that $\mathcal{M}_G = \mathcal{M}_H$ iff $G = H$. Thus, the equivalence question (in the 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. 27) of its cliques. It is known that every factorizable measure is Markovian [66], the converse is true for positive measures [56] but not for all (discrete) measures [77].

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 (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 [132]. 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* [97]. Note that the majority of authors became accustomed to the phrase 'directed acyclic graphs' which is not accurate from the grammatical point of view (because 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.* [66] proposed the *moralization criterion* while the group around Pearl [41] 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 the 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 the d-separation criterion see Figure 3.3. It was shown in [66] that the moralization and the 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 the d-separation criterion is suitable in the opposite case. The third possible equivalent criterion (a compromise between those two criteria) appeared in [75].

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 [97]. These properties 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 [42] that a perfectly Markovian discrete probability measure exists for every acyclic directed graph. The existence of a perfectly Markovian measure with prescribed non-trivial discrete sample space was derived later from that result in [87]. Thus, DAG models are also probabilistic CI models.

Two acyclic directed graphs are *Markov equivalent* if their classes of Markovian measures coincide. An alternative phrase 'faithfully indistinguishable graphs' is used in § 4.2 of [115]. Note that some authors (see § 3.3 of [67]) introduce Markov equivalent graphs as the graphs which induce the same formal independence model - see the explanation in Section 6.1. The classic graphical characterization of Markov equivalent acyclic directed

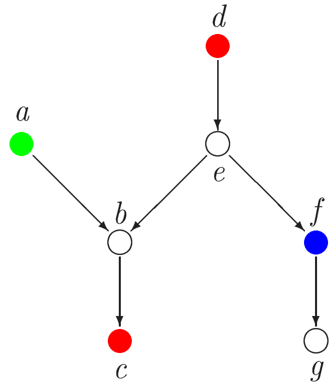


Figure 3.3: The path $a \to b \leftarrow e \to f$ is active with respect to $C = \{c, d\}$.

graphs was given in [142]; but the same result can be found also in later publications [115, 6] and it also follows from the graphical characterization of Markov equivalent chain graphs [35]. 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. 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'.

An alternative (graphical) *transformational characterization* of equivalent acyclic directed graphs was presented in [21]. By a *legal arrow reversal* is understood the change of an acyclic directed graph G into a directed graph H: one replaces an arrow $a \to b$ in G by the arrow $b \to a$ in H provided that the condition $pa_G(b) = pa_G(a) \cup \{a\}$ is fulfilled. This condition ensures that the resulting graph H is again an acyclic directed graph equivalent to G. This observation motivated the terminology here; note that an alternative phrase '$a \to b$ is a covered arc in G' is used in the literature [21, 88]. It was shown in [21] (see also Lemma 3.2 in [55]) that acyclic directed graphs G and H over N are equivalent iff there exists a sequence $G_1, \ldots, G_m$, $m \geq 1$ of acyclic directed graphs over N such that $G_1 = G$, $G_m = H$ and G_{i+1} is obtained from G_i by a legal arrow reversal for $i = 1, \ldots, m-1$.

However, the question of choice of a suitable representative of an 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* [6] or *pattern* [142] were used in the literature to represent uniquely equivalence classes of acyclic directed graphs. The problem of learning of DAG models, more exactly the identification of the 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 [143, 86, 21].

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 [66]. Another specific feature of this case is that an analogue of the 'local Markov property' is equivalent to the 'global Markov property' [66]. This fact can be also derived from the result in [141] saying that the least semi-graphoid containing the following collection of independence statements

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

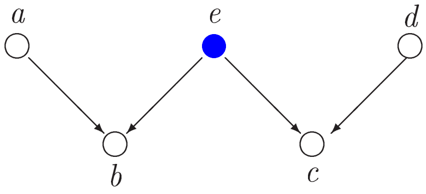


Figure 3.4: Acyclic directed graph with hidden variable e.

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* [13] or 'stratified protocol'. $\triangle$

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 [30] (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, to restriction of DAG models - see Section 3.5.7. $\diamondsuit$

REMARK 3.5 The observation that DAG models are not closed under restriction made in the preceding remark has an important consequence which may not be evident at first sight. The consequence is that DAG models cannot be characterized in an axiomatic way, more exactly, in the way UG models were characterized and discrete probabilistic CI models could possibly be characterized.

To be more precise let me specify what find of axiomatic characterization I have in mind. This is the characterization by means possibly infinite system of inference rules which involve both conditional independence statements and conditional dependence statements. General form of these inference rules is as follows:

$$[\, a_1 \text{ and } \ldots \text{ and } a_n \,] \quad \text{implies} \quad [\, c_1 \text{ or } \ldots \text{ or } c_m \,], \tag{3.1}$$

where $n \geq 0$, $m \geq 1$; that is, the conjunction of a (possibly empty) set of statements, which are called *antecedents*, implies the disjunction of a non-empty set of (other) statements, which are called *consequents*. Let us call by a *context* of a non-trivial independence statement $A \perp\!\!\!\perp B \,|\, C$ the set ABC; the context of a trivial statement is the empty set $\emptyset$ by convention. The context of a dependence statement $A \not\perp\!\!\!\perp B \,|\, C$ coincides with the context of $A \perp\!\!\!\perp B \,|\, C$. The only formal requirement on the considered inference rules is that every inference rule is *context consistent*, by which is meant that at least one statement involved in the rule has the context which is included in the union of the contexts of remaining statements involved in the rule. Observe that all semi-graphoid rules (see p. 19) and all inference rules characterizing UG models (see p. 43) aere context consistent. Note that every rule (3.1) can equivalently be expressed in its canonic form

$$[\,] \quad \text{implies} \quad [\, c_1 \text{ or } \ldots \text{ or } c_m \text{ or } \neg a_1 \text{ or } \ldots \text{ or } \neg a_n \,], \tag{3.2}$$

where the negation of $A \perp\!\!\!\perp B \,|\, C$ is $A \not\perp\!\!\!\perp B \,|\, C$ and conversely. Thus, every involved triplet $\langle A, B|C \rangle \in \mathcal{T}(N)$ can be represented as a consequent, either in the form of an independence statement or in the form of a dependence statement. The remaining involved triplets may be turned into antecedents then which means that every context consistent inference rule can be formulated in such a way that it has only one consequent whose context is included in the union of context of antecedents.

Let us say that a collection $\mathcal{C}$ of statements over N is *full* of for every $\langle A, B|C \rangle \in \mathcal{T}(N)$ exclusively either $A \perp\!\!\!\perp B \,|\, C$ belongs to $\mathcal{C}$ or $A \not\perp\!\!\!\perp B \,|\, C$ belongs to $\mathcal{C}$. The restriction of $\mathcal{C}$ to $\emptyset \neq T$ consists of those statementens in $\mathcal{C}$ which correspond to triplets from $\mathcal{T}(T)$. Observe that every independence model $\mathcal{M}$ can be identified with a full collection of statements: every element of $\mathcal{T}(N) \setminus \mathcal{M}$ will be automatically regarded as a dependence statement. Moreover, the restriction of $\mathcal{M}$ to $\emptyset \neq T$ corresponds to the restriction of $\mathcal{C}$. Every inference rule of the type above can be understood as a condition on the respective full collection of statements. The above mentioned observation about context consisetent inference rules implies the following basic fact: if a full collection of statements is closed under a (possibly infinite) collection of context consistent inference rules then its restriction to a non-empty set $T \subseteq N$ is closed under those rules as well. This fact together with Example 3.1 implies that DAG models cannot be characterized in this way.

On the other hand, I conjecture that discrete probabilistic CI models can be characterized by means of an infinite (but countable) collection of context consistent inference rules. This conjecture is based on Proposition 2 in [123] and the fact that the class of discrete CI structures is closed under restriction and permutation of variables. Well, it may be the case that DAG models can be characterized by means of more general formal properties where 'elementary clauses' are more complex and represent composite statements. For example, one symbol $A \not\perp\!\!\!\perp B \,|\, C^{\uparrow}$ could represent a whole collection of dependence statements $A \not\perp\!\!\!\perp B \,|\, D$ where $C \subseteq D \subseteq N \setminus AB$. According to e-mail communication by T. Verma characterization of this type 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. 186). The class of chain graphs was introduced by Lauritzen and Wermuth in the middle 1980's in the report [61] which became later a basis of a journal paper [64].

Classic interpretation of chain graphs is based on the *moralization criterion for chain graphs* established by Lauritzen [65] and Frydenberg [35]. The main distinction between the moralization criterion for chain graphs and for acyclic directed graphs (see p. 46) 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 [14]. This criterion was

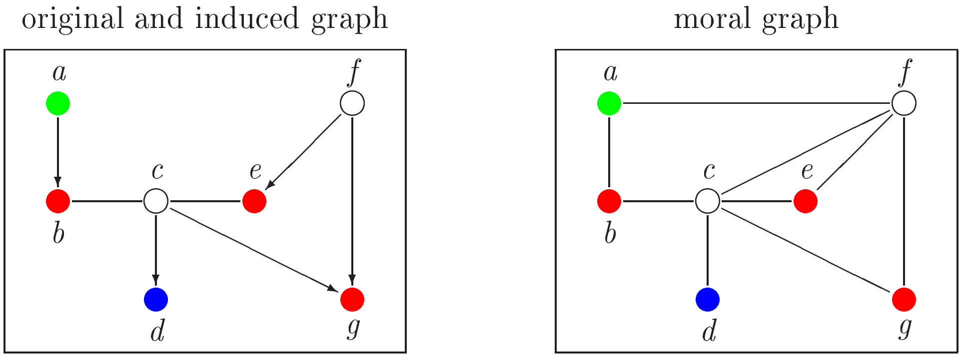


Figure 3.5: Testing $\langle a, d \,|\, \{b, e, g\} \rangle$ according to the moralization criterion for chain graphs.

later simplified as follows [134]. 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 [133] (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 the c-separation criterion 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. 53). Every CG model is a graphoid satisfying the composition property [133]. 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 the 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 superactive 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 the c-separation criterion is finitely implementable for another reason - see Section 5 in [134].

Note that the above mentioned phenomenon was the main reason why the original version of the c-separation criterion [14] 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

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

and *perfectly Markovian* if the converse implication holds as well. The main result of [133] says that a perfectly Markovian positive discrete probability measure exists for every chain graph. In particular, the faithfulness (in the sense of Section 1.1) is ensured in case of CG models 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 [35]. 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 [35] 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 [144]. The paper also describes an algorithm for transformation of every chain graph into the respective largest chain graph. An alternative algorithm is presented in [131] where the problem of finding the largest chain graph on basis of the induced formal independence model is solved. This could be utilized for learning CG models.

Remark 3.7 Lauritzen [67], 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 [35].

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 4. $\triangle$

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 [67], Proposition 2.5) which motivated other alternative name *decomposable graph* (see p. 173). For this reason UG models induced by triangulated graphs are named *decomposable models* [97]. Another equivalent definition (see [67], 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*

$$\forall\, 2 \leq i \leq m \;\; \exists\, 1 \leq k < i \qquad S_i \equiv C_i \cap (\bigcup_{j<i} C_j) \subseteq C_k\,. \tag{3.3}$$

Note that the phrase *acyclic hypergraph* is sometimes used in the literature to name a class of sets admitting an ordering of this type. The sets S_i are called *separators* then 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 [67], 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 [62]). Note that the running intersection property has close connection to the marginal problem within the framework of probabilistic expert systems [49, 52].

One can show (by repeated application of Proposition 3.17 from [67]) 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

$$f_N(x) \;=\; \frac{\prod_{C \in \mathcal{C}} f_C(x_C)}{\prod_{S \in \mathcal{S}} f_S(x_S)^{w(S)}} \tag{3.4}$$

where $\mathcal{C}$ is the class of cliques, $\mathcal{S}$ is 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.

Further related equivalent definition of a triangulated graph is the existence of a *junction tree* ([24], Theorem 4.6) of its cliques (and separators). Junction trees then form a mathematical basis for miscellaneous effective computational methods [108, 24] which originate from the *local computation method* [63]. Thus, decomposable models are very suitable from the point of view of implementation (see Section 1.1, p. 13).

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 [18]. It implies that decomposable models are closed under restriction.

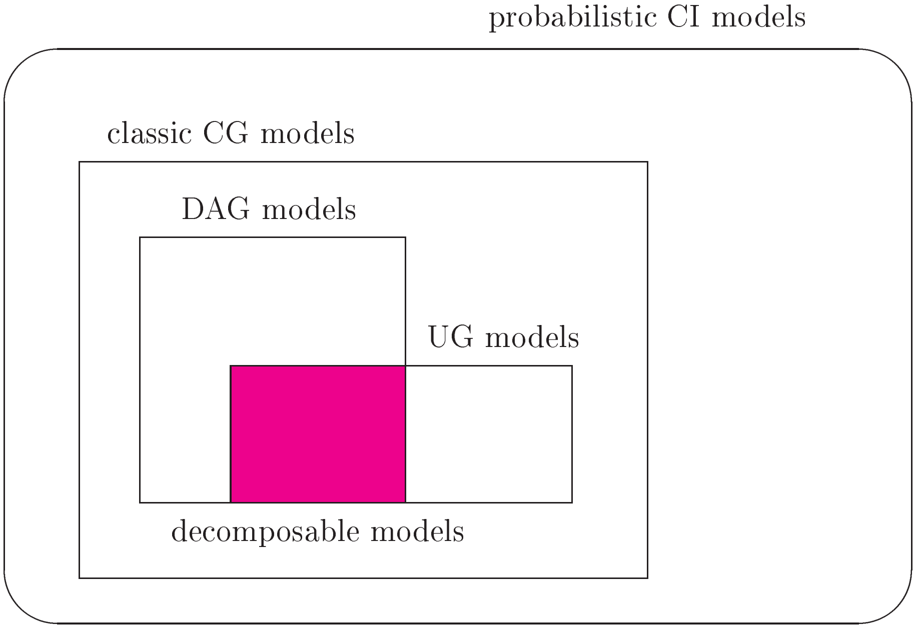


Figure 3.6: Relationships among classic graphical models.

3.4.2 Recursive causal graphs

The concept of *recursive causal graph* [53] seems to precede the concept of chain graph. Such a graph 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 [5] 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 [7] that LCI models coincide with DAG models induced by *transitive acyclic directed graphs* in which $a \to b$ and $b \to c$ implies $a \to c$. 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 [99] 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 [108] 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 the 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 the 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 [115]) mentioned possible use of general directed graphs for description of models allowing feedback. They proposed to use the d-separation criterion (see p. 46) to ascribe a formal independence model to a directed graph (even allowing multiple edges). It was shown in [116] that even in case of general directed graphs, the d-separation criterion is equivalent to the moralization criterion and the criteria are complete (in the sense of Remark 3.2) relative to the class of non-degenerate Gaussian measures. Richardson [102] published a graphical characterization of Markov equivalent directed graphs. It is more complex in comparison with the case of acyclic directed graphs (six independent conditions are involved).

3.5.2 Reciprocal graphs

Koster [57] 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. 49) 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 [116].

Thus, the consistency of reciprocal graphs (see p. 12) 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* [50]). 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 the sense of Remark 3.2).

3.5.3 Joint-response chain graphs

Cox and Wermuth [26] generalized the concept of chain graph by introducing two additional types of edges. A *joint-response chain graph* G is a chain graph (in the 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 [26], 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 which have only solid lines and arrows 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 the 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 [26]), 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 distribution framework 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 [26]) and the question of consistency (see Section 1.1) remains to be solved. △

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

3.5.4 Covariance graphs

However, the consistency was ensured in a special case of undirected graphs made of dashed lines and named *covariance graphs*. Kauermann [51] 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 [51] also showed that the class of Gaussian measures is perfect (in the 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 which have only solid lines and dashed arrows. Led by a specific way of parametrization of non-degenerate Gaussian measures Andersson, Madigan and Perlman [8] introduced 'alternative Markov property' (AMP) for chain graphs. Their *alternative chain graphs* are chain graphs in the 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 [8], § 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 special 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 the 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 [70]. The main result of [70] is the existence of perfectly Markovian non-degenerate Gaussian measure for every alternative chain graph. Thus, the faithfulness is ensured in case of these models. Moreover, Markov equivalent alternative chain graphs were characterized in graphical terms as well [8]. Every class of Markov quivalence can be represented by respective *essential graph* (for details see [8], § 7).

3.5.6 Annotated graphs

Paz proposed in [94] a special fast implementation (that is, 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 $a \to c \leftarrow b$ 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 $\emptyset \neq C \subseteq N \setminus \{a,b\}$ 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 [94].

The point is that this approach has much wider applicability. In [95] 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 [95] 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.

The idea of annotation of edges was also applied in case of acyclic directed graphs. Bouckaert [12] annotated arrows by pairs of disjoint sets of nodes and introduced a special separation criterion for the respective graphs.

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 [45] 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 [45] 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 the faithfulness in case of embedded Bayesian networks is an easy consequence of the faithfullness in case 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 [26], § 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 [103] 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 a 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 [104], 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 $C \subseteq N \setminus \{a, b\}$. 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 9.2.1). 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 case. 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 [104].

3.5.9 MC graphs

Koster [58] 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 the m-separation criterion he 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 [104] the separation criterion for MC graphs generalizes the 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 (c.f., the question of completeness in Section 1.1, p. 12). The 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 [6] and 11519 CG models [144]. 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 [128]! 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 this. 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 2.7 on p. 40 the number of discrete probabilistic CI structures grows with n at least as rapidly as the power of power of n.

Thus, in my opinion, one can hardly achieve the completeness of a graphical approach (see p. 12) relative to the class of discrete measures and this may result in serious methodological errors (see Section 1.1 p. 14). 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 the 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 the completeness in the discrete distribution framework and had some light chance to be acceptable by humans. On the other hand, the mathematical objects which are used for this purpose describe more than it is necesary in the sense that some 'descibed' 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 the 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 that 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 [129] but (later recognized) superfluous details worsened understanding of the message of the original series of papers. This monograph 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. 18) are mainly used in probabilistic reasoning [97] which is an area of artificial intelligence. Positive discrete probability measures are behind the models used in the analysis of contingency tables (see [67], 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 [26]. Consequence 2.6 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 [64] 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 by 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 [67] § 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 [67], 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 the 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 book. $\triangle$

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 [140] 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 term use the fact that the function $y\mapsto(\ln y)^-$ is convex, the inequality $y^-\leq|y|$, Radon-Nikodym theorem and the definition of $\frac{dP}{d\mu}$

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

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

□

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

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

where $f_{\boldsymbol{e}(x),\boldsymbol{\Sigma}(x)}$ is given by (2.19), 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 $P_k = \mathcal{N}(e,\beta)$, $P_l = \mathcal{N}(f,\gamma)$ 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. 32 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 the multiinformation function in terms of parameters of CG-measure remains open (see Theme 1 in Chapter 9).

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.

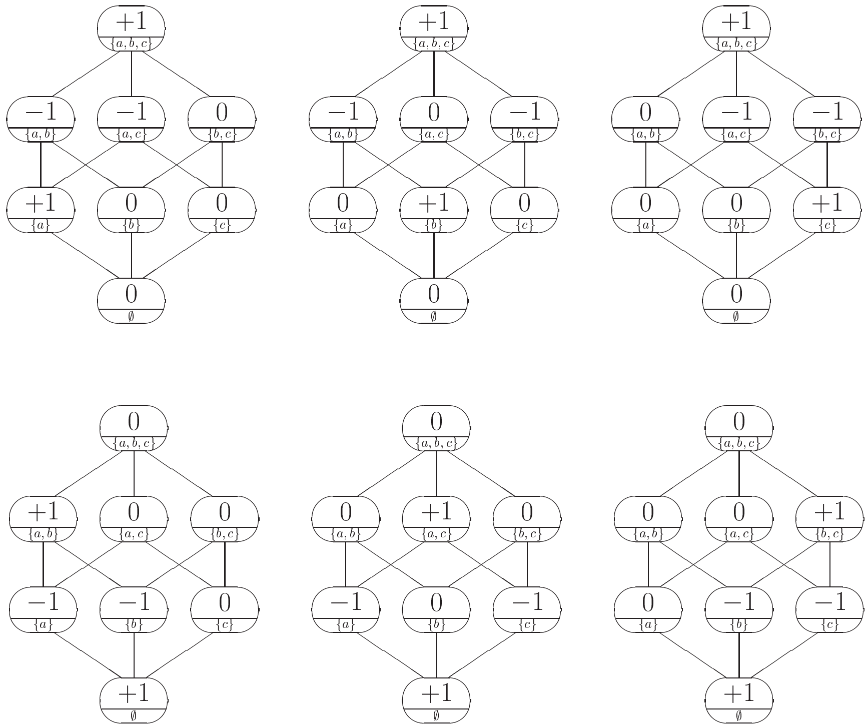


Figure 4.1: Elementary imsets over $N = \{a, b, c\}$.

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 $\langle a, b|K\rangle$ where $K \subseteq N$ and $a, b \in N \setminus K$ are distinct ($a \neq b$) the corresponding *elementary imset* $u_{\langle a,b|K\rangle}$ over N is defined by the formula

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

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 $u_{\langle a,b|K\rangle}$ 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

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

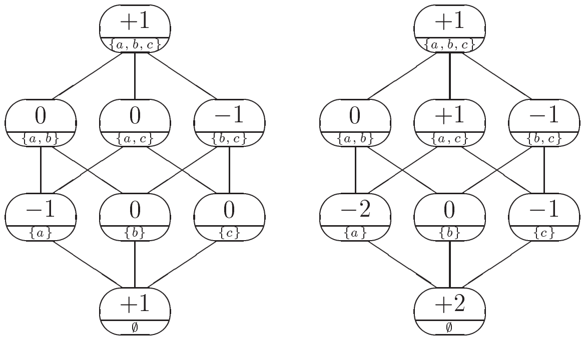


Figure 4.2: Two combinatorial imsets over $N = \{a, b, c\}$.

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 observation 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. 64.

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

Consequence 4.2 If $u \in \mathcal{C}(N)$ then $u = 0 \Leftrightarrow deg(u) = 0$, $u \in \mathcal{E}(N) \Leftrightarrow deg(u) = 1$.

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 [122]. However, the question whether this is true in general remains open (see Question 9 on p. 176).

OBSERVATION 4.4 Every structural imset u over N is o-standardized, $\langle m^{A\uparrow}, u\rangle \geq 0$ and $\langle m^{A\downarrow}, u\rangle \geq 0$ for every $A \subseteq N$ (see p. 41). 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 imset and elementary imsets and can be extended to combinatorial imsets and to structural imsets then. 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 4.4 on p. 73).

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 <u>non-zero</u> 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

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

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 $v = u_{\langle a,b|K\rangle}$ satisfies $\langle s, v\rangle > 0$ iff $\{a, b\} \cup K \in \mathcal{S}_u$ and $\{a\} \cup K, \{b\} \cup K \notin \mathcal{S}_u$. 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 (in 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. 25), 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 case or in purely Gaussian case it coicides with two above mentioned reference systems which I only 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 $\emptyset \neq S \subseteq N$

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

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 the non-uniqueness of marginal densities and the equality is understood in the '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. $\triangle$

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. [119] 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 $x \in \mathsf{X}_N$. $\square$

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 by means of 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. □

Further example is the model induced by an elementary imset.

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

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

Proof: The facts $\langle a,b|K\rangle$, $\langle b,a|K\rangle \in \mathcal{M}_v$ 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 $ABC \subseteq abK$ use Observation 4.4 to derive

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

The fact $\langle m^{ABC\uparrow}, u_{\langle a,b|K\rangle}\rangle > 0$ then implies $ABC \subseteq abK$. 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 $K \subseteq C \subset ABC \subseteq abK$ then implies that $\langle A,B|C\rangle$ coincides either with $\langle a,b|K\rangle$ or with $\langle b,a|K\rangle$. □

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. 21) easily follow from the definition above and the fact that the sum of structural imsets is a structural imset. Indeed, for the triviality property realize $\mu_{\langle A,\emptyset|C\rangle} = 0$, for the symmetry property $\mu_{\langle A,B|C\rangle} = \mu_{\langle B,A|C\rangle}$ and for the remaining properties $\mu_{\langle A,BD|C\rangle} = \mu_{\langle A,B|DC\rangle} + \mu_{\langle A,D|C\rangle}$. □

For the proof of the equivalence result in Section 4.5 a technical lemma is needed. In its proof the following simple observation concerning upper classes (see p. 67) is used.

OBSERVATION 4.8 Supposing $u = w + v$ where $w, v \in \mathcal{S}(N)$ 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 Consequence 4.2 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 the 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 4.3 Let u be a structural imset over N. Then every descending system $\mathcal{E} \subseteq \mathcal{T}(N)$ containing $\mathcal{L}_u$ and satisfying

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

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, the statistical meaning of an 'imsetal' model is completely analogous to the statistical meaning of a graphical model. Every structural imset u over N represents a class of probability measures over N within the respective distribution framework, 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 monograph (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. $\diamondsuit$

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 a Markovian measure is determined by its marginals on the lower class.

CONSEQUENCE 4.4 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. 29) 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 $[\,P^S = Q^S \;\text{ for every } S\, \in \mathcal{D}_i\,]$. $\square$

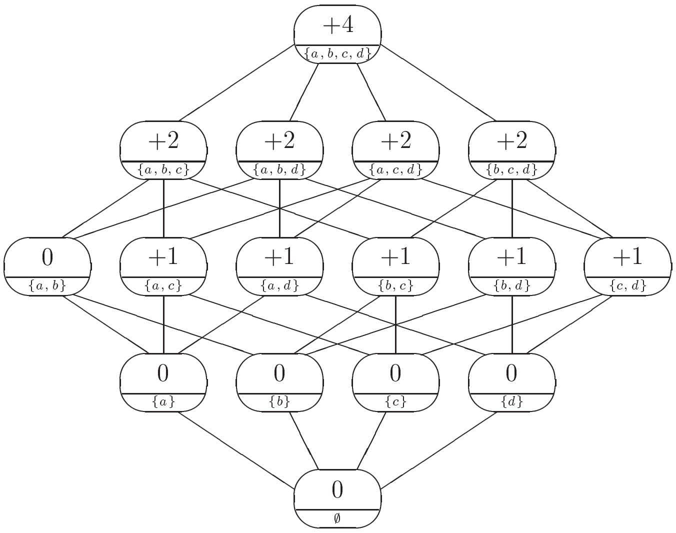


Figure 4.3: The multiset $m_\dagger$ from Example 4.1.

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. 24) $\mu = \prod_{i\in N} \mu_i$ such that $-\infty < H(P^S \mid \prod_{i\in S} \mu_i) < +\infty$ 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:

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

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. 20) with respect to P [78] (in addition to pure conditional independence statements). More specifically, $h_{P,\upsilon}(A) \leq h_{P,\upsilon}(AC)$ 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. △

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, and this holds iff the product formula induced by the imset holds. In the proof below the following simple observation is used.

OBSERVATION 4.9 Suppose the situation from Convention 1 (p. 25). 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 2.9 with the formula (2.4) in the proof of Lemma 2.4. □

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. 68); let us accept Convention 1 on p. 25. 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) ⇒ (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) ⇒ (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) ⇒ (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 4.3 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 $v = u_{\langle a,b|K\rangle} \in \mathcal{E}(N)$ with $k_v > 0$ one has $a \perp\!\!\!\perp b \,|\, K\ [u]$ and therefore by (iv) and (2.3) on p. 26 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. $\square$

Let me note that one can always take the universal reference system (p. 68) 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 4.2. 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

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

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. 66) and the 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). □

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' [105] or even 'convex games' [111]. I followed that terminology in some of my former reports [122, 129]. However, another common term 'supermodular' is used in this book in order to avoid confusion with usual meaning of the adjective 'convex' in mathematics. As mentioned in [16] 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 *independence 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 induced 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 between 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 also used in one of former reports [137] 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 the triviality property realize $\langle m, u_{\langle A,\emptyset|C\rangle}\rangle = \langle m, 0\rangle = 0$, for the symmetry property $\langle m, u_{\langle A,B|C\rangle}\rangle = \langle m, u_{\langle B,A|C\rangle}\rangle$. 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 the contraction property. To verify the decomposition property and the weak union property 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. 31 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. 32). This number can be interpreted (in the discrete case) as a numerical evaluation of the degree of stochastic conditional dependence between A and B given C with respect to P [136]. 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 independence equivalent. Strong equivalence can be equivalently described with the 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 $u_{\langle a,b|K\rangle} \in \mathcal{E}(N)$ such that $S = abK$. The fact $\langle l, u_{\langle a,b|K\rangle}\rangle = 0$ says

$$l(S) = l(aK) + l(bK) - l(K)$$

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. 41) 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 the 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 $s \in \mathcal{K}(N) \cap \mathcal{S}(N)$ coincides 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 book 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. 31). 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 $\mathcal{L}(N)$. $\triangle$

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 independence equivalent to a skeletal function is also skeletal. In particular, strong equivalence has the same property. Of course, independence 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 the 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 $\mathcal{L}(N) \cap \mathcal{K}_\ell(N) = \mathcal{L}(N) \cap \mathcal{S}_\ell(N)$. $\square$

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. 42). 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 just one non-zero imset within the ray is normalized. $\square$

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 the 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. $\square$

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. 71 $A \perp\!\!\!\perp B \,|\, C\ [u]$ 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). $\square$

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, the 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. $\square$

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

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 the 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 $\alpha_r > 0$. The linearity of scalar product with the 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 the 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 the 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 the 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 independence equivalence classes of skeletal imsets is obtained.

Consequence 5.3 Every class of independence equivalence of skeletal supermodular functions over N is characterized by the 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

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

In particular, every skeletal function is independence 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 independence 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 $0 \neq s \in \mathcal{N}(N)$

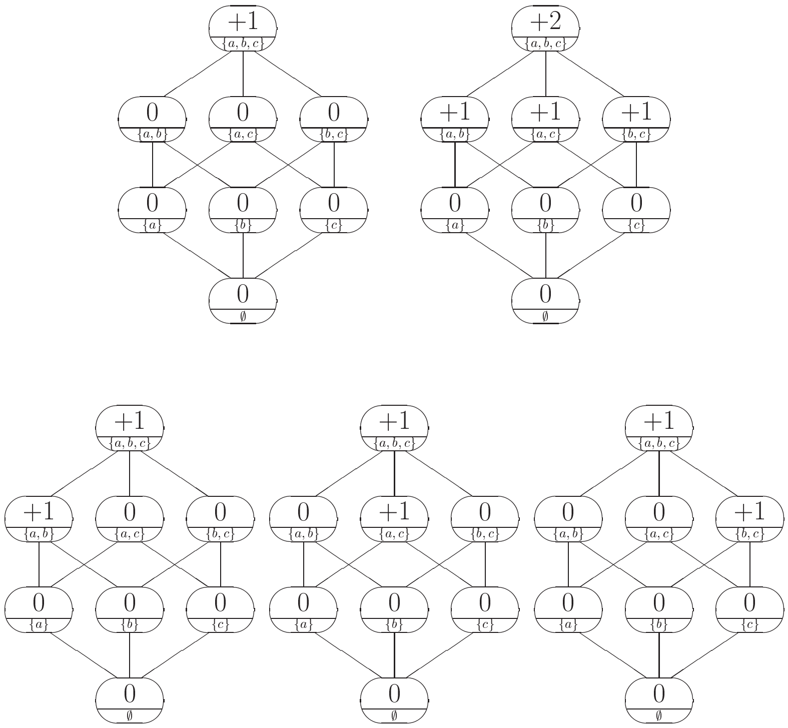


Figure 5.1: The ℓ-skeleton over $N = \{a, b, c\}$.

one has $\langle s, u\rangle > 0$ and therefore there exists $k_s \in \mathbb{N}$ with $\langle s, v_{k_s}\rangle = k_s \cdot \langle s, u\rangle + \langle s, w\rangle \geq 0$. 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. □

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 [122]). 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 [137] - for the proof see [122]. However, the ℓ-skeleton in case $|N| = 5$ was found by means of a computer; it has 117978 imsets - see [137] 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 9.1.2. △

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 [129] to name the least finite set of normalized imsets defining a pointed rational

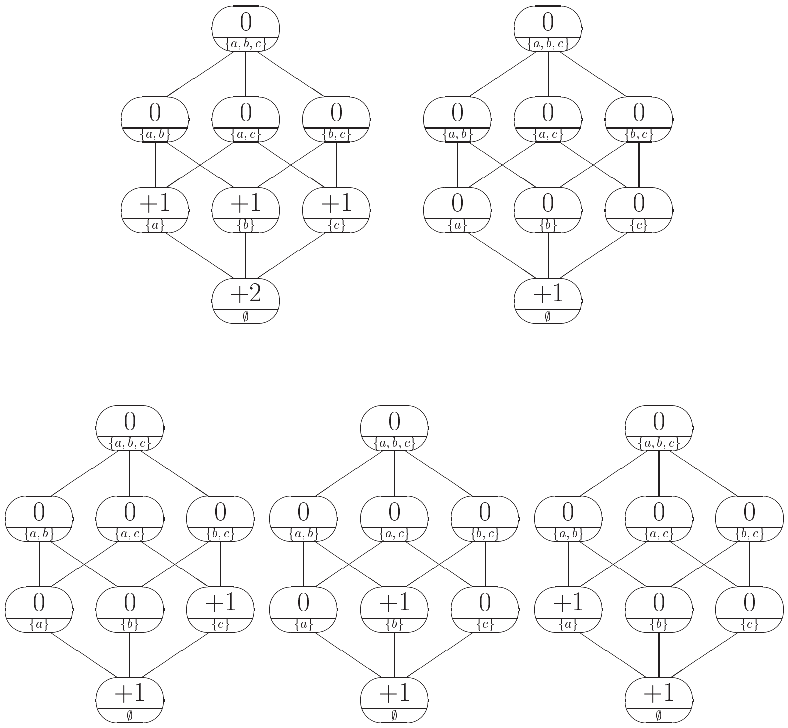


Figure 5.2: The u-skeleton over $N = \{a, b, c\}$.

polyhedral cone as its dual cone; that is, the ℓ-skeleton in case of the conical closure $\mathcal{E}(N)$. Another possible justification is that every independence model produced by a supermodular function is the 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 [6, 8] use the word 'skeleton' to name the underlying graph of a chain graph (see Section 10.3, p. 186). $\triangle$

REMARK 5.6 As explained in Remark 5.3 the 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 independence 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 the 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. $\triangle$

5.3 Description of models by structural imsets

Semi-graphoid induced by a structural imset was already introduced in Section 4.4.1 on p. 71. 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

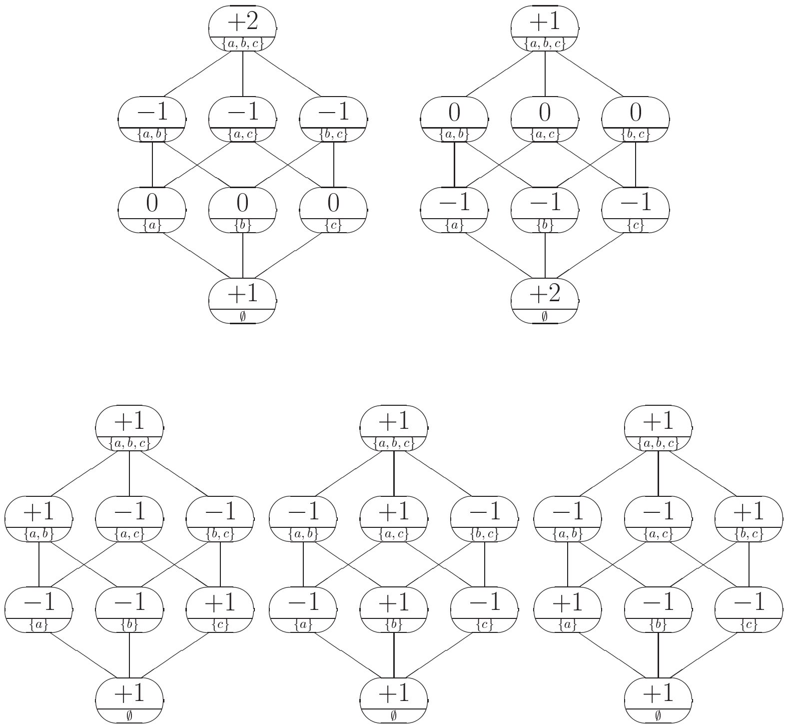


Figure 5.3: The o-skeleton over $N = \{a, b, c\}$.

$\mathcal{M}_u \subseteq \mathcal{M}^m$ 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 $v = u_{\langle a,b|K \rangle} \in \mathcal{E}(N)$ such that $k_v > 0$ observe $\langle a, b|K \rangle \in \mathcal{M}_u$ 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 one.

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 the linearity of scalar product, the converse inclusion can be derived similarly with help of Observation 5.1. $\square$

A 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, u is 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. 80 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 considerations in Section 1.1, Theorem 5.2 means that structural imsets solve satisfactorily the 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. 73 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. 13). 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) [128] 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 for 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 the help of the view of theory of 'formal concept analysis'. This approach, developed in [38], is a specific application of the 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 the 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 [38]. *Formal context* consists of the following items:

- the set of *objects* Œ,
- the set of *attributes* Æ,
- a 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 one writes $x \Im y$ and says that the object x has the attribute y. In general, Galois connection is defined for a pair of posets ([9], § 6 of Chapter IV). However, in the treated special case, *Galois connection* can be introduced as a pair of mappings between the power sets of Œ and Æ (which are posets with respect to inclusion):

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

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 [38]) 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

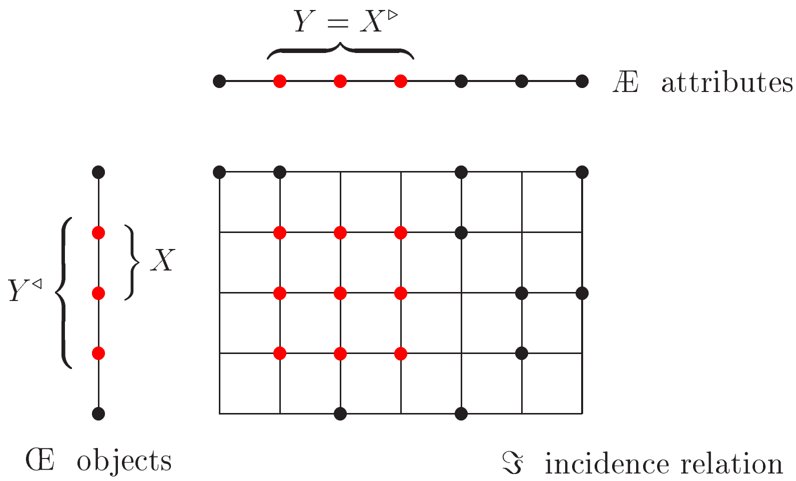


Figure 5.4: Galois connection - an informal illustration.

ordinary inclusion ordering $\subseteq$ on $\{Y \subseteq Æ; Y = Y^{\triangleleft\triangleright}\}$ (see Section 10.2). Thus, the same 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 element of $\mathcal{U}(N)$ is $\mathcal{T}_\emptyset(N)$, that is, the model induced by zero structural imset. The atoms of $\mathcal{U}(N)$ are just the models induced by elementary*

$Æ = \mathcal{K}_\ell^\diamond(N)$

	δ_N	$2 \cdot \delta_N + \delta_{ab} + \delta_{ac} + \delta_{bc}$	$\delta_N + \delta_{ab}$	$\delta_N + \delta_{ac}$	$\delta_N + \delta_{bc}$
$u_{\langle b,c\|a\rangle}$		$\bullet$	$\bullet$	$\bullet$	
$u_{\langle a,c\|b\rangle}$		$\bullet$	$\bullet$		$\bullet$
$u_{\langle a,b\|c\rangle}$		$\bullet$		$\bullet$	$\bullet$
$u_{\langle a,b\|\emptyset\rangle}$	$\bullet$			$\bullet$	$\bullet$
$u_{\langle a,c\|\emptyset\rangle}$	$\bullet$		$\bullet$		$\bullet$
$u_{\langle b,c\|\emptyset\rangle}$	$\bullet$		$\bullet$	$\bullet$	

$Œ = \mathcal{E}(N)$

Figure 5.5: The formal context (5.16) in case $N = \{a, b, c\}$.

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 element of $\mathcal{U}(N)$ is $\mathcal{T}(N)$, that is, the model produced by any modular set function $l \in \mathcal{L}(N)$.

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 $\mathcal{U}(N)$:

$$\forall \mathcal{M} \in \mathcal{U}(N) \quad \mathcal{M} = \sup \{\mathcal{M}_u : u \in \mathcal{Q}\} \;\text{ where } \mathcal{Q} = \{u_{\langle a,b|K\rangle} \in \mathcal{E}(N); \langle a,b|K\rangle \in \mathcal{M}\}\,. \tag{5.17}$$

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 $\mathcal{U}(N)$:

$$\forall \mathcal{M} \in \mathcal{U}(N) \quad \mathcal{M} = \inf \{\mathcal{M}^m; m \in \mathcal{R}\} \;\text{ where } \mathcal{R} = \{r \in \mathcal{K}_\ell^\diamond(N); \mathcal{M} \subseteq \mathcal{M}^r\}\,. \tag{5.18}$$

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 [38]. The theorem says that to show that $\mathcal{U}(N)$ is order-isomorphic to the concept lattice of a formal context $(Œ, Æ, \Im)$ it suffices to show that there exist mappings $\gamma : Œ \to \mathcal{U}(N)$ and $\delta : Æ \to \mathcal{U}(N)$ such that

(a) $\gamma(Œ)$ is supremum-dense in $\mathcal{U}(N)$,

(b) $\delta(Æ)$ is infimum-dense in $\mathcal{U}(N)$,

(c) $u \Im m \Leftrightarrow \gamma(u) \subseteq \delta(m)$ for every $u \in Œ$, $m \in Æ$.

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. 71) and δ ascribes $\mathcal{M}^m$ to every $m \in \mathcal{K}_\ell^\diamond(N)$ (see p. 79). 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 element of $\mathcal{U}(N)$. By Observation 4.7 $\mathcal{T}_\emptyset(N) \in \mathcal{U}(N)$. By Lemma 4.6 any $\mathcal{M} \in \mathcal{U}(N)$ 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 $v = u_{\langle a,b|K\rangle}$ then by Lemma 4.5 obtain $\mathcal{M}_v = \{\langle a,b|K\rangle, \langle b,a|K\rangle\} \cup \mathcal{T}_\emptyset(N)$. As $\mathcal{M}$ is a semi-graphoid $\mathcal{T}_\emptyset(N) \subseteq \mathcal{M}$. If $\mathcal{M} \neq \mathcal{T}_\emptyset(N)$ then either $\langle a,b|K\rangle$ or $\langle b,a|K\rangle$ belongs to $\mathcal{M}$ which implies by the 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 element 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 [38]). 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. $\square$.

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 integral 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 7.9 for illustration). The third option is Œ $= con(\mathcal{E}(N))$ and Æ $= \mathcal{K}_\ell(N)$. Moreover, one can consider an alternative standardization instead of the ℓ-standardization (see Remark 5.3). A special 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 [38]. 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 [38]). $\triangle$

Another point of view on the lattice $(\mathcal{U}(N), \subseteq)$ is the following one. It is 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 the computational point of view since the operation of supremum in the lattice corresponds to the sum 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 implementable by means of summing supermodular functions (imsets).

OBSERVATION 5.7 Let $\mathcal{R}$ be a finite set of supermodular functions over N. Then

$$\inf \{\mathcal{M}^m; m \in \mathcal{R}\} = \bigcap_{m\in\mathcal{R}} \mathcal{M}^m = \mathcal{M}^r \text{ where } r = \sum_{m\in\mathcal{R}} m. \tag{5.19}$$

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 [38]) 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, the original terminology from [129] was motivated by this point of view (see Remark 6.2 on p. 100). △

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 [127]. ♢

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 element 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 2.3. □

REMARK 5.11 This is to explain the relation of the presented approach to polymatroidal description of CI model used in [83]. 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 [78]. 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 [137]. △

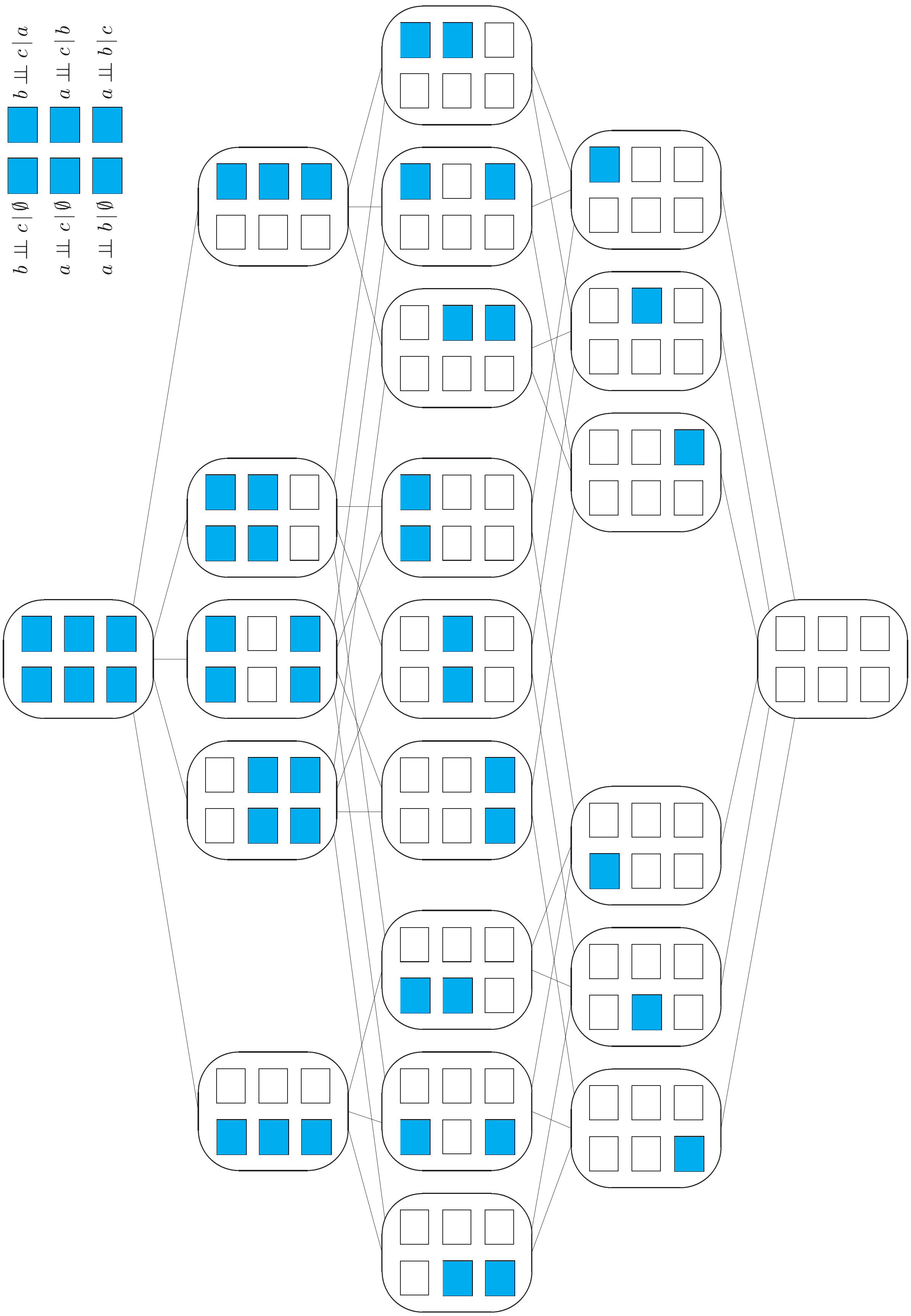


Figure 5.6: The concept lattice of CI models over $N = \{a, b, c\}$ (rotated).

Chapter 6

Markov equivalence

This chapter deals with the implication problem and the equivalence problem for structural imsets. First, the question of 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 characterizations of facial implication, which are analogous to graphical characterizations 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 the *distribution equivalence* which is the requirement that the classes of Markovian measures over N within a certain distribution framework coincide. Thus, the distribution equivalence is always understood relative to a distribution framework, that is, relative to a class Ψ of probability measures over N (see Section 10.10.3).

The second option is the *independence 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, independence equivalence implies distribution equivalence. The converse is true in case of the faithfulness (see Section 1.1 p. 12). That is, if a perfectly Markovian measure within a considered distribution framework Ψ exists for every graph (from the respective universum of graphs) then the distribution equivalence relative to Ψ implies the independence 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 independence equivalence coincide even under weaker assumption that the considered class of measures is perfect for every graph (see Remark 3.2 on p. 44). On the other hand, if the distribution framework is somehow limited then it may happen that independence and distribution equivalence differ. For example, it happens if 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 $a \perp\!\!\!\perp b \,|\, K \; [P]$ for every $P \in \Psi$

and $\langle a,b|K\rangle \in \mathcal{E}(N)$ with $a \in M$ (one can verify this using Lemma 10.1). Thus, 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 [104] and joint-response chain graphs [26]. A specific distribution framework Ψ, usually the class of non-degenerate Gaussian measures over N, is considered. Every edge of a graph of above type represents a real parameter and every collection of edge-parameters determines a unique 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 [104]). Two graphs can be called *parametrization equivalent* if their classes of parametrized distributions coincide. Of course, the parametrization equivalence substantially depends on the considered distribution framework and may not coincide with the distribution (Markov) equivalence; for example, in case of general ancestral graphs [104].

The mentioned point of view motivates a general question of whether (some) structural imsets may lead to a specific way of parametrization of the corresponding class of Markovian distribution (see Direction 6 in Chapter 9). $\triangle$

However, in typical situations distribution and independence 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 to this general equivalence question (see Section 1.1, p. 12) 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 for 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, that is, $\mathcal{M}_u = \mathcal{M}_v$. 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 $\Psi' \subseteq \Psi$. A natural question is whether the converse of Observation 6.1 holds for reasonable classes 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.

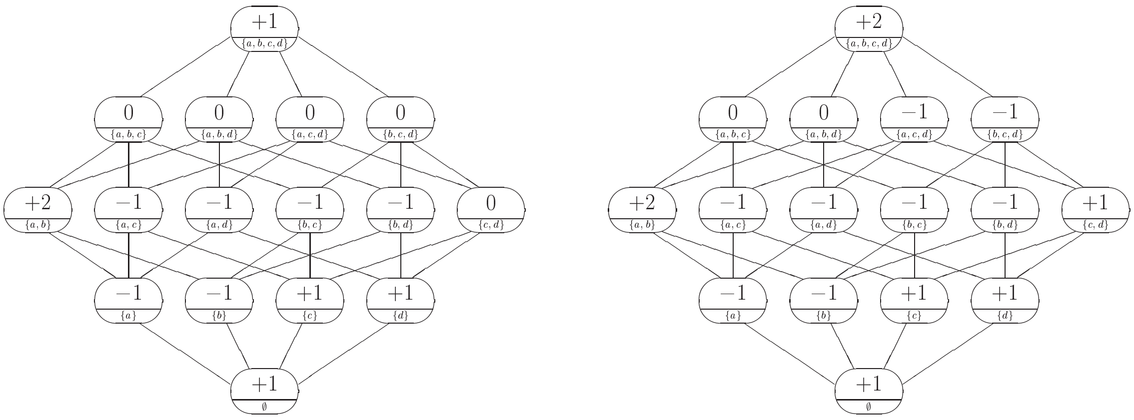


Figure 6.1: Two Markov equivalent structural imsets which are not facially equivalent.

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. 90. Note that, 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 the facial equivalence coincides both with the Markov equivalence relative to the class of discrete measures and with the Markov equivalence relative to the class of measures with finite multiinformation (use Observation 6.1).

In the rest of this chapter the attention is restricted to facial equivalence and related facial implication. One reason for this 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, the facial equivalence represents pure theoretical basis of Markov equivalence. Indeed, it will be shown later (see Lemma 6.6) 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 the framework of Ψ one can limit oneself to structural imsets of this type and the facial equivalence on the considered subclass of stuctural imsets coincides with the 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 the above terminology. The adjective 'facial' was already used in [129] 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 [15]. 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 the facial implication of structural imsets used in [129] was nothing but a 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 [129], 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 on the 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 $w = u_{\langle a,b|K\rangle}$ then $\langle a,b|K\rangle \in \mathcal{M}_v \subseteq \mathcal{M}_u$. 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 The 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 using Observation 4.3. This is the number of

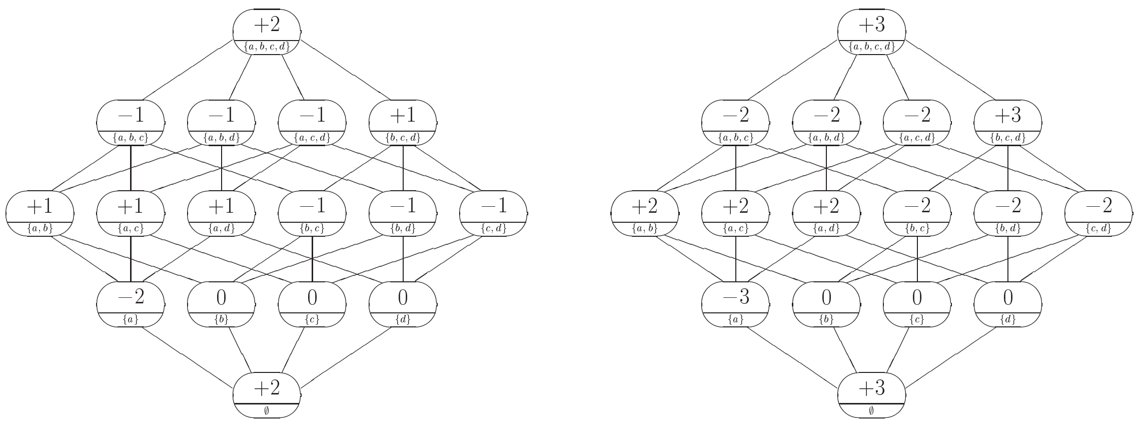


Figure 6.2: Structural imsets u and $2 \cdot u - v$ from Example 6.2.

elementary imsets which have to be summed to obtain w. The only combinatorial imset of 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 $u_{\langle a,b|K \rangle}$ such that $w - u_{\langle a,b|K \rangle}$ is a combinatorial imset of degree $n-1$. Since the class of elementary imsets is known the 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 9. $\triangle$

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

$$u = u_{\langle a,b|\emptyset \rangle} + u_{\langle a,c|\emptyset \rangle} + u_{\langle c,d|b \rangle} + u_{\langle b,d|c \rangle} + u_{\langle a,d|bc \rangle} + u_{\langle b,c|ad \rangle}\,, \quad v = u_{\langle a,bcd|\emptyset \rangle} \tag{6.5}$$

and observe that

$$\begin{aligned} 2 \cdot u - v = 2 \cdot u - u_{\langle a,bcd|\emptyset \rangle} &= u_{\langle a,b|\emptyset \rangle} + u_{\langle a,c|\emptyset \rangle} + u_{\langle a,d|\emptyset \rangle} \\ + \quad & u_{\langle c,d|b \rangle} + u_{\langle b,d|c \rangle} + u_{\langle b,c|d \rangle} + u_{\langle b,c|ad \rangle} + u_{\langle b,d|ac \rangle} + u_{\langle c,d|ab \rangle} \end{aligned}$$

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).

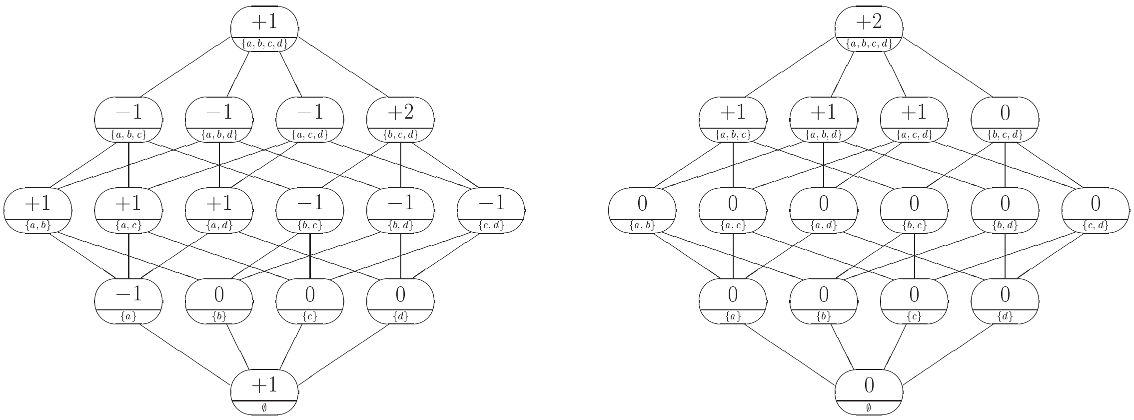


Figure 6.3: The imsets $u-v$ and the supermodular multiset $m_\circ$ from Example 6.2.

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 another case. $\diamondsuit$

Remark 6.4 Example 6.2 shows that the verification of whether a semi-elementary imset is facially implied by a structural imset requires the multiplication of the structural imset by 2 at least. Note that later Consequence 6.3 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 [122] 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, using the arguments similar to those from Section 6.3.2, that 2 is the minimal integer l_* satisfying

$$\forall\, u \in \mathcal{S}(N) \ \ \forall\, v \text{ semi-elementary imset over } N \quad u \rightharpoonup v \ \text{ iff } \ l_* \cdot u - v \in \mathcal{S}(N)\,.$$

Note that one has $l_* = 1$ in case $|N| \leq 3$ for the same reasons. $\triangle$

The following consequence of Lemma 6.1 has already been announced in Section 5.4.2.

Consequence 6.1 Let $\mathcal{Q}$ be a finite set of structural imsets over N. Then

$$\sup \{\mathcal{M}_u; u \in \mathcal{Q}\} = \mathcal{M}_v \quad \text{with } v = \sum_{u \in \mathcal{Q}} u\,, \tag{6.6}$$

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 $l \cdot w - v = \sum_{u \in \mathcal{Q}} (l_u \cdot w - u) \in \mathcal{S}(N)$. $\square$

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. $\triangle$

6.2.2 Skeletal characterization of facial implication

Lemma 6.2 Let u, v are structural imsets over N. Then $u \rightharpoonup v$ iff

$$\forall\, m \in \mathcal{K}_\ell^\diamond(N) \quad \langle m, v \rangle > 0 \;\Rightarrow\; \langle m, u \rangle > 0, \tag{6.7}$$

which is equivalent to the condition

$$\langle m, v \rangle > 0 \;\Rightarrow\; \langle m, u \rangle > 0 \quad \text{for every supermodular function } m \text{ over } N. \tag{6.8}$$

Moreover, the condition (6.7) is also equivalent to the requirement

$$l_* \cdot u - v \in \mathcal{S}(N) \quad \text{whenever } l_* \in \mathbb{N} \text{ such that } l_* \geq \langle r, v \rangle \text{ for every } r \in \mathcal{K}_\ell^\diamond(N). \tag{6.9}$$

Proof: Evidently (6.8) $\Rightarrow$ (6.7). Conversely, if (6.7) 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 (6.8) holds.

By Lemma 6.1 $u \rightharpoonup v$ iff the condition (6.2) holds. However, by Lemma 5.4 this is equivalent to the condition

$$\exists\, l \in \mathbb{N} \quad \forall\, m \in \mathcal{K}_\ell^\diamond(N) \quad l \cdot \langle m, u \rangle \geq \langle m, v \rangle\,,$$

which implies (6.7). The next step is to show that (6.7) implies

$$\forall\, m \in \mathcal{K}_\ell^\diamond(N) \quad l_* \cdot \langle m, u \rangle \geq \langle m, v \rangle \quad \text{for } \; l_* \in \mathbb{N} \; \text{ from (6.9)}\,. \tag{6.10}$$

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 (6.7) 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 (6.10) then implies $l_* \cdot u - v \in \mathcal{S}(N)$ by Lemma 5.4. Thus, (6.7)$\Rightarrow$(6.9). Since $\mathcal{K}_\ell^\diamond(N)$ is finite $l_* \in \mathbb{N}$ satisfying the requirement from (6.9) exists which means that (6.9) implies $u \rightharpoonup v$ by Lemma 6.1. $\square$

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

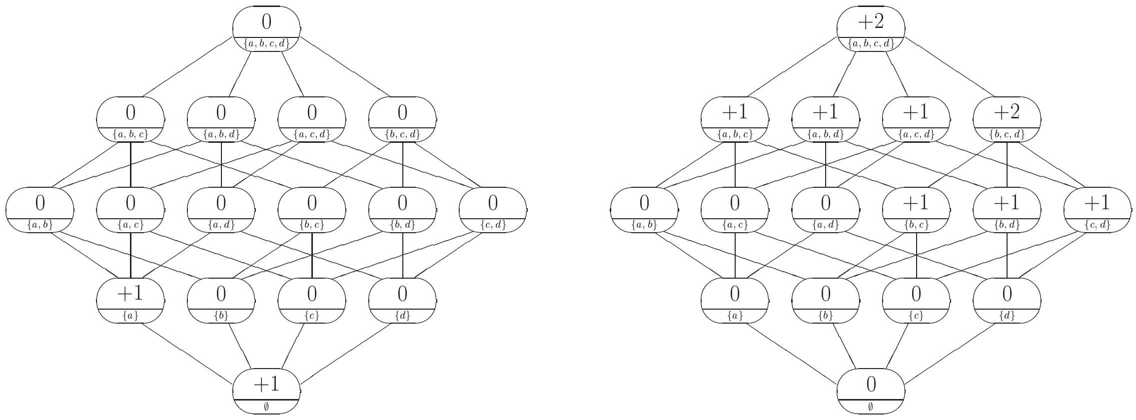


Figure 6.4: Strongly equivalent elements of the u-skeleton and the ℓ-skeleton.

an alternative criterion for 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 (6.7) and (6.8), 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 the moralization criterion and the d-separation criterion in case of DAG models (see Section 3.2).

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 the 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. 41) 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 (6.8), to see that $\neg(u \rightharpoonup w)$. Note that $m^{\{a\}\downarrow}$ belongs to the u-skeleton $\mathcal{K}^{\diamond}_{u}(N)$ 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 for facial equivalence of structural imsets.

Consequence 6.2 Let u, v be structural imsets over N. Then $u \rightleftharpoons v$ iff

$$\forall\, m \in \mathcal{K}^{\diamond}_{\ell}(N) \quad \langle m, u\rangle > 0 \quad \text{iff} \quad \langle m, v\rangle > 0\,, \tag{6.11}$$

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 for testing facial implication and equivalence are effective in particular in case $|N| \leq 4$ since $|\mathcal{K}^{\diamond}_{\ell}(N)|$ is small in that case - see Remark 5.4. They are still implementable in case $|N| = 5$; a computer program which realizes the 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.

6.3 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, the direct definition of structural imset is only 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 6.3 There exists $n \in \mathbb{N}$ such that

$$\forall \text{ imset } u \text{ over } N \quad u \in \mathcal{S}(N) \text{ iff } n \cdot u \in \mathcal{C}(N)\,. \tag{6.12}$$

Proof: One can apply Theorem 16.4 from [109] 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

$$\forall\, u \in \mathcal{S}(N) \quad u = \sum_{v \in \mathcal{H}(N)} k_v \cdot v \ \text{ for some } \ k_v \in \mathbb{Z}^+\,. \tag{6.13}$$

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 (6.13) implies that $n \cdot u \in \mathcal{C}(N)$ for every $u \in \mathcal{S}(N)$. $\square$

A natural question is what is the minimal $n \in \mathbb{N}$ satisfying (6.12). I do not know the answer in case $|N| \geq 5$ (see Theme 10 on p. 176). But if $|N| \leq 4$ then $n = 1$; the main result of [122] can be formulated as follows.

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 6.6 It may be the case that the least $n \in \mathbb{N}$ satisfying (6.12) appears to be too high. An 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 6.3 (see Theme 9). It follows from the proof of Theorem 16.4 of [109] 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. Given a characterization of this kind one can possibly modify the procedure described in Remark 6.3 and test effectively whether an imset can be written as a combination of imsets from $\mathcal{H}(N)$ with non-negative integral coefficients. $\triangle$

6.3.2 Grade

Another natural question arising in connection with Lemma 6.1 is whether there exists $l \in \mathbb{N}$ such that

$$\forall\ u \in \mathcal{S}(N)\ \ v \in \mathcal{E}(N) \quad u \rightharpoonup v \ \text{ iff }\ l \cdot u - v \in \mathcal{S}(N). \tag{6.14}$$

The answer is yes. Evidently, if $l \in \mathbb{N}$ satisfies (6.14) then every $l' \in \mathbb{N}$, $l' \geq l$ satisfies it as well. Therefore one is interested in minimal $l \in \mathbb{N}$ satisfying (6.14) 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 (6.14).

Supposing $|N| \geq 2$ let us call the *grade*, denoted by $gra(N)$, the natural number

$$gra(N) = \max \{\, \langle r, w\rangle;\ r \in \mathcal{K}_\ell^\diamond(N)\ \ w \in \mathcal{E}(N)\,\}\,. \tag{6.15}$$

Evidently, $gra(N)$ depends on $|N|$ only. Lemma 6.2, the condition (6.9), implies this:

CONSEQUENCE 6.3 If $|N| \geq 2$ then $l = gra(N)$ satisfies (6.14).

Consequence 6.3 leads to an effective criterion for 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 6.4 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 [122]. Thus, $gra(N) = 1$ and by Consequence 6.3 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. $\square$

However as shown in [137] $gra(N) = 7$ in case $|N| = 5$. In fact, an example from Section 4.3 of [137] shows that the minimal natural number l for which (6.14) holds is just 7 in case $|N| = 5$. The question of what is the minimal $l \in \mathbb{N}$ satisfying (6.14) is partially answered by the following lemma.

LEMMA 6.4 Suppose that $|N| \geq 2$. Then the minimal $l_* \in \mathbb{N}$ satisfying

$$\forall\, u \in \mathcal{C}(N)\ \forall\, v \in \mathcal{E}(N) \quad u \rightharpoonup v \ \text{ iff } \ l_* \cdot u - v \in \mathcal{S}(N) \tag{6.16}$$

is the upper integer part of

$$gra_*(N) = \max_{m \in \mathcal{K}_\ell^\diamond(N)} \ \frac{\max\left\{\langle m, w\rangle\,;\ w \in \mathcal{E}(N)\right\}}{\min\left\{\langle m, w\rangle\,;\ w \in \mathcal{E}(N)\ \langle m, w\rangle \neq 0\right\}}\,. \tag{6.17}$$

Proof: To show that every $l_* \in \mathbb{N}$ with $l_* \geq gra_*(N)$ satisfies (6.16) 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 one 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 (6.17) is achieved. Then choose $\tilde{w} = u_{\langle a,b|K\rangle} \in \mathcal{E}(N)$ 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 $\langle a, b|K\rangle \in \mathcal{M}_u \setminus \mathcal{M}^m$ 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 6.7 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 independence equivalent element of the ℓ-skeleton. Thus, the ratios maximalized in (6.17) are invariants of classes of independence equivalence of skeletal imsets. Note that if $|N| \leq 5$ then

$$\forall\, m \in \mathcal{K}_\ell^\diamond(N) \quad \min\left\{\langle m, u\rangle;\ u \in \mathcal{E}(N)\ \langle m, u\rangle \neq 0\right\} = 1\,, \tag{6.18}$$

which implies that $gra(N) = gra_*(N)$ in this case. Thus, if the hypothesis (6.18) holds in general (see Question 10) then $gra(N)$ is the least $l \in \mathbb{N}$ satisfying (6.14) by Consequence 6.3 and Lemma 6.4. Note that an analogue of (6.18) holds for $|N| \leq 5$ and the u-skeleton (because of the operation of reflection mentioned on p. 167 or in Section 5.1.3 of [137]) but not for the o-skeleton. This is maybe the main difference between the o-standardization and the 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 the facial equivalence.

Let u be a structural imset over N. By the *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:

$$\mathcal{R}_u \equiv \bigcup_{\langle a,b|K\rangle \in \mathcal{M}_u \cap \mathcal{T}_\epsilon(N)} \{K, aK, bK, abK\} = \bigcup_{\langle A,B|C\rangle \in \mathcal{M}_u \setminus \mathcal{T}_\emptyset(N)} \{C, AC, BC, ABC\}\,. \quad (6.19)$$

Note that the equality of the unions in (6.19) 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. 22.

LEMMA 6.5 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.

Proof: To show $\mathcal{R}_u \subseteq \mathcal{D}_u^*$ consider elementary triplet $\langle a,b|K\rangle \in \mathcal{M}_u$ and find $k \in \mathbb{N}$ with $k\cdot u - u_{\langle a,b|K\rangle} \in \mathcal{S}(N)$. As $\langle m^{abK\uparrow}, u_{\langle a,b|K\rangle}\rangle > 0$ by Observation 5.1 $k\cdot\langle m^{abK\uparrow}, u\rangle > 0$. Hence, $u(T) > 0$ for some $abK \subseteq T \subseteq N$ which means $K, aK, bK, abK \in (\mathcal{D}_u^+)^{\downarrow}$. Analogously, $\langle m^{K\downarrow}, u_{\langle a,b|K\rangle}\rangle > 0$ implies $\langle m^{K\downarrow}, u\rangle > 0$ and $u(S') > 0$ for some $S' \subseteq K$ which means $K, aK, bK, abK \in (\mathcal{D}_u^+)^{\uparrow}$. 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 ($\mathcal{L}_u \subseteq \mathcal{U}_u$). Moreover, given $S \subseteq N$, one can show by reverse induction on $|S|$ that $[\,u(T) = 0$ for every $S \subset T \subseteq N\,]$ iff $[\,\langle m^{T\uparrow}, u\rangle = 0$ for every $S \subset T \subseteq N\,]$.
An analogous arguments allow one 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. An analogous claim about $\mathcal{R}_u$ is evident because of the definition of $\mathcal{R}_u$ 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 $\langle a,b|K\rangle \in \mathcal{M}_u$ whenever $k_v > 0$ for $v = u_{\langle a,b|K\rangle}$. 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 an elementary triplet $\langle a,b|K\rangle \in \mathcal{M}_u$ with $S \in \{K, aK, bK, abK\}$ and observe that $w = u + k\cdot u_{\langle a,b|K\rangle}$ 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.

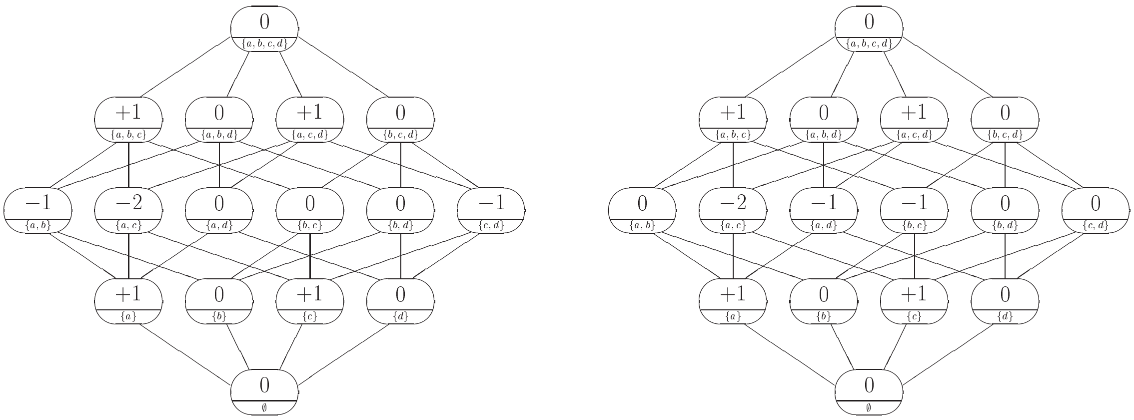


Figure 6.5: Two structural imsets with the same effective domain but different regions.

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, the regions and the effective domains of structural imsets over N coincide in case $|N| \leq 3$. $\diamondsuit$

REMARK 6.8 The significance of the concept of effective domain is that it allows one to restrict the considered class of elementary imsets when one is to test 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 $v = u_{\langle a,b|K\rangle}$ with $k_v > 0$ one has $\langle a, b|K\rangle \in \mathcal{M}_u$ and therefore by Lemma 6.5 $K, aK, bK, abK \in \mathcal{R}_u \subseteq \mathcal{D}^*_u$. Thus, a reduced class of elementary imsets $v = u_{\langle a,b|K\rangle}$ satisfying $K, aK, bK, abK \in \mathcal{D}^*_u$ 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 recognized immediately on basis of sole u. The region is a characteristic of the respective class of facially equivalent structural imsets and can be identified on basis of the imset only partially - see Lemma 6.5.

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 one to recognize immediately imsets which are not facially equivalent. A natural question is whether there exists a complete collection of invariant properties of similar type in the 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 $m \in \mathcal{K}^\diamond_\ell(N)$ is associated with an invariant attribute of a structural imset u over N, namely, the fact whether the scalar product $\langle m, u\rangle$ is zero 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 $|N| \geq 6$.

6.5 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:

$$\text{for every } P \in \Psi \text{ there exists a structural imset } u \text{ over } N \text{ such that } \mathcal{M}_u = \mathcal{M}_P, \quad (6.20)$$

$$\text{for every pair } P, Q \in \Psi \text{ there exists } R \in \Psi \text{ such that } \mathcal{M}_R = \mathcal{M}_P \cap \mathcal{M}_Q. \qquad (6.21)$$

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). On the other hand, the class of non-degenerate Gaussian measures over N considered in § 2.3.6 does not satisfy the condition (6.21) as Example 2.2 shows. The main reason for this is the limitation to fixed sample spaces $\mathsf{X}_i = \mathbb{R}$ for $i \in N$ made there.

The goal of this section is to show that after suitable restriction of the class of structural imsets the facial equivalence and the 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, that is, $\mathcal{M}_u = \mathcal{M}_P$. 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 6.6 Let Ψ be a class of probability measures over N satisfying (6.20) and (6.21) and $u \in \mathcal{S}(N)$. Then the class of structural imsets which are Markov equivalent to u relative to Ψ is the union of a finite collection $\mho$ of facial equivalence classes partially ordered by the 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 the repetitive application of the assumption (6.21) that $R \in \Psi(u)$ such that $\mathcal{M}_R = \mathcal{M}$ exists. By (6.20) 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 $w \rightleftharpoons v$. $\square$

REMARK 6.9 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 6.6. 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 $\mho$ only consists of one class of facial equivalence. 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 [127, 128]. △

The following fact immediately follows from Lemma 6.6.

CONSEQUENCE 6.5 Let Ψ be a class of probability measures over N satisfying (6.20) and (6.21). Consider the collection of Ψ-representable structural imsets $\mathcal{S}_\Psi(N)$ over N. The facial equivalence and the 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 to those problems is at disposal but a practical question of how to recognize imsets from $\mathcal{S}_\Psi(N)$ remains to be solved then. Moreover, the condition (6.21) also implies that the class of CI models induced by the imsets from $\mathcal{S}_\Psi(N)$ is a lattice. This assumption appears to be important in connection with certain learning procedures - for details see Section 8.1, pp. 134 - 137.

REMARK 6.10 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 the 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 (6.21)). The sum $u = \sum_{i=1}^{n} u_i$ then represents the aggregated information about the 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 the CI structure of any $P \in \Psi$.

Nevertheless, the possible inherent complexity of the problem of description of the lattice of CI structures arising within Ψ cannot be avoided. Indeed, the 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. △

Chapter 7

The problem of representative choice

This chapter deals with the problem of choice of a suitable representative within a class of facially equivalent structural imsets. It is an advanced subtask of the 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. Two auxiliary criteria for the choice of a representative are introduced and discussed. The dual way of description of structural independence models is mentioned in the last section of the chapter.

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, $u=3\cdot v+u_{\langle a,b|c\rangle}+u_{\langle a,c|d\rangle}+u_{\langle a,d|b\rangle}$. ♢

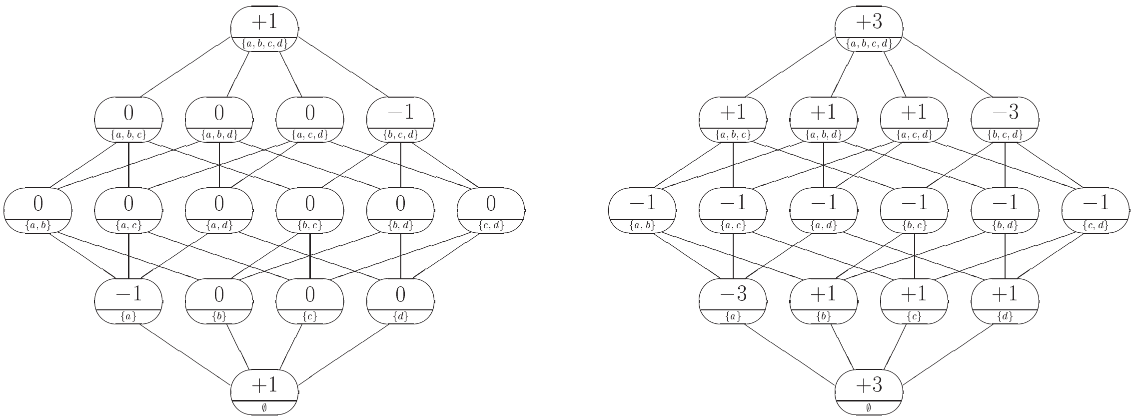


Figure 7.1: Non-proportional respective semi-elementary and baricentral imsets.

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 $u - v \in \mathcal{C}(N)$.

Proof: If $u \rightharpoonup v$ then, for every $w \in \mathcal{E}(N)$, $v \rightharpoonup w$ implies $u \rightharpoonup w$ and (7.1) gives $u - v \in \mathcal{C}(N)$. The converse follows from Lemma 6.1. $\square$

Note that testing combinatorial imsets is straithforward (see Remark 6.3). An analogous result holds if v is replaced by a semi-elementary imset (see Consequence 7.4 below). In particular, the induced model $\mathcal{M}_u$ can be easily identified on basis of a baricentral imset u; that means no multiplication of u is needed like in case of general structural imset (see Section 4.4.1). without 'multiplication').

REMARK 7.1 The terminology 'baricentral imset' was inspired by a 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. 100). Thus, a (minimal) balanced combination of all extreme imsets of this cone forms the 'baricentre' of the cone.

A natural question is what is the number of baricentral imsets over N. A very rough upper bound 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 $Ł(S) \subseteq \mathbb{Z}$ which depends on $k = |S|$ only. Concretely, $Ł(S) = \{-k(n-k), \ldots, \binom{k}{2} + \binom{n-k}{2}\}$ for $2 \leq k \leq n-2$, $Ł(S) = \{0, \ldots, \binom{n}{2}\}$ for $k \in \{0, n\}$ and $Ł(S) = \{-n+1, \ldots, \frac{(n-1)(n-2)}{2}\}$ for $k \in \{1, n-1\}$. Thus, $|Ł(S)| = \binom{n}{2} + 1$ 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 $Ł(S)$ 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 considerationgives 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 the symmetry property (see p. 21) 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 $\text{Ł}(S)$ for $S \subseteq N$ in case $n \geq 6$ which means that the memory demands are slightly lower in case of representation by baricentral imsets. $\triangle$

On the other hand, the actual number of baricentral imsets, that is, 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 to the problem of representative choice from the 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 for the respective baricentral imsets of graphical models are needed (see Theme 3 in Chapter 9). 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 the formula

$$u_G = \delta_N - \delta_\emptyset + \sum_{c \in N} \delta_{pa_G(c)} - \delta_{c \cup pa_G(c)}. \tag{7.3}$$

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.

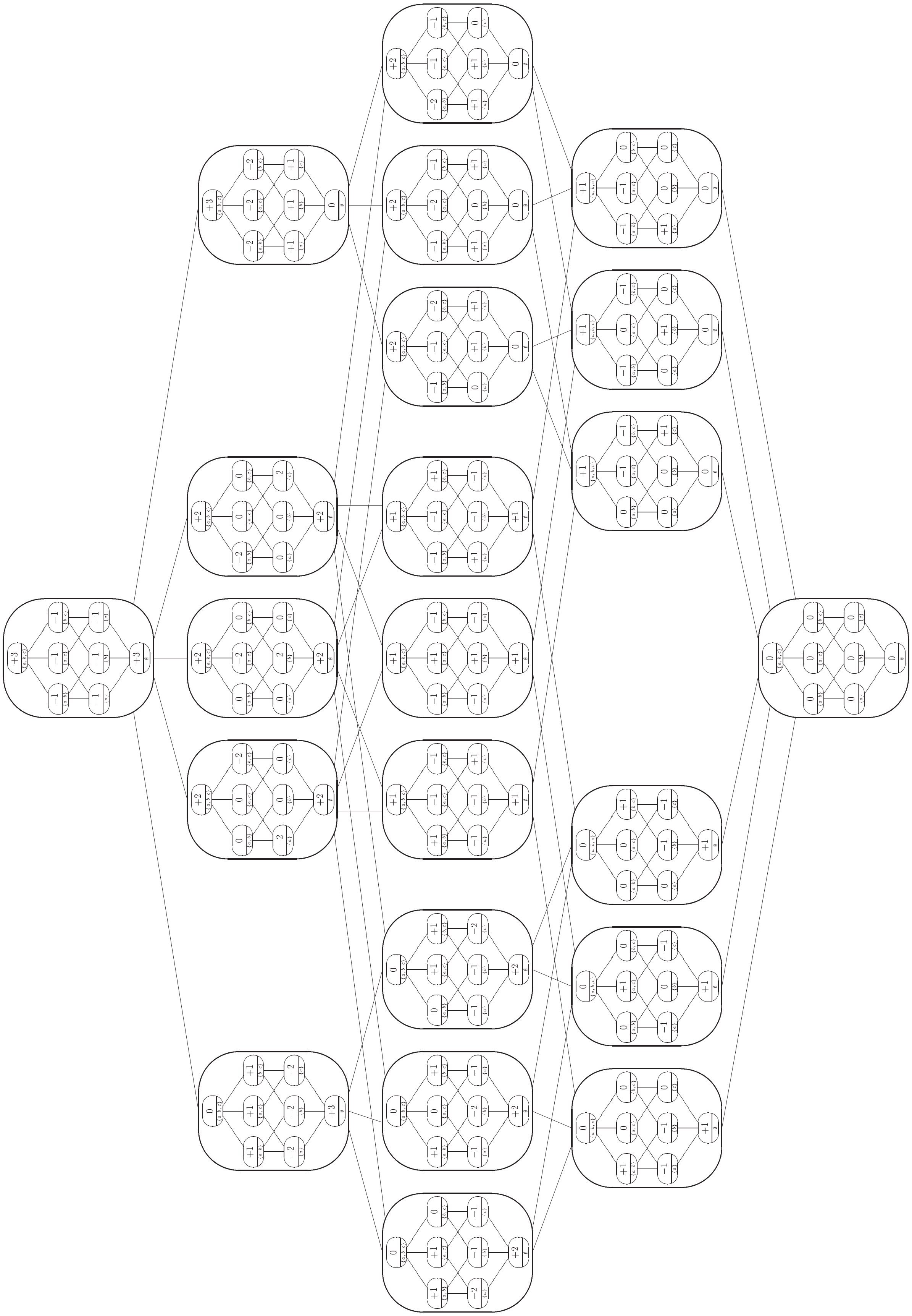


Figure 7.2: Baricentral imsets over $N = \{a, b, c\}$ (rotated).

Proof: Consider a fixed ordering $a_1, \ldots, a_n$, $n \geq 1$ of nodes of G consonant with the 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

$$u \equiv \sum_{j=1}^{n} u_j = \sum_{j=1}^{n} \delta_{\{a_1,\ldots,a_j\}} - \delta_{\{a_1,\ldots,a_{j-1}\}} - \delta_{a_j\, pa_G(a_j)} + \delta_{pa_G(a_j)} = u_G$$

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 a causal input list.

Since $\mathcal{M}_u$ is a semi-graphoid containing (7.4) the result from [141] 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 [42] 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 6.5). △

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 can be verified with the help of transformational characterization of equivalent acyclic directed graphs - see p. 47. Owing to that result it suffices to verify $u_G = u_H$ if H is obtained from G by the legal reversal of an arrow $a \to b$ in G. In that case $pa_G(c) = pa_H(c)$ for any $c \in N \setminus \{a, b\}$. Thus, to show $u_G = u_H$ one has to evidence

$$\sum_{c \in \{a,b\}} \delta_{pa_G(c)} - \delta_{c \,\cup\, pa_G(c)} \;=\; \sum_{c \in \{a,b\}} \delta_{pa_H(c)} - \delta_{c \,\cup\, pa_H(c)}\,.$$

This is easy because by the condition enabling the legal reversal of $a \to b$ one has $pa_G(a) = pa_H(b) = C$, $pa_G(b) = C \cup a$ and $pa_H(a) = C \cup b$ for a certain set $C \subseteq N$. □

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 $N \setminus ABC$. Consider an undirected graph over N in which every pair of distinct nodes is a line except pairs $[a, b]$ where $a \in A$, $b \in B$. Construct a directed graph G which has the above 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 $u_G = u_{\langle A,B|C\rangle}$. △

Note that standard imset representation of DAG models appears to be advantageous from the point of view of learning DAG models - see Section 8.4 for details. The poset of standard imsets over $\{a, b, c\}$ is shown in Figure 7.3.

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 (3.3) on p. 52). 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 the respective separators. In particular, the right-hand side of (7.6) does not depend on the choice of ϱ and u_H 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)

$$u_H = \delta_N - \delta_T + u_{H'} + \sum_{C_m \in \mathcal{B} \subseteq \mathcal{C}} (-1)^{|\mathcal{B}|} \cdot \delta_{\bigcap \mathcal{B}}. \tag{7.8}$$

The running intersection property says $S_m = C_m \cap (\bigcup_{j<m} C_j) \subseteq C_k$ for some $k < m$. This allows one 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). $\square$

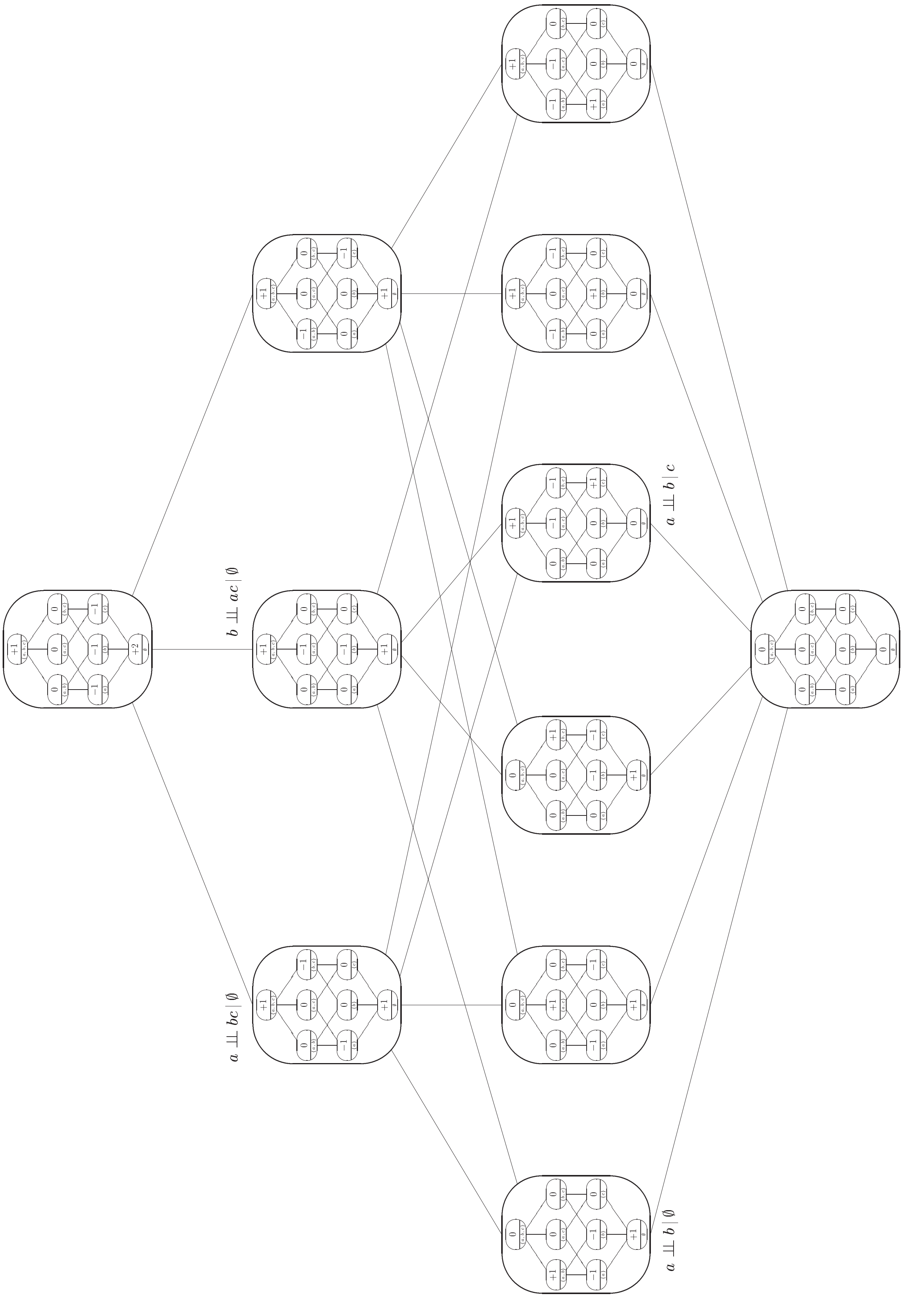


Figure 7.3: Standard imsets for DAG models over $N = \{a, b, c\}$ (rotated).

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 (3.4) 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. The 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 the nodes of R_i precede the 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 $j < l$ 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 (see Section 3.3) with the same underlying graph. To show $\mathcal{M}_G = \mathcal{M}_H$ by well-known graphical characterization [35] (see p. 51) 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 of nodes of H within R_i, $i = 1, \ldots, m$ then (7.3) gives

$$u_G = \delta_N - \delta_\emptyset + \sum_{i=1}^{m} \sum_{d=d^i_*}^{d^i_\dagger} \{\delta_{pa_G(d)} - \delta_{d \cup pa_G(d)}\} = \delta_N - \delta_\emptyset + \sum_{i=1}^{m} \{\, \delta_{S_i} - \delta_{C_i} \,\}$$

(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. △

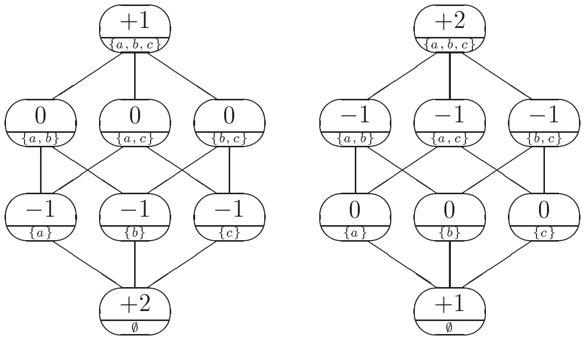


Figure 7.4: Two distinct equivalent imsets of the least degree.

7.3 Imsets of the least degree

One of possible approaches to the choice of a representative of a class $\wp$ of facially equivalent structural imsets is to choose a combinatorial imset of the least degree (for the definition of degree see p. 66). 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 a prescribed degree exists. In particular, the set of combinatorial imsets in $\wp$ of the least degree 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 $v = u_{\langle a,b|c\rangle} + u_{\langle a,c|b\rangle} + u_{\langle b,c|\emptyset\rangle}$, which are shown in Figure 7.4, 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 annonced 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 by Consequence 4.2. □

However, the question of whether the converse implication holds remains open (see Question 6 on p. 165).

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 special closure operation on subsets of $X = \mathcal{T}(N)$ (see Section 10.2 p. 185 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 the 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 [135]. △

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 $\wp = \{v \in \mathcal{C}(N); \mathcal{M}_v = \mathcal{M}\}$. Then every element u of $\wp$ which is minimal 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 $\mathcal{G} = \{\langle a,b|K\rangle \in \mathcal{T}_\epsilon(N)\,;\; k_{u_{\langle a,b|K\rangle}} = 1 \;\; a \prec b\}$ 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 the 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 $\mathcal{G} \subseteq \mathcal{M}'$. Because it is a semi-graphoid, $\mathcal{M}_w \subseteq \mathcal{M}'$ for $w = u_{\langle a,b|K\rangle}$, $\langle a,b|K\rangle \in \mathcal{G}$ 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 $v = \sum_{\langle a,b|K\rangle \in \mathcal{F}} u_{\langle a,b|K\rangle} \in \mathcal{C}(N)$. 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 [11]. 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). The 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 a minimal elementary generator $\mathcal{G} \subseteq \mathcal{M} \cap \mathcal{T}_\epsilon(N)$ such that $v = \sum_{\langle e,f|K\rangle \in \mathcal{G}} u_{\langle e,f|K\rangle}$ is not an imset of the least degree. Let us consider the independence model from Example 3.1 on p. 48, namely, the 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 7.5). 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 the definition of baricentral imset, Observation 7.2, Lemma 7.5; the second part from 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.

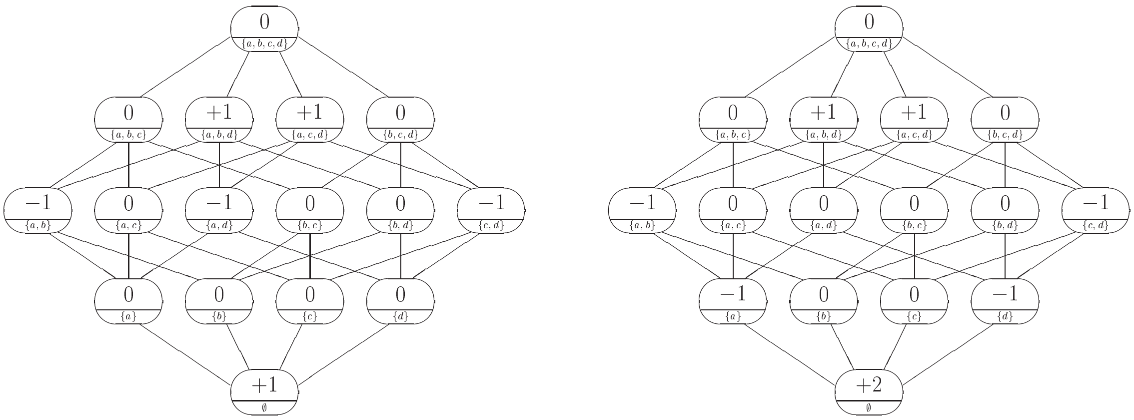


Figure 7.5: An imset of the least degree and the respective imset with the least lower class.

7.4 Span

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. 67) but they differ for a non-zero imset. Moreover, by Consequence 4.4 the 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 7.4 tells more about which marginals determine the whole Markovian measure in comparison with the imset shown in the right-hand picture of Figure 7.4. Thus, facially equivalent imsets need not be equiinformative from this point of view. This consideration motivates an informal concept of *span 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 $\mathcal{U} \equiv \mathcal{U}_u$ and the class of probability measures 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 only determined by $\mathcal{M}$.

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 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 $\mathcal{M}$.

A related question is for which $\mathcal{M} \in \mathcal{U}(N)$ just one minimal, respectively unimarginal, class exists. In other words, the question is for which structural models $\mathcal{M}$ the *least determining class* (respectively the *least unimarginal class*) for $\mathcal{M}$ exists, that is, a determining (respectively unimarginal) class $\mathcal{D}$ such that one has $\mathcal{D} \subseteq \mathcal{D}'$ for every determining (respectively 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 4.4). □

Recall that Consequence 4.3 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 the above implications as follows:

$$\text{lower class} \quad \Longrightarrow \quad \text{determining class} \quad \Longrightarrow \quad \text{unimarginal class.} \tag{7.11}$$

Note that a determining class need not be a lower class (see Example 7.4 below) and the question of whether every unimarginal class is determining remains open - see Question 12 on p. 178. On the other hand, it is known that the concepts of determining and unimarginal classes essentially coincide for DAG models (see Consequence 7.5 in Section 7.4.3).

REMARK 7.7 The concept of unimarginal class can alternatively be 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 distribution 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 14 in Chapter 9). △

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. 120. On the other hand, as the subsequent example shows there are classes of facial equivalence which do not have these imsets but several *imsets with minimal lower class*, that is, 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

$$\mathcal{M} = \{\, \langle A, B|C\rangle \in \mathcal{T}(N);\ |C| \geq 1\,\}, \qquad \mathcal{K} = \mathcal{E}_1(N) \cup \mathcal{E}_2(N)\,.$$

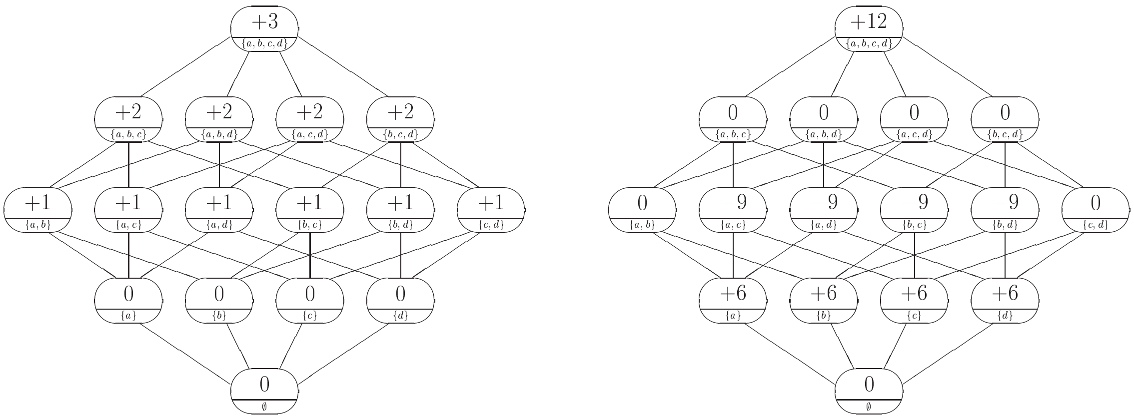


Figure 7.6: A multiset producing $\mathcal{M}$ and the respective imset with a minimal lower class.

Observe that $\mathcal{M} = \mathcal{M}^m$ for the supermodular imset m shown in the left-hand picture of Figure 7.6 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

$$k_w = \begin{cases} 4 & \text{iff } w = u_{\langle a,b|K\rangle} \in \mathcal{K} \text{ or } w = u_{\langle c,d|K\rangle} \in \mathcal{K}\,, \\ 1 & \text{for remaining } w \in \mathcal{K}\,, \end{cases}$$

(u is shown in the right-hand picture of Figure 7.6). By Observation 5.6 $\mathcal{M}_u \cap \mathcal{T}_\epsilon(N) \subseteq \mathcal{M}^m \cap \mathcal{T}_\epsilon(N)$ which implies $\mathcal{M}_u \subseteq \mathcal{M}^m$ by Lemma 2.2. Hence $\mathcal{M}_u = \mathcal{M}^m = \mathcal{M}$.

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 $w \in \mathcal{K}'$. This is 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 $w \in \mathcal{K}'$ and the assumption $l_w = 0$ for $w \in \mathcal{K}'$, therefore, implies $\langle m^{ad\uparrow}, v\rangle = 0$ which meams by Observation 5.6 that $\mathcal{M} = \mathcal{M}_v$ is a subset of the model produced by $m^{ad\uparrow}$. This contradicts the fact $\langle a, d|bc\rangle \in \mathcal{M}$ and $\langle m^{ad\uparrow}, u_{\langle a,d|bc\rangle}\rangle = 1$. Thus, one can put $s = \sum_{S\in\mathcal{S}} \delta_S$ and observe that $\langle s, w\rangle \leq 0$ 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.

An 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 that $v(S) < 0$ for some $S \in \mathcal{S}$ necessitates $v(ac) < 0$. 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 a 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 a minimal determining class and an analogous conclusion can be made for any class obtained by a permutation of variables either from $\mathcal{D}_1$ or from $\mathcal{D}_2 = \{ab, ac, ad\}^\downarrow$. Note that 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 descending 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 $S \in \mathcal{U}$. $\square$

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 $\mathcal{L}_u$ both a determining class and an 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 one to verify that the condition (7.12) for $T = pa_G(c)$ is fulfilled for $\mathcal{M} = \mathcal{M}_G$. $\square$

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 [11] helped to show that in case $|N| \leq 4$ the converse holds, that is, given an acyclic directed graph G, the only imset satisfying these two conditions is the standard imset u_G. The question of whether these two requirements determine standard imsets for acyclic directed graphs in general remains open - see Question 11 on p. 177. $\triangle$

Another interesting feature of the standard imset u for an acyclic directed graph is that it has a lot of zero values, namely, that it vanishes outside $\mathcal{L}_u \cup \mathcal{U}_u^{\max}$.

7.5 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.

7.5.1 Coportraits

This is to explain the dual perspective in more details with the help of the concept of Galois connection from Section 5.4. It was explained there (p. 92) 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

$$\{v \in \mathcal{E}(N)\,;\, v = u_{\langle a,b|K\rangle} \text{ where } \langle a,b|K\rangle \in \mathcal{M} \cap \mathcal{T}_\epsilon(N)\,\}.$$

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 $\text{Œ} = \mathcal{E}(N)$:

$$\begin{aligned} \mathcal{E}_u &\equiv \{v \in \mathcal{E}(N)\,;\, v = u_{\langle a,b|K\rangle},\, \langle a,b|K\rangle \in \mathcal{M}_u\} = \{v \in \mathcal{E}(N)\,;\, u \rightharpoonup v\}, \\ \mathcal{E}^m &\equiv \{v \in \mathcal{E}(N)\,;\, v = u_{\langle a,b|K\rangle},\, \langle a,b|K\rangle \in \mathcal{M}^m\} = \{v \in \mathcal{E}(N)\,;\, \langle m,v\rangle = 0\}. \end{aligned} \tag{7.13}$$

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 also be described by means of its intent, that is, in terms of attributes. In this case the set of attributes is the ℓ-skeleton $\text{Æ} = \mathcal{K}_\ell^\diamond(N)$ 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

$$\mathcal{H}_u = \{r \in \mathcal{K}_\ell^\diamond(N)\,;\, \langle r,u\rangle = 0\}. \tag{7.14}$$

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 7.7.

Remark 7.9 This is to explain the terminology. The idea of dual description of a structural model was presented already in [129] where the concept of *portrait of* $u \in \mathcal{S}(N)$ was introduced as the set of skeletal imsets

$$\{r \in \mathcal{K}_\ell^\diamond(N)\,;\, \langle r,u\rangle > 0\}. \tag{7.15}$$

Thus, the coportrait $\mathcal{H}_u$ is nothing but the relative complement of (7.15) in $\mathcal{K}_\ell^\diamond(N)$ which motivated my terminology here. Provided that the ℓ-skeleton is known (7.15) and (7.14) are equiinformative but the concept of coportrait seems to be more natural from the theoretical point of view (in the light of Galois connection). Despite this fact I decided

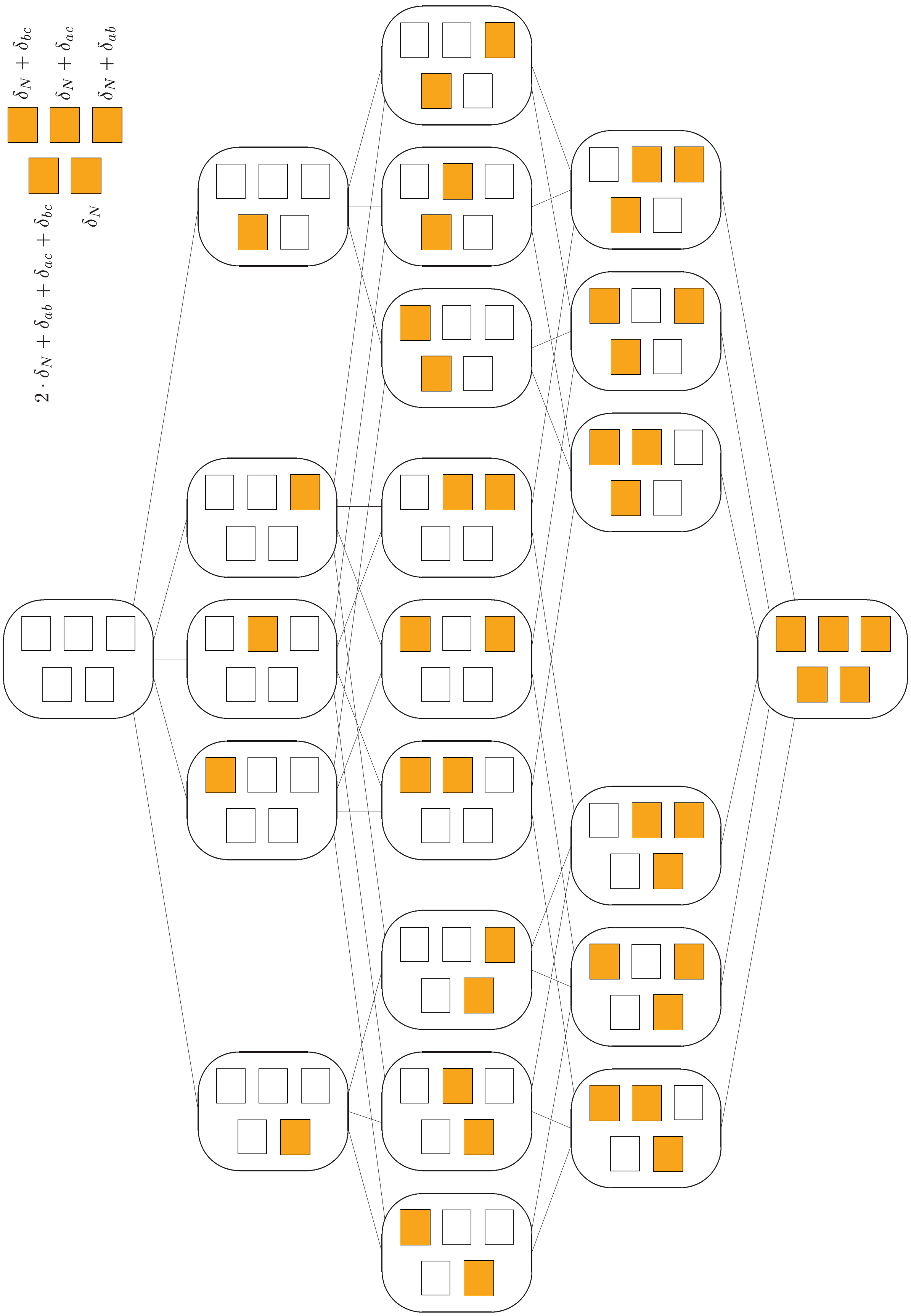


Figure 7.7: Coportraits of structural imsets over $N = \{a, b, c\}$ (rotated).

to keep the former terminology and do not rename things. The reason why I preferred in [129] (7.15) to (7.14) 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 the explored cases). Practical consequence should be that portraits have less cardinality than coportraits for those structural imsets which induce '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.

7.5.2 Dual baricentral imsets and a global view

As mentioned before Observation 5.7 (p. 95) one can take a dual point of view and describe structural models as independence models produced by ℓ-standardized supermodular imsets. One can introduce an analogue of facial equivalence and facial 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 and Fact in Section 10.8.2) so that these imsetes 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.

$$m = \sum_{r \in \mathcal{K}_\ell^\diamond(N),\, \mathcal{M}^m \subseteq \mathcal{M}^r} r\,. \tag{7.16}$$

The corresponding poset of dual baricentral imsets is shown in Figure 7.8.

Global view

Of course, owing to Consequence 5.4 and (7.13) 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 (7.13) 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 7.9. To describe a structural model one can use

1. a set of elementary imsets,
2. a structural imset,
3. a set of ℓ-skeletal imsets,

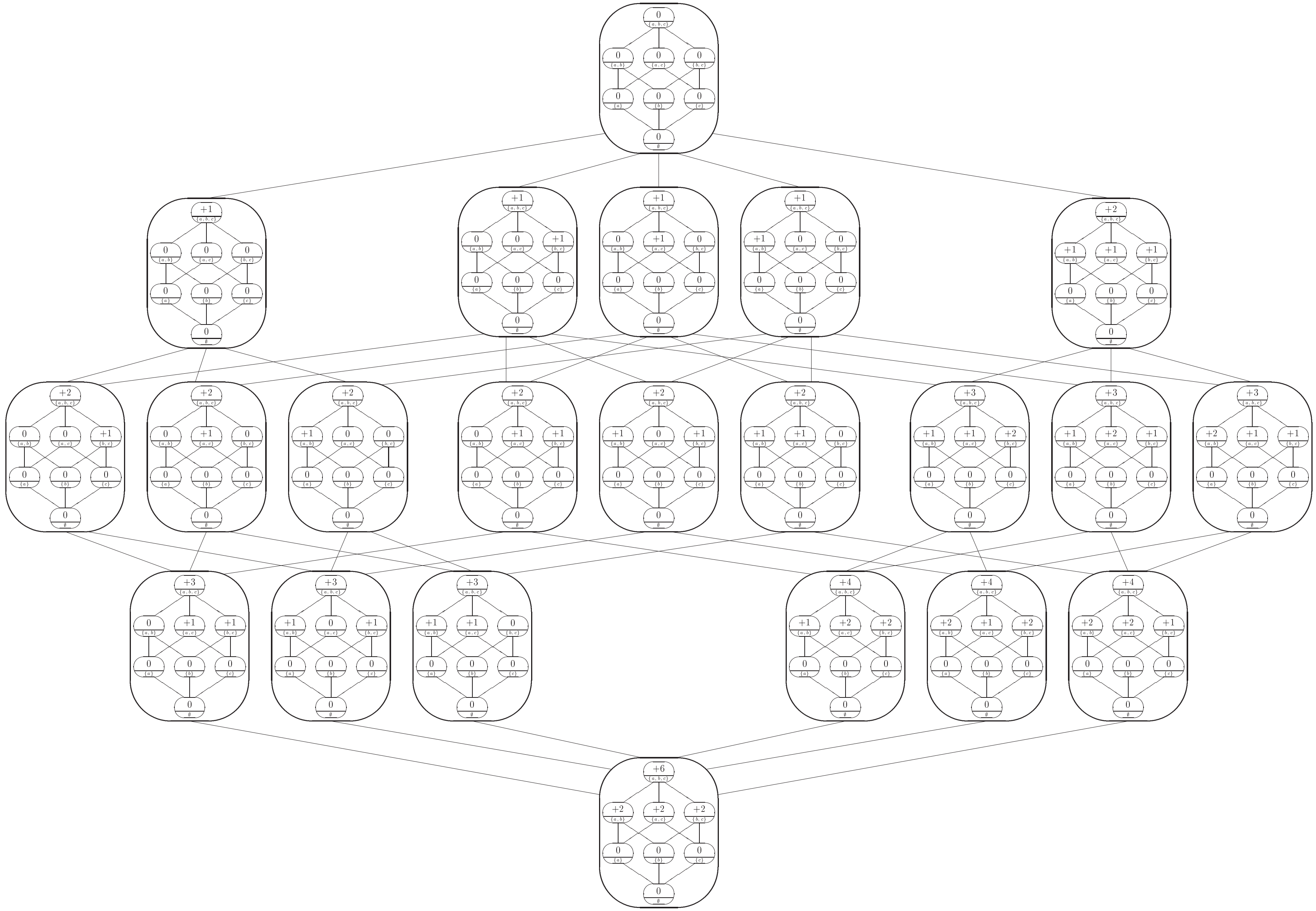


Figure 7.8: Dual baricentral imsets over $N = \{a, b, c\}$ (rotated).

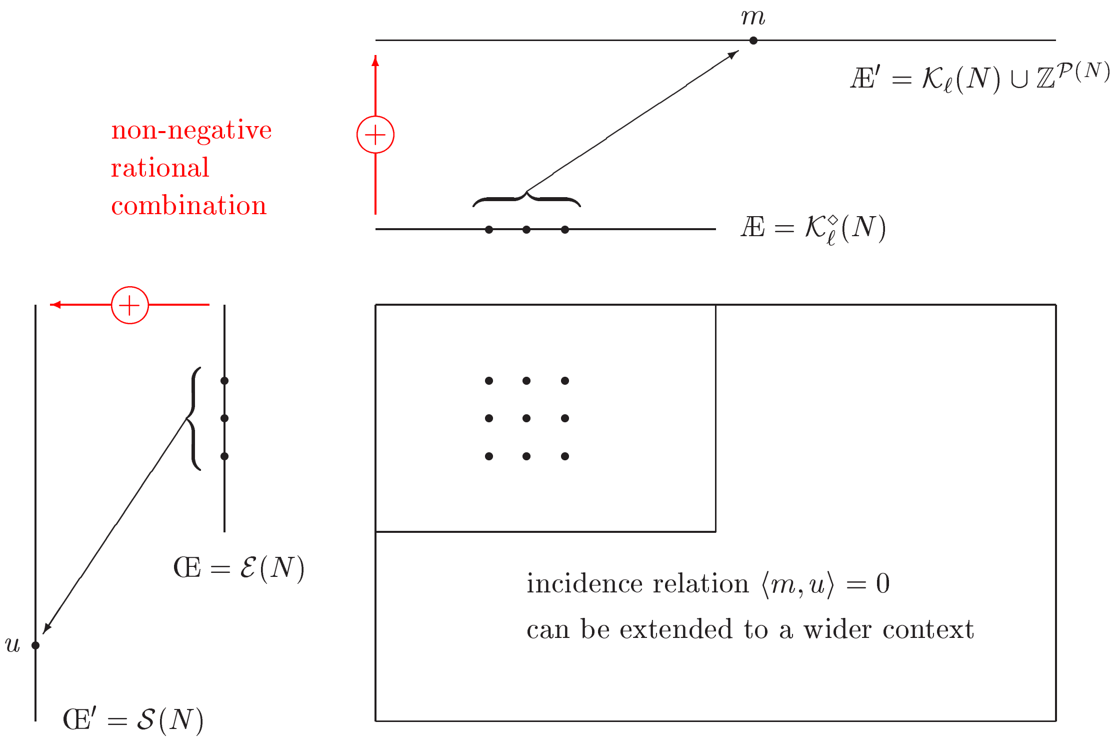


Figure 7.9: Extension of Galois connection for structural models - illustration.

4. an ℓ-standardized supermodular imset.

Recall that the set of elementary imsets can be viewed as the direct translation of the considered structural model, the structural imset is obtained by a non-negative rational combination of elementary imsets, the set of skeletal imsets is obtained by the transfomation given by the Galois connection and the supermodular imset by a non-negative rational combination of skeletal imsets.

Let me emphasize that unlike the case of general Galois connection described in Section 5.4.1 an additional superstructure of summing elementary, respectively summing skeletal imsets, is at disposal. This fact allows one to describe the respective relationships among formal concepts (namely the relation 'be a subconcept') with the help of algebraic operations, more precisely by means of arithmetics of integers! This is the main asset of the described approach.

Remark 7.10 However the dual approach exhibits some different mathematical properties. One can introduce an analogue of the concept of combinatorial imset, that is, 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. 86 equals to the sum of three ℓ-skeletal imsets from the second line of Figure 5.1.

Chapter 8

Learning

This chapter is devoted to the methods of learning CI structures on basis of data. However, it is not the aim of this text to make an overview of numerous methods of learning graphical models from the literature. More likely, the purpose of this chapter is to show how to apply the method of structural imsets to learning DAG models and to indicate that this approach can be extended to learning more general classes of models of CI structure.

Some thoughts about the learning methods which are based on statistical CI tests are in Section 8.1. The next two sections of the chapter contain an analysis of some of those algorithms for of learning DAG models which are based on the maximization of a quality criterion and deal with related questions. It is argued in Section 8.4 that the method of structural imsets can be applied in this area as well.

8.1 Two approaches to learning

There is a plenty of literature about learning graphical models of CI structure - both in the area of statistics and in the area of artificial intelligence. For an overview of some of these methods see § 1.4 and § 4 of [13] and § 3 of [19]. The literature concerns learning UG models, DAG models and DAG models with hidden variables; most attention is devoted to learning DAG models.

In my view, the algorithms for learning graphical models can be divided into two groups on basis of the basic methodological approach.

- Some of these algorithms are based on *significance tests* between two statistical models (of graphical CI structure). These significance tests often correspond to statistical tests for the validity of some CI statements.

- Other algorithms are based on the maximization of a suitable *quality criterion* designed by a statistician. On basis of the way of derivation of the criterion, the algorithms of this kind could be further divided into those based on the Bayesian approach and those which stem from a frequentist point of view.

Nevertheless, some algorithms may be included in both groups because they can be interpreted in both ways and there is a simulation method which is applicable to learning graphical models and which does not belong to either of these two groups.

Data faithfulness assumption

Typical examples of algorithms based on significance tests are the SGS algorithm for learning DAG models and its more effective modifications known as the PC algorithm and the PC* algorithm described in [115]. These procedures stem from the premise called the *data faithfulness assumption* which can be paraphrased as follows (cf. Section 1.1):

> data are 'generated' by a probability measure P which is perfectly Markovian with respect to an object $\boldsymbol{o}$ within the considered universum of objects of discrete mathematics.

In case of the above algorithms from [115] the universum of the objects is the collection of acyclic directed graphs over N. The algorithms are based on statistical CI tests, that is, tests which on basis of given data either reject or do not reject the hypothesis that a certain CI statement $a \perp\!\!\!\perp b \,|\, C \;[P]$ is true. Tests of this kind are usually derived from statistics which are known to be the measures of stochastic conditional dependence. For example, in § 5.5 of [115] two statistics of this kind are mentioned in the discrete case: the X^2-statistics and the G^2-statistics (see Section 10.10.1). The goal of the algorithms is to determine the equivalence class of graphs consisting of acyclic directed graphs with respect to which P is perfectly Markovian. The basic step is an observation that if P is perfectly Markovian with respect to an acyclic directed graph G then the following characterization of edges and immoralities in G holds:

$$[a,b] \text{ is an edge in } G \quad \Leftrightarrow \quad \forall\, C \subseteq N \setminus \{a,b\} \quad a \top\!\!\!\top b \,|\, C \;[P]\,,$$

and if $[a,b]$, $[b,c]$ are edges in G while $[a,b]$ is not an edge in G then

$$a \to c \leftarrow b \text{ in } G \quad \Leftrightarrow \quad \forall\, C \;\; c \in C \subseteq N \setminus \{a,b\} \quad a \top\!\!\!\top b \,|\, C \;[P]$$

- see [142] or [55]. On the condition that data are 'generated' from P the above mentioned statistical CI tests give a criterion for the composite conditional dependence statements on the right-hand side. This allows one to determine the underlying graph of G and all immoralities in G. Thus, a hybrid graph H, called a *pattern*, can be obtained which has the same underlying graph as G and just those arrows which belong to the immoralities in G; the other edges in H are lines. Final step is a purely graphical procedure whose aim is to direct some other edges of the pattern so as to have the direction which is shared by all acyclic directed graphs which are equivalent to G. Nevertheless, the hybrid graph $H_{\ddagger}$ obtained by the algorithm from [115] need not be complete in the sense that every *essential arrow*, that is, an arrow which has the same direction in all acyclic directed graphs equivalent to G, is an arrow with this direction in $H_{\ddagger}$. However, the final step of the SGS algorithm may be replaced by the procedure described in [86]. The resulting hybrid graph is then the essential graph, named also the completed pattern, which can serve as the unique representative of the equivalence class of acyclic directed graphs with respect to which P is perfectly Markovian.

Analogous procedures, sometimes called *recovery algorithms*, were proposed for other classes of graphs. An algorithm of this kind for (classic) chain graphs whose result is the largest chain graph (see p. 51) was presented in [131]. In § 5.6 of [54] a reference is given to an algorithm for learning decomposable models which is based on statistical tests and on the idea of gradual decomposition of cliques of a triangulated undirected graph. Modification of the PC algorithm for learning acyclic directed graphs with hidden variables is described in § 6.2 of [115].

But all above mentioned algorithms are ultimately based on the data faithfulness assumption and the validity of this assumption can hardly be ensured. Indeed, its validity for arbitrary data actually means that, for every probability measure P from the considered distribution framework, there exists an object $\boldsymbol{o}$ of discrete mathematics in the considered universum of objects such that P is perfectly Markovian with respect to $\boldsymbol{o}$. This is nothing but the completeness requirement mentioned in Section 1.1. However, it is explained in Section 3.6 that such a condition cannot be fulfilled for arbitrary class of graphical models and the discrete distribution framework. The point of this consideration is as follows:

> *learning methods based on statistical CI tests are safely applicable only if the data faithfullness assumption is guaranteed.*

Remark 8.1 A boon of the book [115] is that the data faithfulness assumption was made explicit there and the limitation of learning methods based on statistical CI tests was indicated in that way. Unfortunately, some researchers in the area of artificial intelligence do not seem to be aware of this limitation which leads to the methodological error mentioned in the end of Section 1.1. These researchers use the procedures for learning Bayesian networks based on CI tests for 'single edge removals' which have already been criticized in [147]. An example of a similar procedure is the `Construct`-TAN algorithm for 'learning restricted Bayesian networks' from [34] and its modification from [20]. The argument used on p. 8 of [34] to justify the restriction of the class of TAN models is that the model learned by usual method of the maximization of a quality criterion may result in a 'poor' classification. However, the `Costruct`-TAN algorithm is ultimately based on the values of the G^2-statistic for triplets of the form $\langle a, b|c\rangle$, $a, b, c \in N$ which correspond to the CI test for the respective elementary statement $a \perp\!\!\!\perp b \,|\, c$. Owing to the restriction to a special class of TAN models the 'learned' model always involves a fixed-context CI statement $a \perp\!\!\!\perp b \,|\, N \setminus \{a, b\}$. But no evidence for the validity of this CI statement was provided by data and the procedure is, therefore, methodologically incorrect, unless the data faithfulness assumption is somehow ensured!

Another issue is that researchers in the area of artificial intelligence often support the applicability of their various heuristic algorithms for learning Bayesian networks of the above type by computer simulations. The trouble is that in simulations of this type data are 'generated' artificially from those probability measures which are perfectly Markovian with respect to some acyclic directed graphs. Thus, the data faithfulness assumption is 'fulfilled' as a result of the way artificial data were generated! These algorithms may appear worthless in case those real data for which this assumption is not satisfied. I think that real data of this sort are quite common. In my view, very easy example of real data which do not meet the data faithfulness assumption was given already by Whittaker - see Example 8.4.1 and Exercise 1 in § 8.6 in [146]. △

Need for lattice structure

One of the ways to behave in the situation when data do not fit one of the considered models is the change of the objective of the learning procedure. Instead of seeking 'the only true model' the aim may become to find the 'best approximation' in the universum of considered statistical models or an approximation which is 'good enough'. Then, of course, one has to introduce a suitable measure of approximation.

An example of an algorithm whose aim is to find an approximation in the above sense and which can also be interpreted as an algorithm based on significance tests is the *simple two step procedure* for learnig UG models described in § 8.3 of Whittaker's book [146]. The measure which Whittaker uses is the *deviance* of a statistical model $\mathbb{M}_G$ determined by an undirected graph G (see Section 10.10.2). It is defined a multiple by 2 of the difference between the unconstrained maximum and the constrained maximum of the logarithm of the likelihood function over $\mathbb{M}$ (see Section 10.10.4). Whittaker showed that in cases he deals with the deviance has information-theoretical interpretation in terms of relative entropy (see [146] Proposition 6.8.1 in Gaussian case and Proposition 7.4.2 in discrete case). The procedure described by Whittaker is a series of mutual comparisons of two statistical models $\mathbb{M}$ and $\mathbb{M}'$ which are each time nested. Moreover, during the algorithm only 'close' models are compared, that is, models determined by undirected graphs which differ in the presence of one edge. This means, in the statistical interpretation (see Section 10.10.2), that parametrizations differ in the presence of one parameter. To compare statistical models a significance test based on the *deviance difference* is used. Whittaker explicitly wrote on p. 228 of [146] that, if possible, he would like to give the meaning to the deviance difference (between two close models), namely the meaning of a statistic for testing (elementary) CI statements. Indeed, this interpretation is possible in most of the situation Whittaker deals with; for example, for a pair of decomposable models in the discrete distribution framework - see Remark 8.14 on p. 158. The 'simple two stage procedure' (p. 256 in [146]) could be described as follows.

(i) The algorithm starts with the *saturated model* (see Section 10.10.3), that is, the model determined by an undirected graph H^0 in which every pair distinct nodes of N is a line.

(ii) The significance tests for the presence or the absence of each particular edge in H^0 are performed. In the discrete case, these are the statistical tests for CI statements of the form $d \perp\!\!\!\perp e \,|\, N \setminus \{d, e\}$ based on the G^2-statistics (see Section 10.10.1).

(iii) The lines $a - b$ in H^0 for which these tests do not reject the hypothesis about the validity of $a \perp\!\!\!\perp b \,|\, N \setminus \{a, b\}$ are removed from H^0 and a graph H^1 is obtained.

This is the end of so-called *backward phase* of the procedure. Since the deviance of the model $\mathbb{M}_{H^1}$ may be too high the algorithm may continue by its *forward phase.*

(iv) For every pair $[a, b]$ which represents a missing line in H^1 one computes the deviance differences between the model determined by a graph which is created by adding $a - b$ to H^1 and the model determined by H^1. Note that the result is often the value of the G^2-statistics for testing a certain CI statement $a \perp\!\!\!\perp b \,|\, C$ where $C \subseteq N \setminus \{a, b\}$ is a set separating a and b in H^1.

(v) Finally, those lines $a - b$ which correspond to significant deviance differences are added to H^1 and the resulting undirected graph H^2 is obtained.

Whittaker, who also offered similar procedures for other universa of statistical models, mentioned on p. 245 of [146] that the universa of models he considers have often the *lattice* *structure*. In my view, this is the crucial requirement on the universum of statistical models so that a learning method based on significance tests could be applied. Actually, it is a requirement on the respective class of formal independence models. Indeed, one can describe the backward phase of the above procedure from the point of view of CI interpretation as follows. In **(ii)** certain acceptable UG models were generated and in **(iii)** one moves to their infimum (provided one has in mind the order $\mathcal{M}_G \subseteq \mathcal{M}_H$ between UG models). Moreover, in the forward phase, in **(iv)** some suitable UG models are generated, that is, models which seem better than the 'current' UG model $\mathcal{M}_{H^1}$. Then in **(v)** one moves to their supremum.

Thus, the assumption that the class of induced formal independence models has the lattice structure allows one to combine the knowledge obtained as a result of various particular CI tests. Actually, an analogous situation occurs in case of learning CI structure on basis of the knowledge provided by human experts (see Section 1.1). Imagine the situation when two different experts provide the information about CI structure in the form of two objects of discrete mathematics $\boldsymbol{o}_1$ and $\boldsymbol{o}_2$, for example graphs over N. The knowledge can be expressed in the form of the induced formal independence models $\mathcal{M}_{\boldsymbol{o}_1}$ and $\mathcal{M}_{\boldsymbol{o}_2}$ (see Section 2.2.1). If one relies on both experts then one would like to represent the joint knowledge (within the considered class of models) by means of an object $\boldsymbol{o}$ such that $\mathcal{M}_{\boldsymbol{o}}$ involves both $\mathcal{M}_{\boldsymbol{o}_1}$ and $\mathcal{M}_{\boldsymbol{o}_2}$ and $\mathcal{M}_{\boldsymbol{o}}$ is the least possible model of this kind. A natural requirement for the uniqueness of this model is just the requirement of the supremum existence. Well, in case of a suspicious attitude one would only like to accept the pieces of information which are confirmed by both experts and the natural requirement is the requirement of the infimum existence.

The goal of the thoughts above is the following conclusion:

> *if the data faithfulness assumption is not fulfilled and the class of formal independence models induced by objects from the considered universum of objects of discrete mathematics is not a lattice, then a method based on statistical tests of particular CI statements may appear not be applicable.*

Unfortunately, some classes of graphical models do not meet the above condition. For example, the class of DAG models for $|N| \geq 3$ is not a lattice; see Figure 7.3 where the model which corresponds to $a \perp\!\!\!\perp b \,|\, \emptyset$ and the model which corresponds to $a \perp\!\!\!\perp b \,|\, c$ have no join but two minimal upper bounds, namely the models which describe $a \perp\!\!\!\perp bc \,|\, \emptyset$ and $b \perp\!\!\!\perp ac \,|\, \emptyset$. Another example of a class of models which is not a lattice is the class of decomposable models for $|N| \geq 4$ - compare with the example on p. 254 of [146].

Remark 8.2 Whittaker gave examples showing that his simple two step procedure both may and may not lead to the choice of an adequate statistical model, that is, a model with sufficiently low deviance (Examples 8.5.1 and 8.4.2 in [146]). In my view, failing in the aim of the procedure is caused by the fact that the data faithfulness assumption is not 'valid' for UG models in that case. The above mentioned procedures and other procedures mentioned in § 8.3 of [146] can alternatively be understood as modifications of methods based on the maximization of a quality criterion from Section 8.2. Indeed, the deviance of a statistical model $\mathbb{M}_H$ for an undirected graph H has the form $\boldsymbol{k} - 2 \cdot \mathsf{MLL}\,(\mathbb{M}_H, D)$ where $\boldsymbol{k}$ is a constant and $\mathsf{MLL}\,(\mathbb{M}_H, D)$ is the value of maximized log-likelihood criterion for the model $\mathbb{M}_H$ and data D (see Section 10.10.4). Thus, the task to find a model which has the deviance as low as possible is equivalent to the task to find a model which has the value of $\mathsf{MLL}\,(\mathbb{M}_H, D)$ as high as possible.

In Example 8.2.2 [146] Whittaker mentioned an interesting phenomenon: the deviance difference for different pairs of UG models may coincide! Moreover, the deviance difference is then the value of the G^2-statistic for testing a certain CI statement. This feature could be nicely explained as a special case of Observation 8.4 - see Remark 8.14. $\triangle$

Universum of structural imsets

The goal of the considerations in this section, which are, to a certain extent, an extension of the motivation account from Section 1.1, is the moral that in case of learning methods based on statistical CI tests, the universum of considered objects of discrete mathematics should satisfy two natural conditions. The first condition is the *completeness* with respect to the respective distribution framework and the second condition is the requirement that the class of induced formal independence models has the lattice structure.

The universum of structural imsets satisfies both these conditions. The completeness with respect to the discrete and Gaussian distribution framework follows from Theorem 5.2; the fact that the class of structural models is a lattice follows from Theorem 5.3. If one dreams about a learning procedure for structural imsets based on statistical CI tests, then baricentral imsets from Section 7.1 are perhaps suitable representatives of classes of (facially) equivalent structural imsets.

8.2 Quality critera

Most of algorithms for learning graphical models is based on the idea of the maximization of a *quality criterion* which is a function which ascribes to a graph and data a real number which 'evaluates' how the statistical model described by the graph fits the data. Alternative names for a quality criterion are 'quality measure' [13], 'score criterion' [24] and 'score metric' [22]. The method of maximization of a quality criterion is often used in the area of artificial intelligence and seems to be more suitable for learning DAG models and decomposable models than the methods based on significance tests. One can distinguish at least three methodological approaches to the derivation of quality criteria (for details see Section 4.2 in [13]).

- Classic statistical interpretation of a graphical model in terms of a parametrized class of probability measures leads to *information criteria*; for example, the AIC criterion and the BIC criterion - see Section 10.10.4.

- Bayesian approach is based on an additional assumption that a 'prior' probability measure on the respective parameter space is given for each considered graphical model. This approach, initiated by the paper [23], lead to a variety of *Bayesian criteria* (see § 3.2 of [19]).

- *Minimum description length* principle [60] stems from coding theory. Roughly said, the basic idea is to evaluate a graphical model by the overall number of bits needed to encode data within the model. However, the resulting MDL criterion appears to coincide with the classic BIC information criterion (see § 4.2.3 of [13]).

The weak point of this approach is that the class of considered graphs is often too large which makes the direct maximization of a quality criterion infeasible. To avoid this problem various heuristic *local search methods* were developed (see § 4.3 in [13]). Every method of this kind has its specific *search space* which consists of the collection of *states* and the collection of *moves* between the states. The states are typically graphs over N, either acyclic directed graphs or the essential graphs, and the moves are some predefined 'local' grapical operations with them. Thus, in order to apply a local search method one needs to introduce a *neighbourhood structure* in the class of considered graphs. Every graph is assigned a relatively small set of neighbouring graphs, which are typically graphs differing from the considered graph in the presence of one edge. The point is that most quality criteria used in practice have a pleasant property that the difference of their values for neighbouring graphs is a single term which is easy to compute - see Section 8.2.3. Thus, instead of the 'global' maximum of a quality criterion various algorithms allow one to find a 'local' maximum relative to the neighbourhood structure in consideration. On the other hand, some algorithms of this kind are guaranteed to achieve the global maximum provided that certain quite strong conditions are fulfilled (these conditions include the data faithfulness assumption) - see [88].

Remark 8.3 Some of the Bayesian approaches come even with a further assumption that a 'prior' probability measure on the finite class of considered graphs is given. This assumption is then reflected by an additional term in the respective Bayesian criterion. Note that this additional assumption does not seem to prevent one from using local search methods.

Nevertheless, the whole collection of assumptions, that is, every graphical model is understood as a parametrized class of probability measures, a certain reasonable system of 'priors' on parameter spaces is given and a 'prior' on the class of models is specified, allows one to apply a special Bayesian simulation method known as the *Markov chain Monte Carlo* (MCMC) method. From a certain perspective, this method can be viewed as a particular method for learning graphical models which is not based either on significance tests or on the maximization of a quality criterion. The idea is that the assumptions above allow one to define, on basis of data, a 'posterior' probability measure on the class of considered graphs and a system of 'posteriors' on parameter spaces. The posteriors could be interpreted as stochastic estimates of both the graphical models and its parameters. However, their direct computation is again infeasible. To overcome this obstacle a stochastic simulation method is used whose aim is to get the approximation of these 'posteriors'. Of course, to make the method feasible lots of technical assumptions are accepted (which are intentionally omitted here). Nevertheless, various versions of the MCMC method are similar to the local search methods mentioned above; they are also based on the idea of search space and the idea of 'travelling' in the class of considered graphs - for details see § 3.6 in [19]. Some theoretical results ensure the (stochastic) convergence of the MCMC method if mild assumptions are fulfilled. The method was used both for learning decomposable models [39] and DAG models [72]. △

8.2.1 Criteria for learning DAG models

The aim of this section is to provide details about quality criteria for learning DAG models in case of discrete distribution framework. First, let us introduce the symbol $\mathsf{DAGS}(N)$

to denote the collection of all acyclic directed graphs having N as the set of nodes. Let a discrete joint sample space $\mathsf{X}_N = \prod_{i\in N} \mathsf{X}_i$ over N (see Section 10.10.1) is fixed. By a *quality criterion* for learning DAG models adapted to this sample space will be uderstood any function $\mathcal{Q}$ which ascribes a real number $\mathcal{Q}(G, D)$ to a graph $G \in \mathsf{DAGS}(N)$ and a database $D \in \mathsf{DATA}(N, d)$ (see p. 200).

To derive formulas for basic information criteria details about the parametrization of statistical models described by acyclic directed graphs are needed. The statistical model described by $G \in \mathsf{DAGS}(N)$ consists of the class $\mathbb{M}_G$ of probability measures on X_N which are Markovian with respect to G (see p. 46). Recall that a probability measure P on X_N is uniquely detemined by its density f with respect to the counting measure υ on X_N, that is, $f(x) = P(\{x\})$ for every $x \in \mathsf{X}_N$. The respective *marginal densities* of P (see Convention 1 on p. 25) are then

$$f_A(y) = P^A(\{y\}) = \sum_{z\in \mathsf{X}_{N\setminus A}} f(y,z) \quad \text{for } \emptyset \neq A \subset N,\ y \in \mathsf{X}_A\,,$$

and $f_N \equiv f$, $f_\emptyset \equiv 1$ by a convention. The *conditional density* $f_{A|C}$ of P for disjoint $A, C \subseteq N$ can be defined as follows:

$$f_{A|C}(x|z) = \begin{cases} \frac{f_{AC}([x,z])}{f_C(z)} & \text{if } f_C(z) = P^C(\{z\}) > 0, \\ 0 & \text{if } f_C(z) = 0 \text{ by convention,} \end{cases} \qquad \text{for } x \in \mathsf{X}_A,\ z \in \mathsf{X}_C\,.$$

The well-known fact (see Theorem 1 in [66]) is that $P \in \mathbb{M}_G$ iff it is *recursively factorizable* according to G, that is,

$$f(x) = \prod_{i\in N} f_{i|pa_G(i)}(x_i|x_{pa_G(i)}) \qquad \text{for every } x \in \mathsf{X}_N\,. \tag{8.1}$$

Note that this fact can also be derived as a consequence of Theorem 4.1: owing to Lemma 7.1 it suffices to apply it to the standard imset u_G and the standard reference system (in the discrete case) - see p. 68. The definition of conditional density then allows one to rewrite the condition **(i)** in Theorem 4.1 as the formula (8.1).

Convention 3 Let $\mathsf{X}_N = \prod_{i\in N} \mathsf{X}_i$ be a discrete sample space over N. The letter i will be used as *a generic symbol for a variable in* N: $i \in N$.

Denote $r(i) = |\mathsf{X}_i| \geq 1$ and fix an ordering $y_i^1, \ldots, y_i^{r(i)}$ of elements of X_i for every $i \in N$. The letter k will be used as *a generic symbol for a code of an element of* X_i: $k \in \{1, \ldots, r(i)\}$. More specifically, k is the code of the k-th configuration y_i^k in the fixed ordering. Given $i \in N$ and $x \in \mathsf{X}_A$ such that $i \in A \subseteq N$ symbol $k(i, x)$ will denote the code of x_i, that is, the unique $1 \leq k \leq r(i)$ such that $x_i = y_i^k$.

Let $G \in \mathsf{DAGS}(N)$. Denote by $q(i, G) \equiv |\mathsf{X}_{pa_G(i)}| = \prod_{l\in pa_G(i)} r(l) \geq 1$ the number of *parent configurations* for the variable $i \in N$ and accept the convention that $q(i, G) = 1$ if $pa_G(i) = \emptyset$. Fix an ordering $z_i^1, \ldots, z_i^{q(i,G)}$ of parent configurations, that is, elements of $\mathsf{X}_{pa_G(i)}$, for every $i \in N$. If $pa_G(i) = \emptyset$ then it consists of the only parent configuration. The letter j will be used as *a generic symbol for a code of a parent configuration*: $j \in \{1, \ldots, q(i, G)\}$. Thus, j is the code of the j-th configuration z_i^j in the fixed ordering. Given $i \in N$ and $x \in \mathsf{X}_A$ such that $pa_G(i) \subseteq A \subseteq N$ the symbol $j(i, x)$ will denote the code of $x_{pa_G(i)}$, that is, the unique $1 \leq j \leq q(i,G)$ such that $x_{pa_G(i)} = z_i^j$ if $pa_G(i) \neq \emptyset$ and $j = 1$ otherwise.

Moreover, let $D \in \mathsf{DATA}(N,d)$, $d \in \mathbb{N}$ be data over N (see p. 200), more specifically $D : x^1, \ldots, x^d$. Introduce the following special notation for the numbers of configuration occurences in the database D:

$$\begin{aligned}
d_{ij} &= |\{1 \leq \ell \leq d;\ x^\ell_{pa_G(i)} = z_i^j\}| \quad \text{for } i \in N,\, j \in \{1,\ldots,q(i,G)\},\\
d_{ijk} &= |\{1 \leq \ell \leq d;\ x^\ell_{\{i\}\cup pa_G(i)} = (y_i^k, z_i^j)\}| \quad \text{for } i \in N,\, j \in \{1,\ldots,q(i,G)\},\, k \in \{1,\ldots,r(i)\},\\
d_{\boldsymbol{y}} &= |\{1 \leq \ell \leq d;\ x^\ell_A = \boldsymbol{y}\}| \quad \text{for } \emptyset \neq A \subseteq N,\, \boldsymbol{y} \in \mathsf{X}_A\,.
\end{aligned}$$

Of course, $d_{i1} = d$ if $pa_G(i) = \emptyset$.

Given $G \in \mathsf{DAGS}(N)$, the respective 'standard' parametrization of $\mathbb{M}_G$ is based on the recursive factorization (8.1). The set of parameters Θ_G consists of vectors

$$\begin{gathered}
\boldsymbol{\theta} \,:\, \theta_{ijk} \in [0,1] \quad \text{for } i \in N,\, j \in \{1,\ldots,q(i,G)\},\, k \in \{1,\ldots,r(i)\}\\
\text{such that } \sum_{k=1}^{r(i)} \theta_{ijk} = 1 \text{ for every } i \in N,\ 1 \leq j \leq q(i,G)\,.
\end{gathered} \tag{8.2}$$

Actually, every parameter θ_{ijk} can be interpreted as the value of the conditional density $f_{i|pa(i)}(y_i^k|z_i^j)$. Given a vector parameter $\boldsymbol{\theta}$, the respective probability measure on X_N is given by its density

$$f^{\boldsymbol{\theta}}(x) \;=\; \prod_{i\in N} \theta_{i\,j(i,x)\,k(i,x)} \qquad \text{for } x \in \mathsf{X}_N\,. \tag{8.3}$$

Lemma 8.1 Let us accept Convention 3 suppose that a set of parameters $\boldsymbol{\theta}$ satisfying (8.2) is given. Then the formula (8.3) defines a density of a probability measure $P^{\boldsymbol{\theta}}$ on X_N (with respect to the counting measure). Moreover, $P^{\boldsymbol{\theta}} \in \mathbb{M}_G$ and $\forall\, i \in N$ $\forall j \in \{1,\ldots,q(i,G)\}$ $\forall\, k \in \{1,\ldots,r(i)\}$

$$f^{\boldsymbol{\theta}}_{i|pa_G(i)}(y_i^k|z_i^j) \;=\; \theta_{ijk} \qquad \text{if } f^{\boldsymbol{\theta}}_{pa_G(i)}(z_i^j) > 0\,. \tag{8.4}$$

The mapping $\boldsymbol{\theta} \mapsto P^{\boldsymbol{\theta}}$ is a mapping onto $\mathbb{M}_G$.

Proof: Given $A \subseteq N$ let us denote the collection of paramaters θ_{ijk} with $i \in A$ by $\boldsymbol{\theta}[A]$. Let $t \in N$ be a terminal node in G. The first observation is that

$$f^{\boldsymbol{\theta}}([y,z]) \;=\; (\prod_{i\in N\setminus t} \theta_{i\,j(i,z)\,k(i,z)}) \cdot \theta_{t\,j(t,z)\,k(t,y)} \quad \text{for every } z \in \mathsf{X}_{N\setminus t},\, y \in \mathsf{X}_t\,. \tag{8.5}$$

Indeed, given $x = [y,z] \in \mathsf{X}_N$ by Convention 3 and the fact that $pa_G(i) \subseteq N \setminus \{t\}$ for $i \in N$ one has $j(i,x) = j(i,z)$ for $i \in N$, $k(t,x) = k(t,y)$ and $k(i,x) = k(i,z)$ for $i \in N \setminus \{t\}$. It suffices to substitute these equalities into (8.3).
The second observation is that

$$f^{\boldsymbol{\theta}}_{N\setminus t}(z) \equiv \sum_{y\in\mathsf{X}_t} f^{\boldsymbol{\theta}}([y,z]) = \prod_{i\in N\setminus t} \theta_{i\,j(i,z)\,k(i,z)} = f^{\boldsymbol{\theta}[N\setminus t]}(z) \quad \text{for every } z \in \mathsf{X}_{N\setminus t}\,. \tag{8.6}$$

Indeed, substitute (8.5) in $f^{\boldsymbol{\theta}}_{N\setminus t}(z) = \sum_{y\in\mathsf{X}_t} f^{\boldsymbol{\theta}}([y,z])$ and get $f^{\boldsymbol{\theta}}_{N\setminus t}(z) = (\prod_{i\in N\setminus t} \theta_{i\,j(i,z)\,k(i,z)}) \cdot \sum_{y\in\mathsf{X}_t} \theta_{t\,j(t,z)\,k(t,y)}$. The last sum is 1 by (8.2) and one can apply (8.3) to $G_{N\setminus t}$ and $\boldsymbol{\theta}[N \setminus t]$.

To verify that $f^{\boldsymbol{\theta}}$ is a probability density choose an ordering $t_1, \ldots, t_{|N|}$ of the nodes of G which is consonant with the direction of arrows, that is, $t_n \rightarrow t_m$ in G implies $n < m$. Put $N(n) = \{t_1, \ldots, t_n\}$ for $n = 1, \ldots, |N|$, observe that t_n is a terminal node in $G_{N(n)}$ and prove by induction on $n = 1, \ldots, |N|$ that $f^{\boldsymbol{\theta}[N(n)]}$ is a probability density on $\mathsf{X}_{N(n)}$. Indeed, it is evident that $f^{\boldsymbol{\theta}[N(n)]}(x) \geq 0$ for $x \in \mathsf{X}_{N(n)}$. The fact $\sum_{x \in \mathsf{X}_{N(n)}} f^{\boldsymbol{\theta}[N(n)]}(x) = 1$ for $n = 1$ is implied by (8.2) and the induction step follows from (8.6).
Thus, $f^{\boldsymbol{\theta}}$ defines a probability measure $P^{\boldsymbol{\theta}}$ on X_N. The next observation is that provided t is a terminal node in G one has

$$f^{\boldsymbol{\theta}}_{t\,|\,N\setminus t}(y|z) \;=\; \theta_{t\,j(t,z)\,k(t,y)} \quad \text{for } z \in \mathsf{X}_{N\setminus t} \text{ with } f^{\boldsymbol{\theta}}_{N\setminus t}(z) > 0,\ y \in \mathsf{X}_t\,. \tag{8.7}$$

Indeed, substitute (8.5) and (8.6) into the definition of $f^{\boldsymbol{\theta}}_{t\,|\,N\setminus t}(y|z)$. The term $\prod_{i \in N\setminus t} \theta_{i\,j(i,z)\,k(i,z)}$ is ensured to be non-zero in this case. Thus, it can be cancelled.
It follows from (8.5) by means of Lemma 2.5 that $t \perp\!\!\!\perp N \setminus t \cup pa_G(t) \,|\, pa_G(t)\ [P^{\boldsymbol{\theta}}]$. An analogous argument applied to t_n and G_n gives $t_n \perp\!\!\!\perp N(n-1) \setminus pa_G(t_n) \,|\, pa_G(t_n)\ [P^{\boldsymbol{\theta}}]$ for $n = 1, \ldots, |N|$. This property, known as the 'local Markov property' (see Remark 3.4) implies by the result from [141] that $P^{\boldsymbol{\theta}}$ is (globally) Markovian with respect to G, that is, $P^{\boldsymbol{\theta}} \in \mathbb{M}_G$ (see also Theorem 1 in [66]). Further observation is that if t is a terminal node in G then

$$f^{\boldsymbol{\theta}}_{t\,|\,pa_G(t)}(y^k_t|z^j_t) \;=\; \theta_{tjk} \quad \text{for } j \in \{1, \ldots, q(t,G)\} \text{ with } f^{\boldsymbol{\theta}}_{pa_G(t)}(z^j_t) > 0,\ k \in \{1, \ldots, r(t)\}\,.$$

Indeed, choose $z \in \mathsf{X}_{N\setminus t}$ such that $f^{\boldsymbol{\theta}}_{N\setminus t}(z) > 0$ and $z_{pa(t)} = z^j_t$, that is, $j = j(t,z)$. Then take $y \in \mathsf{X}_t$ such that $k = k(t,y)$. Thus, one is entitled to apply (8.7) to z. Write $f^{\boldsymbol{\theta}}_{t\,|\,pa(t)}(y^k_t|z^j_t) = f^{\boldsymbol{\theta}}_{t\,|\,pa(t)}(y|z_{pa(t)}) = (f^{\boldsymbol{\theta}}_{pa(t)}(z_{pa(t)}))^{-1} \cdot f^{\boldsymbol{\theta}}_{t \cup pa(t)}([y, z_{pa(t)}])$. However, $t \perp\!\!\!\perp N \setminus t \cup pa(t) \,|\, pa(t)\ [P^{\boldsymbol{\theta}}]$ implies by Lemma 2.4 that the latter term equals $(f^{\boldsymbol{\theta}}(z))^{-1} \cdot f^{\boldsymbol{\theta}}([y,z])$. By (8.7) it equals to $\theta_{t\,j(t,z)\,k(t,y)} = \theta_{tjk}$.
The same argument applied to t_n and $G_{N(n)}$ for $n = 1, \ldots, |N|$ gives (8.4). Finally, given $P \in \mathbb{M}_G$ let us put $\theta_{ijk} = f_{i\,|\,pa_G(i)}(y^k_i|z^j_i)$ for every $i \in N$, $j \in \{1, \ldots, q(i,G)\}$, $k \in \{1, \ldots, r(i)\}$. Observe that (8.2) is valid which allows one to define $P^{\boldsymbol{\theta}}$. By (8.1) and (8.3) get $f(x) = \prod_{i \in N} \theta_{i\,j(i,x)\,k(i,x)} = f^{\boldsymbol{\theta}}(x)$ for every $x \in \mathsf{X}_N$ which implies $P = P^{\boldsymbol{\theta}}$. □

Remark 8.4 The correspondence $\boldsymbol{\theta} \mapsto P^{\boldsymbol{\theta}}$ is not a one-to-one mapping. The reason is that if t is a terminal node in G and $P^{\boldsymbol{\theta}}_{N\setminus t}(x_{N\setminus t}) = 0$ then $P^{\boldsymbol{\theta}}(x) = 0$ no matter what is the value of $\theta_{t\,j(t,x)\,k(t,x)}$. However, it is a one-to-one mapping on the set of parameters $\boldsymbol{\theta}$ such that $\theta_{ijk} \in (0,1)$ for every i, j, k. These parameters correspond to strictly positive probability measures on $\mathbb{M}_G$. △

To derive a maximum likelihood estimate in $\mathbb{M}_G$ and the formula for the MLL quality criterion the following technical lemma is needed.

Lemma 8.2 Let $d_k \geq 0$, $k \in K$ be a non-empty finite collection of non-negative numbers such that $d = \sum_{k \in K} d_k > 0$. The the vector function f with values in $[-\infty, \infty)$ given by

$$f([\theta_k]_{k \in K}) \;=\; \sum_{k \in K} d_k \cdot \ln \theta_k\,,$$

defined on the domain $\mathcal{D}(f) = \{[\theta_k]_{k \in K}\,;\ \theta_k \geq 0\ \sum_{k \in K} \theta_k = 1\,\}$ attains its maximum in $[\hat{\theta}_k]_{k \in K}$ where $\hat{\theta}_k = \frac{d_k}{d}$ for $k \in K$. In particular, the value of the maximum of f on $\mathcal{D}(f)$ is $\sum_{k \in K, d_k > 0} d_k \cdot \ln \frac{d_k}{d}$.

Proof: The first step is the following auxiliary observation.
I. Suppose $0 \leq s \leq 1$, $d_i, d_j \geq 0$ and $d_i + d_j > 0$. Then the function

$$g(\theta_i, \theta_j) \;=\; d_i \cdot \ln \theta_i + d_j \cdot \ln \theta_j \qquad \text{defined for } \theta_i, \theta_i \geq 0,\; \theta_i + \theta_j = 1 - s\,,$$

taking values in $[-\infty, \infty)$ attains the maximum in $(\hat{\theta}_i, \hat{\theta}_j) = (\frac{d_i}{d_i+d_j} \cdot (1-s), \frac{d_j}{d_i+d_j} \cdot (1-s))$. Indeed, the claim is trivial in case $s = 1$. If $s < 1$ then $\theta_j = 1 - s - \theta_i$ and the task is equivalent to the maximization of the function $\tilde{g}(\theta) = d_i \cdot \ln \theta + d_j \cdot \ln(1 - s - \theta)$ for $\theta \in [0, 1-s]$. The rest is elementary computation.
Lemma 8.2 can be verified by induction on $|K|$. It is trivial if $|K| = 1$; if $|K| = 2$ then it is implied by Step **I.** with $K = \{i, j\}$ and $s = 0$. To verify the induction step suppose $|K| \geq 3$, choose distinct $i, j \in K$ such that $d_i + d_j > 0$ and put $L = K \setminus \{i, j\}$. Introduce notation $\theta_L = [\theta_l]_{l \in L}$, $s(\theta_L) = \sum_{l \in L} \theta_l$ and put

$$\varphi_i(\theta_L) = \frac{d_i}{d_i + d_j} \cdot (1 - s(\theta_L)), \quad \varphi_j(\theta_L) = \frac{d_j}{d_i + d_j} \cdot (1 - s(\theta_L))\,,$$

for every θ_L with $\theta_l \geq 0$, $l \in L$ and $s(\theta_L) \leq 1$. Observe that $[\theta_L, \varphi_i(\theta_L), \varphi_j(\theta_L)] \in \mathcal{D}(f)$. Moreover, for $\hat{\theta}_L = [\hat{\theta}_l]_{l \in L} = [\frac{d_l}{d}]_{l \in L}$ one has $\varphi_i(\hat{\theta}_L) = \hat{\theta}_i$ and $\varphi_j(\hat{\theta}_L) = \hat{\theta}_j$. Further steps are as follows.
II. Observe $f(\theta_K) \leq f(\theta_L, \varphi_i(\theta_L), \varphi_j(\theta_L))$ for every $\theta_K = [\theta_k]_{k \in K} \in \mathcal{D}(f)$.
Indeed, apply Step **I.** to d_i, d_j and $s = s(\theta_L)$ which gives $g(\theta_i, \theta_j) \leq g(\varphi_i(\theta_L), \varphi_j(\theta_L))$. Then add $\sum_{l \in L} d_l \cdot \ln \theta_l \in [-\infty, 0]$ to both sides of the obtained inequality to get what is desired.
III. Observe that $f(\theta_L, \varphi_i(\theta_L), \varphi_j(\theta_L)) \leq f(\hat{\theta}_L, \varphi_i(\hat{\theta}_L), \varphi_j(\hat{\theta}_L))$ for every θ_L with $\theta_l \geq 0$ and $s(\theta_L) \leq 1$.
Let us introduce a new artificial index $r \notin K$ and put $\bar{K} = L \cup \{r\}$, $d_r = d_i + d_j$. Observe that $d = \sum_{k \in \bar{K}} d_k$ and introduce a function $\bar{f}$ taking values in $[-\infty, \infty)$ by

$$\bar{f}([\theta_k]_{k \in \bar{K}}) \;=\; \sum_{k \in \bar{K}} d_k \cdot \ln \theta_k \quad \text{on } \mathcal{D}(f) = \{\, [\theta_k]_{k \in \bar{K}}\,;\, \theta_k \geq 0,\; \sum_{k \in \bar{K}} \theta_k = 1 \,\}\,.$$

By the induction hypothesis $\bar{f}$ attains its maximum in $\hat{\theta}_{\bar{K}} = [\hat{\theta}_k]_{k \in \bar{K}} = [\frac{d_k}{d}]_{k \in \bar{K}}$. Because of $(\theta_L, 1 - s(\theta_L)) \in \mathcal{D}(\bar{f})$ one has $\bar{f}(\theta_L, 1 - s(\theta_L)) \leq \bar{f}(\hat{\theta}_{\bar{K}})$. But, owing to $d = \sum_{k \in \bar{k}} d_k$, $\hat{\theta}_{\bar{K}} = (\hat{\theta}_L, 1 - s(\hat{\theta}_L))$ which gives $\bar{f}(\theta_L, 1 - s(\theta_L)) \leq \bar{f}(\hat{\theta}_L, 1 - s(\hat{\theta}_L))$. By simple computation observe that

$$f(\theta_L, \varphi_i(\theta_L), \varphi_j(\theta_L)) = \bar{f}(\theta_L, 1 - s(\theta_L)) + d_i \cdot \ln \theta_i + d_j \cdot \ln \theta_j - d_r \cdot \ln \theta_r$$

for any θ_L in the respective domain, in particular for $\hat{\theta}_L$. Since the constant $d_i \cdot \ln \theta_i + d_j \cdot \ln \theta_j - d_r \cdot \ln \theta_r$ is a finite number one can add it to the last inequality to get $f(\theta_L, \varphi_i(\theta_L), \varphi_j(\theta_L)) \leq f(\hat{\theta}_L, \varphi_i(\hat{\theta}_L), \varphi_j(\hat{\theta}_L))$.
IV. Because $\varphi_i(\hat{\theta}_L) = \hat{\theta}_i$ and $\varphi_j(\hat{\theta}_L) = \hat{\theta}_j$ one has $(\hat{\theta}_L, \varphi_i(\hat{\theta}_L), \varphi_j(\hat{\theta}_L)) = \hat{\theta}_K \in \mathcal{D}(f)$ and using **II.** and **III.** get

$$f(\theta_K) \leq f(\theta_L, \varphi_i(\theta_L), \varphi_j(\theta_L)) \leq f(\hat{\theta}_L, \varphi_i(\hat{\theta}_L), \varphi_j(\hat{\theta}_L)) = f(\hat{\theta}_K)$$

for every $\theta_L \in \mathcal{D}(f)$ which concludes the proof of the induction hypothesis. □

Consequence 8.1 Let us accept Convention 3. The the formulas for parameters

$$\hat{\theta}_{ijk} \;=\; \begin{cases} \frac{d_{ijk}}{d_{ij}} & \text{if } d_{ij} > 0, \\ 0 & \text{if } d_{ij} = 0, \end{cases} \qquad \text{for } i \in N,\; 1 \leq j \leq q(i, G),\; 1 \leq k \leq r(i), \tag{8.8}$$

define a maximum likelihood estimate $\hat{P} = P^{\hat{\boldsymbol{\theta}}}$ in $\mathbb{M}_G$ on basis of data $D \in \mathsf{DATA}(N,d)$ (see Section 10.10.4). Moreover, the maximized log-likelihood criterion has the form

$$\mathsf{MLL}\,(\mathbb{M}_G, D) \;=\; \sum_{i\in N}\sum_{j=1}^{q(i,G)}\sum_{k=1}^{r(i)} d_{ijk}\cdot \ln\frac{d_{ijk}}{d_{ij}} \quad \text{for } G\in\mathsf{DAGS}(N),\ D\in\mathsf{DATA}(N,d), \quad (8.9)$$

where the convention $0\cdot\ln\frac{0}{*}\equiv 0$ is accepted.

Proof: Let $D: x^1, \dots x^d$, $d\geq 1$. Consider the task of maximization of the logarithm of the likelihood function $l(\theta) = \ln L(\theta, D)$. By the definition of the likelihood function and (8.3) write

$$l(\theta) = \ln L(\theta, D) = \sum_{\ell=1}^{d} \ln f^{\boldsymbol{\theta}}(x^\ell) = \sum_{\ell=1}^{d}\sum_{i\in N} \ln\theta_{i\,j(i,x^\ell)\,k(i,x^\ell)}\,.$$

To get a better formula introduce auxiliary notation for $i\in N$, $x\in \mathsf{X}_N$, $1\leq j\leq q(i,G)$ and $1\leq k\leq r(i)$:

$$\delta_i(j,k,x) = \begin{cases} 1 & \text{if } x_{i\cup pa_G(i)} = (y_i^k, z_i^j),\\ 0 & \text{otherwise.}\end{cases}$$

This allows one to write

$$l(\theta) \;=\; \sum_{\ell=1}^{d}\sum_{i\in N}\sum_{j=1}^{q(i,G)}\sum_{k=1}^{r(i)} \ln\theta_{ijk}\cdot\delta_i(j,k,x^\ell) \;=\; \sum_{i\in N}\sum_{j=1}^{q(i,G)}\sum_{k=1}^{r(i)} \ln\theta_{ijk}\cdot\sum_{\ell=1}^{d}\delta_i(j,k,x^\ell)\,.$$

By Convention 3 observe that the latter sum is nothing but d_{ijk}. Therefore

$$l(\theta) = \sum_{i\in N}\sum_{j=1}^{q(i,G)}\sum_{k=1}^{r(i)} d_{ijk}\cdot\ln\theta_{ijk}\,. \quad (8.10)$$

Clearly, the task of maximization of $l(\theta)$ is equivalent to the task to maximize, for every $i\in N$ and $j\in\{1,\dots,q(i,G)\}$, the function $\sum_{k=1}^{r(i)} d_{ijk}\cdot\ln\theta_{ijk}$ defined on the set $\{[\theta_{ijk}]_{k=1}^{r(i)}\,;\ \theta_{ijk}\geq 0\ \sum_{k=1}^{r(i)}\theta_{ijk}=1\,\}$. Since $\sum_{k=1}^{r(i)} d_{ijk} = d_{ij}$ by Lemma 8.2, if $d_{ij}>0$ then the maximum is attained in $\hat{\theta}_{ijk} = \frac{d_{ijk}}{d_{ij}}$. If $d_{ij}=0$ then the function is constant and its maximum 0 is attained in any point, in particular, in $\hat{\theta}_{ijk}=0$. This gives (8.8). Finally, substitute this into (8.10) to get (8.9). $\square$

Consequence 8.2 Let us accept Convention 3. The the effective dimension of the statistical model $\mathbb{M}_G$ (see Section 10.10.4) is

$$\mathsf{DIM}\,(G) = \mathsf{DIM}\,(\mathbb{M}_G) = \sum_{i\in N}(r(i)-1)\cdot q(i,G) \;=\; \sum_{i\in N}(r(i)-1)\cdot\prod_{l\in pa_G(i)} r(l)\,. \quad (8.11)$$

In particular, Akaike's information criterion (AIC) and the Bayesian information criterion (BIC) are given by the formulas

$$\mathsf{AIC}\,(G,D) \;=\; \sum_{i\in N}\sum_{j=1}^{q(i,G)}\sum_{k=1}^{r(i)} d_{ijk}\cdot\ln\frac{d_{ijk}}{d_{ij}} - (r(i)-1)\cdot q(i,G) \quad (8.12)$$

$$\mathsf{BIC}\,(G,D) \;=\; \sum_{i\in N}\sum_{j=1}^{q(i,G)}\sum_{k=1}^{r(i)} d_{ijk}\cdot\ln\frac{d_{ijk}}{d_{ij}} - \frac{\ln d}{2}\cdot(r(i)-1)\cdot q(i,G) \quad (8.13)$$

for every $G\in\mathsf{DAGS}(N)$ and every $D\in\mathsf{DATA}(N,d)$.

Proof: It follows from (8.2) that, for fixed $i \in N$, $1 \leq j \leq q(i,G)$, the number of linearly independent parameters is $r(i)-1$. This gives (8.11). The other formulas follow from Consequence 8.1 - see Section 10.10.4. □

Remark 8.5 Bayesian criteria for learning DAG models in discrete distribution framework could be introduced in the following manner. In general, for every $G \in \mathsf{DAGS}(N)$ a prior probability measure π_G on the respective parameter space Θ_G (see (8.2)) is given. This allows one to define the *marginal likelihood* as the integral of the likelihood function with respect to π_G. The respective Bayesian criterion is the *logarithm of the marginal likelihood* given by

$$\mathsf{LML}\,(G,D) \;=\; \ln \int_{\Theta_G} L(\theta,D)\, d\pi_G(\theta)\,. \tag{8.14}$$

However, to get more suitable formula for this criterion the assumption that π_G is a product measure $\prod_{i\in N}\prod_{j=1}^{q(i,G)} \pi_{(ij)}$ where $\pi_{(ij)}$ is a 'prior' probability measure on the 'local' parameter space $\Theta_{(ij)} = \{[\theta_{ijk}]_{k=1}^{r(i)}\,;\, \theta_{ijk} \geq 0\ \sum_{k=1}^{r(i)} \theta_{ijk}\,\}$ is accepted - see the assumptions of 'global' and 'local' parameter independence from [114]. These assumptions allow one to derive a more explicit formula using (8.10) and Fubini theorem:

$$\mathsf{LML}\,(G,D) \;=\; \sum_{i\in N}\sum_{j=1}^{q(i,G)} \ln \int_{\Theta_{(ij)}} \prod_{k=1}^{r(i)} \theta_{ijk}^{d_{ijk}}\, d\pi_G(\theta_{ijk})\,. \tag{8.15}$$

Typically, there are lots of technical assumptions on the priors $\pi_{(ij)}$ which allow one to obtain even more convenient formulas for the LML criterion. The flexibility in the choice of priors leads to a great variety of particular Bayesian criteria for learning DAG models. However, because it is not the aim of this monograph to explain pecularities of the Bayesian approach to learning DAG models the reader is referred to § 3.2 of [19]. △

8.2.2 Score equivalent criteria

A quality criterion $\mathcal{Q}$ for learning DAG models is *score equivalent* if for every pair $G,H \in \mathsf{DAGS}(N)$ and every $D \in \mathsf{DATA}(N,d)$

$$\mathcal{Q}(G,D) \;=\; \mathcal{Q}(H,D) \qquad \text{whenever } G \text{ and } H \text{ are Markov equivalent.} \tag{8.16}$$

This requirement is quite natural from the statistical point of view. Indeed, if acyclic directed graphs G and H are Markov equivalent then they represent the same statistical model and data 'generated' by $P \in \mathbb{M}_G = \mathbb{M}_H$ do not distinguish between G and H. Thus, provided that the aim is to learn a statistical model the respective quality criterion should not distinguish Markov equivalent acyclic directed graphs as well.

Remark 8.6 The concept of score equivalent criterion was pinpointed in § 4.2.5 of [13]. Note that most criteria for learning DAG models are score equivalent - see Observation 8.2 for an example. However, there are quality criteria which are not score equivalent. A well-known example, which is mentioned on p. 71 of [13], is a particular Bayesian criterion named the K2 metric in [54] after the algorithm in which it was used first [23]. A possible argument in favour of quality criteria which are not score equivalent is that if one has causal interpretation of acyclic directed graphs in mind (see [115]) then a criterion of this kind allow one to distinguish between different causal interpretations. However, the point is that the difference in causal interpretation cannot be identified on basis of statistical data and, therefore, it is not a good idea to use such a criterion in the learning phase which is based on data. $\triangle$

8.2.3 Decomposable criteria

Given a database $D \in \mathsf{DATA}(N,d) : x^1,\ldots,x^d,\ d \geq 1$ over N and $\emptyset \neq A \subseteq N$ by the *projection* of D onto A will be understood the database $D_A \in \mathsf{DATA}(A,d) : x^1_A,\ldots,x^d_A$ over A. A quality criterion $\mathcal{Q}$ for learning DAG models will be called *decomposable* if there exists a collection of functions $q_{i|B} : \mathsf{DATA}(\{i\} \cup B, d) \to \mathbb{R}$ where $i \in N,\ B \subseteq N \setminus \{i\}$ and $d \in \mathbb{N}$ such that for every $G \in \mathsf{DAGS}(N)$ and every $D \in \mathsf{DATA}(N,d)$

$$\mathcal{Q}(G,D) \,=\, \sum_{i\in N} q_{i|pa_G(i)}(D_{i\cup pa_G(i)})\,. \tag{8.17}$$

An important fact is that the functions $q_{i|B}$ do not depend on G; the graph G is only represented in the right-hand side of (8.17) by the collection of sets $pa_G(i),\ i \in N$. The criterion $\mathcal{Q}$ will be called *strongly decomposable* if, moreover, $q_{i|B}$ only depends on $D_{i\cup B}$ through the respective marginal contingency table $ct_{i\cup B}(D)$ (see p. 200). More precisely, $\mathcal{Q}$ is strongly decomposable if a collection of functions $\bar{q}_{i|B} : \mathsf{CONT}(\{i\} \cup B, d) \to \mathbb{R}$, $i \in N,\ B \subseteq N \setminus \{i\}$, where $d \in \mathbb{N}$, exists such that

$$\mathcal{Q}(G,D) \,=\, \sum_{i\in N} \bar{q}_{i|pa_G(i)}(ct_{i\cup pa_G(i)}(D))\,. \tag{8.18}$$

for every $G \in \mathsf{DAGS}(N)$ and $D \in \mathsf{DATA}(N,d)$ (see Section 10.10.1).

Remark 8.7 The concept of decomposable criterion was pinpointed in [22]. Actually, all criteria for learning DAG models which are used in practice are decomposable owing to the way of their construction - see [19]. This is caused by the intended application of the local search method (see p. 138). Indeed, if acyclic directed graphs G and H differ in the presence of one arrow then there exists $t \in N$ such that $pa_G(i) = pa_H(i)$ for every $i \in N \setminus t$ and by (8.17)

$$\mathcal{Q}(G,D) - \mathcal{Q}(H,D) \,=\, q_{t|pa_G(t)}(D_{t\cup pa_G(t)}) - q_{t|pa_H(t)}(D_{t\cup pa_H(t)})\,.$$

Thus, the decomposability of a criterion for learning DAG models is a requirement brought by the researchers in the area of artificial intelligence in order to make computation feasible. However, the definitions of the concept of decomposable criterion I found in the literature are slightly vague. The authors [19, 22] seem to repeat a sketchy phrase from [47]: a criterion is decomposable if it can be written as a product of 'measures', each of which is a function only of one node and its parents. What is not clear is how data affect the value of the criterion and what type of data is supposed to be given. Indeed, one can either consider data in the form of an ordered sequence of elements of the sample space or in the form of a contingency table. The distinction can be explained by the following simple example: if $y, z \in \mathsf{X}_N,\ y \neq z$ then data $D : x^1 = y,\ x^2 = z$ of the length 2 differ from data $D' : x^1 = z,\ x^2 = y$ if one accepts the interpretation of data in terms of an ordered sequence but they coincide of one accepts the other interpretation! One can certainly imagine quality criteria whose values depend on the order of items in a database. In particular, it really matters what type of input data one has in mind because the other understanding of data confines the class of quality criteria.

Some explanation was only given by Chickering [22] who made an additional remark saying that data are implicit in the 'measures' and it is meant the data on which a 'measure' depends are restricted to those columns which correspond to the respective node and its parents. This suggests that Chickering has perhaps in mind data in the form of an ordered sequence of items. On the other hand, in the formulas for various decomposable criteria [13, 19] data are represented in the form of a contingency table. Because two different understandings of data lead to two different concepts of decomposable quality criterion I decided to include both versions and to distinguish betwen them in the terminology: the concept of strong decomposability corresponds to the situation when data are supposed in the form of a contingency table. The distinction is also reflected in consequent concepts and results. △

8.2.4 Regular criteria

Two traditional requirements on a quality criterion can be joined in the following concept. A quality criterion for learning DAG models will be called *regular* if there exists a collection of functions $\boldsymbol{t}_A : \mathsf{DATA}(A,d) \to \mathbb{R}$, $\emptyset \neq A \subseteq N$ and a constant depending on the sample space and d, by convention denoted by $\boldsymbol{t}_\emptyset(D_\emptyset)$, such that for every $G \in \mathsf{DAGS}(N)$ and every $D \in \mathsf{DATA}(N,d)$

$$\mathcal{Q}(G,D) \;=\; \sum_{i\in N} \boldsymbol{t}_{i\cup pa_G(i)}(D_{i\cup pa_G(i)}) - \boldsymbol{t}_{pa_G(i)}(D_{pa_G(i)})\,. \qquad (8.19)$$

A criretion $\mathcal{Q}$ will be called *strongly regular* if $\boldsymbol{t}_A$ depend on D_A through $ct_A(D)$, that is, if there exists a collection of functions $\bar{\boldsymbol{t}}_A : \mathsf{CONT}(A,d) \to \mathbb{R}$, $\emptyset \neq A \subseteq N$ and a constant $\bar{\boldsymbol{t}}_\emptyset(-)$ depending on the sample space and d such that for every $G \in \mathsf{DAGS}(N)$ and every $D \in \mathsf{DATA}(N,d)$

$$\mathcal{Q}(G,D) \;=\; \sum_{i\in N} \bar{\boldsymbol{t}}_{i\cup pa_G(i)}(ct_{i\cup pa_G(i)}(D)) - \bar{\boldsymbol{t}}_{pa_G(i)}(ct_{pa_G(i)}(D))\,. \qquad (8.20)$$

Observe that it follows immediately from the definition that a linear combination of (strongly) regular criteria is again a (strongly) regular criterion.

Lemma 8.3 A quality criterion $\mathcal{Q}$ for learning DAG models is regular iff it is decomposable and score equivalent. Moreover, it is strongly regular iff it is strongly decomposable and score equivalent.

Proof: If $\mathcal{Q}$ is regular then put

$$q_{i|B}(D) \;=\; \boldsymbol{t}_{i\cup B}(D) - \boldsymbol{t}_B(D_B) \qquad \text{for } i \in N,\ B \subseteq N\setminus\{i\},\ D \in \mathsf{DATA}(\{i\}\cup B, d), \quad (8.21)$$

and observe that (8.19) implies (8.17). Transformational characterization of equivalent acyclic directed graphs (see p. 47) can simplify the proof of score equivalence of $\mathcal{Q}$: it suffices to verify for every $D \in \mathsf{DATA}(N,d)$ that $\mathcal{Q}(G,D) = \mathcal{Q}(H,D)$ whenever H is obtained from $G \in \mathsf{DAGS}(N)$ by a legal reversal of an arrow $a \to b$ in G. The proof of this fact is analogous to the proof of Consequence 7.1: owing to (8.19) and the fact $pa_G(i) = pa_H(i)$ for $i \in N \setminus \{a,b\}$ one needs to evidence

$$\sum_{i\in\{a,b\}} \boldsymbol{t}_{i\cup pa_G(i)}(D_{i\cup pa_G(i)}) - \boldsymbol{t}_{pa_G(i)}(D_{pa_G(i)}) \;=\; \sum_{i\in\{a,b\}} \boldsymbol{t}_{i\cup pa_H(i)}(D_{i\cup pa_H(i)}) - \boldsymbol{t}_{pa_H(i)}(D_{pa_H(i)})\,.$$

This makes no problem since there exists $C \subseteq N$ such that $pa_G(a) = pa_H(b) = C$, $pa_G(b) = C \cup a$ and $pa_H(a) = C \cup b$.

To verify the converse implication suppose that $\mathcal{Q}$ is decomposable and score equivalent. The first observation is that $\forall\, C \subseteq N$, $\forall\, a,b \in N\setminus C$, $a \neq b$ and $\forall\, D \in \mathsf{DATA}(N,d)$

$$q_{a|bC}(D_{abC}) + q_{b|C}(D_{bC}) \;=\; q_{b|aC}(D_{abC}) + q_{a|C}(D_{aC})\,. \tag{8.22}$$

Indeed, let us construct $G, H \in \mathsf{DAGS}(N)$ as follows. Both graph have all arrows from C to $\{a,b\}$. Moreover, $b \to a$ is an arrow in G while $a \to b$ in an arrow in H. Clearly, G and H are equivalent acyclic directed graphs over N and $\forall\, i \in N \setminus \{a,b\}\quad pa_G(i) = pa_H(i)$. It suffices to substitute (8.17) into (8.16); after cancellation of the terms which correspond to $i \in N \setminus \{a,b\}$ (8.22) is obtained.
The required system of functions $\boldsymbol{t}_A$, $A \subseteq N$ can be constructed recursively. First, put $\boldsymbol{t}_\emptyset(D_\emptyset) = 0$. For $\emptyset \neq A \subseteq N$ define $\boldsymbol{t}_A(D_A)$ on basis of $\boldsymbol{t}_B(D_B)$ with $B \subset A$. Choose $a \in A$ and put

$$\boldsymbol{t}_A(D_A) \;=\; q_{a|A\setminus a}(D_A) + \boldsymbol{t}_{A\setminus a}(D_{A\setminus a}) \qquad \text{for } D \in \mathsf{DATA}(N,d)\,. \tag{8.23}$$

However, one has to show that this definition does not depend on the choice of $a \in A$, that is, (8.23) holds for any other $b \in A$ in place of a. This can be by induction on $|A|$. It is trivial if $|A| = 1$. If $|A| \geq 2$ observe that $\forall\, D \in \mathsf{DATA}(N,d)$, $\forall\, a,b \in A$, $a \neq b$

$$q_{a|A\setminus a}(D_A) + \boldsymbol{t}_{A\setminus a}(D_{A\setminus a}) \;=\; q_{b|A\setminus b}(D_A) + \boldsymbol{t}_{A\setminus b}(D_{A\setminus b})\,.$$

Indeed, put $C = A\setminus\{a,b\}$ in (8.22) and add the term $\boldsymbol{t}_{A\setminus ab}(D_{A\setminus ab})$ to both sides of the equality

$$q_{a|A\setminus a}(D_A) + \{\, q_{b|A\setminus ab}(D_{A\setminus a}) + \boldsymbol{t}_{A\setminus ab}(D_{A\setminus ab})\,\} \;=\; q_{b|A\setminus b}(D_A) + \{\, q_{a|A\setminus ab}(D_{A\setminus b}) + \boldsymbol{t}_{A\setminus ab}(D_{A\setminus ab})\,\}\,.$$

By the induction hypothesis the expressions in braces are $\boldsymbol{t}_{A\setminus a}(D_{A\setminus a})$ and $\boldsymbol{t}_{A\setminus b}(D_{A\setminus b})$ which yields the desired equality.
The validity of (8.23) for every $A \subseteq N$ implies (8.21): put $i = a$ and $B = A \setminus a$. This in combination with (8.17) gives (8.19).

The proof of the claim about strongly regular criterion is omitted. It is analogous: one simply writes $ct_A(D)$ instead of D_A, $\bar{q}$ instead of q and $\bar{\boldsymbol{t}}$ instead of $\boldsymbol{t}$. □

The functions $\boldsymbol{t}_A$, $A \subseteq N$ inducing a regular criterion are not uniquely determined. To characterize those collections of functions which induce the same quality criterion the following lemma characterizing special modular functions (see p. 80) is needed.

Lemma 8.4 Let $\mathcal{L}_*(N)$ denote the set of modular functions l such that $l(N) = l(\emptyset)$. Then

$$\mathcal{L}_*(N) \;=\; \{l \in \mathbb{R}^{\mathcal{P}(N)} :\; \langle l, \delta_N - \delta_\emptyset\rangle = 0 \text{ and } \forall\, G \in \mathsf{DAGS}(N)\ \langle l, u_G\rangle = 0\}\,. \tag{8.24}$$

Moreover, $\mathcal{L}_*(N)$ is a linear subspace of $\mathbb{R}^{\mathcal{P}(N)}$ of dimension $|N|$ and

$$l \in \mathcal{L}_*(N) \;\Leftrightarrow\; l = w_\emptyset \cdot m^{\emptyset\uparrow} + \sum_{i\in N} w_i \cdot m^{\{i\}\uparrow} \quad \text{where } w_\emptyset, w_i \in \mathbb{R},\ \sum_{i\in N} w_i = 0\,. \tag{8.25}$$

Proof: To show (8.24) it suffices to show that $l \in \mathbb{R}^{\mathcal{P}(N)}$ is modular iff $\langle l, u_G \rangle = 0$ for every $G \in \mathsf{DAGS}(N)$.
Indeed, by Observation 5.1 $l \in \mathcal{L}(N)$ iff $\langle l, u \rangle = 0$ for every $u \in \mathcal{E}(N)$, respectively for every $u \in \mathcal{S}(N)$. However, by Remark 7.3 and Lemma 7.1 $\mathcal{E}(N) \subseteq \{u_G;\, G \in \mathsf{DAGS}(N)\} \subseteq \mathcal{S}(N)$.
By Lemma 5.2, $\mathcal{L}(N)$ is a linear subspace of dimension $|N|+1$. Thus, $\mathcal{L}_*(N) \subset \mathcal{L}(N)$ implies that $\mathcal{L}_*(N)$ is a linear subspace whose dimension does not exceed $|N|$. To show that its dimension is $|N|$ is suffices to construct a linearly independent subset of $\mathcal{L}_*(N)$ of cardinality $|N|$. To this end fix a total ordering $b_1, \ldots, b_{|N|}$ of N and put

$$l_1 = m^{\emptyset\uparrow}, \quad l_i = m^{\{b_{i-1}\}\uparrow} - m^{\{b_i\}\uparrow}, \quad \text{for } i-2, \ldots, |N|\,. \tag{8.26}$$

By Lemma 5.2 $l_1, \ldots, l_{|N|} \in \mathcal{L}_*(N)$. Observe that they form a linearly independent set.
Indeed, suppose that for some $\alpha_i \in \mathbb{R}$ one has $\sum_{i=1}^{|N|} \alpha_i \cdot l_i(A)$ for any $A \subseteq N$. Substitute $A = \{b_i\}$ for $i = 1, \ldots, |N|$ and obtain

$$0 = \alpha_1 + \alpha_2, \quad 0 = \alpha_1 - \alpha_i + \alpha_{i+1} \text{ for } i = 2, \ldots, |N|-1 \quad \text{and} \quad 0 = \alpha_1 - \alpha_{|N|}\,.$$

Sum these equalities to get $0 = |N| \cdot \alpha_1$. Then substitute $\alpha_1 = 0$ into the equalities and show by induction that $\alpha_i = 0$ for $i = 2, \ldots, |N|$. Thus, $\alpha_1 = \ldots = \alpha_{|N|} = 0$ which concludes the proof of the fact that $l_1, \ldots, l_{|N|}$ are linearly independent.
The above observation means that $l_1, \ldots, l_{|N|}$ is a linear basis of $\mathcal{L}_*(N)$. Therefore, $l \in \mathcal{L}_*(N)$ iff $l = \sum_{i=1}^{|N|} \alpha_i \cdot l_i$ for some $\alpha_i \in \mathbb{R}$. Substitute (8.26) in this formula and get (8.25) where

$$w_\emptyset = \alpha_1, \; w_{\{b_1\}} = \alpha_2, \; w_{\{b_i\}} = -\alpha_i + \alpha_{i+1} \text{ for } i = 2, \ldots, |N|-1 \text{ and } w_{\{b_{|N|}\}} = -\alpha_{|N|}\,.$$

Because the above correspondence between w's and α's is a one-to-one correspondence, a general expresion for $l \in \mathcal{L}_*(N)$ is given by (8.25). □

Consequence 8.3 Let $\mathcal{Q}$ be a regular quality criterion for learning DAG models given by (8.19) for some $\boldsymbol{t}_A$, $A \subseteq N$. Then a collection of functions $\tilde{\boldsymbol{t}}_A : \mathsf{DATA}(A, d) \to \mathbb{R}$, $A \subseteq N$ (where $\tilde{\boldsymbol{t}}_\emptyset$ is a constant depending on d and the sample space) defines $\mathcal{Q}$ by (8.19) iff there exists constant $w_\emptyset(d)$ and $w_i(d)$, $i \in N$ such that $\sum_{i \in N} w_i(d) = 0$ and

$$\tilde{\boldsymbol{t}}_A(D_A) \;=\; \boldsymbol{t}_A(D_A) + w_\emptyset(d) + \sum_{i \in A} w_i(d) \quad \text{for } A \subseteq N,\, D \in \mathsf{DATA}(N, d)\,. \tag{8.27}$$

Proof: Let us fix the sample space and the length d of data. Then the collection of functions $\boldsymbol{t}_A$, $A \subseteq N$ can be interpreted as a real function on $\mathsf{DATA}(N, d) \times \mathcal{P}(N)$ which assigns the value $\boldsymbol{t}_A(D_A) \equiv t_D(A)$ to $D \in \mathsf{DATA}(N, d)$ and $A \subseteq N$. Moreover, the function, for every $\emptyset \neq A \subseteq N$, depends on D_A and it is constant on $\mathsf{DATA}(N, d)$ if $A = \emptyset$. The collection $\tilde{\boldsymbol{t}}_A$, $A \subseteq N$ can be represented in an analogous way. Now, by (8.19) and the definition of standard imset (see p. 114) derive that $\boldsymbol{t}_A$, $A \subseteq N$ and $\tilde{\boldsymbol{t}}_A$, $A \subseteq N$ both induce $\mathcal{Q}$ iff

$$\forall\, D \in \mathsf{DATA}(N, d)\; \forall\, G \in \mathsf{DAGS}(N) \quad \langle t_D - \tilde{t}_D, \delta_N - \delta_\emptyset - u_G \rangle = 0\,.$$

Observe that this condition is equivalent to the requirement that $l_D \equiv t_D - \tilde{t}_D \in \mathcal{L}_*(N)$ for every $D \in \mathsf{DATA}(N, d)$.

Indeed, for fixed $D \in \mathsf{DATA}(N,d)$, choose $G \in \mathsf{DAGS}(N)$ with $u_G = 0$ and observe $\langle l_D, \delta_N - \delta_\emptyset \rangle = 0$. Hence $\langle l_D, u_G \rangle = 0$ for every $G \in \mathsf{DAGS}(N)$. These two conditions are equivalent to $l_D \in \mathcal{L}_*(N)$ by (8.24) in Lemma 8.4.
Now, let us use (8.25) to every $D \in \mathsf{DATA}(N,d)$ to see that the requirement is equivalent to the existence of functions $w_\emptyset, w_i : \mathsf{DATA}(N,d) \to \mathbb{R}$, $i \in N$ such that $\sum_{i\in N} w_i(D) = 0$ and

$$l_D(A) = w_\emptyset(D) \cdot m^{\emptyset\uparrow}(A) + \sum_{i\in N} w_i(D) \cdot m^{\{i\}\uparrow}(A) \qquad \text{for } D \in \mathsf{DATA}(N,d),\ A \subseteq N. \quad (8.28)$$

The point is that these functions have to be constant.
Indeed, the substitution $A = \emptyset$ in (8.28) gives $w_\emptyset(D) = l_D(\emptyset) = \boldsymbol{t}_\emptyset(D_\emptyset) - \tilde{\boldsymbol{t}}_\emptyset(D_\emptyset)$ which is a constant. Analogously, by putting $A = \{i\}$ for $i \in N$ derive that w_i depends on D_i. Suppose for contradiction that w_i is not constant for some $i \in N$. Then databases $D^1 : x^1, \ldots, x^d$ and $D^2 : y^1, \ldots, y^d$ exist such that $w_i(D^1) \neq w_i(D^2)$. Construct a database $D^3 : z^1, \ldots, z^d$ by putting $z^\ell = [x^\ell_{N\setminus i}, y^\ell_i]$ for $\ell = 1, \ldots d$. Thus, $D^3_{N\setminus i} = D^1_{N\setminus i}$ and $D^3_i = D^2_i$. As $w_{i'}$, $i' \in N$ depend on $D_{i'}$ one has $\sum_{i'\in N} w_{i'}(D^3) - \sum_{i'\in N} w_{i'}(D^1) = w_i(D^2) - w_i(D^1) \neq 0$ which contradicts the assumption $\sum_{i\in N} w_i(D) = 0$ for any $D \in \mathsf{DATA}(N,d)$.
On the other hand, the validity of (8.28) for every $D \in \mathsf{DATA}(N,d)$ and constant functions $w_\emptyset$ and w_i, $i \in N$ is nothing but (8.27). □

Remark 8.8 The class $\mathcal{L}_*(N)$ induces an equivalence relation $t - \tilde{t} \in \mathcal{L}_*(N)$ for $t, \tilde{t} \in \mathbb{R}^{\mathcal{P}(N)}$. One can introduce a reasonable 'standard' representative of every equivalence class then, namely the only t_* within the class such that $t_*(A) = 0$ for $A \subseteq N$ with $|A| = 1$. However, the idea is not applicable in the situation one is looking for a 'standardized' collection of functions $\boldsymbol{t}_A$, $A \subseteq N$ inducing a regular criterion. The reason is that $\boldsymbol{t}_i$ need not be constant for $i \in N$ while the available 'standardization' operation (8.27) works with constants. Thus, one can achieve by 'standardization' $\boldsymbol{t}_i(D_i) = 0$ for every $i \in N$ and any chosen database $D \in \mathsf{DATA}(N,d)$ but then another database D' may exist for which this condition is not valid. △

The formulas derived in Section 8.2.1 allow one to verify easily that basic information criteria are strongly regular.

Observation 8.1 The maximized log-likelihood criterion (MLL) is strongly regular criterion for learning DAG models.

Proof: Suppose the situation described by Convention 3. Put

$$\bar{\boldsymbol{t}}_A(\boldsymbol{d}) = \begin{cases} \sum_{x\in \mathsf{X}_A} \boldsymbol{d}(x) \cdot \ln \boldsymbol{d}(x) & \text{if } \emptyset \neq A \subseteq N,\ \boldsymbol{d} \in \mathsf{CONT}(A,d), \\ d \cdot \ln d & \text{if } A = \emptyset, \end{cases} \quad (8.29)$$

where the convention $0 \cdot \ln \frac{0}{*}$ is accepted. Write by (8.9), the formula $d_{ij} = \sum_{k=1}^{r(i)} d_{ijk}$ and Convention 3 for $G \in \mathsf{DAGS}(N)$, $D \in \mathsf{DATA}(N,d)$:

$$\mathsf{MLL}\,(G,D) \;=\; \sum_{i\in N} \sum_{j=1}^{q(i,G)} \sum_{k=1}^{r(i)} d_{ijk} \cdot \ln \frac{d_{ijk}}{d_{ij}} \;=$$

$$= \sum_{i\in N}\sum_{j=1}^{q(i,G)}\sum_{k=1}^{r(i)} d_{ijk}\cdot \ln d_{ijk} - \sum_{i\in N}\sum_{j=1}^{q(i,G)} \ln d_{ij}\cdot \sum_{k=1}^{r(i)} d_{ijk} =$$

$$= \sum_{i\in N}\{ \underbrace{\sum_{j=1}^{q(i,G)}\sum_{k=1}^{r(i)} d_{ijk}\cdot \ln d_{ijk}}_{\bar{\boldsymbol{t}}_{i\cup pa_G(i)}(ct_{i\cup pa_G(i)}(D))} \quad - \quad \underbrace{\sum_{j=1}^{q(i,G)} d_{ij}\cdot \ln d_{ij}}_{\bar{\boldsymbol{t}}_{pa_G(i)}(ct_{pa_G(i)}(D))} \}\,.$$

□

The mapping which assigns the effective dimension of $\mathbb{M}_G$ to every $G \in \mathsf{DAGS}(N)$ can be viewed as a special quality criterion for learning DAG models, namely a criterion which does not depend on data.

Observation 8.2 The effective dimension (DIM) is a strongly regular criterion for learning DAG models. Akaike's information criterion (AIC) and the Bayesian information criterion (BIC) are strongly regular criteria as well. In particular, they are score equivalent and strongly decomposable.

Proof: Supposing Convention 3 put

$$\bar{\boldsymbol{t}}'_A(\boldsymbol{d}) = \begin{cases} \prod_{l\in A} r(l) & \text{if } \emptyset \neq A \subseteq N, \quad \text{for every } \boldsymbol{d} \in \mathsf{CONT}(A,d), \\ 1 & \text{if } A = \emptyset. \end{cases} \tag{8.30}$$

Then write by (8.11) and Convention 3 for every $G \in \mathsf{DAGS}(N)$ and $D \in \mathsf{DATA}(N,d)$:

$$\mathsf{DIM}\,(G,D) = \sum_{i\in N}(r(i)-1)\cdot q(i,G) = \sum_{i\in N} r(i)\cdot q(i,G) - \sum_{i\in N} q(i,G) =$$

$$= \sum_{i\in N}\; \{ \underbrace{\prod_{l\in\{i\}\cup pa_G(i)} r(l)}_{\bar{\boldsymbol{t}}'_{i\cup pa_G(i)}(ct_{i\cup pa_G(i)}(D))} \quad - \quad \underbrace{\prod_{l\in pa_G(i)} r(l)}_{\bar{\boldsymbol{t}}'_{pa_G(i)}(ct_{pa_G(i)}(D))} \}\,.$$

Thus, the DIM criterion is strongly regular. This fact together with Obervation 8.1 and the definition of the AIC criterion and the BIC criterion implies the same conclusion for the these criteria. Finally, use Lemma 8.3. □

Remark 8.9 In my view, the logarithm of the marginal likelihood (LML) mentioned in Remark 8.5 can also be shown to be strongly regular if relevant assumptions on the priors π_G, $G \in \mathsf{DAGS}(N)$ are accepted. More exactly, I have in mind the assumptions explicated in [28], namely technical assumption that π_G are Dirichlet measures and the assumption that the priors are compatible (see § 6 of [28]). Actually, as mentioned on p. 54 of [19] there are examples of Bayesian criteria which are both decomposable and score equivalent. Therefore, they are regular by Lemma 8.3. △

8.3 Inclusion neighbourhood

The binary relation $\mathcal{M}_K \subseteq \mathcal{M}_L$ for $K, L \in \mathsf{DAGS}(N)$ defines the *inclusion quasiordering* on the set of acyclic directed graphs over N. It can also be viewed as a partial ordering on the set of Markov equivalence classes of acyclic directed graphs over N respectively on the set of DAG models over N. In this chapter, every $G \in \mathsf{DAGS}(N)$ is also interpreted as a statistical model $\mathbb{M}_G$ on a (fixed) discrete sample space $\mathsf{X}_N = \prod_{i \in N} \mathsf{X}_i$ (see p. 140). Clearly, $\mathcal{M}_K \subseteq \mathcal{M}_L$ implies $\mathbb{M}_K \supseteq \mathbb{M}_L$ and the converse is true if $|\mathsf{X}_i| \geq 2$ for every $i \in N$ because of the existence of a perfectly Markovian measure on X_N for every $G \in \mathsf{DAGS}(N)$ [87]. In particular, the *strict inclusion* of formal independence models $\mathcal{M}_K \subset \mathcal{M}_L$ for $K, L \in \mathsf{DAGS}(N)$ is equivalent to strict inclusion $\mathbb{M}_K \supset \mathbb{M}_K$ of statistical models then.

We say that $K, L \in \mathsf{DAGS}(N)$ are *inclusion neighbours* and write $\mathcal{M}_K \sqsubset \mathcal{M}_L$ if $\mathcal{M}_K \subset \mathcal{M}_L$ and there is no $G \in \mathsf{DAGS}(N)$ such that $\mathcal{M}_K \subset \mathcal{M}_G \subset \mathcal{M}_L$. More exactly, we say then that $\mathcal{M}_L$ is the *upper neighbour* of $\mathcal{M}_K$ or, dually, that $\mathcal{M}_K$ is the *lower neighbour* of $\mathcal{M}_L$. The *inclusion neighbourhood* of $\mathcal{M}_G, G \in \mathsf{DAGS}(N)$ consists of the union of the collection of upper neighbours and the collection of lower neighbours.

Remark 8.10 The inclusion ordering can be understood in two ways, that is, either in the sense $\mathcal{M}_K \subseteq \mathcal{M}_L$ or in the sense $\mathbb{M}_K \supseteq \mathbb{M}_L$. Thus, it is the matter of taste how this asymmetric binary relation of graphs K and L is reflected in the terminology, notation and pictures. Whittaker [146] had in mind the interpretation of graphs in terms of statistical models. Thus, in his view, if $\mathbb{M}_K \supseteq \mathbb{M}_L$ then K represents a bigger statistical model than L. Therefore, he called the saturated model, which is represented by a complete graph, the 'maximal' graphical model (see p. 185 of [146]) and placed the pictorial representatives of this model always at the top of his lattice diagrams. Chickering [22] used symbol $L \leq K$ to denote $\mathcal{M}_K \subseteq \mathcal{M}_L$. I think that he was mainly led by graphical pictorial representation. Indeed, $\mathcal{M}_K \subseteq \mathcal{M}_L$ implies that K has at least those edges which are present in L. Similarly, in [54, 19], the graphs with maximal numbers of edges are at the top of the respective Hasse diagrams.

However, CI interpretation is the main issue in this monograph. Therefore I prefer the interpretation in terms of formal independence model, that is, $\mathcal{M}_K \subseteq \mathcal{M}_L$. One of my reasons is that this point of view admits a wider perspective: graphical models can be interpreted as the elements of the lattice of CI structures. This approach is naturally reflected in terminology and pictures. Therefore, if $\mathcal{M}_K \sqsubset \mathcal{M}_L$ then $\mathcal{M}_L$ is called the upper neighbour of $\mathcal{M}_K$ and the pictorial representative of $\mathcal{M}_L$ is put above the pictorial representative of $\mathcal{M}_K$ in all Hasse diagrams in this book. In particular, the saturated model, which represents the largest class of probability measures but no CI information, is always placed at the bottom of a Hasse diagram. Note that in order to avoid confusion I intentionally do not use notation like $L \leq K$. $\triangle$

The concept of inclusion neighbourhood is certainly important from the theoretical point of view. Nevertheless, there are practical reasons for which it deserves serious attention. These reasons are strongly related to the local search method mentioned on p. 138. It is claimed on p. 66 of [19] that a necessary condition for a neighbourhood structure to avoid an unpleasant situation that a local search procedure finds a local maximum of a quality criterion which is not the global maximum is the *inclusion boundary condition* [54]. This condition requires, for every $G \in \mathsf{DAGS}(N)$, that the inclusion neighbourhood of $\mathcal{M}_G$ should be covered by $\bigcup\{\mathcal{M}_H; H \in nei\,(G)\}$ where $nei\,(G)$ is the class of neighbouring graphs for G.

Actually, the significance of this condition was made out somehow implicitly already in [88] in connection with a particular local search procedure named *greedy equivalence* *search* (GES) algorithm. The goal of the GES algorithm is learning DAG models. Every DAG model over N can be represented graphically by any member of an equivalence class of acyclic directed graphs over N. Therefore, the quality criteria used in this procedure are supposed to be both decomposable and score equivalent. The states of the procedure are in fact equivalence classes of graphs although it formally works with individual elements of $\mathsf{DAGS}(N)$. The algorithm has two phases. In the first phase edges are added and in the second phase edges are removed. Thus, in each phase a different part of the overall neighbourhood structure of a 'current' graph is considered.

(i) The first phase starts with an acyclic directed graph G over N which has no arrows. In this phase, the set of neighbouring graphs of a current $G \in \mathsf{DAGS}(N)$ consists of $K \in \mathsf{DAGS}(N)$ such that there exists $L \in \mathsf{DAGS}(N)$ which is equivalent to G and K is obtained from L by adding one arrow. Provided there exists a neighbouring graph K which has strictly positive increase $\mathcal{Q}(K, D) - \mathcal{Q}(G, D)$, $D \in \mathsf{DATA}(N, d)$ in the value of the quality criterion one chooses K with maximal increase and moves from G to K which becomes a new current graph. The procedure is repeated until the value of the quality criterion in the current graph cannot be increased in this way.

(ii) The second phase starts with the result of the first phase. It employs the complementary part of the overall neighbourhood structure. In the second phase, the set of neighbouring graphs of a current $G \in \mathsf{DAGS}(N)$ consists of $L \in \mathsf{DAGS}(N)$ such that there exists $K \in \mathsf{DAGS}(N)$ which is equivalent to G and L is obtained from K by the removal of an arrow. Again, one is trying to maximize the increase $\mathcal{Q}(L, D) - \mathcal{Q}(G, D)$ in the value of the quality criterion and the procedure stops when the value of $\mathcal{Q}$ in the current graph cannot be increased in this way. The output of the GES algorithm is the equivalence class of acyclic directed graphs over N which contains the last current graph $G \in \mathsf{DAGS}(N)$.

Meek [88] formulated a conjecture about transformational graphical characterization of inclusion quasiordering for acyclic directed graphs (see Lemma 8.5 below). As reported in [22], Meek also showed that if his conjecture is true, the data faithfulness assumption (see p. 133) together with several other technical statistical assumptions are fulfilled, and the length of the database D approaches to infinity then the GES algorithm should find the "optimal" solution, that is, the equivalence class of acyclic directed graphs with respect to which the probability measure which "generates" data is perfectly Markovian. The validity of the conjecture is needed to show that the GES algorithm cannot terminate with a model which is not the optimal solution - for details see the proof of Lemma 4 in [22]. The conjecture about tranformational characterization of inclusion quasiordering have been recently confirmed in [22] with help of a method used in [55] to verify the conjecture in a special case. The characterization is based on the operation of legal arrow reversal (see p. 47) and the operation of *legal arrow adding* by which is meant the adding of an arrow into an acyclic directed graph such that the resulting directed graph remains acyclic. However, I prefer a simpler formulation of the result which uses the inverse operation. By an *arrow removal* is understood a change of an acyclic directed graph K over N into a directed graph L over N: one simply removes an arrow $a \to b$ in K. Evidently, L is again an acyclic directed graph. Moreover, one has $\mathcal{M}_K \subset \mathcal{M}_L$ (which follows from the following lemma).

Lemma 8.5 Suppose that $K, L \in \mathsf{DAGS}(N)$. Then $\mathcal{M}_K \subseteq \mathcal{M}_L$ iff there exists a sequence $K = G_1, \ldots, G_n = L$, $n \geq 1$ of acyclic directed graphs over N such that for every $i = 1, \ldots, n-1$

- either G_{i+1} is obtained from G_i by a legal arrow reversal, or
- G_{i+1} is obtained from G_i by an arrow removal.

In particular, $\mathcal{M}_K \subset \mathcal{M}_L$ if a sequence of graphs of this type exists such that $n \geq 2$ and the arrow removal operation is applied at least once in the sequence.
Finally, $\mathcal{M}_K \sqsubset \mathcal{M}_L$ if there exists $K', L' \in \mathsf{DAGS}(N)$ such that K' is equivalent to K, L' is equivalent to L and L' is obtained from K' by an arrow removal.

Proof: The first claim of the lemma is proved in [22]. However, it is formulated in "converse" ordered; namely $\mathcal{M}_K \subseteq \mathcal{M}_L$ if there exists a sequence $L = H_1, \ldots, H_n = K$, $n \geq 1$ such that, for every $j = 1, \ldots, n-1$, H_{j+1} is obtained from H_j either by a legal arrow reversal or by a legal arrow adding.

The second claim of the lemma can be verified as follows: the necessity of the condition follows from the first claim and the tranformational characterization of equivalence of acyclic directed graphs mentioned on p. 47. The sufficiency follows from the equivalence characterization and the observation that if L is obtained from $K \in \mathsf{DAGS}(N)$ by the removal of an edge then $\mathcal{M}_K \subset \mathcal{M}_L$.
Indeed, $\mathcal{M}_K \subseteq \mathcal{M}_L$ follows from the definition of the induced DAG model (p. 46): every route which is active in L is an active route in K which means that every route blocked in K has to be blocked in L. The fact $\mathcal{M}_K \neq \mathcal{M}_L$ follows from graphical characterization of Markov equivalence of acyclic directed graphs (p. 46): equivalent graphs have the same underlying graph.
The condition in the third claim is equivalent to the requirement that there exists a sequence $K = G_1, \ldots, G_n = L$, $n \geq 2$ of the considered type such that the arrow removal operation is applied exactly once (use transformational characterization of equivalence). Thus, the third claim can be derived from the second one.
Indeed, if there are two arrow removals in the sequence then one has $\mathcal{M}_K \subset \mathcal{M}_G \subset \mathcal{M}_L$ for any graph G in the sequence which is after the first removal and before the last removal. Conversely, if $\mathcal{M}_K \subset \mathcal{M}_G \subset \mathcal{M}_L$ for an acyclic directed graph G then by the second claim of the lemma $|\mathcal{A}(K)| > |\mathcal{A}(G)| > |\mathcal{A}(L)|$ which implies that K has at least two more edges than L. □

Thus, transformational characterization of inclusion quasiordering for acyclic directed graphs is available. However it does not seem to be very suitable for testing inclusion of DAG models which is the task to decide, given $K, L \in \mathsf{DAGS}(N)$, whether $\mathcal{M}_K \subseteq \mathcal{M}_L$ holds or not. Indeed, one has to try to construct a sequence of graphs between K and L and it is not clear how to do it. The question of whether a suitable direct graphical characterization of inclusion quasiordering for acyclic directed graphs exists remain open - see Theme 2. On the other hand, a special algebraic criterion is proposed in the next section.

8.4 Standard imsets and learning

The method of structural imsets can be applied in the area of learning DAG models. It brings a certain simplifying perspective.

8.4.1 Inclusion neighbourhood characterization

Standard imsets for acyclic directed graphs (see p. 114) allow one to characterize the inclusion quasiordering in an algebraic way.

Observation 8.3 Let $K, L \in \mathsf{DAGS}(N)$ such that L is obtained from K by the removal of an arrow $a \to b$ in K. Then $u_L - u_K = u_{\langle a,b|C\rangle}$ where $C = pa_K(b) \setminus \{a\} = pa_L(b)$.

Proof: As $pa_L(c) = pa_K(c)$ for $c \in N \setminus b$, $pa_L(b) = C$ and $pa_K(b) = aC$ one has by (7.3) $u_L - u_K = \{\delta_{pa_L(b)} - \delta_{bpa_L(b)}\} - \{\delta_{pa_K(b)} - \delta_{bpa_K(b)}\} = \delta_C - \delta_{bC} - \delta_{aC} + \delta_{abC}$. $\square$

Lemma 8.6 Suppose $K, L \in \mathsf{DAGS}(N)$. Then $\mathcal{M}_K \subseteq \mathcal{M}_L$ ($\equiv$ $\mathbb{M}_K \supseteq \mathbb{M}_L$) iff $u_L - u_K$ is a combinatorial imset over N. Moreover, $\mathcal{M}_K \subset \mathcal{M}_L$ ($\equiv$ $\mathbb{M}_K \supset \mathbb{M}_L$) iff $u_L - u_K$ is a non-zero combinatorial imset.

Proof: If $\mathcal{M}_K \subseteq \mathcal{M}_L$ then apply Lemma 8.5 and consider the respective sequence $K = G_1, \ldots, G_n = L$, $n \geq 1$. If G_{i+1} is obtained from G_i by a legal arrow reversal then G_{i+1} is Markov equivalent to G_i and $u_{G_{i+1}} = u_{G_i}$ by Consequence 7.1. If G_{i+1} is obtained from G_i by an arrow removal then $u_{G_{i+1}} - u_{G_i} \in \mathcal{E}(N)$ by Observation 8.3. Thus $u_L - u_K = \sum_{i=1}^{n-1} u_{G_{i+1}} - u_{G_i}$ is a combinatorial imsets. Conversely, if $u_L - u_K \in \mathcal{C}(N)$ then by Lemma 6.1 $u_L \rightharpoonup u_K$, that is, $\mathcal{M}_{u_K} \subseteq \mathcal{M}_{u_L}$. By Lemma 7.1 this means $\mathcal{M}_K \subseteq \mathcal{M}_L$.

Thus, the proved fact implies that $\mathcal{M}_K \subset \mathcal{M}_L$ iff $\{\, u_L - u_K \in \mathcal{C}(N)\ \&\ u_K - u_L \notin \mathcal{C}(N)\,\}$. However, it follows from Observation 4.4 for $w \in \mathcal{C}(N)$ that $-w \in \mathcal{C}(N) \Leftrightarrow w = 0$. Hence, the latter condition is that $u_L - u_K$ is a non-zero combinatorial imset. $\square$

Note that the above characterization gives a simple algebraic criterion for testing inclusion of DAG models. Indeed, as mentioned in Remark 6.3 the question whether an imset is a combinatorial imset is decidable.

Consequence 8.4 Suppose $K, L \in \mathsf{DAGS}(N)$. Then $\mathcal{M}_K \sqsubset \mathcal{M}_L$ ($\equiv \mathbb{M}_K \sqsupset \mathbb{M}_L$) iff $u_L - u_K$ is an elementary imset.

Proof: To prove the necessity of the condition use the third claim of Lemma 8.5: by Consequence 7.1 and Observation 8.3 get $u_L - u_K = u_{L'} - u_{K'} \in \mathcal{E}(N)$. To prove the sufficiency use Lemma 8.6 to derive $\mathcal{M}_K \subset \mathcal{M}_L$ and suppose for contradiction that $\mathcal{M}_K \subset \mathcal{M}_G \subset \mathcal{M}_L$ for some $G \in \mathsf{DAGS}(N)$. Then observe again by Lemma 8.6 that $u_L - u_K = (u_L - u_G) + (u_G - u_K)$ is the sum of two non-zero combinatorial imsets. Thus, $deg(u_L - u_K) \geq 2$ and $u_L - u_K \notin \mathcal{E}(N)$ by Consequence 4.2 which contradicts the assumption. $\square$

Of course, given a pair $K, L \in \mathsf{DAGS}(N)$ such that $\mathcal{M}_K \sqsubset \mathcal{M}_L$, the elementary imset $u_L - u_K$ is determined uniquely. It will be called the *differential imset* for K and L.

Remark 8.11 The concept of differential imset brings interesting interpretation of the edges in the Hasse diagram of the poset of DAG models over N. Every link of this diagram can be annotated by the respective elementary triplet and one can interpret the links of the diagram in terms of elementary conditional (in)dependence statements. An example of the diagram of this type for N with $|N| = 3$ is given in Figure 8.1. Note that the DAG models are represented by standard imsets in the figure and that graphical representatives, namely the essential graphs which coincide with the largest chain graphs in case $|N| \leq 3$ are attached for readers convenience.

The above mentioned interpretation becomes relevant in connection with the local search method for learning DAG models (see p. 138). It was explained in Section 8.3 that, both from the theoretical and practical point of view, it is natural to consider the search space (of unique representatives) of DAG models endowed with the neighbourhood stucture given by the concept of inclusion neighbourhood. This search space is nothing but the Hasse diagram mentioned above. The point is that the 'moves' in the search space between neighbouring states have natural interpretation in terms of elementary conditional (in)dependence statements. This fact becomes even more easy to see when one considers a particular regular criterion for learning DAG models - see Remark 8.13. Note, however, that this particular CI interpretation of moves between states of the search space in only possible in the context of DAG models - see Example 9.4.

8.4.2 Regular criteria and standard imsets

LEMMA 8.7 Let $\mathcal{Q}$ be a regular quality criterion for learning DAG models. Then there exist a function $s : \mathsf{DATA}(N,d) \to \mathbb{R}$ and a mapping $t :\; D \in \mathsf{DATA}(N,d) \mapsto t_D \in \mathbb{R}^{\mathcal{P}(N)}$ such that

- $\forall\, A \subseteq N\; |A| \leq 1 \quad t_D(A) = 0$ for every $D \in \mathsf{DATA}(N,d)$,
- $\forall\, A \subseteq N\; |A| \geq 2$ the coordinate mapping $D \mapsto t_D(A)$ depends on D_A,

and for every $G \in \mathsf{DAGS}(N)$ and $D \in \mathsf{DATA}(N,d)$

$$\mathcal{Q}(G,D) \;=\; s(D) - \langle t_D, u_G \rangle\,. \tag{8.31}$$

The function s and the mapping t are uniquely detemined by these two requirements. Moreover, if $\mathcal{Q}$ is strongly regular, then

- $\forall\, A \subseteq N\; |A| \geq 2$ the mapping $D \mapsto t_D(A)$ depends on the contingency table $ct_A(D)$.

Proof: Supposing (8.19) holds for a collection $\boldsymbol{t}_A : \mathsf{DATA}(A,d) \to \mathbb{R}$, $A \subseteq N$ let us put

$$t'_D(A) \;=\; \boldsymbol{t}_A(D_A) \quad \text{for } D \in \mathsf{DATA}(N,d),\; A \subseteq N. \tag{8.32}$$

Thus, $t'_D \in \mathbb{R}^{\mathcal{P}(N)}$ for every data D. Introduce for every $D \in \mathsf{DATA}(N,d)$

$$s(D) \;=\; t'_D(N) - t'_D(\emptyset) = \langle t'_D, \delta_N - \delta_\emptyset \rangle \text{ and} \tag{8.33}$$

$$t_D \;=\; t'_D - t'_D(\emptyset)\cdot m^{\emptyset\uparrow} - \sum_{i\in N} \{\, t'_D(\{i\}) - t'_D(\emptyset)\,\} \cdot m^{\{i\}\uparrow} \tag{8.34}$$

(see p. 41). This definition ensures $t_D(A) = 0$ for every D and $|A| \leq 1$. Moreover, for every D and $|A| \geq 2$ one has

$$t_D(A) \;=\; \boldsymbol{t}_A(D_A) - \sum_{i\in A} \boldsymbol{t}_{\{i\}}(D_i) + (|A|-1)\cdot \boldsymbol{t}_\emptyset(D_\emptyset),$$

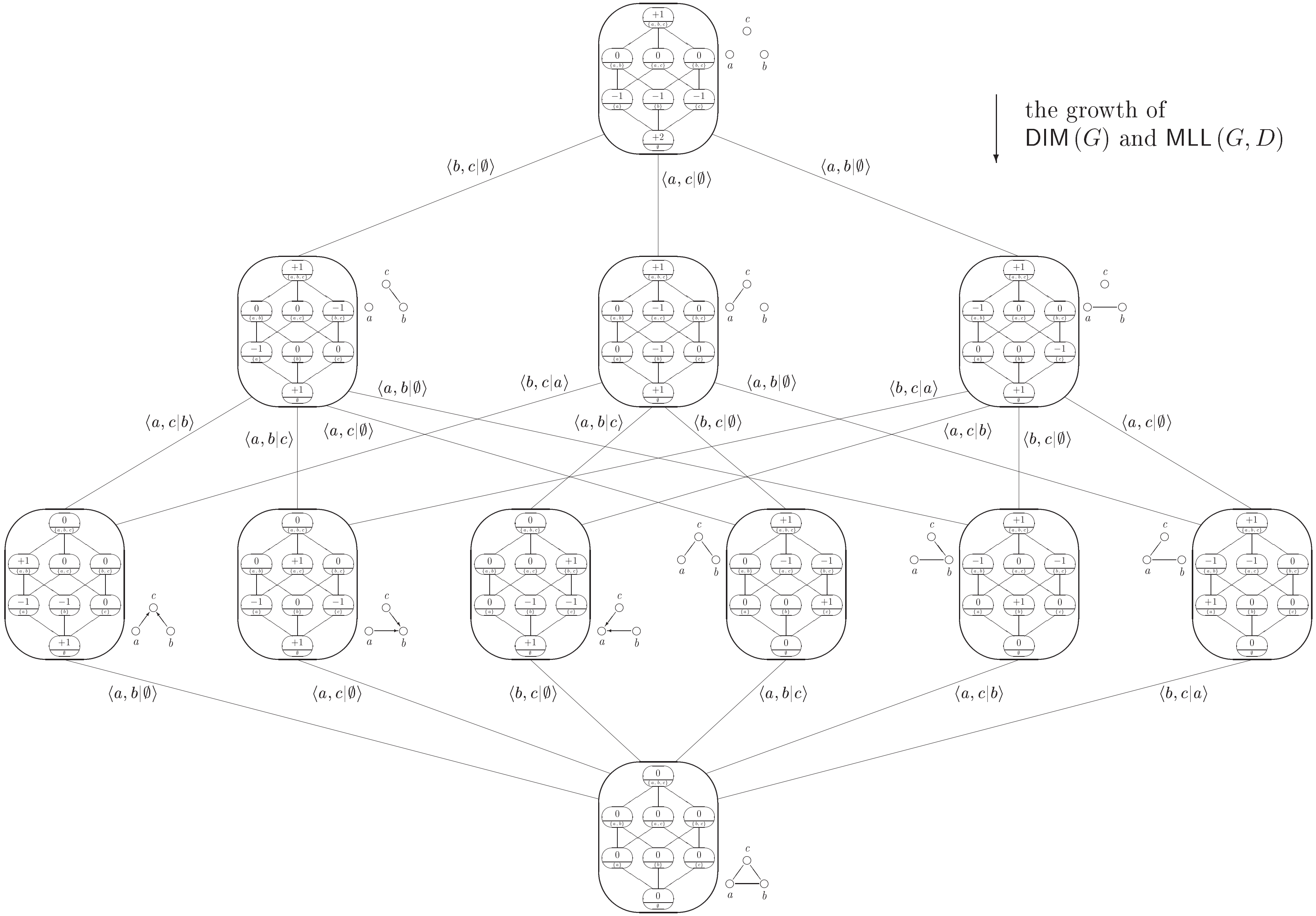


Figure 8.1: The search space of standard imsets over $N = \{a, b, c\}$ (rotated).

which implies that the mapping $D \mapsto t_D(A)$ depends on D_A. Observe that if $\mathcal{Q}$ is strongly regular then $t_D(A)$ depends on $ct_A(D)$ since $ct_B(D)$ is a function of $ct_A(D)$ if $B \subseteq A$.

By Lemma 5.2, for every D, $t_D - t'_D \in \mathcal{L}(N)$ which implies by Observation 5.1 that $\langle t_D - t'_D, u \rangle = 0$ for any $u \in \mathcal{S}(N)$. Hence, $\langle t_D, u_G \rangle = \langle t'_D, u_G \rangle$ for every $G \in \mathsf{DAGS}(N)$, $D \in \mathsf{DATA}(N,d)$. Thus, (8.19) takes the form

$$\mathcal{Q}(G,D) \;=\; \langle t'_D, \delta_N - \delta_\emptyset \rangle - \langle t'_D, u_G \rangle \;=\; s(D) - \langle t_D, u_G \rangle \,.$$

To evidence the uniqueness of s take $H \in \mathsf{DAGS}(N)$ with $u_H = 0$ and by (8.31) observe $s(D) = \mathcal{Q}(H,D)$ for every D. To verify the uniqueness of t suppose that another mapping $\tilde{t} : D \in \mathsf{DATA}(N,d) \mapsto \tilde{t}_D \in \mathbb{R}^{\mathcal{P}(N)}$ satisfies (8.31). Then $\langle t_D - \tilde{t}_D, u_G \rangle = 0$ for every $D \in \mathsf{DATA}(N,d)$ and $G \in \mathsf{DAGS}(N)$. By Remark 7.3 $\mathcal{E}(N) \subseteq \{u_G;\, G \in \mathsf{DAGS}(N)\}$ and therefore $t_D - \tilde{t}_D \in \mathcal{L}(N)$ for every D (use Observation 5.1). As $t_D - \tilde{t}_D$ is ℓ-standardized $t_D - \tilde{t}_D \equiv 0$ by Remark 5.3. □

Remark 8.12 The condition from Lemma 8.7 is necessary for the regularity of a quality criterion but it is not sufficient. To evidence this in case of non-trivial sample spaces, that is, if $|\mathsf{X}_i| \geq 2$ for every $i \in N$, consider a statistic $s : \mathsf{DATA}(N,d) \to \mathbb{R}$ which cannot be written in the form of the sum $s(D) = \sum_{i \in N} s_i(D_i)$ where $s_i : \mathsf{DATA}(\{i\},d) \to \mathbb{R}$, $i \in N$. The existence of s follows easily from the dimension comparison of both linear spaces of functions. Then put $t_D = 0$ for any D and define $\mathcal{Q}$ by (8.31): $\mathcal{Q}(G,D) = s(D)$ for any $G \in \mathsf{DAGS}(N)$, $D \in \mathsf{DATA}(N,d)$. However, $\mathcal{Q}$ is not regular because s was constructed not be decomposable (consider $G \in \mathsf{DAGS}(N)$ which has no arrows).

Another warning is that even if $\mathcal{Q}$ is a regular criterion given by (8.31) the mapping $A \mapsto t_D(A)$ from Lemma 8.7 need not induce $\mathcal{Q}$ in the sense of (8.19). △

Given a particular quality criterion $\mathcal{Q}$ for learning DAG models the symbols $s^{\mathcal{Q}}$ and $t^{\mathcal{Q}}$ will be used to denote those unique function s and mapping t which satisfy the requirement from Lemma 8.7. The function $s^{\mathcal{Q}}$ will be called the *saturating function* (of the criterion $\mathcal{Q}$) and the mapping $t^{\mathcal{Q}}$ the ℓ-standardized *transformation of data* (relative to $\mathcal{Q}$). Moreover, given a database $D \in \mathsf{DATA}(N,d)$ the vector $[t^{\mathcal{Q}}_D(A)]_{A \subseteq N} \in \mathbb{R}^{\mathcal{P}(N)}$ will be called the (ℓ-standardized) *data vector* (relative to $\mathcal{Q}$).

Consequence 8.5 Let $\mathcal{Q}$ be a regular quality criterion for learning DAG models and $K, L \in \mathsf{DAGS}(N)$ such that $\mathcal{M}_K \sqsubset \mathcal{M}_L$. Moreover, let $t^{\mathcal{Q}}_D$ denote the data vector given by $D \in \mathsf{DATA}(N,d)$ and $u_{\langle a,b|C \rangle}$ is the differential imset for K and L. Then

$$\mathcal{Q}(K,D) - \mathcal{Q}(L,D) \;=\; \langle t^{\mathcal{Q}}_D, u_{\langle a,b|C \rangle} \rangle \,. \tag{8.35}$$

Proof: This immediately follows from (8.31) and the definition of the differential imset $u_{\langle a,b|C \rangle} = u_L - u_K$. □

Remark 8.13 The method of structural imsets leads to the following proposal of how to modify and implement the local search method for learning DAG models.

- The states of the search space can be represented by *standard imsets*.
- The moves betveen states of the space can be represented by *differential imsets*.

- Given a regular criterion $\mathcal{Q}$ for learning DAG models data be represented by the respective ℓ-standardized *data vector*.

The formula (8.31) says that $\mathcal{Q}$ is a linear function of those state representatives and data representatives. Moreover, owing to (8.35) the change in the value of $\mathcal{Q}$ after a move can be interpreted in terms of CI. Indeed, the value $\langle t_D^{\mathcal{Q}}, u_{\langle a,b|C\rangle}\rangle$ can be viewed as the qualitative evaluation of the step in which the hypothesis $a \perp\!\!\!\perp b\,|\,C$ is rejected. This interpretation is indeed possible in case of the MLL criterion - see Remark 8.14. $\triangle$

OBSERVATION 8.4 The saturating function and data vector relative to the MLL criterion are as follows:

$$\begin{array}{rcl} s^{\mathsf{MLL}}(D) &=& d\cdot H(\hat{P}|\upsilon)\,,\\ t_D^{\mathsf{MLL}}(A) &=& d\cdot m_{\hat{P}}(A)\,, \end{array} \qquad \text{for } D\in\mathsf{DATA}(N,d),\ A\subseteq N,, \tag{8.36}$$

where $\hat{P}$ is the empirical measure computed from D (see Section 10.10.1), υ the counting measure on X_N and $m_{\hat{P}}$ the multiinformation function (see Section 2.3.4).

Proof: Use the procedure from the proof of Lemma 8.7. By (8.32) and (8.29) write for $D\in\mathsf{DATA}(N,d)$ using Convention 3

$$\boldsymbol{t}'_D(A) \;=\; \sum_{\boldsymbol{x}\in\mathsf{X}_A} d_{\boldsymbol{x}}\cdot\ln d_{\boldsymbol{x}} \qquad \text{if } \emptyset\neq A\subseteq N,\quad \boldsymbol{t}'_D(\emptyset)=d\cdot\ln d\,.$$

Hence, by (8.33)

$$\begin{array}{rcl} s^{\mathsf{MLL}}(D) &=& \displaystyle\sum_{\boldsymbol{x}\in\mathsf{X}_N} d_{\boldsymbol{x}}\cdot\ln d_{\boldsymbol{x}} - d\cdot\ln d \;=\; \sum_{\boldsymbol{x}\in\mathsf{X}_N} d_{\boldsymbol{x}}\cdot\ln d_{\boldsymbol{x}} - \sum_{\boldsymbol{x}\in\mathsf{X}_N} d_{xmx}\cdot\ln d \;=\\ &=& \displaystyle\sum_{\boldsymbol{x}\in\mathsf{X}_N} d_{\boldsymbol{x}}\cdot\ln\frac{d_{\boldsymbol{x}}}{d} \;=\; d\cdot\sum_{\boldsymbol{x}\in\mathsf{X}_N}\frac{d_{\boldsymbol{x}}}{d}\cdot\ln\frac{d_{\boldsymbol{x}}}{d} \;=\; d\cdot\sum_{\boldsymbol{x}\in\mathsf{X}_N}\hat{p}_N(\boldsymbol{x})\cdot\ln\hat{p}_N(\boldsymbol{x})\,. \end{array}$$

The formula for t_D^{MLL} follows from (8.34) using Convention 3:

$$\begin{array}{rcl} t_D^{\mathsf{MLL}}(A) &=& \displaystyle\sum_{\boldsymbol{x}\in\mathsf{X}_A} d_{\boldsymbol{x}}\cdot\ln d_{\boldsymbol{x}} - \sum_{i\in A}\sum_{\boldsymbol{y}\in\mathsf{X}_i} d_{\boldsymbol{y}}\cdot\ln d_{\boldsymbol{y}} + (|A|-1)\cdot d\cdot\ln d \;=\\ &=& \displaystyle\sum_{\boldsymbol{x}\in\mathsf{X}_A} d_{\boldsymbol{x}}\cdot\ln d_{\boldsymbol{x}} - \sum_{i\in A}\sum_{\boldsymbol{x}\in\mathsf{X}_A} d_{\boldsymbol{x}}\cdot\ln d_{\boldsymbol{x}_i} + (|A|-1)\cdot\sum_{\boldsymbol{x}\in\mathsf{X}_A} d_{\boldsymbol{x}}\cdot\ln d \;=\\ &=& \displaystyle\sum_{\boldsymbol{x}\in\mathsf{X}_A} d_{\boldsymbol{x}}\cdot\ln\frac{d_{\boldsymbol{x}}\cdot d^{|A|-1}}{\prod_{i\in A} d_{\boldsymbol{x}_i}} \;=\; d\cdot\sum_{\boldsymbol{x}\in\mathsf{X}_A}\frac{d_{\boldsymbol{x}}}{d}\cdot\ln\frac{d_{\boldsymbol{x}}\cdot d^{-1}}{\prod_{i\in A} d_{\boldsymbol{x}_i}\cdot d^{-1}} \;=\\ &=& \displaystyle d\cdot\sum_{\boldsymbol{x}\in\mathsf{X}_A}\hat{p}_A(\boldsymbol{x})\cdot\ln\frac{\hat{p}_A(\boldsymbol{x})}{\prod_{i\in A}\hat{p}_i(\boldsymbol{x}_i)} \;=\; d\cdot m_{\hat{P}}(A)\,. \end{array}$$

$\square$

REMARK 8.14 It follows from Observation 8.4 and Consequence 2.2 that t_D^{MLL} is a supermodular function for every $D\in\mathsf{DATA}(N,d)$. More specifically, supposing $K,L\in$

$\mathsf{DAGS}(N)$ such that $\mathcal{M}_K \sqsubset \mathcal{M}_L$ and $u_{\langle a,b|C\rangle}$ is the differential imset for K and L the formula (8.35) in Consequence 8.5 implies

$$\mathsf{MLL}\,(K,D) - \mathsf{MLL}\,(L,D) \;=\; \langle t_D^{\mathsf{MLL}}, u_{\langle a,b|C\rangle}\rangle = d \cdot \langle m_{\hat{P}}, u_{\langle a,b|C\rangle}\rangle\,.$$

Thus, the MLL criterion is descreasing with respect to the inclusion quasiordering. No matter what data are given it attains its maximum for graphs whose underlying graph is complete. (see Figure 8.1 for illustration). Moreover, it follows from the proof of Consequence 2.2 that the number $\langle m_{\hat{P}}, u_{\langle a,b|C\rangle}\rangle$ is just the relative entropy of $\hat{P}^{abC}$ with respect to the conditional product $\hat{Q}$ of $\hat{P}^{aC}$ and $\hat{P}^{bC}$ introduced in Remark 2.11. Therefore,

$$\mathsf{MLL}\,(K,D) - \mathsf{MLL}\,(L,D) \;=\; d \cdot \langle m_{\hat{P}}, u_{\langle a,b|C\rangle}\rangle \;=\; d \cdot H(\hat{P}^{abC}|\hat{Q})\,.$$

Because $\hat{Q}$ is nothing but the fitted empirical measure defined in Section 10.10.1 the later expression is one half of the value of the G^2-statistics for testing $a \perp\!\!\!\perp b\,|\,C$. This observation is interesting in connection with Whittaker's approach [146] to learning UG models by the minimization of the deviance (see Section 8.1). Well, if one is interested in learning decomposable models which are known to be special cases of DAG models (see Section 3.4.1) then by Remark 8.2 the value of the deviance of a statistical model described by an undirected graph H over N can be expressed as $\boldsymbol{k} - 2 \cdot \mathsf{MLL}\,(G,D)$ where $\boldsymbol{k}$ is a constant which does not depend on the model and $G \in \mathsf{DAGS}(N)$ induces the same model as H. Thus, the deviance difference for a pair of triangulated undirected graphs which differ in the presence of one line is nothing but the multiple by two of the difference in the value of the MLL criterion for a pair of acyclic directed graphs which are inclusion neighbours. In particular, as explained above, the deviance difference is the value of the G^2-statistic for testing the respective elementary CI statement. This explains the phenomenon observed by Whittaker [146] and mentioned on p. 135 and in Remark 8.2. △

CONSEQUENCE 8.6 The saturating function and the data vector relative to the DIM criterion are as follows:

$$\begin{array}{rcl} s^{\mathsf{DIM}}(D) &=& \prod_{i\in N} r(i) - 1 \\ t_D^{\mathsf{DIM}}(A) &=& \prod_{i\in A} r(i) - \sum_{i\in A} r(i) + |A| - 1 \end{array} \qquad \text{for any } D \in \mathsf{DATA}(N,d),\ A \subseteq N\,. \qquad (8.37)$$

The formulas for data vectors relative to the AIC criterion and the BIC criterion are here:

$$\begin{array}{rcl} t_D^{\mathsf{AIC}}(A) &=& d \cdot m_{\hat{P}}(A) - \prod_{i\in A} r(i) + \sum_{i\in A} r(i) - |A| + 1\,, \\ t_D^{\mathsf{BIC}}(A) &=& d \cdot m_{\hat{P}}(A) - \frac{\ln d}{2} \cdot \{\prod_{i\in A} r(i) - \sum_{i\in A} r(i) + |A| - 1\,\}\,. \end{array} \qquad (8.38)$$

Proof: The formula (8.37) can be obtained by substituting (8.30) into (8.33) and (8.34). The other formulas follow from Observation 8.4 using the definition of information criteria (see Section 10.10.4). □

Note that is follows from Consequence 8.6 that

$$\langle t_D^{\mathsf{DIM}}, u_{\langle a,b|C\rangle}\rangle \;=\; (r(a)-1)\cdot(r(b)-1)\cdot\prod_{c\in C} r(c) \qquad \text{for every } u_{\langle a,b|C\rangle} \in \mathcal{E}(N)\,.$$

In particular, t_D^{DIM} is also supermodular and the DIM criterion is decreasing with respect to inclusion quasiordering (see Figure 8.1 for illustration). On the other hand, this is not true for the AIC criterion and for the BIC criterion: each of them is defined as the difference of two decreasing criteria. Thus, both the AIC criterion and the BIC criterion can attain its maximal values inside the Hasse diagram of the inclusion ordering for DAG models.

As indicated in Remark 8.9, I believe that analogous formulas like in Consequence 8.6 can be derived for some of the Bayesian criteria for learning DAG models.

REMARK 8.15 One of possible objections against the use of imsets is that memory demands for their representation in a computer are too high. However, if one is only interested in learning DAG models then standard imsets for acyclic directed graphs can be represented in memory of a computer effectively. Indeed, any standard imset over N has at most $2|N|+2$ non-zero values - see (7.3). Thus, to represent it in memory of a computer one can assign to every subset $A \subseteq N$ a numerical code and represent an imset u by the list of at most $2 \cdot |N|+2$ items where each item is a pair $[code(A), u(A)]$ for a set $A \subseteq N$ with $u(A) \neq 0$. The number of bits needed to represent a standard imset is a polynomial function of $|N|$ which means that memory demands are equivalent to those in case of the use of graphs over N.

As concerns computer representation of data, a data vector t_D (see p. 157) can be represented by the list of $2^{|N|}$ items of the form $[code(A), t_D(A)]$ for $A \subseteq N$. This may appear to be even more effective than the traditional way of representing data in the form of a contingency table because $2^{|N|} \leq |X_N|$. On the other hand, a contingency table is an integral function on X_N while t_D is a real function on $\mathcal{P}(N)$. One can also keep a contingency table in the memory of a computer and compute any value of $t_D(A)$, $A \subseteq N$ each time when this is needed. Thus, the memory demands for representing data are equal to those in case of the use of graphs for representing models. △

Chapter 9

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 the involved concepts. *Directions* (of research) are 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.

9.1 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.

9.1.1 Miscellaneous topics

Multiinformation

There are some open problems related to the concept of multiinformation.

QUESTION 1 Let P and Q are probability measures over N defined on the 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. 29). Has their convex combination $\alpha \cdot P + (1-\alpha)\cdot Q$, $\alpha \in [0,1]$ finite multiinformation as well?

QUESTION 2 Let $\mathcal{K}'(N)$ denote the cone generated by multiinformation functions induced by discrete probability measures over N (see p. 18). Is $\mathcal{K}'(N)$ a rational polyhedral cone?

The answer to Question 2 is positive in case $|N| \leq 3$; but I do not know the answer if $|N| = 4$. The significance of this question consists in the fact that discrete CI models can be characterized properly if the answer is positive.

Observation 9.1 If the answer to Question 2 is positive then there exists a non-empty finite set $\mathbb{S} \subseteq \mathbb{Z}^{\mathcal{P}(N)} \setminus \{0\}$ such that every $s \in \mathbb{S}$ generates an exteme ray of $\mathcal{K}'(N)$ and $\mathcal{K}'(N) = con(\mathbb{S})$. Then the following conditions are equivalent for $\mathcal{M} \subseteq \mathcal{T}(N)$:

(i) $\mathcal{M}$ is the CI model induced by a discrete probability measure over N,

(ii) $\mathcal{M}$ is produced by an element of $\mathcal{K}'(N)$,

(iii) $\mathcal{M}$ has the form $\mathcal{M} = \bigcap_{t \in \mathbb{T}} \mathcal{M}^t$ where $\mathbb{T} \subseteq \mathbb{S}$.

Proof: Because $\mathcal{K}'(N) \subseteq \mathcal{K}_\ell(N)$ and $\mathcal{K}_\ell(N)$ is a pointed cone (Lemma 5.3) $\mathcal{K}'(N)$ is a pointed rational polyhedral cone then. As mentioned in Section 10.8.2 this implies that $\mathcal{K}'(N)$ has finitely many extreme rays and every extreme ray is generated by a non-zero integral vector.

The implication (i) $\Rightarrow$ (ii) follows directly from Consequence 2.2. To evidence (ii) $\Rightarrow$ (iii) suppose $\mathcal{M} = \mathcal{M}^t$ where $t = \sum_{t \in \mathbb{T}} \alpha_s \cdot s$ with $\alpha_s \geq 0$ and put $\mathbb{T} = \{s \in \mathbb{S}; \alpha_s > 0\}$. Using the fact $\mathcal{K}'(N) \subseteq \mathcal{K}_\ell(N)$ and Observation 5.1 then derive $\mathcal{M} = \bigcap_{s \in \mathbb{T}} \mathcal{M}^s$. To evidence (iii) $\Rightarrow$ (i) observe that every $s \in \mathbb{S}$ has the form $\alpha \cdot m_P$ for a discrete probability measure over N (since s generates an extreme ray of $\mathcal{K}'(N)$). Thus, for every $s \in \mathbb{S} \cup \{0\}$, $\mathcal{M}^s$ is a discrete CI model and one can use Lemma 2.9 to derive (i). $\square$

Moreover, it seems that if Question 2 has a positive solution then discrete CI models and implications between them can be characterized in terms of an algebraic relation between certain special imsets over N. Well, what follows is more likely an intuitive plan than the list of verified claims. Roughly said, the plan is to do with the cone $\mathcal{K}'(N)$ an analogue of what was done with the cone $\mathcal{K}_\ell(N)$ in Chapters 5 and 6. However, it may be the case that some of the steps indicated below cannot be made.

The first observation should be that the cone $\mathcal{K}'(N)$ has finitely many faces and each of them is generated by a finite subset $\mathbb{T} \subseteq \mathbb{S}$. The second step is to establish a one-to-one correspondence between discrete CI models and faces of $\mathcal{K}'(N)$: every $\mathcal{M} \subseteq \mathcal{T}(N)$ is assigned the face $\{t \in \mathcal{K}'(N)\,;\, \mathcal{M} \subseteq \mathcal{M}^t\}$ and every $\mathbb{F} \subseteq \mathcal{K}'(N)$ is assigned the model $\bigcap_{t \in \mathbb{T}} \mathcal{M}^t$. The conjecture is that it defines Galois connection in the sense of Section 5.4. The third possible step is to introduce a suitable pointed rational polyhedral cone $\mathcal{K}^*(N)$ which should correspond to the dual cone of $\mathcal{K}'(N)$. The cone $\mathcal{K}^*(N)$ should be an analogue of the cone $con(\mathcal{E}(N))$ in case of $\mathcal{K}'(N)$; perhaps it can also be introduced as the dual cone of the cone $\mathcal{K}'(N) \oplus \mathcal{L}(N)$. The faces of $\mathcal{K}^*(N)$ should correspond to faces of $\mathcal{K}'(N)$ - the respective Galois connection should be given by the incidence relation $\langle m, u \rangle = 0$ for $m \in \mathcal{K}'(N)$ and $u \in \mathcal{K}^*(N)$. Moreover, every face of $\mathcal{K}^*(N)$ should be generated by an element of $\mathcal{K}^*(N) \cap \mathbb{Z}^{\mathcal{P}(N)}$. The fourth step should be to characterize extreme rays of $\mathcal{K}^*(N)$ and choose a non-zero normalized imset from every exteme ray of the cone. These imsets should be analogues of elementary imsets while the imsets from $\mathcal{K}^*(N) \cap \mathbb{Z}^{\mathcal{P}(N)}$ should be analogues of structural imsets. The last step should be to characterize the inclusion of faces of $\mathcal{K}^*(N)$ as an algebraic relation of (generating) elements of $\mathcal{K}^*(N) \cap \mathbb{Z}^{\mathcal{P}(N)}$ - this should be an analogue of the direct characterization of facial implication from Section 6.2.1. The definition of the formal independence model induced by an element of $\mathcal{K}^*(N) \cap \mathbb{Z}^{\mathcal{P}(N)}$ can be obtained as a special case of this characterization - like the definition from Section 4.4.1.

The expected advantage of that approach should be that the obtained universum of special imsets from $\mathcal{K}^*(N) \cap \mathbb{Z}^{\mathcal{P}(N)}$ achieves both the completeness and the faithfulness relative to the discrete distribution framework (see Section 1.1). Moreover, the conjectured algebraic characterization of inclusion of faces should offer an algebraic way of computer implementation. Note that Question 2 was formulated for a discrete distribution framework (see Section 10.10.3), but an analogous question can be raised for any other distribution framework Ψ which satisfies the condition (6.21).

THEME 1 Is there any (direct) formula for the multiinformation function of a non-degenerate CG measure (see p. 61) in terms of their canonical or moment characteristics? Alternatively, is there any (iterative) method of its computing?

Note that owing to Lemma 2.7 an equivalent formulation of Theme 1 can be as follows: to find a formula for the 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. 69). I am more likely sceptical about the existence of a direct formula of this kind.

Formal independence models

Two open problems concern the concept of formal independence model (see Section 2.2.1).

QUESTION 3 It it true that every formal independence model induced by a binary probability measure over N, that is, a measure on $\mathsf{X}_N = \prod_{i\in N}\mathsf{X}_i$ with $|\mathsf{X}_i| = 2$, is a formal independence model induced by a Gaussian measure over N (see Section 2.3.6).

A minor modification of Question 3 is the question of whether every formal independence model induced by a positive binary probability measure over N (see Section 2.2.2) is induced by a non-degenerate Gaussian measure over N. Note that the converse implication does not hold. The reason is that the weak transitivity property (see p. 37) hold for arbitrary conditioning set C in the Gaussian case but it only holds for $C=\emptyset$ in the binary case as the following example shows.

EXAMPLE 9.1 There exists a binary measure P over the set $N=\{a,b,c,d\}$ such that $a \perp\!\!\!\perp b\,|\,\{c\}\ [P]$, $a \perp\!\!\!\perp b\,|\,\{c,d\}\ [P]$ but $a \top\!\!\!\top d\,|\,\{c\}\ [P]$ and $d \top\!\!\!\top b\,|\,\{c\}\ [P]$. Indeed, put $\mathsf{X}_i = \{0,1\}$ for $i\in N$ and assign the probability $\frac{1}{8}$ to every one of the following configuration of values (the order of variables is $a,b,,c,d$): $(0,0,0,0)$, $(0,1,0,0)$, $(1,0,0,1)$, $(1,1,0,1)$, $(0,0,1,0)$, $(1,0,1,0)$, $(0,1,1,1)$ and $(1,1,1,1)$.

DIRECTION 1 Let us consider an abstract system of formal independence models, that is, a mapping which assigns a class $\{\mathcal{M}_\alpha \subseteq \mathcal{T}(N)\,;\, \alpha\in I(N)\}$ of formal independence models over N to every non-empty finite set of variables N. Find out what conditions on the abstract system ensure that there exists an axiomatic characterization of formal independence models in the system by means of (possibly infinite) collection of context consistent inference rules (see Remark 3.5). Try to characterize those abstract systems which have an axiomatic characterization of this kind.

The basic conjecture is that if an abstract system of formal independence models admits an axiomatic characterization then it has to be closed under the operation of restriction (see p. 20) and under the operation which is attributed to a permutation of variables as follows. Every permutation of variables $\pi : N \to N$ induces a transformation of a formal independence model $\mathcal{M}\subseteq\mathcal{T}(N)$ into the model $\{\langle \pi(A),\pi(B)|\pi(C)\rangle\,;\,\langle A,B|C\rangle\in\mathcal{M}\}$ where $\pi(A)=\{\pi(i)\,;\, i\in A\}$ for $A\subseteq N$. Well, it may be the case that these two conditions already characterize abstract systems of formal independence models admitting an axiomatic characterization of the above kind.

Graphs

Further open problems are related to Chapter 3. The following problem, named the "inclusion problem" in [55], can be viewed as an advanced subtask of the equivalence task (see Section 1.1).

THEME 2 Let K, L be acyclic directed graphs over N (see p. 186). Is there any direect graphical characterization of the inclusion $\mathcal{M}_K \subseteq \mathcal{M}_L$ (see Section 3.2)?

Note that suitable graphical characterization of Markov equivalence of acyclic directed graphs is known - see p. 47. By direct characterization of inclusion $\mathcal{M}_K \subseteq \mathcal{M}_L$ is meant a generalization of that characterization of equivalence, that is, a characterization by means of conditions on induced subgraphs K and L. More specifically, I have in mind a collection of conditions on $K, L \in \mathsf{DAGS}(N)$ each of which says something like this: if L has a certain induced subgraph for $T \subseteq N$ then K has a certain induced subgraph for a subset of T. A concrete conjecture of this kind is formulated on p. 35 of [55]. Note that an indirect transformational characterization of the inclusion of DAG models is mentioned in Lemma 8.5 and a characterization in terms of standard imsets is given by Lemma 8.6.

QUESTION 4 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 K, L be (classic) chain graphs over N (p. 186) such that $\mathcal{M}_K = \mathcal{M}_L$ (see p. 50). Does the class of factorizable measures on $(\mathsf{X}_N, \mathcal{X}_N)$ with respect to K (see Remark 3.7 on p. 51) coincide with the class of factorizable measures with respect to L?

Semi-graphoids

The concept of semi-graphoid (see Section 2.2.2) can be viewed as a special concept of discrete mathematics which has already its own significance. Several recent papers devoted to the topic of semi-graphoid inference [130, 31, 85] indicate that it is a challenging topic of research. As the intersection of any class od semi-graphoids is a semi-graphoid the set of disjoint semi-graphoids over N is a (finite) complete lattice. A natural question is what is the relation of this lattice and the lattice of structural independence models (see Section 5.4.2).

QUESTION 5 Is it true that every coatom of the lattice of disjoint semi-graphoids over N is a structural independence model over N?

Note that the condition from Question 5 is equivalent to the condition that the set of coatoms of the lattice of semi-graphoids coincides with the set of coatoms of the lattice of structural models - this can be verified using the fact that every structural model is a semi-graphoid (see Lemma 4.6). An analogous question for atoms of the lattice of semi-graphoids has a positive answer - use Lemma 4.6 and Lemma 4.5 to show that every $\mathcal{M}_v$, $v \in \mathcal{E}(N)$ is both an atom of the lattice of semi-graphoids and a structural model, and Lemma 2.2 to show that every atom of that lattice has this form.

The above conjecture is true if $|N| \leq 4$. The computer program [10] made it possible to find all 37 coatoms of the lattice of semi-graphoids for $|N| = 4$. These semi-graphoids can be shown to be structural models over N with help of the results of the analysis made in [126]. However, I do not know whether the conjecture is true for $|N| = 5$.

Remark 9.1 If $|N| \leq 3$, then the lattice of semi-graphoids and the lattice of structural models coincide. If $|N| = 4$, then they are similar: they have the same unit element, the same null element, the same set of atoms and the same set of coatoms. The basic difference is a follows. While the lattice of structural models is both atomistic and coatomistic (see Theorem 5.3) the lattice of semi-graphoids is only atomistic (use Lemma 2.2 to verify this fact). However, it is not coatomistic as the following example shows. Consider the semi-graphoid generated by the following list of disjoint triplets over $N = \{a, b, c, d\}$:

$$\langle a, b|\{c, d\}\rangle,\ \langle a, b|\{c\}\rangle,\ \langle a, b|\{d\}\rangle,\ \langle c, d|\{a\}\rangle,\ \langle c, d|\{b\}\rangle,\ \langle a, b|\emptyset\rangle,\ \langle c, d|\emptyset\rangle.$$

This semi-graphoid is not the intersenction of coatomistic semi-graphoids over N because each of those is a structural model and, therefore, it satisfies the formal property mentioned in Section 2.2.4 (for details see [128]). Actually, it can be shown that the above mentioned semi-graphoid is a meet-irreducible element of the lattice of semi-graphoids.

Thus, if the answer to Question 5 is yes, then the set of structural independence models can equivalently be introduced as follows: these are the semi-graphoids over N which can be obtained as the intersection of coatoms of the lattice of semi-graphoids provided that the empty intersection is the unit element $\mathcal{T}(N)$. $\triangle$

Structural imsets

There are some open problems related to Chapter 7. The first group of these problems concerns the concept of baricentral imset (see p. 112).

Theme 3 Let G be a chain graph over N (see p. 186 and Section 3.3). Is there any direct formula for the baricentral imset u over N with $\mathcal{M}_u = \mathcal{M}_G$? Can every supermodular function m over N (see p. 78 and 79) be effectively 'translated' into a baricentral imset u over N with $\mathcal{M}_u = \mathcal{M}^m$? Is there any effective criterion which decides whether a given structural imset is a baricentral imset.

Note that a positive solution to the first question can have an impact on the methods of learning graphical models. Chain graph models can be represented by baricentral imsets then - see the thoughts about the universum of structural imsets on p. 137 and the note after Direction 9.

Question 6 Let $\wp$ be a class of facially equivalent structural imsets over N (see p. 98) and $u \in \wp$ be a combinatorial imset which is minimal element of $\wp$ with respect to the strong facial implication $\rightsquigarrow$ (see p. 121). Is u an imset of the least degree in $\wp$?

Note that the converse implication is true - see Observation 7.2.

9.1.2 Classification of skeletal imsets

Basic problem related to the concept of skeletal imset (see Section 5.2) is the following one.

Theme 4 Is there any suitable characterization of skeletal imsets which allows to find the skeleton $\mathcal{K}^{\diamond}_{\ell}(N)$ for any finite non-empty set of variables N? How does $|\mathcal{K}^{\diamond}_{\ell}(N)|$ depend on $|N|$?

Note that the paper [105] offers a characterization of extreme supermodular functions but that result more likely gives a criterion whether a given ℓ-standardized supermodular function is skeletal; more precisely, the result can be utilized for this purpose. However, that criterion does not seem suitable for the purpose of computer implementation. Therefore, the result of [105] does not solve the problem of finding the skeleton for every N. A promising idea of how to tackle the problem is indicated in the rest of this section. A related task is the task to classify coatomic structural models. One can fix a way of standardization of skeletal imsets (see Remark 5.6) as coatomic (= submaximal) structural models are in a 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 [137]). Of course, every permutation of a skeletal imset defines a transformation of respective produced independence model. The basic way of classification is the division of the class of (standardized) skeletal imsets into the classes of permutable equivalence. Every equivalence class represents a *type of a skeletal imset* then. For example, 5 standardized skeletal imsets break into 3 types in case $|N| = 3$, while 37 imsets break into 10 types in case $|N| = 4$ and 117978 imsets break into 1319 types in case $|N| = 5$.

Level equivalence

Nevertheless, perhaps even a 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 independence equivalent element of the ℓ-skeleton, the u-skeleton and the o-skeleton obtained by formulas from Remark 5.6 (p. 87-88). Thus, m, more precisely, the respective class of independence equivalent skeletal imsets, defines a certain equivalence on the class of subsets of N:

$$\forall\, S,T \subseteq N \quad S \sim_m T \;\Leftrightarrow\; [\, m_o(S) = m_o(T),\; m_\ell(S) = m_\ell(T) \text{ and } m_u(S) = m_u(T)\,]\,. \qquad (9.1)$$

The equivalence classes of $\sim_m$ could be interpreted as the areas in which these standardized skeletal imsets have the same values; in other words, they correspond to some *value levels*. In fact, I would like to know whehter the following hypothesis is true.

QUESTION 7 Let $m \in \mathcal{K}_\ell^\diamond(N)$, m' is an independence equivalent skeletal imset over N (see p. 79) and $S,T \subseteq N$ such that $S \sim_m T$. Is it true that $m'(S) = m'(T)$ necessarily?

Two skeletal imsets over N will be called *level equivalent* if they induce the same equivalence on $\mathcal{P}(N)$.

OBSERVATION 9.2 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 the 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. □

REMARK 9.2 Further interesting operation with supermodular functions can be introduced with the help of a special 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 [137]) 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 9.2 below). In particular, m and z are level equivalent in this case. $\triangle$

Nevertheless, the question of whether the above hypotheses holds in general is open.

QUESTION 8 Let $m \in \mathcal{K}_o^\diamond(N)$ and $S, T \subseteq N$ such that $S \sim_m T$. Is it true that $N \setminus S \sim_m N \setminus T$ hold then?

Supertypes

A natural consequence of Observation 9.2 is that the concept of permutable equivalence can be extended to classes of level equivalence. Then every class of this extended permutable equivalence breaks into several classes of level equivalence which break 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 if $|N| = 4$ then every equivalence (9.1) on $\mathcal{P}(N)$, 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 (= value 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 9.2 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 9.1. 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 9.1. 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 the ℓ-standardized version and the u-standardized version is shown.

3. The criterion $|S \cap \{a, b, c, d\}|$ divides $\mathcal{P}(N)$ into 5 levels - see the upper picture of Figure 9.2. 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. The composed criterion $[\,|S\cap\{a,b,c\}|\,,\,|S\cap\{d\}|\,]$ divides $\mathcal{P}(N)$ into 8 levels - see the lower picture of Figure 9.2. 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. The composed criterion $[\,|S\cap\{a,b\}|\,,\,|S\cap\{c,d\}|\,]$ divides $\mathcal{P}(N)$ into 9 levels - see Figure 9.3. 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. $\diamondsuit$

The endeavour described in Section 9.1.2 can be summarized in the following open problem.

Theme 5 Can the classification of supertypes of skeletal imsets by cardinal criteria be extended to the case of general N?

9.2 Operations with structural models

There are various basic operations with structural models - some of them are realized by means of operations with structural imsets and some of them by means of operations with supermodular functions. The overview is given in § 8.2 of [138]. The aim of this section is to recall some of these operations and to formulate relevant open problems.

9.2.1 Reductive operations

The operations of this kind assign a formal independence model over T, $\emptyset\neq T\subseteq N$ to a structural model over N. A typical example of an operation of this kind is the operation of restriction to a set $\emptyset\neq T\subseteq N$ mentioned on p. 20 which assings the model $\mathcal{M}_T\equiv\mathcal{M}\cap\mathcal{T}(T)$ to a formal independence model $\mathcal{M}\subseteq\mathcal{T}(N)$. The basic observation is that the retsriction of a structural model is a structural model.

Observation 9.3 If $\mathcal{M}\in\mathcal{U}(N)$ and $\emptyset\neq T\subseteq N$ then $\mathcal{M}_T\in\mathcal{U}(T)$.

Proof: Given $m\in\mathbb{R}^{\mathcal{P}(N)}$ consider its restriction m_T to $\mathcal{P}(T)$. By Observation 5.1 $m\in\mathcal{K}(N)$ implies $m_T\in\mathcal{K}(T)$. Moreover, $\langle m,u_{\langle A,B|C\rangle}\rangle=\langle m_T,u_{\langle A,B|C\rangle}\rangle$ for every $\langle A,B|C\rangle\in\mathcal{T}(T)$. Thus, the model produced by m_T coincides with the restriction of the model produced by m to T. Hence, Observation 9.3 follows from (5.15). $\square$

Theme 6 Let u be a baricentral imset over N and $\emptyset\neq T\subseteq N$. Is there any direct formula for the baricentral imset inducing $(\mathcal{M}_u)_T$ is terms of u?

One can consider an alternative version of the above problem: provided $u\in\mathcal{S}(N)$ is there any algebraic formula which always gives a structural imset over T inducingthe restriction of $\mathcal{M}_u$ to T?

Remark 9.3 In § 8.2.1 of [138] two other reductive operations with structural models are defined and shown to yield structural models. One of them is an operation which corresponds to the concept of *minor* of a semi-graphoid introduced in [83]. $\triangle$

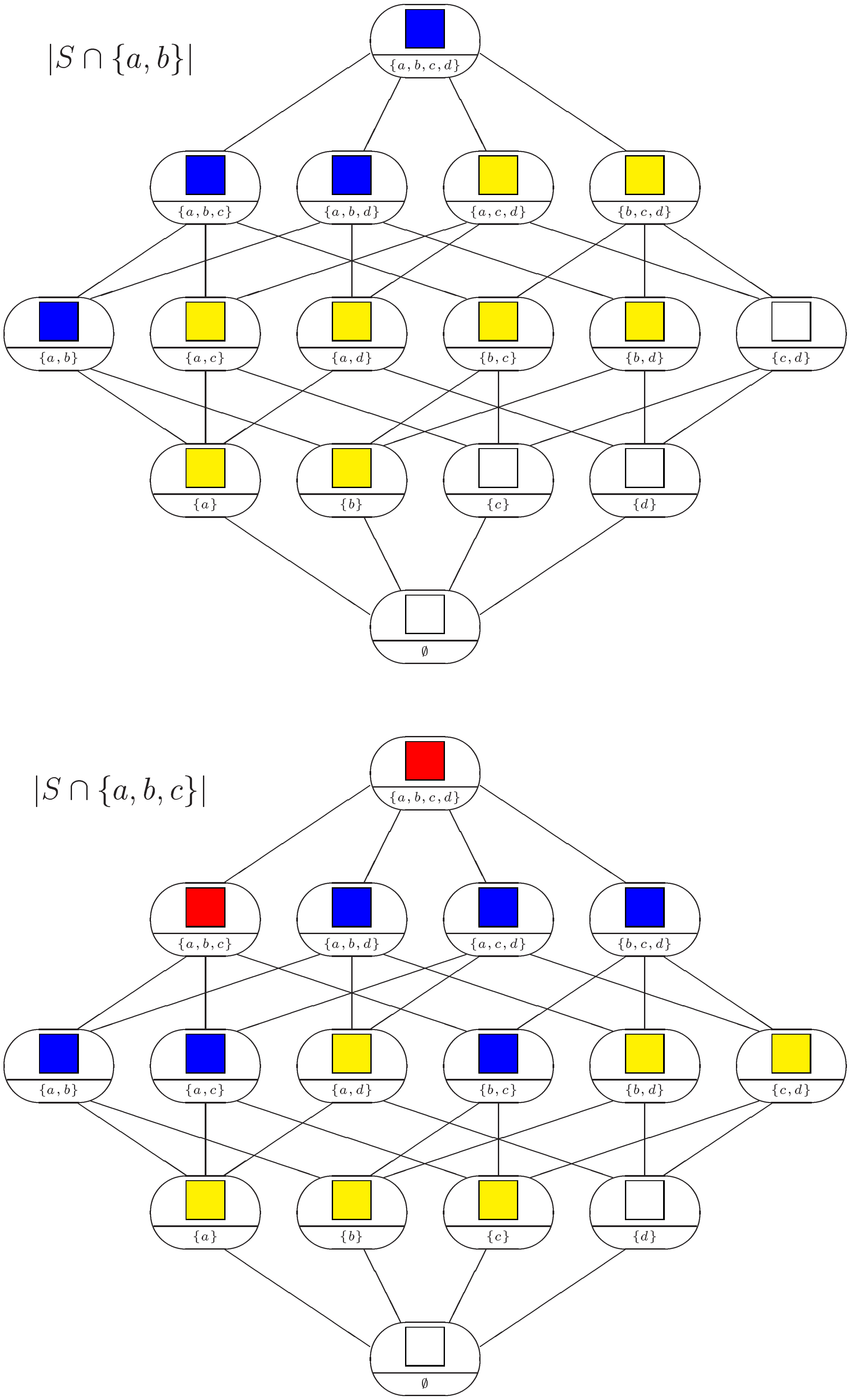
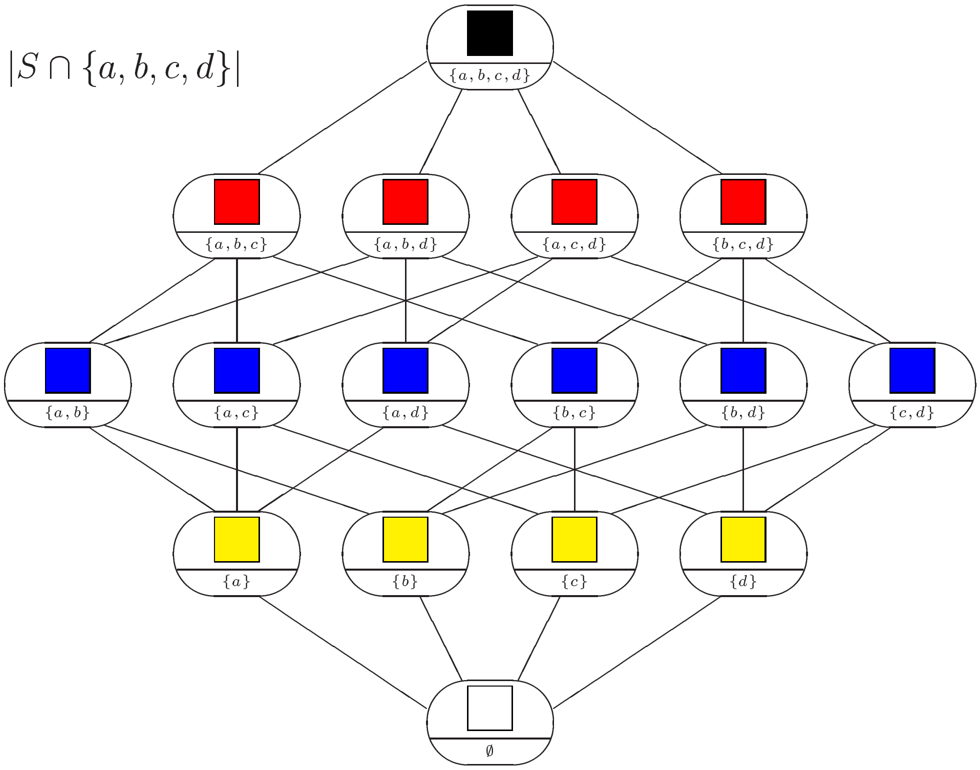


Figure 9.1: Cardinality criteria and respective levels for $N = \{a, b, c, d\}$.



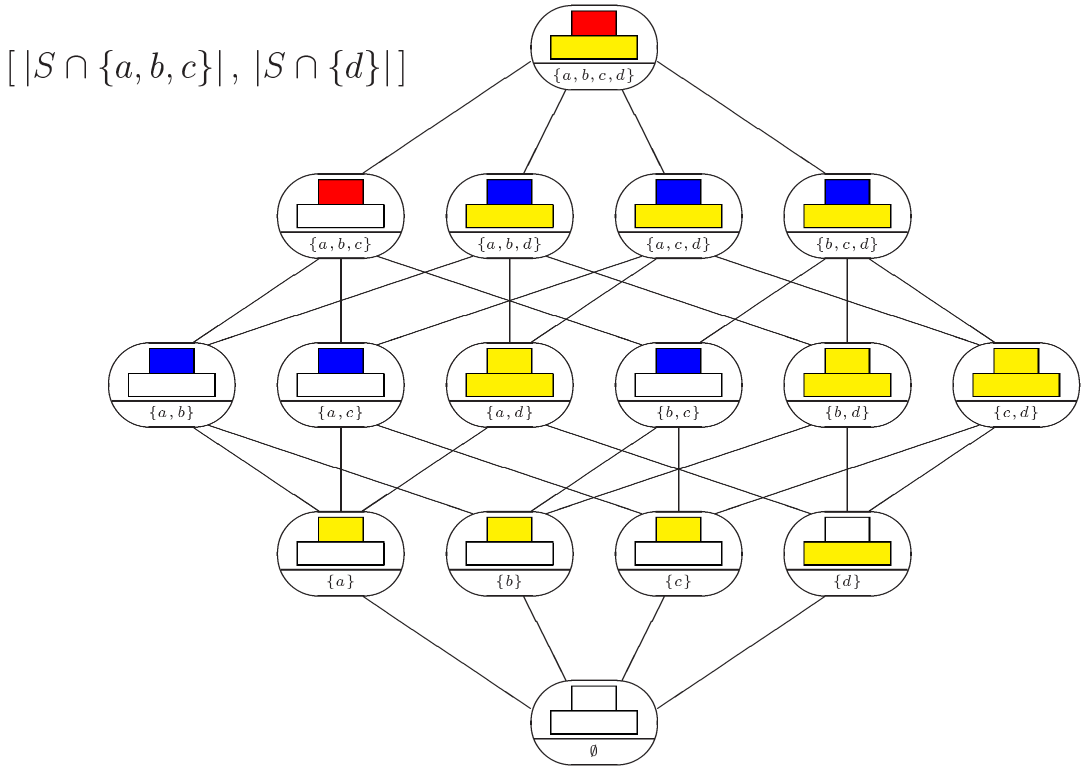


Figure 9.2: Further cardinality criteria and respective levels for $N = \{a, b, c, d\}$.

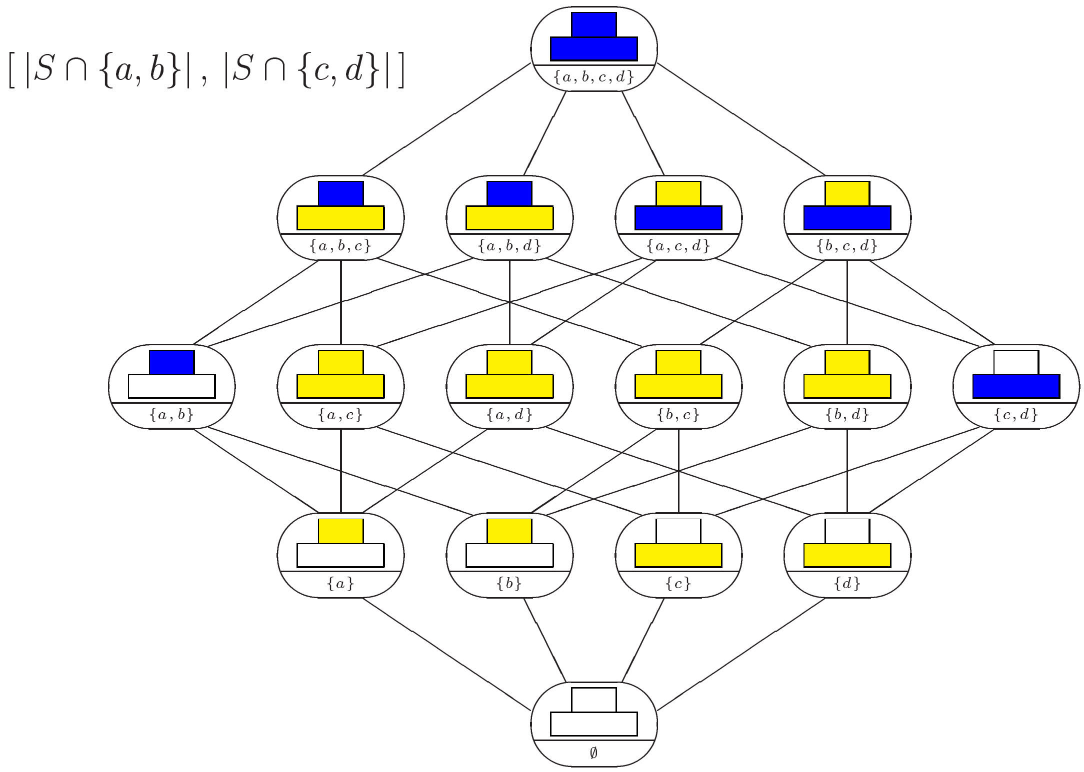


Figure 9.3: The last cardinality criterion and respective levels for $N = \{a, b, c, d\}$.

9.2.2 Expansive operations

The operations of this kind assign a formal independence model over N to a formal independence model over T, $\emptyset \neq T \subseteq N$. In § 8.2.2 of [138] four examples of such operations are given. It is also shown there that those four operations ascribe structural models to structural models. One of the operations is the *ascetic extension* $as(\mathcal{M}, N) \equiv \mathcal{T}_{\emptyset}(N) \cup \mathcal{M}$ of $\mathcal{M} \subseteq \mathcal{T}(T)$ with $\emptyset \neq T \subseteq N$. It is a natural counterpart of the concept of restriction mentioned in Section 9.2.1. Actually, it is the least extension $\mathcal{M}'$ of $\mathcal{M} \subseteq \mathcal{T}(T)$ such that $\mathcal{M}'_T = \mathcal{M}$.

Observation 9.4 If $\mathcal{M} \in \mathcal{U}(T)$, $\emptyset \neq T \subseteq N$ then $as(\mathcal{M}, N) \in \mathcal{U}(N)$ and $as(\mathcal{M}, N)_T = \mathcal{M}$.

Proof: Let $\mathcal{M}$ is induced by $v \in \mathcal{S}(T)$. Consider the zero extension $u \in \mathcal{S}(N)$ of v given by

$$u(S) = \begin{cases} v(S) & \text{if } S \subseteq T\,, \\ 0 & \text{otherwise}\,, \end{cases} \qquad \text{for } S \subseteq N\,,$$

and observe that $as(\mathcal{M}, N)$ is induced by u. □

9.2.3 Accumulative operations

The operations of this kind ascribe a formal independence model to a pair of formal independence models. More specifically, given $U, V \subseteq N$ such that $UV = N$, an accumulative operation should ascribe a model $\mathcal{M}^3$ over N to a pair $\mathcal{M}^1 \subseteq \mathcal{T}(U)$, $\mathcal{M}^2 \subseteq \mathcal{T}(V)$ which is consonant, that is, $\mathcal{M}^1_{U\cap V} = \mathcal{M}^2_{U\cap V}$. The intended interpretationis that the knowledge represented by $\mathcal{M}^1$ and $\mathcal{M}^2$ is gathered in $\mathcal{M}^3$. Note that the motivation source for the idea to introduce some accumulative operations with structural independence models is an abstract concept of *conditional product* from [30] and the intention to utilize this operation for the purpose of decomposition of a structural model (see Section 9.2.4). A simple example of an accumulative operation with structural imsets could be the following concept. 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 *weak 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

$$\mathcal{M}^1 \otimes \mathcal{M}^2 = cl_{\mathcal{U}(ABC)}(as(\mathcal{M}^1, ABC) \cup as(\mathcal{M}^2, ABC) \cup \{\langle A, B|C\rangle\})\,. \qquad (9.2)$$

In words, both $\mathcal{M}^1$ and $\mathcal{M}^2$ are embedded into $\mathcal{U}(ABC)$ by the ascetic extension, then the triplet $\langle A, B|C\rangle$ is added and the structural closure operation (see p. 121) is applied. It follows directly from the definition that $\mathcal{M}^1 \otimes \mathcal{M}^2 \in \mathcal{U}(ABC)$.

Theme 7 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 (9.2) be restricted suitably so that the axioms of conditional product from [30] are fulfilled for it? Is there any algebraic formula for the baricentral imset of $\mathcal{M}^1 \otimes \mathcal{M}^2$ on basis of baricentral imsets inducing $\mathcal{M}^1$ and $\mathcal{M}^2$?

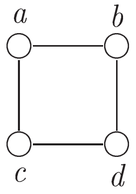


Figure 9.4: A prime graph over $N = \{a, b, c, d\}$ which is not complete.

9.2.4 Decomposition of structural models

The final aim of the considerations about operations with structural models is to develop an effective method of 'local' representation of structural independence models. More exactly, one would like to represent a structural model over a large set of variables in memory of a computer by means of a list of representatives of structural models over small sets of variables. To this end one needs a suitable concept of decomposition of a structural model. The first attempt at the definition of this concept is the following concept which is strongly related to the weak decomposition introduced in the previous section. Let $\mathcal{M}$ be a structural model over N and $U, V \subseteq N$ such that $UV = N$. One says that the pair $(\mathcal{M}_U, \mathcal{M}_V)$ is a *weak decomposition of* $\mathcal{M}$ if $\mathcal{M} = \mathcal{M}_U \otimes \mathcal{M}_V$. This weak decomposition is *proper* if $U \setminus V \neq \emptyset \neq V \setminus U$. Note that $(\mathcal{M}_U)_{U \cap V} = (\mathcal{M}_V)_{U \cap V}$ which means that the weak composition $\mathcal{M}_U \otimes \mathcal{M}_V$ is always defined. Clearly, a necessary condition for the existence of a weak decomposition $(\mathcal{M}_U, \mathcal{M}_V)$ is the CI statement $U \setminus V \perp\!\!\!\perp V \setminus U \,|\, U \cap V \; [\mathcal{M}]$ (see p. 20).

Motivation thoughts about the concept of decomposition

The motive of open problems formulated below is the concept of decomposition of UG model from [67, 24]. Let us say that an undirected graph G over N has a *proper decomposition* if there exists $\langle A, B|C \rangle \in \mathcal{T}(N) \setminus \mathcal{T}_{\emptyset}(N)$ such that $A \perp\!\!\!\perp B \,|\, C \; [G]$ (see p. 43) and G_C is a complete graph. The graphs G_{AC} and G_{BC} form a (proper) decomposition of G then.

The above concept allows one to classify subsets of N. A set $T \subseteq N$ is called a *prime set* relative to G if G_T has no proper decomposition. An undirected graph G over N is called a *prime graph* if the set N is a prime set relative to G. It is evident that every complete set in an undirected graph H is a prime set relative to H but the converse implication does not hold as the following example shows.

Example 9.3 There exists a prime graph over $N = \{a, b, c, d\}$ which is not a complete graph. Let us consider the graph shown in Figure 9.4. $\diamondsuit$

The above example also shows that a subset of a prime set relative to an undirected graph need not be a prime set. Indeed, the set $\{a, b, c\}$ is not a prime set in the considered graph. The maximal set (with respect to inclusion) which are prime set relative to an undirected graph G over N will be called *prime components* of G. The collection of prime components will be denoted by $\mathcal{P}_{pri}^{\max}(G)$.

It is clear that one can try to apply the operation of proper decomposition to any undirected graph G over N and if one succeeds then one can try to apply it to one of the induced subgraphs which were obtained by the operation. Thus, one can successively decompose G into a system of prime graphs. Nevertheless, the result of the successive decomposition need not be unique. Indeed, let us consider the graph G over $N = \{a, b, c\}$ which only has the line $b - c$. The first option is that one decomposes G into $G_{\{a,b\}}$ and $G_{\{b,c\}}$ using $a \perp\!\!\!\perp c \,|\, b\ [G]$ and then decomposes $G_{\{a,b\}}$ into $G_{\{a\}}$ and $G_{\{b\}}$ because of $a \perp\!\!\!\perp b \,|\, \emptyset\ [G]$. The resulting list of prime graphs is $G_{\{a\}}$, $G_{\{b\}}$ and $G_{\{b,c\}}$ then. The second option is to decompose G into $G_{\{a\}}$ and $G_{\{b,c\}}$ using $a \perp\!\!\!\perp \{b, c\} \,|\, \emptyset\ [G]$. The graph $G_{\{b\}}$ in the first list of prime graphs seems to be superfluous because $\{b\}$ is not a prime component of G. Thus, one is interested in a successive decomposition of an undirected graph G into subgraphs which leads to prime components.

In my view, the result from [69, 80] can be paraphrased as follows: for every undirected graph G over N, the class of set $\mathcal{P}_{pri}^{\max}(G)$ satisfies the running intersection property (see p. 52) and, therefore, it is a system of cliques of a triangulated undirected graph H over N which can be called the *canonical triangulization of* G (evidently G is a subgraph of H). Then one can say that the collection of induced subgraphs $\{G_T; T \in \mathcal{P}_{pri}^{\max}(G)$ together with the triangulated graph H defines the *canonical decomposition of* G into prime components.

The concept of (proper) decomposition of an undirected graph G into its induced subgraphs G_{AC} and G_{BC} has a special significance. The existence of the decomposition allows one to simplify the task to find the maximum likelihood estimate (see p. 203) in the statistical model $\mathbb{M}_G$ (= the class of Markovian probability measures with respect to G) to the task to find the maximum likelihood estimates in the statistical models $\mathbb{M}_{G_{AC}}$ and $\mathbb{M}_{G_{BC}}$ - see § 4.3 in [67]. The desired 'global' estimate in $\mathbb{M}_G$ can be obtained from 'local' estimates in $\mathbb{M}_{G_T}$, $T \in \mathcal{P}_{pri}^{\max}(G)$ by means of the product formula (3.4) which corresponds to the triangulated graph H. The point is as follows: if prime components of G are sets of small cardinality then the task of computing the maximum likelihood estimate is manageable unlike the task of finding the maximum likelihood estimate in $\mathbb{M}_G$ which involves too many variables.

The same idea is utilized in the paper [80] where the existence of a canonic decomposition of an undirected graph is utilized to tranform the task to compute the maximum entropy extension of a collection of consonant discrete probability measures for a large set of variables to the task to compute several maximum entropy extensions for small sets of variables. Moreover, an analogous idea is behind the method of *local computation* [24] which is a classic method used to overcome computational problems in case of DAG models. The essential idea of this method is that the original representation of a discrete probability measure over N in the form of a recursive factorization (8.1) is transformed by a series of graphical procedures into a representation in the form of potentials $\{g_T; T \in \mathcal{C}\}$ where $\mathcal{C}$ is the collection of cliques of a triangulated graph over N. These potentials could be the marginals of the original probability measure P which can be expressed in terms of its marginals for $T \in \mathcal{C}$ by means of the above mentioned product formula (3.4).

Remark 9.4 One of equivalent definitions of a triangulated graph is as follows (see [80]). An undirected graph is triangulated iff its prime components are complete sets in the graph. In other words, a graph is triangulated iff complete graphs can be obtained by successive application of the operation of proper decomposition (see Proposition 2.5 in [67]). That's why some authors name triangulated graphs decomposable. In my view, this terminology is slightly misleading: the adjective decomposable usually means that a graph can be decomposed which in the considered situation means that it is not a prime graph. However, the above terminology is generally accepted. $\triangle$

The concept of canonical decomposition of an undirected graph G can be interpreted as follows. The graph G can be viewed as a mathematical object which somehow represents the knowledge about the UG model $\mathcal{M}_G$, respectively the knowledge about the statistical model $\mathbb{M}_G$. The decomposition can be viewed as a way of parsimonious representation of the knowledge about the 'big' model $\mathcal{M}_G$ in the form of a

- a system of knowledge representatives of small 'local' models $\mathcal{M}_{G_T}$, and
- a way of 'computing' of a knowledge representative of the 'big' global model on basis of 'local' representatives.

The way of 'computing' of the global knowledge is highly effective - it is nothing but the application of the method of local computation. The above mentioned interpretation leads to a natural conjecture that undirected graph can possibly be replaced by other objects of discrete mathematics which describe the models of CI structure. This step should enlarge the applicability of the method of local computation to more general classes of models of CI structure. I guess that the concept of weak decomposition defined above should have some connection to the hypothetical concept of decomposition of a structural model. This motivated to the following open problem.

Theme 8 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 weak decomposition of $\mathcal{M}_G$? Is there an analogue of the concept of prime graph within the framework of (classic) chain graph models?

However, the goal of the motivation thoughts above is the following open problem.

Direction 2 Is there any concept of decomposition of a structural model which generalizes the concept of decomposition of an UG model such that an analogue of the result about the existence of a canonical decomposition holds? Try to find sufficient conditions for the existence of decomposition of this type which can be verified by statistical tests of CI or on basis of the knowledge provided by human experts.

If there is an analogy of the result on canonical decomposition of an UG model, then every structural model over N should have canonical decomposition into a decomposable model $\mathcal{M}_H$ over N and a series of structural models $\mathcal{M}_T$ over $T \in \mathcal{C}$ where $\mathcal{C}$ is the class of cliques of H. This should make it possible to develop an analogy of the method of local computation which should be applicable to some structural models, namely those for which the sets $T \in \mathcal{C}$ in the hypothetical decomposition of the structural model have small cardinality. The next step should be an analogy of Shenoy's pictorial method of valuation networks [113] for local representation of structural imsets and Markovian measures in memory of a computer.

9.3 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 9 Is every structural imset u over N already a combinatorial imset over N (see p. 66)?

If the answer to Question 9 is negative then the following two problems may become a topic of immediate interest.

Theme 9 Given a finite non-empty set of variables N, find the least finite class $\mathcal{H}(N)$ of structural imsets 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 the minimal integral Hilbert basis of $con(\mathcal{E}(N))$ follows from Theorem 16.4 in [109]. One has $\mathcal{E}(N) = \mathcal{H}(N)$ iff $\mathcal{S}(N) = \mathcal{C}(N)$.

Theme 10 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, that is,

$$\forall\, u \in \mathbb{Z}^{\mathcal{P}(N)} \quad u \in \mathcal{S}(N) \;\Leftrightarrow\; n_* \cdot u \in \mathcal{C}(N).$$

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 the existence of n_* and n_{**} follows from Lemma 6.3 and $n_{**} \leq n_*$. At this stage, 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 10 Let $\mathcal{K}_\ell^\diamond(N)$ be the ℓ-skeleton over N (see p. 83) and $\mathcal{E}(N)$ the class of elementary imsets over N (p. 64). Is the condition

$$\min \{\, \langle m, u \rangle\,;\; u \in \mathcal{E}(N) \;\; \langle m, u \rangle \neq 0 \,\} = 1$$

fulfilled for every $m \in \mathcal{K}_\ell^\diamond(N)$?

Note that the condition from Question 10 implies that $gra(N) = gra_*(N)$ (see p. 106). The following problem becomes relevant if both Question 9 and Question 10 have negative answers.

Theme 11 What is the dependence of the value of the least $l \in \mathbb{N}$ satisfying the condition

$$\forall\, u \in \mathcal{S}(N) \;\; \forall\, v \in \mathcal{E}(N) \quad u \rightharpoonup v \;\Leftrightarrow\; l \cdot u - v \in \mathcal{S}(N) \tag{9.3}$$

(see Section 6.3.2) on $|N|$? Can one determine $gra(N)$ directly without finding the skeleton, that is, without solving Theme 4?

Recall that if either Question 9 or Question 10 has a positive answer then the least $l \in \mathbb{N}$ satisfying (9.3) is $gra(N)$. The following open problem is a combination of Question 9 and the preceding problem.

THEME 12 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, then find out, for a given $u \in \mathcal{S}(N)$, what is the structure of the set $k \in \mathbb{N}$ satisfying

$$\forall\, v \in \mathcal{E}(N) \quad u \rightharpoonup v \;\Leftrightarrow\; k \cdot u - v \in \mathcal{C}(N)\,.$$

Is there a structural imset $u \in \mathcal{S}(N)$ such that the condition (6.4) from Remark 6.3 is not fulfilled?

The last open problem of this section also concerns facial implication.

DIRECTION 3 Is there any method of testing facial implication which combines direct and skeletal criteria (see Lemma 6.1 on p. 100 and Lemma 6.2 on p. 103) 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. The 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, the plenty of 'transient' imsets obtained during the '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 testing facial implication. That can save superfluous steps of the recursive 'decomposition' procedure. The observation that a 'transient' imset v is not combinatorial could be made on the 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. 41) or m_l for $l = 0, \ldots, |N| - 2$ (p. 64). The point is that one need not to have the whole skeleton at disposal to do that. In fact, Remark 6.8 is based just on the observations of this type.

9.4 Interpretation and learning tasks

In this section, open problems loosely motivated by 'practical' questions of interpretation and learning from Section 1.1 are gathered.

9.4.1 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 11 Let G be an acyclic directed graph over N (p. 186). Is it true that the standard imset for G (see p. 114) is the only imset from the class of combinatorial imsets inducing $\mathcal{M}_G$ which is simultaneously an imset of the least degree (p. 120) and an imset with the least lower class (p. 124)?

In my view, the concept of standard imset for an acyclic directed graph emphasizes some interpretable aspects of the respective DAG model. Thus, a natural question is whether the concept of standard imset for a DAG model can be generalized.

Direction 4 Is there any consistent principle of a unique choice of representatives of classes of facial equivalence (see p. 98) such that, for every acyclic directed graph G over N, the standard imset u_G is chosen from the class $\wp = \{u \in \mathcal{S}(N)\,;\, \mathcal{M}_u = \mathcal{M}_G\}$?

Further open problem is motivated by the ideas from Section 6.4.

Direction 5 Look for the necessary conditions for the facial equivalence of structural imsets which are formulated in terms of invariants of facial equivalence and which are easy to verify and offer a 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.

The desired complete set of interpretable invariants could then become a basis of an alternative way of description of structural models which should be suitable from the point of view of interpretation.

9.4.2 Tasks concerning the distribution frameworks

Below mentioned open problems concern more or less the concept of distribution framework (see Section 10.10.3 and Section 6.5).

Theme 13 Let Ψ be a class of probability measures over N satisfying the conditions (6.20) and (6.21) from Section 6.5. Let $\Psi(u)$ denote the class of Markovian measures with respect to $u \in \mathcal{S}(N)$ given by (6.1) on p. 98 and $\mathcal{S}_\Psi(N)$ the class of Ψ-representable structural imsets over N (p. 110). Is the condition

$$\forall\, u, v \in \mathcal{S}_\Psi(N) \quad u \rightharpoonup v \;\Leftrightarrow\; \Psi(u) \subseteq \Psi(v)$$

fulfilled then? More exactly, what additional assumptions on Ψ are needed to ensure its validity?

Question 12 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. 123). 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. 124).

Theme 14 Let Ψ_1, Ψ_2 be classes of probability measures satisfying (6.20) 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, what is the answer if the class of discrete measures (p. 18) is in place of Ψ_1 and the class of non-degenerate Gaussian measures (p. 34) is in place of Ψ_2?

The last two open problems are related to mathematical statistics. The first one is the 'parametrization problem'.

DIRECTION 6 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 can be defined.

Note that one is interested in parametrization by means of 'independent' parameters, that is, parametrizations in which elements of $\Psi(u)$ correspond to parameters belonging to a polyhedron in $\mathbb{R}^n$ for some $n \in \mathbb{N}$ - see, for example, the parametrization (8.2) of $\mathbb{M}_G$ for $G \in \mathsf{DAGS}(N)$. Another typical example is the parametrization of non-degenerate Gaussian measures which are Markovian with respect to an acyclic directed graph [8, 104].

DIRECTION 7 Can an informal concept of distribution framework (see Section 10.10.3) be formalized and defined in an axiomatic way? Try to clarify what are the operations under which a general class of probability measures should be closed to constitute what is meant by a distribution framework.

9.4.3 Learning tasks

The open problems gathered in this section are motivated by Chapter 8. The first two of them concern the concept of regular criteria from Section 8.2.4. The following problem was indicated in Remark 8.9.

THEME 15 Clarify what assumptions on the priors π_G, $G \in \mathsf{DAGS}(N)$ (see Remark 8.9) ensure that the respective LML criterion for learning DAG models is strongly regular. Derive the respective formula for the data vector relative to the LML criterion (see p. 157) in terms of hyperparameters of Dirichlet measures.

The conjectured formula for the data vector in terms of hyperparameters should be an analogue of the formula (8.38) and could be a basis for comparison of different criteria for learning DAG models. The next open problem is motivated by the observations made in Section 8.4.

DIRECTION 8 Can the fact that a regular criterion for learning DAG models is linear derived in Lemma 8.7 be utilized in an alternative method of the maximization of the value of the criterion by means of procedures of integer programming?

Well, one of typical tasks of integer programming is to find the maximum of a linear function on the set of integral vectors from a certain bounded polyhedron. Perhaps some of the algorithms developed in that area can be successfully applied to a similar problem mentioned in Section 8.4.2. The other motive of Direction 8 is that the segments in the Hasse diagram of a DAG model correspond to elementary imsets (see Remark 8.11) and the number of elementary imsets is limited. Perhaps this obsevation can be utilized somehow.

The next open problem concerns the local search method (see p. 138) and the idea to apply the method of structural imsets in this area (see Remark 8.12). It can also be viewed as an extension of the problem of characterization of inclusion neighbourhood (see Section 8.3 and Theme 2).

Theme 16 Let u_G be a standard imset for an acyclic directed graph G over N (see p. 114). Is it possible to characterize inclusion neighbours of $\mathcal{M}_G$ (see p. 151) directly in terms of u_G? In other words, is there any algebraic criterion to find on basis of u_G all differential imsets (see p. 154) which correspond to the moves from $\mathcal{M}_G$ to its lower neighbours and to its upper neighbours in the sense of inclusion neighbourhood?

Motivation thoughts about learning general CI structures

The last open problem concerns a particular phenomenon mentioned in Remarks 8.11 and 8.14, namely the fact that the moves between neighbouring DAG models have CI interpretation if one uses a regular quality criterion for learning DAG models. This is a very pleasant property because it is advantageous both from the point of view of computer implementation of the local search method (see p. 138) and from the point of view of interpretation.

The consequence of the above fact about the universum of DAG models is that the same statement is true for the universum of decomposable models which is embedded in the universum of DAG models (see Figure 3.6). Note that this statement is valid irrespectively which regular criterion is taken into consideration. Thus, the local search method combined with CI interpretation of moves can also be successfully applied within the universum of decomposable (UG) models if classic quality criteria for learning decomposable models are used.

A natural question is whether these classic criteria can be extended to a wider universum of models of CI structure in such a way that the interpretation of moves to neighbouring models is kept. Unfortunately, the answer is negative. The following example shows that the classic criteria cannot be extended from the universum of decomposable models to the universum of UG models in that way.

Example 9.4 The concept of standard imset for a triangulated undirected graph (see Section 7.2.2) cannot be extended to general undirected graphs in such a way that the difference between imsets representing models which are in inclusion is a structural imset. In particular, the consequence is that usual quality criteria like the MLL criterion, the AIC criterion and the BIC criterion, which have the form (8.31), cannot be extended to the universum of UG models in such a way that CI interpretation of moves between neighbouring models in the sense of Consequence 8.5 is possible.

To show this consider the universum of UG models over $N = \{a, b, c, d\}$ and undirected graphs K, L, G, H shown in Figure 9.5. Observe that $\mathcal{M}_K \subseteq \mathcal{M}_H, \mathcal{M}_G \subseteq \mathcal{M}_L$, and that K, L and H are triangulated. Suppose that there exist $u \in \mathcal{S}(N)$ such that $u_L - u \in \mathcal{S}(N)$, $u - u_K \in \mathcal{S}(N)$ and u induces $\mathcal{M}_G$. It suffices to show that these assumptions lead to contradiction. Indeed, by Observation 6.2 derive that $v \equiv u - u_K \in \mathcal{C}(N)$ and $w \equiv u_L - u \in \mathcal{C}(N)$. Thus, $u \equiv u_{\langle a,d|c\rangle} + u_{\langle a,b|d\rangle} = u_L - u_K = v + w$. The next step is to show that one of the following four options occurs: $v = 0$, $v = u$, $v = u_{\langle a,b|d\rangle}$ and $v = u_{\langle a,d|c\rangle}$.
To this end observe that the level-degrees of u (see p. 66) are as follows: $deg(u, 0) = deg(u, 2) = 0$ and $deg(u, 1) = 2$. Hence, $deg(v, 0) = deg(w, 0) = 0$ and $deg(v, 2) = deg(w, 2) = 0$ and $0 \leq deg(v, 1) \leq 2$. If $deg(v, 1) = 0$ then $deg(v) = 0$ which means $v = 0$. If $deg(v, 1) = 2$ then $deg(w, 1) = 0$ implies $deg(w) = 0$ which means $w = 0$ and $v = u$. If $deg(v, 1) = 1$ then $deg(w, 1) = 1$ and one can conclude $v, w \in \mathcal{E}_1(N)$. Since $1 = \langle m^{abd\uparrow}, u\rangle = \langle m^{abd\uparrow}, v\rangle + \langle m^{abd\uparrow}, w\rangle$

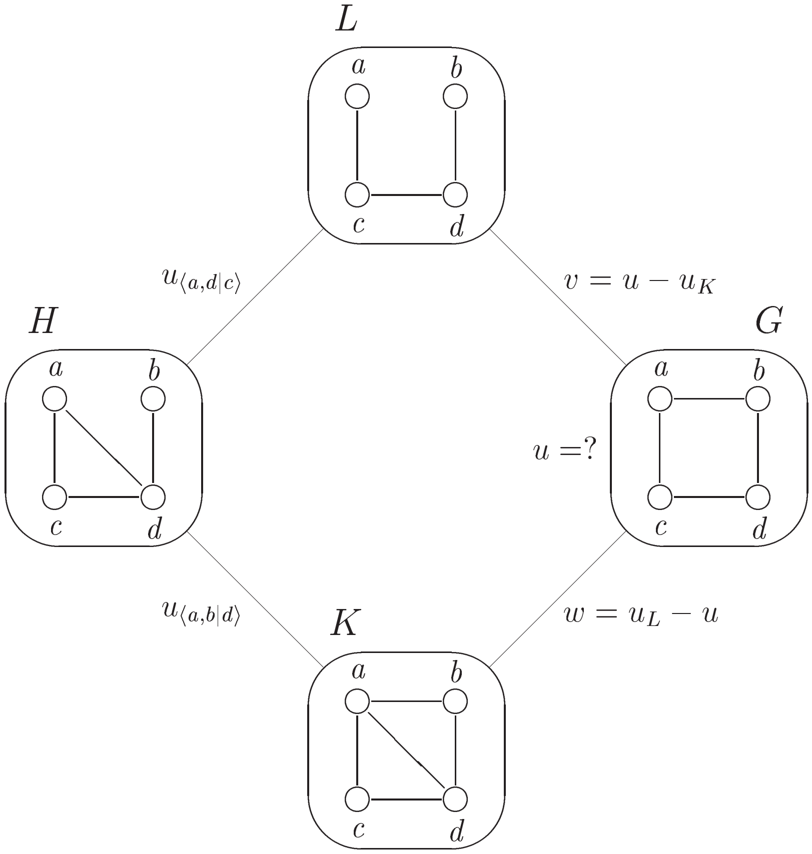


Figure 9.5: Undirected graphs from Example 9.4

by Observation 5.1 $\langle m^{abd\uparrow}, v\rangle \in \{0,1\}$. If $\langle m^{abd\uparrow}, v\rangle = 1$ then observe that the only imsets from $\mathcal{E}_1(N)$ which satisfy this equality are $u_{\langle a,b|d\rangle}$, $u_{\langle a,d|b\rangle}$ and $u_{\langle b,d|a\rangle}$. However, $\langle m^{a\downarrow}, u\rangle = 0 = \langle m^{b\downarrow}, u\rangle$ together with Observation 5.1 implies $\langle m^{a\downarrow}, v\rangle = 0 = \langle m^{b\downarrow}, v\rangle$ which means that one has necessarily $v = u_{\langle a,b|d\rangle}$. If $\langle m^{abd\uparrow}, v\rangle = 0$ then $\langle m^{abd\uparrow}, w\rangle = 1$ and the same consideration gives $w = u_{\langle a,b|d\rangle}$ which means $v = u_{\langle a,d|c\rangle}$.
Now, the cases $v = 0$, $v = u$ and $v = u_{\langle a,b|d\rangle}$ are excluded because G differs from K, L and H. The only remaining option is $v = u_{\langle a,d|c\rangle}$. However, the imset $u + v = u_{\langle b,c|ad\rangle} + u_{\langle a,d|c\rangle}$ does not induce $\mathcal{M}_G$ because $\langle a, d|\{b,c\}\rangle \in \mathcal{M}_G \setminus \mathcal{M}_{u+v}$. To evidence this observe that, for every $k \in \mathbb{N}$, $\langle m^{abd\uparrow}, k \cdot (u+v) - u_{\langle a,d|\{b,c\}\rangle}\rangle = -1$ which means by Observation 5.1 that $k \cdot (u+v) - u_{\langle a,d|\{b,c\}\rangle} \notin \mathcal{S}(N)$. $\diamondsuit$

The preceding consideration and counterexample therefore lead to the following open problem.

Direction 9 Is there any other way of derivation of quality criteria for learning models of CI structure such that the moves between neighbouring models (in the sense of inclusion neighbourhood) have CI interpretation? Propose a suitable method of learning structural models on basis of data.

One of possible methods of derivation of quality criteria which admits CI interpretation of 'moves' is as follows. One chooses a suitable representative of every structural model over N such that the difference between representatives of neighbouring models is always a structural imset. An example of such a choice is the choice of the baricentral imset (see Observation 7.1). One can also choose a special way of representation of data and penalization of the complexity of a model by giving a formula for a data vector, which should be a real vector whose components correspond to subsets of N. The criterion can be then introduced as the scalar product of the data vector and the imsetal representation of the model (c.f., Section 8.4.2).

An alternative methodological approach to learning statistical models of CI structure could be as follows. One can introduce a 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 from the respective statistical model of CI structure, what is the set of Markovian measures with respect to a prospective structural imset.

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, in particular in computer science and statistics, may not be familiar with all basic facts. Thus, to avoid misunderstanding and to facilitate reading I decided to recall these concepts here. The aim is 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) and $A, B \subseteq N$ the juxtaposition AB will be used as a shorthand for $A \cup B$. A class $\mathcal{D} \subseteq \mathcal{P}(N)$ is called *ascending* if it is closed under supersets, that is,

$$\forall\, S,T \subseteq N \quad S \in \mathcal{D},\ S \subseteq T \ \Rightarrow\ T \in \mathcal{D}\,.$$

Given $\mathcal{D} \subseteq \mathcal{P}(N)$, the *induced ascending class*, denoted by $\mathcal{D}^{\uparrow}$, is the least ascending class containing $\mathcal{D}$,, that is,

$$\mathcal{D}^{\uparrow} = \{\, T \subseteq N;\ \exists\, S \in \mathcal{D} \quad S \subseteq T\}.$$

Analogously, a class $\mathcal{D} \subseteq \mathcal{P}(N)$ is called *descending* if it is closed under subsets, that is,

$$\forall\, S,T \subseteq N \quad S \in \mathcal{D},\ T \subseteq S \quad \Rightarrow\ T \in \mathcal{D}\,,$$

and given $\mathcal{D} \subseteq \mathcal{P}(N)$ the *induced descending class* $\mathcal{D}^{\downarrow}$ consists of subsets of sets in $\mathcal{D}$,, that is,

$$\mathcal{D}^{\downarrow} = \{\, T \subseteq N;\ \exists\, S \in \mathcal{D} \quad T \subseteq S\}.$$

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) transitive: $\forall\, x, y, z \in L \quad x \preceq y,\ y \preceq z \ \Rightarrow\ x \preceq z\,,$

(iii) antisymmetric: $\forall\, x, y \in L \quad x \preceq y,\ y \preceq x \ \Rightarrow\ x = y.$

The phrase *total ordering* is used if, moreover, $\forall\, x, y \in L$ either $x \preceq y$ or $y \preceq x$. By *quasiordering* is meant a binary relation which satisfies (i) and (ii). 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 $z \succ x$. A finite poset L can be represented by a special diagram, sometimes named *Hasse diagram* (see [9], p. 6). In a diagram of this kind, the elements of L are represented by ovals so that the oval representing y is higher than the oval representing x whenever $y \succ x$. Moreover, a segment is drawn between ovals representing x and y if y is an upper neighbour of 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 the 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.

A *complete lattice* is a poset $(L, \preceq)$ such that every subset $M \subseteq L$ has the supremum and the 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 element* 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*, that is, for every $y \in L$, there exists $M' \subseteq M$ with $y = \inf M'$. A 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 only atoms. It is *coatomistic* if the set of its coatoms is infimum dense in L, that is, 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 [38]). 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 [38].

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 book, 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 the corresponding links beween them. Note that an explicit requirement $a \neq b$ excludes any *loop*, that is, an edge connecting a node with itself (but loops are allowed 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 shortly says that $[a, b]$ is an *edge* in G. Note explicitly that this definition still 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 G with mixed edges G which has no *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. The adjectives undirected and (strictly) directed are used for paths as well. 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. An *undirected cycle* is a cycle which is an undirected route, that is, it consists of lines.

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. An undirected graph H over N is *complete* if N is a complete set in H.

An *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. In particular, every acyclic directed graph G has at least one *terminal node*, that is, a node which has no child in G.

A *chain* for a hybrid graph G over N is a partition of N into an ordered sequence of non-empty disjoint 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 [131] Lemma 2.1). Clearly, 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 also be introduced as either ordered or unordered pairs of nodes, but their meaning is different. Thus, because of their 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) $\rho(x,z)\le\rho(x,y)+\rho(y,z)$.

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 the center x and the 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, a set $\mathsf{S}\subseteq\mathsf{X}$ such 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 $x_1,x_2,\dots$ of elements of X, that is, a sequence satisfying $\forall\,\varepsilon>0\ \ \exists\, n\in\mathbb{N}$ such that $\forall\, k,l\ge n\quad \rho(x_k,x_l)<\varepsilon$, converges to an element $x\in\mathsf{X}$, that is, $\forall\,\varepsilon>0\ \ \exists\, n\in\mathbb{N}$ such that $\forall\, k\ge n\quad \rho(x_k,x)<\varepsilon$.

A 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 $\mathbb{R}^n$, $n\ge 1$ endowed with the *Eucledian distance*

$$\varrho(\boldsymbol{x},\boldsymbol{y})=\sqrt{\sum_{i=1}^{n}(x_i-y_i)^2}\qquad\text{for }\ \boldsymbol{x}=[x_1,\dots,x_n],\ \boldsymbol{y}=[y_1,\dots y_n]\,.$$

The set of real numbers $\mathbb{R}$ with $\varrho(x,y)=|x-y|$ is a special case of that.

A *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 the *Eucledian topology* induced by the 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), $i\in N$, $|N|\ge 2$ of topological spaces is defined analogously. For example, $\mathbb{R}^n\,(n\ge 2)$ endowed with the topology induced by the Eucledian distance can be viewed as the product of topological spaces $\mathsf{X}_i=\mathbb{R}$, $i\in\{1,\dots,n\}$.

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

A *measurable space* $(\mathsf{X},\mathcal{X})$ is a non-empty set X endowed with a *σ-algebra* $\mathcal{X}$ (= sigma-algebra) 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 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 *infimum* $\mathcal{S}\wedge\mathcal{T}$ is simply the intersection $\mathcal{S}\cap\mathcal{T}$ while their *supremum* $\mathcal{S}\vee\mathcal{T}$ is the σ-algebra generated by $\mathcal{S}\cup\mathcal{T}$. 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 $|f|=f^++f^-$.

A 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 the *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,

$$\lambda\,(\,\prod_{i=1}^{n}[r_i,s_i)\,)\;=\;\prod_{i=1}^{n}(s_i-r_i)\qquad\text{whenever } r_i,s_i\in\mathbb{R},\ r_i<s_i,\ i=1,\ldots,n\,.$$

A *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 the sense of Lebesgue. Given a non-negative measure μ on $(\mathsf{X},\mathcal{X})$ this construction (described, for example, in [106] 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 only be defined on A). 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. A 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 also that in case $(\mathsf{X}, \mathcal{X}) = (\prod_{i\in N} \mathsf{X}_i, \prod_{i\in N} \mathcal{X}_i)$ this is equivalent to apparently weaker requirement that the equality of integrals only holds for every measurable rectangle A. 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 (the concept of 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 $[-\infty, +\infty]$. The reader is referred to [106], Chapter 1 for elementary properties of Lebesgue integral.

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 [106], 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 the *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 [106]) 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, the 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) the Radon-Nikodym derivative of ν with respect to μ.

The *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:

$$(\mu_1 \times \mu_2)\,(\mathsf{A} \times \mathsf{B}) = \mu_1(\mathsf{A}) \cdot \mu_2(\mathsf{B}) \quad \text{whenever } \mathsf{A} \in \mathcal{X}_1,\ \mathsf{B} \in \mathcal{X}_2\,.$$

The reader is referred to [106] (Chapter 7, Example 7 and Sections 7.6, 7.7) for the proof of existence and uniqueness of (necessarily σ-finite) product measure $\mu_1 \times \mu_2$. The 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 [106], 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 of this: if 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 the 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 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

$$\int_{\mathsf{A} \cap \mathsf{B}} g(x)\; dP(x) = \int_{\mathsf{A}} g(x) \cdot h(x)\; dP(x) \equiv \int_{\mathsf{A}} g(x) \cdot h(x)\; dP^{\mathcal{A}}(x)\,.$$

Note that the proof of equivalence of three above mentioned defintions of conditional probability can be done by standard 'extension' procedure used tacitly in probability theory. The argument, which is based on so-called 'monotone class theorems', is that if one shows that the class of sets satisfying a particular property is closed under certain operation then the fact that the property is valid for a class of sets implies that it is valid for the σ-algebra generated by the class. An analogous argument is used for a class of functions satisfying a particular property. The reader is referred to § 0.2.4 of [33] or to § I.1 of [119] for basic monotone class theorems.

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 anarbitrary version of conditional probability of B given $\mathcal{A}$ (w.r.t. P).

Remark Having fixed 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 [71], § 26.1 and that a regular version of conditional probability may not exist in general (e.g. Example VI.1.35 in [119]). However, on 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 [119], Theorem VI.1.21 or [93], Consequence of Theorem V.4.4). $\triangle$

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

$$\int\limits_{\mathsf{X}_1\times\mathsf{X}_2} h(x_1)\;d\mu([x_1,x_2])=\int\limits_{\mathsf{X}_1} h(x_1)\;d\nu(x_1)\,.$$

A set $\Theta\subseteq\mathbb{R}^n$, $n\geq 1$ is called *convex* if $\forall\,\boldsymbol{x},\boldsymbol{y}\in\Theta$ and $\alpha\in[0,1]$ one has $\alpha\cdot\boldsymbol{x}+(1-\alpha)\cdot\boldsymbol{y}\in\Theta$. An example of a convex set in $\mathbb{R}$ is the interval $[0,\infty)$. A real function $\varphi:\Theta\to\mathbb{R}$ is called *convex* if for all $\boldsymbol{x},\boldsymbol{y}\in\Theta$ and $\alpha\in[0,1]$

$$\varphi\,(\alpha\cdot\boldsymbol{x}+(1-\alpha)\cdot\boldsymbol{y})\;\leq\;\alpha\cdot\varphi(\boldsymbol{x})+(1-\alpha)\cdot\varphi(\boldsymbol{y})\,,$$

and it is called *concave* if the converse inequality holds instead. Moreover, φ is called *strictly convex*, respectively *stricly concave*, if the inequality is strict whenever $\boldsymbol{x}\neq\boldsymbol{y}$ and $\alpha\in(0,1)$. Further basic result is *Jensen's inequality* (one can modify the proof from [106], 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

$$\varphi(\int\limits_{\mathsf{X}} f(x)\,d\mu(x)\,)\leq\int\limits_{\mathsf{X}}\varphi\,(f(x))\,d\mu(x).$$

In case φ is strictly convex the equality occurs if and only if f is constant μ-a.e., more exactly, if $f(x)=k$ 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 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 $\sigma(\tilde{\mathcal{B}})=\mathcal{B}$. The proof of equivalence of these two definitions can again be done by standard extension arguments based on monotone class theorems [33, 119].

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 thedefinition 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). □

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 $\mathcal{A} \perp\!\!\!\perp \mathcal{B} \vee \mathcal{D} \,|\, \mathcal{C}\ [P]$. □

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 the decomposition use (iii) with $\mathcal{E} = \mathcal{B} \vee \mathcal{D}$, $\mathcal{F} = \mathcal{D}$, $\mathcal{G} = \mathcal{C}$; for the weak union put $\mathcal{E} = \mathcal{B} \vee \mathcal{D}$, $\mathcal{F} = \mathcal{B}$, $\mathcal{G} = \mathcal{D} \vee \mathcal{C}$ instead. □

10.7 Relative entropy

Suppose that P is a finite measure and μ a σ-finite measure on a measurable space $(\mathsf{X}, \mathcal{X})$. If $P \ll \mu$ choose a version of the 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 let us call the integral the *relative entropy of P with respect to μ*. Of course, the 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 if $H(P|\mu)$ is defined, then 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 $P = \mu$. □

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 the 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, that is,

$$\forall \boldsymbol{x}, \boldsymbol{y} \in \mathsf{L} \quad \forall \alpha, \beta \in \mathbb{R} \qquad \alpha\cdot\boldsymbol{x} + \beta\cdot\boldsymbol{y} \in \mathsf{L}\,.$$

Every linear subspace of $\mathbb{R}^n$ is a closed set with respect to the 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, that is,

$$\forall\, \boldsymbol{x} \in \mathsf{L} \quad \exists\, \mathsf{B} \subseteq \mathsf{A} \text{ finite such that } \boldsymbol{x} = \sum_{\boldsymbol{y}\in\mathsf{B}} \alpha_{\boldsymbol{y}} \cdot \boldsymbol{y} \ \text{ for some } \alpha_{\boldsymbol{y}} \in \mathbb{R},\ \boldsymbol{y} \in \mathsf{B}.$$

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 $\mathsf{L} = (\mathsf{L}^{\perp})^{\perp}$.

By a *linear manifold* in $\mathbb{R}^n$ is understood a set $\mathsf{M} \subseteq \mathbb{R}^n$ of the form $\{\boldsymbol{x}\} + \mathsf{L}$ where $\boldsymbol{x} \in \mathbb{R}^n$ and L is a linear subspace of $\mathbb{R}^n$. The space L is then determined uniquely and the *dimension* of M is defined as teh dimension of L. The linear manifold generated by a set G is the least linear manifold containing G.

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, that is,

$$\forall\, \boldsymbol{x}, \boldsymbol{y} \in \mathsf{K} \quad \forall\, \alpha, \beta \geq 0 \quad \alpha\cdot\boldsymbol{x} + \beta\cdot\boldsymbol{y} \in \mathsf{K}\,.$$

By *closed convex cone* is understood a convex cone which is a closed set with respect to the 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 [126]). A 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 [126]). Any 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 $\mathsf{B} \subseteq \mathbb{Q}^n$. A 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 [126]). 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 a usual definition of a face of a closed convex set from [15] 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 [15]).

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 [126]). 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 a 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 [126]). 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 [126]). Basic result of Section 5.2 are based on the following specific property of pointed rational polyhedral cones (see Consequence 8 in [126]).

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 [126]).

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 [32], Chapters 1 and 2. The proofs of basic facts from Section 10.9.3 are in any reasonable textbook of statistics, see e.g. [4], 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: $\mathbf{\Sigma} \cdot \boldsymbol{v}$. The *null matrix* or a vector having all components zero is denoted by $\mathbf{0}$; the *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 $i, j \in N$; it is *regular* if there exists an *inverse* $N \times N$-matrix $\mathbf{\Sigma}^{-1}$ such that $\mathbf{\Sigma} \cdot \mathbf{\Sigma}^{-1} = \boldsymbol{I} = \mathbf{\Sigma}^{-1} \cdot \mathbf{\Sigma}$ (which is uniquely determined then). The *determinant* of $\mathbf{\Sigma}$ will be denoted by $\det(\mathbf{\Sigma})$, the *transpose* of $\mathbf{\Sigma}$ by $\mathbf{\Sigma}^\top$: $\mathbf{\Sigma}^\top = (\bar{\sigma}_{ij})_{i,j \in N}$ where $\bar{\sigma}_{ij} = \sigma_{ji}$ for $i, j \in N$.

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 [101], Section 1b.5). However, the expressions in which generalized inverses are commonly used do not usually 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 the 'conditional' matrix $\mathbf{\Sigma}_{A|C}$ even in case $\mathbf{\Sigma}_{C\cdot C}$ is not regular. Thus, supposing $\mathbf{\Sigma}$ is a positive semi-definite matrix one can introduce $\mathbf{\Sigma}_{A|C}$ by means of the 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 [101], 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 measurable spaces, e.g., in finite sets instead of $\mathbb{R}$, are also often considered.

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

$$e_i = \int\limits_{\mathbb{R}^N} x_i \; dP(\boldsymbol{x}) = \int\limits_{\mathsf{X}_i} y \; dP^{\{i\}}(y) \quad \text{for } i \in N.$$

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 $\mathbf{\Sigma} = (\sigma_{ij})_{i,j\in N}$ with elements

$$\sigma_{ij} = \int\limits_{\mathbb{R}^N} (x_i - e_i) \cdot (x_j - e_j) \; dP(\boldsymbol{x}) = \int\limits_{\mathsf{X}_i \times \mathsf{X}_j} (y - e_i) \cdot (z - e_j) \; dP^{\{i,j\}}(y, z) \quad \text{for } i, j \in N.$$

An alternative names are 'variance matrix', 'dispersion matrix' [101], or even 'variance-covariance matrix' [146]. An elementary fact is that the 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 a 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 the Radon-Nikodym derivative with respect to the 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 the 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 it is determined uniquely by the requirement above. Moreover, P has then the expectation $\boldsymbol{e}$ and the covariance matrix $\boldsymbol{\Sigma}$. This indicates why the parameters were designed in this way and shows that every positive semi-definite matrix is the covariance matrix of a Gaussian measure.

A 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, a 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 linear manifold 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 the whole $\mathbb{R}^N$ and $P = \mathcal{N}(\boldsymbol{e},\boldsymbol{\Sigma})$ can be introduced directly by defining the Radon-Nikodym derivative with respect to the Lebesgue measure on $\mathbb{R}^N$

$$f(x) = \frac{1}{\sqrt{(2\pi)^{|N|}\cdot\det(\boldsymbol{\Sigma})}} \cdot \exp^{-\frac{(x-\boldsymbol{e})^\top\cdot\boldsymbol{\Sigma}^{-1}\cdot(x-\boldsymbol{e})}{2}} \qquad \text{for } x \in \mathbb{R}^N\,.$$

The respective Gaussian measure on $\mathbb{R}^n$ is then called *non-negative*. This version of the Radon-Nikodym derivative is strictly positive and continuous with respect to the Eucledian topology on $\mathbb{R}^N$. Moreover, it is the unique continuous version within the class of all possible versions of the Radon-Nikodym derivative of P with respect to the Lebesgue measure λ. This simple fact motivates an implicit convention used commonly in the statistical literature; exclusively continuous versions, called *densities* are taken into consideration. The convention is in concordance with the usual way of 'marginalizing' since, for $\emptyset \neq A \subset N$, by integrating a continuous density f, that is,

$$f_A(x) = \int_{\mathsf{X}_{N\setminus A}} f(x,y) \;\, d\lambda(y) \qquad \text{for } \; x \in \mathbb{R}^A \,,$$

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 the conditional measure for every $z \in \mathbb{R}^C$

$$P_{A|C}(\mathsf{A}|z) = \int_{\mathsf{A}} f_{A|C}(x|z) \;\, d\lambda_A(x) \qquad \text{for every Borel set } \mathsf{A} \subseteq \mathbb{R}^A,$$

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 [101], Section 8a.2(v)), called sometimes the *conditional Gaussian measure*. An 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 $z \in \mathbb{R}^C$ which belongs to respective linear manifold mentioned in (10.7). It is again a Gaussian measure, given by (10.8) but $(\boldsymbol{\Sigma}_{C \cdot C})^{-1}$ is replaced by a generalized inverse $(\boldsymbol{\Sigma}_{C \cdot C})^{-}$. As shown in [101], section 8a.2(v), this conditional measure does not depend on the choice of generalized inverse.

The last important fact is that if $\boldsymbol{\Sigma}$ is positive definite then the measure $P = \mathcal{N}(\boldsymbol{e}, \boldsymbol{\Sigma})$ has finite relative entropy with respect to the Lebesgue measure λ on $\mathbb{R}^N$, namely

$$H(P\,|\,\lambda) \;=\; \frac{-|N|}{2} \cdot \ln(2\pi) - \frac{|N|}{2} - \frac{1}{2}\, \ln(\det(\boldsymbol{\Sigma})), \tag{10.9}$$

see [101], section 8a.6 (note that Rao's entropy is nothing but minus relative entropy).

10.10 Elementary statistical concepts

The task of statistics is to examine data. The aim is often to extract structural information about the relations among variables of one's interest. To this purpose, quite complicated collection of mathematical ideas, concepts and assumptions was established in order to substitute a desired relationship between reality and theory. What follows is an overview of some of these elementary statistical concepts which is, however, adapted to the goals of this monograph. Intentionally, a distinction is made between empirical concepts, that is, concepts which can be introduced in terms of sole observed data, and the concepts which are based on statistical assumptions.

10.10.1 Empirical concepts

In general, (joint) *sample space over* a non-empty finite set (of variables) N can be any Cartesian product $(\prod_{i\in N}\mathsf{X}_i, \prod_{i\in N}\mathcal{X}_i)$ of some measurable spaces $(\mathsf{X}_i, \mathcal{X}_i)$, $i\in N$, that is, of the sample spaces for individual variables. Nevertheless, two typical cases occur in practice:

- X_i is a finite non-empty set and $\mathcal{X}_i = \mathcal{P}(\mathsf{X}_i)$ for every $i\in N$, which is the case of a *discrete sample space* over N,
- $\mathsf{X}_i = \mathbb{R}$ and $\mathcal{X}_i$ is the σ-algebra of Borel sets for every $i\in N$, which is the case of the *continuous sample space* over N.

Data over N are expected in the form of a finite sequence of elements of a fixed joint sample space over N. If the sample space is fixed then the symbol $\mathsf{DATA}(N,d)$ where $d\in\mathbb{N}$ will denote the collection of all ordered sequences

$$\boldsymbol{x}^1,\ldots,\boldsymbol{x}^d \qquad \text{where } \ \boldsymbol{x}^\ell\in\prod_{i\in N}\mathsf{X}_i \quad \text{for every} \quad \ell=1,\ldots,d\,,$$

that is, the collection of all possible *databases* of the length d. The vector $\boldsymbol{x}^\ell=[x_i^\ell]_{i\in N}$ represents the ℓ-th observation, respectively the result of the ℓ-th measurement. In this monograph, the case of *complete database* is only considered unlike the case of data with *missing values*, sometimes called the case of *missing data* in the literature, in which case $\boldsymbol{x}^\ell\in\prod_{i\in N}\mathsf{X}_i$ is replaced by $\tilde{\boldsymbol{x}}^\ell\in\prod_{i\in A(\ell)}\mathsf{X}_i$ where $\emptyset\neq A(\ell)\subseteq N$ for every $\ell=1,\ldots,d$.

Any measurable mapping from $\mathsf{DATA}(N,d)\equiv(\prod_{i\in N}\mathsf{X}_i)^d$ to a measurable space $(\mathsf{T},\mathcal{T})$ is called a *statistic*. Statistics are mostly real functions - they form a basis of various statistical tests and estimates. Simple examples of statistics used in case of the continuous sample space over N are the *sample expectation* $\hat{\boldsymbol{e}}=d^{-1}\cdot\sum_{\ell=1}^d\boldsymbol{x}^\ell$ (with $\boldsymbol{x}^\ell\in\mathbb{R}^N$) and the *sample covariance matrix* defined by

$$\hat{\boldsymbol{\Sigma}}=(\hat{\sigma}_{ij})_{i,j\in N} \quad \text{where} \ \ \hat{\sigma}_{ij}=d^{-1}\cdot\sum_{\ell=1}^{d}(\boldsymbol{x}_i^\ell-\hat{\boldsymbol{e}}_i)\cdot(\boldsymbol{x}_j^\ell-\hat{\boldsymbol{e}}_j)\,.$$

To give two specific examples of statistics which form a basis of statistical conditional independence tests in case of a discrete sample space some auxiliary concepts are needed. One of them is the concept of *contingency table* over N which is a function induced by a discrete database over N which ascribes, to every element $\boldsymbol{x}$ of the sample space, the number of its counts, that is, the number of occurences of $\boldsymbol{x}$ in the database. More formally, if a discrete joint sample space is fixed the the symbol $\mathsf{CONT}(N,d)$ where $d\in\mathbb{N}$ will denote the collection of all functions $\boldsymbol{d}:\prod_{i\in N}\mathsf{X}_i\to\mathbb{Z}^+$ such that $\sum\{\boldsymbol{d}(\boldsymbol{y})\,;\,\boldsymbol{y}\in\prod_{i\in N}\mathsf{X}_i\}=d$. Given $D\in\mathsf{DATA}(N,d)$ and $\emptyset\neq A\subseteq N$ by the *marginal contingency table for* A will be understood the function

$$ct_A[D]: \quad \boldsymbol{y}=[y_i]_{i\in A}\in\prod_{i\in A}\mathsf{X}_i\mapsto|\{\ell\,;\,\forall\, i\in A \ \ x_i^\ell=y_i\}|\,.$$

The convention that $ct_\emptyset[D]\equiv d$ for $D\in\mathsf{DATA}(N,d)$ is sometimes advantageous. Note that the order of the items in a database if often unimportant in which case data are supposed directly in the form of an element of $\mathsf{CONT}(N,d)$.

A derived concept is the concept of *empirical measure* on $\prod_{i\in N}\mathsf{X}_i$ which is given by the formula

$$\hat{P}(\mathsf{A})=d^{-1}\cdot|\{\ell\,;\,\boldsymbol{x}^\ell\in\mathsf{A}\}|=d^{-1}\cdot\sum_{\boldsymbol{y}\in\mathsf{A}}ct_N[D](\boldsymbol{y}) \quad \text{for } \mathsf{A}\subseteq\prod_{i\in N}\mathsf{X}_i\,.$$

Observe that the density of the marginal measure of $\hat{P}$ on $\prod_{i\in A}\mathsf{X}_i$ with respect to the corresponding counting measure is given by

$$\hat{p}_A(\boldsymbol{y}) \;=\; d^{-1}\cdot ct_A[D](\boldsymbol{y}) \quad \text{for} \;\; \boldsymbol{y}\in\prod_{i\in A}\mathsf{X}_i,\;\; A\subseteq N\,.$$

Given a triplet $\langle a,b|C\rangle$ where $a,b\in N$ are distinct and $C\subseteq N\setminus\{a,b\}$ the respective *fitted empirical measure* on $\prod_{i\in\{a,b\}\cup C}\mathsf{X}_i$ is given by its density with respect to the counting measure:

$$\hat{p}_{\langle a,b|C\rangle}(\boldsymbol{y}) = \begin{cases} \frac{\hat{p}_{\{a\}\cup C}([y_i]_{i\in\{a\}\cup C})\cdot\hat{p}_{\{b\}\cup C}([y_i]_{i\in\{b\}\cup C})}{\hat{p}_C([y_i]_{i\in C})} & \text{if} \;\; \hat{p}_C([y_i]_{i\in C})>0,\\ 0 & \text{if} \;\; \hat{p}_C([y_i]_{i\in C})=0, \end{cases}$$

for $\boldsymbol{y}=[y_i]_{i\in\{a,b\}\cup C}\in\prod_{i\in\{a,b\}\cup C}\mathsf{X}_i$. Note that the marginal of the empirical measure $\hat{P}$ on $\prod_{i\in\{a,b\}\cup C}\mathsf{X}_i$ is absolutely continuous with respect to the fitted empirical measure. The G^2-*statistics* (see [115] p. 129) is defined as the $2d$-multiple of the respective relative entropy:

$$G^2 \;=\; 2d\cdot\sum\{\,\hat{p}_{\{a,b\}\cup C}(\boldsymbol{y})\cdot\ln\frac{\hat{p}_{\{a,b\}\cup C}(\boldsymbol{y})}{\hat{p}_{\langle a,b|C\rangle}(\boldsymbol{y})}\;;\;\boldsymbol{y}\in\prod_{i\in\{a,b\}\cup C}\mathsf{X}_i\,,\quad \hat{p}_{\{a,b\}\cup C}(\boldsymbol{y})>0\,\}\,.$$

Finally, Pearson's X^2-*statistics* (see [146] p. 216) is defined by

$$X^2 \;=\; d\cdot\sum\{\,\frac{(\hat{p}_{\{a,b\}\cup C}(\boldsymbol{y})-\hat{p}_{\langle a,b|C\rangle}(\boldsymbol{y}))^2}{\hat{p}_{\langle a,b|C\rangle}(\boldsymbol{y})}\;;\;\boldsymbol{y}\in\prod_{i\in\{a,b\}\cup C}\mathsf{X}_i\,,\quad \hat{p}_{\langle a,b|C\rangle}(\boldsymbol{y})>0\,\}\,.$$

Clearly, these two statistics take values in $[0,\infty)$. The respective statistical tests have the following form: one chooses a critical value $t\in(0,\infty)$ and rejects the corresponding conditional independence hypothesis if the value of the statistics exceeds t. However, the exact specification of the *critical value* t depends on mysterious statistical assumptions and considerations (see § 10.10.5).

10.10.2 Statistical conceptions

Basic statistical idea is that data are supposed to be values of random variables with a shared unknown distribution. The idea that a probability measure P 'generates' data, sometimes also expressed by an alternative phrase 'data are sampled from P', is usually formalised by the following concept. By a *random sample* from a probability measure P on $(\prod_{i\in N}\mathsf{X}_i,\prod_{i\in N}\mathcal{X}_i)$ of the length $d\in\mathbb{N}$ is understood a series $\boldsymbol{\xi}^1,\ldots,\boldsymbol{\xi}^d$ of generalized random variables which take values in $\prod_{i\in N}\mathsf{X}_i$, have a shared distribution P and are independent each other, that is,

$$\boldsymbol{P}(\{\omega\in\Omega\,;\;\forall\,\ell=1,\ldots,d\;\;\boldsymbol{\xi}^\ell(\omega)\in\mathsf{A}_\ell\,\}\;=\;\prod_{\ell=1}^{d}\boldsymbol{P}(\{\omega\in\Omega\,;\;\boldsymbol{\xi}^\ell(\omega)\in\mathsf{A}_\ell\,\}\;\equiv\;\prod_{\ell=1}^{d}P(\mathsf{A}_\ell)$$

holds for every sequence $\mathsf{A}_1,\ldots,\mathsf{A}_d\in\prod_{i\in N}\mathsf{X}_i$. The idea that the probability measure P is only partially unknown is formalised by the assumption that P belongs to a family $\mathbb{M}$ of probability measures on $(\prod_{i\in N}\mathsf{X}_i,\prod_{i\in N}\mathcal{X}_i)$ which could be called a *statistical model* then. Typically, a statistical model of this kind consists of a parametrized class of probability measures which are absolutely continuous with respect to a given σ-finite product measure μ on the joint sample space. The parameters belong to a convex subset Θ of the Eucledian space $\mathbb{R}^n$ for some $n\in\mathbb{N}$, mostly to an n-dimensional interval (but it can be a polytope too). Thus, a statistical model $\mathbb{M}=\{P_\theta\,;\,\theta\in\Theta\}$ is typically determined by a collection of densities

$$\{\,f(\boldsymbol{x},\theta)\,;\;\;\boldsymbol{x}\in\prod_{i\in N}\mathsf{X}_i,\quad\theta\in\Theta\,\}\quad\text{where}\quad\frac{dP_\theta}{d\mu}(\boldsymbol{x})=f(\boldsymbol{x},\theta)\;\mu\text{-a.e. on}\;\prod_{i\in N}\mathsf{X}_i\,.$$

Remark Perhaps the above description is too crude simplification of what is meant by a model in the statistical literature. In elementary statistical textbooks, a general phrase 'model' is often used to name the collection of assumptions about functional relationships among considered random variables in which some unknown parameters occur; for example, the linear regression model in Chapter VI of [4]. However, the respective complex collection of assumptions about a model of this kind can usually be reformulated as the requirement that the joint distribution of the 'observed' random variables belongs to a parametrized class of probability measures. Therefore, a lot of models from the statistical literature can be conceived in the way described above. The understanding of a statistical model as a parametrized class of probability measures also allows one to introduce further classic statistical concepts in § 10.10.4. Moreover, traditional graphical models, like graphical Gaussian (covariance selection) models and graphical log-linear models (see Chapters 6 and 7 of [146]), certainly fall within this scope. △

10.10.3 Distribution framework

However, in the area of graphical models one meets even more complex situation. A whole family of statistical models of above mentioned kind with a fixed joint sample space is at disposal and the task is to select one of the possible candidate models as the best explanation of data. That task is sometimes called the problem of *model choice* - see [24]. An important difference from the classic statistical task of testing hypothesis [68], in which case two competing classes of probability measures are disjoint, is that the competing models could intersect in case of graphical models. Actually, a pair of competing models is typically a pair of nested models, which means that one class of probability measures is contained in the other one. Thus, the whole collection of considered candidate models is a poset: the natural ordering is dictated by the inclusion (of classes of probability measures). Typically, there exists the greatest element in this poset of candidate models which is called the *saturated model.*

To compare different classes of probability measures one needs to have some common distribution framework, that is, a sufficiently comprehensive class of probability measures. Of course, some implicit consistency is desired: a distribution framework for a set of variables N should be 'of the same type' as a distribution framework over another set of variables N'. For example, one should certainly avoid of a strange combination like that in case of 3 variables all considered probability measures have discrete sample spaces but they have a continuous sample space in case 4 variables. In this monograph, the concept of *distribution framework*, that is, a rule which ascribes to every finite non-empty set of variables the respective class of probability measures (together with their sample spaces) is understood in slightly vague way. The reason is that it has not been completely clarified under which operations the ascribed class of measures should be closed (e.g. the operation of permutation of variables) and what should be the relation of ascribed classes for different sets of variables. Thus, instead of giving either descriptive or axiomatic definition of this concept four concrete examples of distribution frameworks are given:

- *Discrete distribution framework* includes, for each particular non-empty set of variables N, all probability measures on arbitrary discrete sample space over N.
- *Positive discrete distribution framework* consists of the class of probability measures on any discrete sample space $\prod_{i\in N}\mathsf{X}_i$ for which $P(\{x\})>0$ for every $x\in\prod_{i\in N}\mathsf{X}_i$ and N.
- General *Gaussian distribution framework* has, for every N, a fixed continuous sample space over N: it consists of the class of Gaussian measures on $\mathbb{R}^N$.
- *Non-degenerate Gaussian distribution framework* has also the sample space $\mathbb{R}^N$ for every

N, but it involves only those Gaussian measures whose covariance matrix is positive definite.

10.10.4 Likelihood function

Let $\mathbb{M} = \{P_\theta; \theta \in \Theta\}$ be a statistical model determined by $\{f(\boldsymbol{x}, \theta)\,;\, \boldsymbol{x} \in \prod_{i \in N} \mathsf{X}_i\,,\ \theta \in \Theta\}$. The *likelihood function* is a function which ascribes a parameter θ and data D the likelihood of the data provided that they are 'generated' from P_θ:

$$L(\theta, D) = \prod_{\ell=1}^{d} f(\boldsymbol{x}^\ell, \theta) \qquad \text{where } \ \theta \in \Theta\,,\ d \in \mathbb{N},\ D \in \mathsf{DATA}(N,d) : \boldsymbol{x}^1, \ldots, \boldsymbol{x}^d\,.$$

Indeed, in case of discrete sample space over N the value $L(\theta, D)$ is nothing but the probability of the occcurence of the database D in a random sample from P_θ. The likelihood function is a theoretical basis for classic statistical estimates and tests.

Remark Some statistical textbooks introduce the likelihood function as follows: given a single observation $\boldsymbol{x} \in \prod_{i \in N} \mathsf{X}_i$ it is a function on Θ which ascribes $f(\boldsymbol{x}, \theta)$ to $\theta \in \Theta$. However, this elementary definition is usually later extended to the case of d observations: $\prod_{i \in N} \mathsf{X}_i$ is replaced by $(\prod_{i \in N} \mathsf{X}_i)^d$ and P_θ by $\prod_{\ell=1}^{d} P_\theta$. In this monograph, like in some advanced statistical books, the extended definition of the likelihood function is accepted. $\triangle$

By the *maximum likelihood estimate* in $\mathbb{M}$ on basis of data $D \in \mathsf{DATA}(N,d)$ is understood any probability measure $P_{\hat{\theta}}$, $\hat{\theta} \in \Theta$ such that $L(\theta, D) \leq L(\hat{\theta}, D)$ for every $\theta \in \Theta$. Note that the existence and the uniqueness of the value $\hat{\theta}$ which achieves the maximum of L is not ensured in general. However, the logarithm of the likelihood function for fixed data is very often a strictly concave function on a convex set Θ and, morever, its maximum is typically attained in a unique point of Θ.

The likelihood function also allows one to derive various criteria to tackle the problem of model choice. If one has a fixed distribution framework and considers a finite family of statistical models within this framework such that the maximum likelihood estimate exists in every model from the family then the *maximized log-likelihood criterion* given by

$$\mathsf{MLL}\,(\mathbb{M}, D) = \max_{\theta \in \Theta} \ln L(\theta, D) \qquad \text{where } \ \mathbb{M} = \{P_\theta; \theta \in \Theta\},\ D \in \mathsf{DATA}(N,d)\,,$$

can be regarded as a basic information criterion to be maximized. However, if the family of competing models is a poset with the greatest element then the maximum of the MLL criterionis always achieved in the saturated model, although it may not be the only such model. Since it is not desirable to choose the most complex saturated model as the explanation of data this basic criterion is modified by subtracting a penalization term which reflects the complexity of the model. A classic criterion of complexity is the *effective dimension* $\mathsf{DIM}\,(\mathbb{M})$ of the model $\mathbb{M} = \{P_\theta; \theta \in \Theta\}$ which is typically the dimension of the linear manifold generated by Θ. The most popular criteria are *Akaike's information criterion* [2] given by

$$\mathsf{AIC}\,(\mathbb{M}, D) = \mathsf{MLL}\,(\mathbb{M}, D) - \mathsf{DIM}\,(\mathbb{M}) \qquad \text{for } \ \mathbb{M} = \{P_\theta; \theta \in \Theta\},\ D \in \mathsf{DATA}(N,d)\,, \tag{10.10}$$

and Jeffreys-Schwartz criterion [110] also named *Bayesian information criterion*:

$$\mathsf{BIC}\,(\mathbb{M}, D) = \mathsf{MLL}\,(\mathbb{M}, D) - \frac{1}{2}\mathsf{DIM}\,(\mathbb{M}) \cdot \ln d \quad \text{for } \ \mathbb{M} = \{P_\theta; \theta \in \Theta\},\ D \in \mathsf{DATA}(N,d). \tag{10.11}$$

10.10.5 Testing statistical hypotheses

Given a statistical model $\mathbb{M} = \{P_\theta; \theta \in \Theta\}$, a *statistical hypothesis* $\mathbb{H} = \{P_\theta; \theta \in \Theta_{\mathbb{H}}\}$ is specified by a non-empty subset $\Theta_{\mathbb{H}} \subset \Theta$ of the set of parameters. The respective *alternative* $\mathbb{A} = \{P_\theta; \theta \in \Theta_{\mathbb{A}}\}$ is given by the complement $\Theta_{\mathbb{A}} = \Theta \setminus \Theta_{\mathbb{H}}$. If one deals with the problem of model choice the hypothesis $\mathbb{H}$ often represents a submodel of a temporalily considered model $\mathbb{M}$. Then testing of $\mathbb{H}$ on basis of data is to answer the question whether a simpler model $\mathbb{H}$ can possibly replace $\mathbb{M}$.

As mentioned in § 10.10.1 an usual statistical test is based on a suitable statistic S. Given a critical value $t \in \mathbb{R}$ the respective *critical region* $\{D \in \mathsf{DATA}(N,d)\,;\, S(D) \geq t\}$ is the set of databases on basis of which $\mathbb{H}$ is rejected. The statistic S is designed in the way that, on condition data are 'generated' from P_θ, its value reflects the case $\theta \in \Theta_{\mathbb{H}}$. If the assumption that $D \in \mathsf{DATA}(N,d)$ is a random sample from P_θ, $\theta \in \Theta$ is accepted then the statistic S becomes a random variable with a distribution $Q_\theta^{S,d}$ and the probability $\boldsymbol{P}\{\omega \in \Omega; S(\boldsymbol{\xi}^1(\omega), \ldots, \boldsymbol{\xi}^d(\omega)) \geq t\}$ of the rejection of the hypothesis can be expressed:

$$(\prod_{\ell=1}^{d} P_\theta)\{\,[\boldsymbol{x}^1, \ldots, \boldsymbol{x}^d]\,;\, S(\boldsymbol{x}^1, \ldots, \boldsymbol{x}^d) \geq t\} \;=\; Q_\theta^{S,d}([t,\infty))\,.$$

The *error of the first kind* occurs when the hypothesis is rejected even if it holds. Its least upper bound $\sup\{Q_\theta^{S,d}([t,\infty))\,;\, \theta \in \Theta_{\mathbb{H}}\}$ is called the *size* of the test. The *error of the second kind* occurs when the hypothesis is not rejected while it should be. It is characterized by the *power function* of the test which acribes to every $\theta \in \Theta_{\mathbb{A}}$ the probability of rejection $Q_\theta^{S,d}([t,\infty))$, that is, the probability that the error of the second kind is not made. A classic method of derivation of a critical value t is to choose a *significance level* $\alpha \in (0,1)$, typically $\alpha = 0.05$, and among tests based on S whose size does not exceed α look for the most powerful test. This typically leads to the value

$$t_d^*(\alpha) = \inf\{t \in \mathbb{R};\; Q_\theta^{S,d}([t,\infty)) \leq \alpha \;\text{ for every }\; \theta \in \Theta_{\mathbb{H}}\}\,,$$

However, in practice it is often infeasible to get an analytic expression for $Q_\theta^{S,d}([t,\infty))$ for every value of d and θ. Therefore, the 'exact' critical values can hardly be obtained. The usual trick how to avoid this complicacy is to use the idea of *asymptotic distribution of* S. To this end one usually needs a theoretical result saying that there exists a probability measure Q^S on $\mathbb{R}$ which is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}$ such that

$$\lim_{d\to\infty} Q_\theta^{S,d}([t,\infty)) = Q^S([t,\infty)) \qquad \text{for every } \theta \in \Theta_{\mathbb{H}},\; t \in \mathbb{R}\,.$$

An interesting fact is that the asymptotic distribution Q^S does not usually depend on the choice of $\theta \in \Theta_{\mathbb{H}}$ and its analytic expression is known. Then an approximate critical value $t_\infty^*(\alpha) = \inf\{t \in \mathbb{R};\; Q^S([t,\infty)) \leq \alpha\}$ can be taken instead of the 'exact' value $t_d^*(\alpha)$. Typically, the asymptotic distribution of a non-negative statistic is χ^2*-distribution with* r *degrees of freedom*. Its density with respect to the Lebesgue measure on $\mathbb{R}$ is

$$f_r(x) = \begin{cases} \frac{1}{2^{\frac{r}{2}}\cdot\Gamma(\frac{r}{2})} \cdot x^{\frac{r}{2}-1} \cdot e^{-\frac{x}{2}} & \text{if } x \geq 0, \\ 0 & \text{if } x < 0, \end{cases} \qquad \text{for } r \in \mathbb{N},\; x \in \mathbb{R}\,,$$

where $\Gamma(a) = \int_0^\infty z^{a-1} \cdot e^{-z}\,dz$, $a > 0$ is the value of gamma function in a. For example, if one considers a discrete framework, the G^2-statistic, the saturated statistical model $\mathbb{M}$ and the hypothesis $\mathbb{H}$ specified by conditional independence restriction given by a triplet $\langle a, b|C\rangle$ then Whittaker [146] on p. 216 refers to [25] for a general method which allows one to verify that the asymptotic distribution of the G^2-statistics is χ^2-distribution with $r = \mathsf{DIM}\,(\mathbb{H})$ of degrees of freedom (in that special case). An analogous claim for X^2-statistics is stated on page 129 in [115].

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

H

I

J

L

M

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 189
$\perp\!\!\!\perp$ $\top\!\!\!\top$	symbols for conditional independence and dependence 19
$\otimes$	symbol for conditional product **??**
$\oplus$	symbol for direct sum (of linear spaces 194)
$\emptyset$	symbol for the empty set 183
$\rightleftharpoons$	symbol for facial equivalence 98
$\rightharpoondown$	symbol for facial implication 100
∞	symbol for infinity
$\int$	symbol for Lebegue integral 188
$\sim_m$	symbol for level equivalence induced by a skeletal imset m 166
$\wedge$ $\vee$	symbols for meet (infimum) and join (supremum) 184 (188)
$\cdot$	symbol for multiplication of matrices 196, (scalar) multiple 194
$\neg$	symbol for negation
$\preceq$ $\succeq$ $\prec$ $\succ$	symbols for partial ordering 184
$\prod$ $\times$	symbols for product (in general)
$\setminus$	symbol for set difference, e.g. $A \setminus B$
$\subseteq$ $\supseteq$ $\subset$ $\supset$	symbols for set inclusion 183
$\bigcup$ $\cup$ $\bigcap$ $\cap$	symbols for set union and intersection 183
$\rightsquigarrow$	symbol for strong facial implication 121
0	zero, zero imset 41
$\mathbf{0}$	zero vector 194, null matrix 196

Composite conventional symbols

$\vert * \vert$	absolute value, cardinality 17
$\lfloor * \rfloor$	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 185
$* \to *$	arrow (directed edge): $a \to b$ is an arrow from a to b 185
$(*, *)$	open interval, ordered pair 185
$[*, *]$	closed interval, edge in a graph 185
$\langle *, * \rangle$	scalar product 42 194
$\{*, \ldots, *\}$	set containing elements of the list $*, \ldots, *$

Symbols from other languages

Æ	the set of attributes of a formal context 91
χ_A	indicator of a set A 188
δ_A	identificator of a set A 41
Δ	generic symbol for a set of discrete variables 61
Γ	generic symbol for a set of continuous variables 61
$\Im$	incidence relation of a formal context 91
λ	(occasionally) Lebesgue measure 188
Œ	the set of objects of a formal context 91

π	Ludolf's constant, generic symbol for a permutation 183
$\wp$	generic symbol for a single class of facial equivalence 110
Ψ	the class of measures over N (distribution framework) 44 97-98 110
$\Psi(u)$	the class of Markovian measures with respect to u in Ψ 98
σ-	sigma = a generic symbol for countable infinite operation
$\sigma(\mathcal{A})$	σ-algebra generated by a class of set $\mathcal{A}$ 187
$\boldsymbol{\Sigma}^{-1}$, $\boldsymbol{\Sigma}^{-}$	inverse and generalized inverse of a matrix $\boldsymbol{\Sigma}$ 196
$\boldsymbol{\Sigma}_{A\cdot B}$, $\boldsymbol{\Sigma}_{A\mid B}$	submatrix of a matrix $\boldsymbol{\Sigma}$, Schur complement 196
$\boldsymbol{\Sigma}^{\top}$, $\boldsymbol{v}^{\top}$	transpose of a matrix $\boldsymbol{\Sigma}$ and of a vector $\boldsymbol{v}$ 196
$\mho$	generic symbol for a set of classes of facial equivalence 110
υ	counting measure 188

Symbols in alphabetic order

$-\mathsf{A}$	the set $\{-x\,;\, x \in \mathsf{A}\}$ 194
$\mathsf{A}^{\perp}$, A^{*}	orthogonal complement of A, dual cone to A 194
$\langle A, B\vert C\rangle$	disjoint triplet over N 19
$A \perp\!\!\!\perp B \mid C\ [\boldsymbol{o}]$	(conditional) independence statement 19 18 43 46 50 71 79
$A \top\!\!\!\top B \mid C\ [\boldsymbol{o}]$	(conditional) dependence statement 19
$\mathcal{A} \perp\!\!\!\perp \mathcal{B} \mid \mathcal{C}\ [P]$	conditional independence for σ-algebras 191
$an_G(A)$	the set of ancestors of a set A in a graph G 186
$as\,(\mathcal{M}, N)$	ascetic extension of a model $\mathcal{M}$ to N 172
$\mathcal{C}$	(occasionally) the set of cliques of an undirected graph 52
$\mathcal{C}(N)$	the class of combinatorial imsets over N 66
cl	generic symbol for a closure operation 185
$cl_{\mathcal{U}(N)}$	structural closure 121
$\mathcal{D}^{\downarrow}$, $\mathcal{D}^{\uparrow}$	induced descending and ascending class 183
$\mathcal{D}_u^{+}$, $\mathcal{D}_u^{-}$	positive and negative domain of an imset u 41
$\mathcal{D}_u^{*}$	the effective domain of a structural imset u 108
$\frac{d\nu}{d\mu}$	Radon-Nikodym derivative (density) of ν with respect to μ 189
$deg(u,l)$, $deg(u)$	(level-) degree of a combinatorial imset u 66
$\det(\boldsymbol{\Sigma})$	determinant of a matrix $\boldsymbol{\Sigma}$ 196
$\mathcal{E}(N)$, $\mathcal{E}_l(N)$	the class of elementary imsets over N (of degree l) 64
exp	the symbol for exponential function
f^{+}, f^{-}	positive and negative part of a function f 188
f_A	generic symbol for marginal density 25 199
$f^{\downarrow A}$	projection of a density f to A 26
$f_{A\mid C}$	generic symbol for conditional density 199
$f_{\boldsymbol{e},\boldsymbol{\Sigma}}$	density of non-degenerate Gaussian measure $\mathcal{N}(\boldsymbol{e},\boldsymbol{\Sigma})$ 34
G_T	induced subgraph of G for T 186
G^T	marginal undirected graph 45
$gra_*(N)$, $gra(N)$	(generalized) grade of the set of variables N 106
$\mathcal{H}(N)$	Hilbert basis of the cone $con(\mathcal{E}(N))$ 105
$\mathcal{H}_u$	coportrait of a structural imset 127
$H(P\vert\mu)$	relative entropy of P with respect to μ 193
$H(P\vert\mu : Q)$	Q-perturbated relative entropy of P with respect to μ 61
$h_{P,\mu}$	entropy function 74

$\boldsymbol{I}$	unit matrix 196
$\langle i,j\|K\rangle$	elementary triplet over N 22
$i \perp\!\!\!\perp j \mid K$	elementary independence statement over N 22
inf	infimum, the greatest lower bound 184
$\mathcal{K}(N)$	the class of supermodular functions over N 78
$\mathcal{K}_{\ell}(N)$	the class of ℓ-standardized supermodular functions over N 82
$\mathcal{K}_{\ell}^{\diamond}(N)$	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 42
$\mathcal{L}(N)$	the class of modular functions over N 80
$\mathcal{L}_u$	the lower class of a structural imset u 67
ln	symbol for (natural) logarithm
$m^{A\uparrow}$, $m^{A\downarrow}$	identificators of supersets and subsets of a set A 41
m_l, m_*	(level-) degree identifiers 64 (66)
m_ℓ, m_u, m_o	elements of the ℓ-, u- and o-skeleton corresponding to m 87-88
m_P	multiinformation function induced by P 31
m_T	restriction of a supermodular function m to T **??**
$m_{\dagger}$, $m_{\circ}$	special supermodular functions 74 (Figure 4.3) 102 (Figure 6.3)
$\mathcal{M}^m$	independence model produced by a supermodular function m 79
$\mathcal{M}_{\boldsymbol{o}}$ ($\mathcal{M}_P$, $\mathcal{M}_G$, $\mathcal{M}_u$)	independence model induced by $\boldsymbol{o}$ (P, G, u) 19 43 46 50 71
$\mathcal{M}_T$	restriction of a model $\mathcal{M}$ to T 20 **??**
max, $\mathcal{D}^{\max}$	maximum, the class of maximal sets in $\mathcal{D}$ 183
min, $\mathcal{D}^{\min}$	minimum, the class of minimal sets in $\mathcal{D}$ 183
N	generic symbol for a non-empty finite set of variables 17
$\mathbb{N}$	the set of natural numbers 17
$\mathcal{N}(r,s)$	one-dimensional Gaussian measure 198
$\mathcal{N}(\boldsymbol{e},\boldsymbol{\Sigma})$	Gaussian measure with expectation $\boldsymbol{e}$ and covariance matrix $\boldsymbol{\Sigma}$ 198 (34)
o-	generic symbol for 'orthogonal' standardization 42 82
P	generic symbol for a probability measure over N 17
P^A (μ^A)	marginal of a measure P (μ) for a set A 17
$P^{\mathcal{A}}$	restriction of P to a σ-algebra $\mathcal{A}$ 190
$\mathcal{P}(N)$ $\mathcal{P}(X)$	the power set 183 (40)
$P_{A\|C}$	conditional probability on X_A given C 17 (199)
$\bar{P}_{A\|C}$	its regular version (in Gaussian case 35)
$P(\mathsf{B}\|\mathcal{A})$	conditional probability of B given $\mathcal{A}$ with respect to P 190
P-a.e. (μ-a.e.)	almost everywhere with respect to P (μ) 188
$pa_G(b)$	the set of parents of a node b in a graph G 186
$\mathbb{Q}$, $\mathbb{Q}^n$	the set of rational numbers 17, rational vectors 195
$\mathbb{R}$, $\mathbb{R}^n$	the set of real numbers 17, real vectors 187
$\mathbb{R}^{\mathcal{P}(N)}$	the class of real functions on the power set $\mathcal{P}(N)$
$\mathcal{R}_u$	region of a structural imset u 108
R_u	range of a structural imset u 67
$R_{\boldsymbol{x}}$	ray generated by a vector $\boldsymbol{x}$ 195
$(\mathbb{R},\mathcal{B})$, $(\mathbb{R}^n,\mathcal{B}^n)$	the space of real numbers (vectors) with Borel σ-algebra 188 (34)
$\mathcal{S}$	(occasionally) the class of separators of a triangulated graph 52

$\mathcal{S}(N)$	the class of structural imsets over N 66
$\mathcal{S}_\Psi(N)$	the class of Ψ-representable structural imsets over N 110
sup	supremum, the least upper bound 184
$\mathcal{T}(N)$	the class of disjoint triplets over N 19
$\mathcal{T}_\emptyset(N)$, $\mathcal{T}_\epsilon(N)$	the classes of trivial and elementary triplets 22
u-	generic symbol for 'upper' standardization 42 81
u^+, u^-	positive and negative part of an imset u 41
u_G, u_H	standard imsets for graphs G and H 114 117
$u_{\langle A,B\|C\rangle}$ $u_{\langle i,j\|K\rangle}$	(semi-) elementary imset 65 64
$\mathcal{U}_u$	the upper class of a structural imset u 67
$\mathcal{U}(N)$	the class of structural models over N 92
$U(x,\varepsilon)$	ball with center x and radius ε 187
x_A	projection of a configuration x onto a set A 25
X_A	generic symbol for a sample space for A 17
$\mathcal{X}_A$, $\bar{\mathcal{X}}_A$	coordinate σ-algebra for A 17, its isomorphic representation 18
$(\mathsf{X},\mathcal{X})$ $(\mathsf{X}_A,\mathcal{X}_A)$	generic symbol for a measurable space 187 conventional notation 17
$\mathcal{X}_\emptyset$	trivial σ-algebra 17 (187)
$X^\triangleright$, $Y^\triangleleft$	Galois connection 91
$\mathbb{Z}$, $\mathbb{Z}^+$	the set of integers, the set of non-negative integers 17
$\mathbb{Z}^{\mathcal{P}(N)}$	the class of imsets over N 40

Bibliography

[1] M. Aigner: Combinatorial Theory, Springer-Verlag 1979.

[2] H. Akaike: A new look at the statistical model identification, IEEE Transactions on Automatic Control 19 (1974), pp. 716-722.

[3] Z. An, D. A. Bell, J. G. Hughes: On the axiomatization of conditional independence, Kybernetes 21 (1992), n. 7, pp. 48-58.

[4] J. Anděl: Mathematical Statistics (in Czech), SNTL (Prague) 1985.

[5] S. A. Andersson, M. D. Perlman: Lattice models for conditional independence in multivariate normal distributions, Annals of Statistics 21 (1993), pp. 1318-1358.

[6] 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.

[7] 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.

[8] 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.

[9] G. Birkhoff: Lattice Theory, AMS Colloquim Publications 1951.

[10] P. Boček: SGPOKUS, a computer program, Institute of Information Theory and Automation, June 1994.

[11] P. Boček: GENERATOR, a computer program, Institute of Information Theory and Automation, March 2001.

[12] 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.

[13] R. R. Bouckaert: Bayesian belief networks -from construction to inference, PhD thesis, University of Utrecht (Netherland) 1995.

[14] 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.

[15] A. Brøndsted: An Introduction to Convex Polytopes, Springer-Verlag 1983 (Russian translation Mir 1988).

[16] 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.

[17] 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.

[18] L. M. de Campos: Characterization of decomposable dependency models, Journal of Artificial Intelligence Research 5 (1996), pp. 289-300.

[19] R. Castelo: The discrete acyclic digraph Markov model in data mining, PhD thesis, University of Utrecht 2002.

[20] J. Cheng, R. Greiner: Comparing Bayesian network classifiers, in Uncertainty in Artificial Intelligence 15 (K. B. Laskey, H. Prade eds.), Morgan Kaufmann 1999, pp. 101-107.

[21] 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.

[22] D. M. Chickering: Optimal structure identification with greedy search, to appear in Journal of Machine Learning Research.

[23] G. F. Cooper, E. Herskovits: A Bayesian method for induction of probabilistic networks from data, Machine Learning 9 (1992), pp. 309-341.

[24] R. G. Cowell, A. P. Dawid, S. L. Lauritzen, D. J. Spiegelhalter: Probabilistic Networks and Expert Systems, Springer-Verlag 1999.

[25] D. R. Cox, D. V. Hinkley: Theoretical Statistics, Chapman and Hall 1982.

[26] D. R. Cox, N. Wermuth: Multivariate Dependencies - Models, Analysis and Interpretation, Chapman and Hall 1996.

[27] A. P. Dawid: Conditional independence in statistical theory, Journal of the Royal Statistical Society series B 41 (1979), n. 1, pp. 1-31.

[28] A. P. Dawid, S. L. Lauritzen: Hyper Markov laws in the statistical analysis of decomposable graphical models, Annals of Statistics 21 (1993), pp. 1272-1317.

[29] 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.

[30] 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.

[31] A. P. Dawid: Separoids - a general framework for conditional independence and irrelevance, submitted to Annals of Mathematics and Artificial Intelligence.

[32] M. Fiedler: Special Matrices and Their Use in Numerical Mathematics (in Czech), SNTL (Prague) 1981.

[33] J.-P. Florens, M. Mouchart, J.-M. Rolin: Elements of Bayesian Statistics, Marcel Dekker 1990.

[34] N. Friedman, D. Geiger, M. Goldszmidt: Bayesian network classifier, Machine Learning 29 (1997), pp. 131-163.

[35] M. Frydenberg: The chain graph Markov property, Scandinavian Journal of Statistics 17 (1990), n. 4, pp. 333-353.

[36] M. Frydenberg: Marginalization and collapsibility in graphical interaction models, Annals of Statistics 18 (1990), n. 2, pp. 790-805.

[37] 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.

[38] B. Ganter, R. Wille: Formal Concepts Analysis - Mathematical Foundations, Springer-Verlag 1999.

[39] P. Giudici, P. J. Green: Decomposable graphical Gaussian determination, Biometrika 86 (1999), pp. 785-801.

[40] D. Geiger: Towards the formalization of informational dependencie, technical report CSD-850053, R-102, UCLA Los Angeles, December 1987.

[41] D. Geiger, T. Verma, J. Pearl: Identifying independence in Bayesian networks, Networks 20 (1990), n. 5, pp. 507-534.

[42] 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.

[43] D. Geiger, A. Paz, J. Pearl: Axioms and algorithms for inferences involving probabilistic independence, Information and Computation 91 (1991), n. 1, pp. 128-141.

[44] D. Geiger, J. Pearl: Logical and algorithmic properties of conditional independence and graphical models, Annals of Statistics 21 (1993), n. 4, pp. 2001-2021.

[45] 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.

[46] P. Hájek, T. Havránek, R. Jiroušek: Uncertain Information Processing in Expert Systems, CRC Press 1992.

[47] D. Heckerman: A tutorial on learning Bayesian networks, technical report MSR-TR-95-06, Microsoft Research, Redmont, March 1995.

[48] R. A. Horn, Ch. R. Johnson: Matrix Analysis, Cambridge University Press 1986 (Russian translation, Mir 1989).

[49] R. Jiroušek: Solution of the marginal problem and decomposable distributions, Kybernetika 27 (1991), pp. 403-412.

[50] K. G. Jöreskog, D. Sörbom: LISREL 7 - A Guide to the Program and Application, SPSS Inc. 1989.

[51] G. Kauermann: On a dualization of graphical Gaussian models, Scandinavian Journal of Statistics 23 (1996), n. 1, pp. 105-116.

[52] H. G. Kellerer: Verteilungsfunktionen mit gegebenem Marginalverteilungen (in German), Z. Wahrscheinlichkeitstheorie 3 (1964), pp. 247-270.

[53] H. Kiiveri, T. P. Speed, J. B. Carlin: Recursive causal models, Journal of Australian Mathematical Society series A 36 (1984), pp. 30-52.

[54] T. Kočka: Graphical models - learning and applications. PhD thesis, University of Economics Prague (Czech Republic) 2001.

[55] T. Kočka, R. R. Bouckaert, M. Studený: On the inclusion problem, research report n. 2010, Institute of Information Theory and Automation, Prague, February 2001.

[56] J. T. A. Koster: Gibbs factorization and the Markov property, unpublished manuscript.

[57] J. T. A. Koster: Markov properties of nonrecursive causal models, Annals of Statistics 24 (1996), n. 5, pp. 2148-2177.

[58] J. T. A. Koster: Marginalizing and conditioning in graphical models, submitted to Bernoulli.

[59] I. Kramosil: A note on non-axiomatizability of independence relations generated by certain probabilistic structures, Kybernetika 24 (1988), n. 2, pp. 439-446.

[60] W. Lam, F. Bacchus: Learning Bayesian belief networks, an approach based on the MDL principle, Computational Intelligence 10 (1994), pp. 269-293.

[61] 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 [64].

[62] 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.

[63] 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.

[64] 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.

[65] S. L. Lauritzen: Mixed graphical association models, Scandinavian Journal of Statistics 16 (1989), n. 4, pp. 273-306.

[66] 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.

[67] S. L. Lauritzen: Graphical Models, Clarendon Press 1996.

[68] E. L. Lehman: Testing Statistical Hypothesis, John Wiley 1957.

[69] H.-G. Leimer: Optimal decomposition by clique separators, Discrete mathematics 113 (1993), pp. 99-123.

[70] M. Levitz, M. D. Perlman, D. Madigan: Separation and completeness properties for AMP chain graph Markov models, submitted to Annals of Statistics.

[71] M. Loéve: Probability Theory, Foundations, Random Processes, D. van Nostrand 1955.

[72] D. Madigan, J. York: Bayesian graphical models for discrete data, International Statistical Review 63 (1995), pp. 215-232.

[73] 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.

[74] F. M. Malvestuto: Theory of random observables in relational data bases, Information Systems 8 (1983), n. 4, pp. 281-289.

[75] J. L. Massey: Causal interpretation of random variables (in Russian), Problemy Peredachi Informatsii 32 (1996), n. 1, pp. 112-116.

[76] 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.

[77] F. Matúš: On equivalence of Markov properties over undirected graphs, Journal of Applied Probability 29 (1992), n. 3, pp. 745-749.

[78] F. Matúš: Probabilistic conditional independence structures and matroid theory, backgrounds, International Journal of General Systems 22 (1994), n. 2, pp. 185-196.

[79] F. Matúš: Stochastic independence, algebraic independence and abstract connectedness, Theoretical Computer Science A 134 (1994), n. 2, pp. 445-471.

[80] 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.

[81] F. Matúš, M. Studený: Conditional independences among four random variables I., Combinatorics, Probability and Computing 4 (1995), n. 4, pp. 269-278.

[82] F. Matúš: Conditional independences among four random variables II., Combinatorics, Probability and Computing 4 (1995), n. 4, pp. 407-417.

[83] F. Matúš: Conditional independence structures examined via minors, Annals of Mathematics and Artificial Intelligence 21 (1997), pp. 99-128.

[84] F. Matúš: Conditional independences among four random variables III., final conclusion, Combinatorics, Probability and Computing 8 (1999), n. 3, pp. 269-276.

[85] F. Matúš: On the length of semigraphoid inference, Annals of Mathematic and Artificial Intelligence....

[86] C. 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.

[87] C. 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.

[88] C. Meek: Graphical models, selecting causal and statistical models, PhD thesis, Carnegie Melon University 1997.

[89] E. Mendelson: Introduction to Mathematical Logic, 2nd edition, D. van Nostrand 1979.

[90] M. Mouchart, J.-M. Rolin: A note on conditional independence with statistical applications, Statistica 44 (1984), n. 4, pp. 557-584.

[91] M. Mouchart, J.-M. Rolin: Letter to the editor, Statistica 45 (1985), n. 3, pp. 427-430.

[92] J. Moussouris: Gibbs and Markov properties over undirected graphs, Journal of Statistical Physics 10 (1974), n. 1, pp. 11-31.

[93] J. Neveu: Bases Mathématiques du Calcul des Probabilités (in French), Masson et Cie 1964.

[94] 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.

[95] A. Paz, R. Y. Geva, M. Studený: Representation of irrelevance relations by annotated graphs, Fundamenta Informaticae 42 (2000), pp. 149-199.

[96] 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.

[97] J. Pearl: Probabilistic Reasoning in Intelligent Systems - Networks of Plausible Inference, Morgan Kaufmann 1988.

[98] A. Perez: ε-admissible simplifications of the dependence structure of a set of random variables, Kybernetika 13 (1979), pp. 439-449.

[99] 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.

[100] 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 [91] the majority of the results of this paper is almost identical with the results of [90].

[101] C. R. Rao: Linear Statistical Inference and Its Application, John Wiley 1965.

[102] 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.

[103] 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.

[104] 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.

[105] J. Rosenmüller, H. G. Weidner: Extreme convex set functions with finite carrier - general theory, Discrete Mathematics 10 (1974), n. 3-4, pp. 343-382.

[106] W. Rudin: Real and Complex Analysis, McGraw-Hill 1974.

[107] 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.

[108] G. Shafer: Probabilistic Expert Systems, CBMS-NSF Regional Conference Series in Applies Mathematics 67, SIAM 1996.

[109] A. Schrijver: Theory of Linear and Integer Programming, John Wiley 1986.

[110] G. Schwartz: Estimating the dimension of a model, Annals of Statistics 6 (1978), pp. 461-464.

[111] L. S. Shapley: Cores of convex games, International Journal of Game Theory 1 (1971/1972), pp. 11-26.

[112] P. P. Shenoy: Conditional independence in valuation-based systems, International Journal of Approximate Reasoning 10 (1994), n. 3, pp. 203-234.

[113] P. P. Shenoy: Representing conditional independence relations by valuated networks, International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 2 (1994), pp. 143-165.

[114] D. J. Spiegelhalter, S. L. Lauritzen: Sequential updating of conditional probabilities on directed graphical structures, Networks 20 (1990), n. 5, pp. 579-605.

[115] P. Spirtes, C. Glymour, R. Scheines: Causation, Prediction and Search, Lecture Notes in Statistics 81, Springer-Verlag 1993.

[116] 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.

[117] W. Spohn: Stochastic independence, causal independence and shieldability, Journal of Philosophical Logic 9 (1980), n. 1, pp. 73-99.

[118] 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.

[119] J. Štěpán: Probability Theory - Mathematical Foundations (in Czech), Academia (Prague) 1987.

[120] M. Studený: Asymptotic behaviour of empirical multiinformation, Kybernetika 23 (1987), n. 2, pp. 124-135.

[121] 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.

[122] M. Studený: Convex set functions I. and II., research reports n. 1733 and n. 1734, Institute of Information Theory and Automation, Prague, November 1991.

[123] 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.

[124] M. Studený: Multiinformation and conditional independence II., research report n. 1751, Institute of Information Theory and Automation, Prague, September 1992.

[125] 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.

[126] M. Studený: Convex cones in finite-dimensional real vector spaces, Kybernetika 29 (1993), n. 2, pp. 180-200.

[127] M. Studený: Structural semigraphoids, International Journal of General Systems 22 (1994), n. 2, pp. 207-217.

[128] 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.

[129] 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.

[130] M. Studený: Semigraphoids and structures of probabilistic conditional independence, Annals of Mathematics and Artificial Intelligence 21 (1997), n. 1, pp. 71-98.

[131] M. Studený: A recovery algorithm for chain graphs, International Journal of Approximate Reasoning 17 (1997), n. 2-3, pp. 265-293.

[132] 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.

[133] 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.

[134] 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.

[135] M. Studený: Complexity of structural models, in Prague Stochastics'98, August 23-28, Prague 1998, pp. 521-528.

[136] 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.

[137] 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.

[138] M. Studený: On mathematical description of probabilistic conditional independence structures, DrSc thesis, Institute of Information and Automation, Prague, PerezMay 2001.

[139] 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.

[140] I. Vajda: Theory of Statistical Inference and Information, Kluwer 1989.

[141] 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.

[142] 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.

[143] 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.

[144] M. Volf, M. Studený: A graphical characterization of the largest chain graphs, International Journal of Approximate Reasoning 20 (1999), n. 3, pp. 209-236.

[145] S. Watanabe: Information theoretical analysis of multivariate correlation, IBM Journal of Research and Development 4 (1960), pp. 66-81.

[146] J. Whittaker: Graphical Models in Applied Multivariate Statistics, John Wiley 1990.

[147] 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.