Information Science and Statistics

Milan Studený

On Probabilistic Conditional Independence Structures

With 42 Illustrations

Springer

Milan Studený, RNDr, DrSc
Institute of Information Theory and Automation,
Academy of Sciences of the Czech Republic, CZ18208 Pod vodárenskou věží 4,
Prague 8, Libeň Czech Republic

Cover illustration: Details

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

Library of Congress Cataloging-in-Publication Data
A catalog record for this book is available from the Library of Congress

Mathematics Subject Classification (1991): 62-02, 68-02, 62H05, 68R01, 68T30, 94A17, 06A17, 90C10, 15A99, 52B99

ISBN 1-85233-891-1 Springer London Berlin Heidelberg
Springer is a part of Springer Science+Business Media
springeronline.com

Printed in the United States of America

The use of registered names, trademarks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant laws and regulations and therefore free for general use.

The publisher makes no representation, express or implied, with regard to the accuracy of the information contained in this book and cannot accept any legal responsibility or liability for any errors or omissions that may be made.

Typesetting: Camera-ready by authors
12/3830-543210 Printed on acid-free paper SPIN 10943397

Special acknowledgements in Czech

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

Preface

The book is devoted to the mathematical description of probabilistic conditional independence structures. The topic of conditional independence, which falls within both the scope of statistics and of artificial intelligence, has been at the center of my research activity for many years – since the late 1980s. 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 for describing probabilistic conditional independence structures which, in my view, overcomes an 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 1990s in a series of papers [137]. 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 perhaps not pinpointed clearly. Third, the motivation was not explained in detail. I also think that the original series of papers was difficult for researchers in the area of artificial intelligence to read because "practical" implementation aspects of the presented approach were suppressed there. Another point is that the pictorial representation of considered mathematical objects, to which researchers interested in graphical models are accustomed, was omitted.

Within the next six years, further mathematical results were achieved which amended, supplemented and gave more precision to the original idea. I have also deliberated about suitable terminology and the way to present the method of structural imsets which would be acceptable to statisticians and researchers in the area of artificial intelligence, as well as exact from the mathematical point of view. I wrote it up in my DrSc thesis [146], which became the basis of this monograph. After finishing the thesis, I realized the 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 this monograph is to present the method of structural imsets in its full (present) extent: the motivation; the mathematical foundations, which I tried to present in a didactic form; indication of the area of application; and open problems. The motivation is explained in the first chapter. The second chapter recalls basic concepts in the area of probabilistic conditional independence structures. The third chapter is an overview of classic graphical methods for describing conditional independence structures. The core of the method of structural imsets is presented in the next four chapters. The eighth chapter shows application of the method to learning graphical models. Open problems are gathered in the ninth chapter and necessary elementary mathematical notions are provided in the Appendix for the reader's convenience. Then the List of Notation follows. As there are many cross-references to elementary units of the text, like Lemmas, Remarks etc., they are listed with page numbers afterwards. The text is concluded by the References and the Index.

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 understanding of the mathematical basis of the theory of (graphical) models of conditional independence structures;
- advanced 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.

In particular, I have in mind those PhD students who are thinking about an academic career. They are advised to read the book starting with the Appendix and to utilize the lists at the end of the book.

Many people deserve my thanks for 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 and pointers to the literature and discussion which helped me clarify the view on the topic of the book. I have also profited from 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 of my colleagues mentioned above involved joint theoretical research as well. A preliminary version of the book was read by my PhD student Petr Šimeček, who gave me several useful comments and recommendations including an important example. I also made minor changes in response to comments given by Tomáš Kroupa and Helen Armstrong, who read some parts of the manuscript. As concerns the technical help I would like to thank Václav Kelar for making special LaTeX fonts for me and to Jarmila Pánková for helping me to prepare several pages with special pictures. I am likewise grateful to Cheri Dohnal and Antonín Otáhal for correcting my (errors in) English. I was very pleased by the positive attitude of Stephanie Harding, who is the Mathematics and Statistics Editor at Springer London; the cooperation with her was smooth and effective. She found suitable reviewers for the book and they gave me further useful comments, which helped me to improve the quality of the book.

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, many other colleagues influenced me in addition to those who were already mentioned above. I will name some of them here: 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. Of course, the above list is not exhaustive; I apologize to anyone whose name may have been omitted.

Let me emphasize that I profited from meeting several colleagues who gave me inspiration during the seminar, “Conditional Independence Structures”, which was held from September 27 to October 17, 1999 in the Fields Institute for Research in Mathematical Sciences, University of Toronto, Canada, and during several events organized within the framework of the ESF program, “Highly Structured Stochastic Systems” in the years 1997–2000. In particular, I wish to thank Helène Massam and Steffen Lauritzen, who gave me a chance to participate actively in these wonderful events. For example, I remember the stimulating atmosphere of the HSSS research kitchen “Learning conditional independence models”, held in Třešť, Czech Republic, in October 2000.

Finally, this 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 was supported by the projects GA AVČR n. K1019101 and GAČR n. 201/01/1482. It is a result of long-term research performed in the institute, which has provided a suitable environment for my work since 1983.

Prague,
March 2004

Milan Studený

Contents

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 found its motivation in artificial intelligence and statistics. In fact, these two fields are the main areas where the concept of conditional independence has been successfully applied. More specifically, graphical models of conditional independence structure are widely used in:

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

A (non-probabilistic) concept of conditional independence was also 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 a multidisciplinary flavor. Nevertheless, it certainly falls within the scope of *informatics* or *theoretical cybernetics*, and the main emphasis is put on mathematical fundamentals.

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 Motivational thoughts

The following "methodological" considerations are meant to explain my motivation. In this section six general questions of interest are formulated which may arise in connection with any particular method for describing conditional independence structures. I think these questions should be answered in order to judge fairly and carefully the quality and suitability of every considered method.

To be more specific, one can assume a general situation, illustrated by Figure 1.1. One would like to describe *conditional independence structures* (in short, CI structures) induced by probability distributions from a given fixed class of distributions over a set of variables N. For example, we can consider the class of discrete measures over N (see p. 11), the class of regular Gaussian measures over N (see p. 30), the class of conditional Gaussian (CG) measures over N (see p. 66) or any specific parameterized class of distributions. In other words, a certain *distribution framework* is specified (see Section A.9.5). 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 an even 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.).

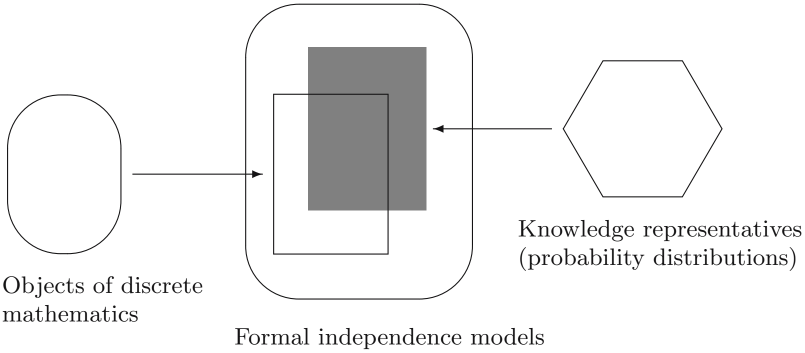
Objects of discrete mathematics
Formal independence models
Knowledge representatives (probability distributions)

Fig. 1.1. Theoretical fundamentals (an informal illustration).

Every knowledge representative of this kind induces a formal independence model over N (for definition see p. 12). Thus, the class of induced conditional independence models is defined; in other words, the class of CI structures to be described is specified (the shaded area in Figure 1.1). One has in mind a method for describing CI structures in which objects of discrete mathematics – for example, graphs, finite lattices and discrete functions – are used to describe CI structures. Thus, a certain *universum of objects of discrete mathematics* is specified. Typical examples are classic graphical models widely used in multivariate analysis and probabilistic reasoning (for details, see Chapter 3). It is supposed that every object of this type induces a formal independence model over N. The intended interpretation is that the object thus "describes" an induced independence model so that it can possibly describe one of the CI structures that should be described.

The definition of the induced formal independence model depends on the type of considered objects. Every particular universum of objects of discrete mathematics has its respective 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 evolution of miscellaneous Markov properties (see Remark 3.1 in Section 3.1). The evolution has led to the concept of "global Markov property" which establishes a graphical criterion to determine the maximal set of conditional independence statements represented in a given graph. This set is the ascribed formal independence model. The above-mentioned implicit assumption of the existence of the respective criterion is a basic requirement of *consistency*, that is, the requirement that every object in the considered universum of objects has unambiguously ascribed a certain formal independence model. Note that some recently developed graphical approaches (see Section 3.5.3) still need to be developed up to the concept of a global Markov property so that they will comply with the basic requirement of consistency. Under the above situation I can formulate the first three questions of interest which, in my opinion, are the most important theoretical questions in this general context.

- The *faithfulness* question is whether every object from the considered universum of objects of discrete mathematics indeed describes one of the CI structures.
- The *completeness* question is whether every CI structure can be described by one of the considered objects. If this is not the case an advanced subquestion occurs, namely the task to characterize conveniently those formal independence models which can be described by the objects from the considered universum.
- The *equivalence* question involves the task of characterizing 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 used by Spirtes et al. [122], 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 a discrete distribution framework (= relative to the class of discrete measures) and faithfulness in a 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 models, including myself, also used a traditional term "(strong) completeness of a separation graphical criterion" [44, 90, 141, 73]. However, according to the above classification, results of this type are among the results gathered under the label "faithfulness" (customary reasons for traditional terminology are explained in Remark 3.2 on p. 45). Thus, I distinguish between the "completeness of a criterion" on the one hand and the "completeness of a universum of objects" (for the description of a class of CI structures) on the other hand.

Now I will 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).

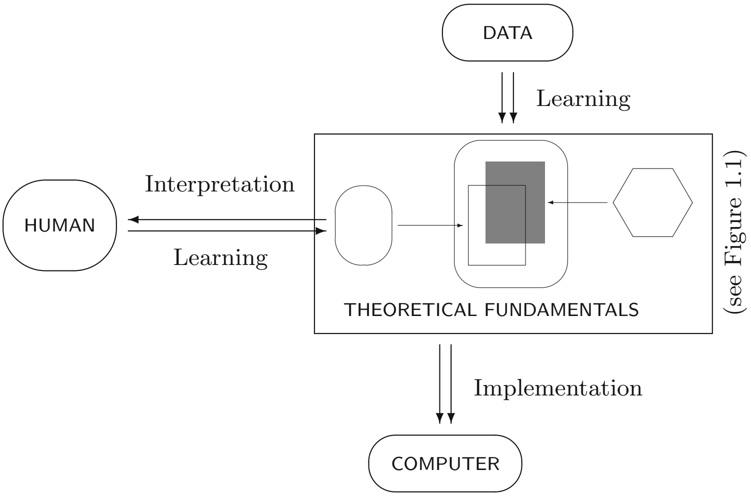

Fig. 1.2. Practical questions (an informal illustration).

- The *interpretability* question is whether considered objects of discrete mathematics can be conveyed to humans in an acceptable way. That usually means whether or not they can be visualized so that they are understood easily and interpreted correctly as CI structures.
- The *learning* question is how to determine the most suitable CI structure either on the basis of statistical data (= testing problem) or on the basis of expert knowledge provided by human experts. An advanced statistical subquestion is the task to determine even a particular probability distri-

bution inducing the CI structure, which is equivalent to the problem of "estimation" of parameters of a statistical model.

- The *implementation* question is how to manage the corresponding computational tasks. An advanced subquestion is whether or not the acceptance of a particular CI structure allows one to do respective subsequent calculation with probability distributions effectively, namely whether the considered objects of discrete mathematics give guidance in the calculation.

Classic graphical models are easily accepted by humans; however, their pictorial representation may sometimes lead to another interpretation. For example, acyclic directed graphs can either be interpreted as CI structures or one can prefer "causal" or "deterministic" interpretation of their edges [122], which is different. Concerning computational aspects, an 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 the basis of a well-known "local computation method" [66] which is at the core of several working probabilistic expert systems [49, 26]. Of course, the presented questions of interest are connected to each other. For example, structure learning from experts certainly depends on interpretation while (advanced) distribution learning is closely related to the "parameterization problem" (see p. 210), which also has a strong computational aspect.

The goal of these motivational thoughts is the idea that the practical questions are ultimately connected with the theoretical basis. Before inspection of practical questions one should first solve the related theoretical questions, in my opinion. Regrettably, some researchers in artificial intelligence (and to a lesser degree, those in statistics) do not pay enough attention to the theoretical grounds and concentrate mainly on practical issues like simplicity of accepted models, either from the point of view of computation or visualization. They usually settle on a certain class of "nice" graphical models (e.g. Bayesian networks – see p. 46) 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 that is my main source of motivation. Consider a hypothetical situation where one is trying to learn the CI structure induced by a discrete distribution on the basis of statistical data. Suppose, moreover, that one is limited to a certain class of graphical models – say, Bayesian networks. It is known that this class of models is not complete in the discrete distribution framework (see Section 3.6). Therefore one searches for the "best approximation". Some of the learning algorithms for graphical models browse through the class of possible graphs as follows. One starts with a graph with the maximum 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. This is a correct procedure in the case where the underlying probability distribution indeed induces the CI structure that 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 that are represented in the “new” graph but which *may not be valid with respect to the underlying distribution.* Again I emphasize that this erroneous acceptance of additional conditional independence statements is made on the 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 has already been criticized within the community of researchers in artificial intelligence (see [159] and Remark 8.1). In my opinion, these recurring problems in solving practical questions of learning are inevitable consequences of the omission of theoretical grounds, namely the question of completeness. This may have motivated several recent attempts to introduce wider and wider classes of graphs which, however, lose easy interpretation and do not achieve completeness. Therefore, in this book, I propose a non-graphical method for describing probabilistic CI structures which primarily solves the completeness problem 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 provide an overview of traditional methods for describing probabilistic CI structures. These methods mainly use graphs whose nodes correspond to variables as basic tools for visualization and interpretation. The overview involves basic results about conditional independence, including those published in my earlier papers.

The second goal is to present the mathematical basis of an alternative method for describing 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 may perhaps seem longer and more detailed than necessary from an expert’s perspective. The reason for this is that not only top experts in the field and mathematicians are the expected audience. The intention was to write a report which can be read and understood by advanced PhD students in computer science and statistics. This was the main stimulus which compelled me to resolve the dilemma of “understandability” versus “conciseness” in favor of precision and potential 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 an *imset*, which is a certain integer-valued discrete function, are supposed to form the theoretical basis for the rest of the book.

Chapter 3 is an overview of *graphical methods* for describing CI structures. Both classic approaches (undirected graphs, acyclic directed graphs and chain graphs) and recent attempts are included. The chapter argues for a conclusion that a non-graphical method achieving the completeness (in the sense mentioned on p. 3) 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 a 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 that 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 relationship between those two ways of describing CI structures (by imsets) is established. A unifying point of view provided by the theory of formal concept analysis is offered.

Chapter 6 is devoted to an advanced question of equivalence (in the sense mentioned on p. 3) 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 *independence implication* of structural imsets is characterized in two different ways. One of them allows one to transform the task of computer implementation of independence 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 choosing a suitable representative of a class of equivalent structural imsets. Two approaches to this problem are offered. The concept of a *baricentral imset* seems to be a good solution from a theoretical point of view in the general context while the concept of a *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 an analytic review of methods for 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 characterizing inclusion quasi-ordering 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 questions (which were mentioned on pp. 4–5).

The Appendix 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.

For the reader's convenience two lists are included after the Appendix: the List of Notation and the List of Lemmas, Propositions etc. The text is concluded by the References and the Index.

2

Basic Concepts

Throughout the book the symbol N will denote a non-empty finite set of *variables*. The 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.).

CONVENTION 1. The following conventions will be used throughout the book. Given sets $A, B \subseteq N$ the juxtaposition AB will denote their union $A \cup B$. 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 excluding 0). The symbol $|A|$ will be used to denote the number of elements of a finite set A, that is, its *cardinality*. The symbol $|x|$ will also denote the *absolute value* of a real number x, that is, $|x| = \max\{x, -x\}$. ♢

2.1 Conditional independence

A basic notion of the monograph is a *probability measure over* N. This phrase will be used to describe the situation in which 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

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

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, a 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 a 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 a certain general understanding of the concept of conditional probability. Given a probability measure P over N and disjoint sets $A, C \subseteq N$, *conditional probability on* X_A *given* C (more specifically given $\mathcal{X}_C$) will be understood as a function of two arguments $P_{A|C} : \mathcal{X}_A \times \mathsf{X}_C \to [0,1]$ which ascribes an $\mathcal{X}_C$-measurable function $P_{A|C}(\mathsf{A}|\star)$ to every $\mathsf{A} \in \mathcal{X}_A$ such that

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

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 A.6.4, Remark A.1) 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 (P^C-a.e.). 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 accordance 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 the coordinate σ-algebra representing the set A, namely

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

Thus, $A \subseteq B \subseteq N$ is reflected by $\bar{\mathcal{X}}_A \subseteq \bar{\mathcal{X}}_B$ and it is natural to require 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 A.6.4. △

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 a 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-a.e.

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 $\mathcal{X}_i = \mathcal{P}(\mathsf{X}_i)$ is the class of all its subsets for every $i \in N$. Then $P_{A|C}$ is determined uniquely exactly on the set $\{x \in \mathsf{X}_C\,;\ P^C(\{x\}) > 0\}$ by means of the formula

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

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

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

for every $\mathsf{A} \subseteq \mathsf{X}_A$, $\mathsf{B} \subseteq \mathsf{X}_B$ and $x \in \mathsf{X}_C$ with $P^C(\{x\}) > 0$. Of course, A and B can be replaced by singletons. Note that the fact that the equality P^C-a.e. coincides 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 [133, 117, 31]. 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 the above-mentioned basic formal properties of conditional independence. In modern statistics, they were first accentuated by Dawid [29], then mentioned by Mouchart and Rolin [93], and van Putten and van Shuppen [103]. Spohn [124] 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 [99]. Their terminology [100] 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 always be 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), a possibility distribution over N [18, 149], a relational database over N [112] and a structural imset over N (see Section 4.4.1). The notation $A \perp\!\!\!\perp B \,|\, C \; [\boldsymbol{o}]$ will be used in those cases, but the symbol $[\boldsymbol{o}]$ can be omitted if 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, written after the straight line, is the conditioning set while the two former components are independent areas, usually interchangeable. The 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 triplets 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 a 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 a *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. On the other hand, 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

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

The *restriction* of a formal independence model $\mathcal{M}$ over N to a non-empty set $\emptyset \neq T \subseteq N$ will be understood as 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 (induced by the marginal).

Remark 2.3. I should explain my limitation to disjoint triplets over N, since some authors, e.g. Dawid [33], do not make this restriction at all. For simplicity of explanation consider a discrete probabilistic framework. Indeed, given a discrete probability measure P over N, the statement $A \perp\!\!\!\perp B \,|\, C \; [P]$ can also be defined for non-disjoint triplets $A, B, C \subseteq N$ in a reasonable way [41, 81]. However, then the statement $A \perp\!\!\!\perp A \,|\, C \; [P]$ has specific interpretation, namely that the variables in A are functionally dependent on the variables in C (with respect to P), so that it can be interpreted as a *functional dependence statement*. Let us note (cf. § 2 in [81]) that one can easily derive that

$$A \perp\!\!\!\perp B \,|\, C \; [P] \;\;\Leftrightarrow\;\; \left\{ \begin{array}{c} A \setminus C \perp\!\!\!\perp B \setminus AC \mid C \; [P] \quad \text{and} \\ (A \cap B) \setminus C \perp\!\!\!\perp (A \cap B) \setminus C \mid C \cup (B \setminus A) \; [P] \end{array} \right\}.$$

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

Remark 2.4. To avoid misunderstanding, the reader should be aware that the noun *model* may have any of 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 Pearl [100]. 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 A.9.2, that is, a class of probability measures. Note that this terminology is often used in statistics – see Remark A.3 for more detailed explanation.

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 A.9.5) 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 a graphical model [157, 70]. 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 that the class $\mathbb{M}_G$ is often introduced in another way – see Section 8.2.1 for equivalent definitions in case of acyclic directed graphs in terms of recursive factorization and in terms of parameterization. △

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 [100], 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 over N to $\mathcal{T}(T)$ for non-empty $T \subseteq N$ is a semi-graphoid over T. The following fact is important.

Lemma 2.1. Every conditional independence model $\mathcal{M}_P$ induced by a probability measure P over N is a disjoint semi-graphoid over N.

Proof. This can be derived easily from Corollary A.2 proved in the Appendix (see p. 235). 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 coordinate σ-algebra $\bar{\mathcal{X}}_A \subseteq \mathcal{X}_N$ as 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 A.7. 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 Corollary A.2. □

Note that the above mentioned fact is not a special feature of a probabilistic framework. Conditional independence models occurring within other uncertainty calculi (in artificial intelligence) mentioned at the end of Section 2.1 are also (disjoint) semi-graphoids. 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 a semi-graphoid is not substantial. One can introduce an *abstract semi-graphoid* on a join semi-lattice $(\mathcal{S}, \vee)$ as a ternary relation $\star \perp\!\!\!\perp \star \,|\, \star$ 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 $[A \perp\!\!\!\perp B \,|\, D \vee C$ and $A \perp\!\!\!\perp D \,|\, C]$.

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 for σ-algebras with respect to a probability measure on $(\mathsf{X}, \mathcal{X})$ (see Corollary A.2). Note that the above concept of an abstract semi-graphoid is essentially equivalent to the concept of a *separoid* introduced by Dawid [33], which is a mathematical structure unifying a variety of notions of "conditional independence" arising in probability, statistics, artificial intelligence, and other fields.

Let me conclude this remark by a note which indicates the obstacles that authors in mathematics meet if they want to establish new terminology. Pearl and Paz [99] decided to use the word "graphoid" to name a new concept they introduced (see p. 29 for this concept). However, it appeared that this word had already been "occupied": it was used to name one of equivalent definitions of a matroid [155]. One of the motives which led Dawid [33] to use the word "separoid" to name his general concept was to avoid a terminological clash. However, it appeared that this word had also been used independently by Strausz [128] to name a certain abstract binary relation between sets whose aim is to generalize geometric separation of sets in $\mathbb{R}^n$ by hyperplanes. An interesting observation is that, by coincidence, there is a weak connection between two concepts of a separoid. For example, an undirected graph G and the relation of separation for sets of nodes in G, which is defined as in Section 3.1 but non-disjoint sets are allowed, can give an example of both separoids. The difference is that Dawid's separoid is a ternary relation $A \perp\!\!\!\perp B \,|\, C \; [G]$ while a binary relation $A \perp\!\!\!\perp B \,|\, \emptyset \; [G]$ can serve as an example of Strausz's separoid. △

2.2.3 Elementary independence statements

To store a semi-graphoid over N in the memory of a computer it is not necessary to allocate all $|\mathcal{T}(N)| = 4^{|N|}$ bits. A more economic way of their representation is possible. For example, one can 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 *trivial disjoint triplets* over N by $\mathcal{T}_{\emptyset}(N)$.

However, independence statements of principal importance are *elementary statements*, which correspond to *elementary triplets*, that is, disjoint triplets $\langle A, B|C\rangle$ over N where both A and B are singletons (cf. [3, 79]). A simplifying convention will be used in this case: braces in singleton notation will be omitted so that $\langle a, b|K\rangle$ or $a \perp\!\!\!\perp b \,|\, K$ 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 disjoint 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

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

In particular, every semi-graphoid is determined by its "trace" within the class of elementary triplets, that is, by the intersection with $\mathcal{T}_{\epsilon}(N)$. Moreover, if $\mathcal{M}_1, \mathcal{M}_2$ are semi-graphoids over N then $\mathcal{M}_1 \cap \mathcal{T}_{\epsilon}(N) \subseteq \mathcal{M}_2 \cap \mathcal{T}_{\epsilon}(N)$ is equivalent to $\mathcal{M}_1 \subseteq \mathcal{M}_2$.

Proof. (see also [79]) 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 (otherwise 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 statement mode* of representing a semi-graphoid, that is, by the list of contained elementary triplets, is more suitable. The characterization of those collections of elementary triplets which represent semi-graphoids is given in Proposition 1 of Matúš [79].

Remark 2.6. Another reduction of memory demands for semi-graphoid representation follows from the symmetry property. 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 elementary imset 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 (cf. Lemmas 4.6 and 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. These observations influenced the terminology: the adjective "elementary" is used to indicate the respective disjoint triplets and independence statements. $\triangle$

2.2.4 Problem of axiomatic characterization

Pearl and Paz [99, 100] formulated a conjecture that semi-graphoids coincide with conditional independence models induced by discrete probability measures. However, this conjecture was refuted in Studený [130] by finding a further formal property of these models, which is not derivable from semi-graphoid properties, namely

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

Another formal property of this sort was later derived in An et al. [3]. Consequently, a natural question occurred. Can conditional independence models arising in a 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 Studený [132], then it would have been possible to construct a deductive system that is an analog of the notion of a "formal axiomatic theory" from Mendelson [92]. The considered 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 Studený [132] (for a more didactic proof see [144]) that, for every $n \in \mathbb{N}$, there exists a formal property of (discrete) probabilistic conditional independence models which applies to a set of variables N with $|N| = n$ but which cannot be revealed on a set of smaller 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 probabilistic conditional independence models over N suggests that they can be characterized in terms of a finite number of formal properties of semi-graphoid type. Thus, a related task is, for a small cardinality of N, to characterize them in that way. It is no problem to verify that they coincide with semi-graphoids in the case $|N| = 3$ (see Figure 5.6 for illustration). Discrete probabilistic conditional independence models over N with $|N| = 4$ were characterized in a series of papers by Matúš [84, 85, 87]; for an overview see Studený and Boček [136] where the respective formal properties of these models are explicitly formulated – one has 18300 different models of this kind and these 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 Geiger et al. [45] and independently in Matúš [82] 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. An 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 by Geiger and Pearl [46] and by Malvestuto [77]. The result from Studený [138] can be interpreted as a specific relative-completeness result, 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 Kramosil [62]. △

2.3 Classes of probability measures

There is no uniformly accepted conception of the notion of a *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., elements of the 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 its general sense, mainly in vague motivational parts. Moreover, terminological distinction is made between those two above-mentioned approaches. The concept of a *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.

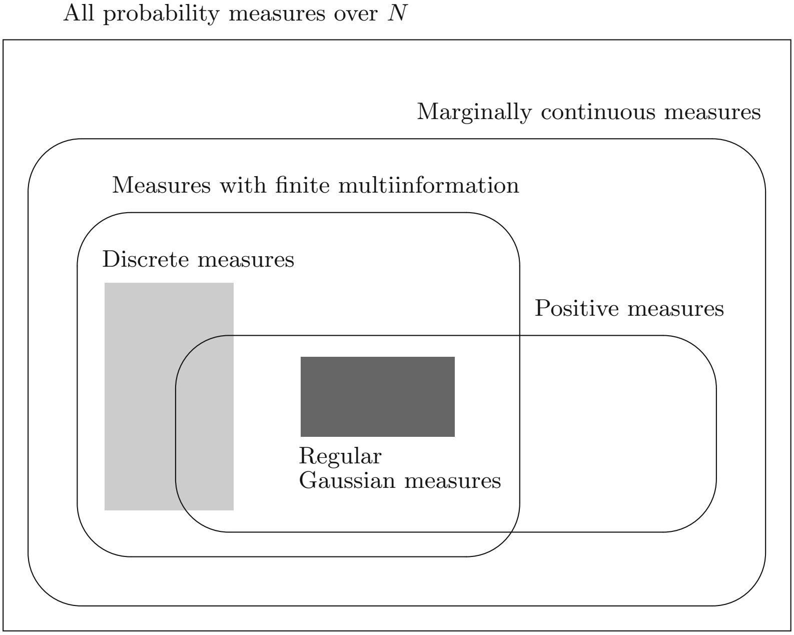
All probability measures over N
Marginally continuous measures
Measures with finite multiinformation
Discrete measures
Positive measures
Regular
Gaussian measures

Fig. 2.1. A comparison of basic classes of probability measures over N.

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 continuous measures", are introduced. One of them, the class of measures with finite multiinformation, is strongly related to the method of structural imsets described in later chapters. The information-theoretical methods are applicable to measures belonging to this class which, fortunately, involves typical measures used in practice. Inclusion 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 an 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. (see also § 1.2.2 in [37]) It was shown in [130], Proposition 1, that in the 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 non-zero 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 for 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, the 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 as its dominating measure. Note that nearly all multidimensional measures used in practice are marginally continuous (see Sections 2.3.5, 2.3.6 and 4.1.3 for other examples). However, there are probability measures over N which are not marginally continuous; in particular, some singular Gaussian measures – see Example 2.3 on p. 35.

Having fixed a dominating measure μ for a marginally continuous measure P over N 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 explaining details (see Remark 1 in [130]) 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 the sense of Loéve [74]. The regularity of conditional probability is usually derived as a consequence of special topological assumptions on $(\mathsf{X}_i, \mathcal{X}_i)$, $i \in N$ (see the Appendix, Remark A.1). Thus, the 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 a *dominated experiment* in Bayesian statistics – see § 1.2.2 and § 1.2.3 in the book by Florens et al. [37]. △

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) conventions.