1

Conditional Independence and Markov Properties for Basic Graphs

Milan Studený

Institute of Information Theory and Automation of the CAS

CONTENTS

The aim of the chapter 4
1.1 Introduction: historical overview and an example 4
1.1.1 Stochastic conditional independence 4
1.1.2 Graphs and local computation method 4
1.1.3 Conditional independence in other areas 5
1.1.4 Geometric approach and methods of modern algebra 5
1.1.5 A motivational example 6
1.2 Notation and elementary concepts 8
1.2.1 Discrete probability measures 8
1.2.2 Continuous distributions 10
1.3 The concept of conditional independence 10
1.3.1 Conditional independence in the discrete case 11
Factorization and other equivalent definitions 12
1.3.2 More general CI concepts 13
1.4 Basic properties of conditional independence 15
1.4.1 Conditional independence structure 15
Side-remark about relational databases 16
1.4.2 Statistical model of a CI structure 17
1.5 Semi-graphoids, graphoids, and separoids 17
1.5.1 Elementary and dominant triplets 19
1.6 Elementary graphical concepts 21
1.7 Markov properties for undirected graphs 22
1.7.1 Global Markov property for an UG 22
1.7.2 Local and pairwise Markov properties for an UG 23
1.7.3 Factorization property for an UG 24
1.8 Markov properties for directed graphs 24
1.8.1 Directional separation criteria 24
Straightforward criterion in terms of walks 24
D-separation criterion 26
Moralization criterion 27
The third option 28
Equivalence of directional criteria 28
1.8.2 Global Markov property for a DAG 29
1.8.3 Local Markov property for a DAG 29
1.8.4 Factorization property for a DAG 29
1.8.5 Markov equivalence for DAGs 30
1.9 Remarks on chordal graphs 31
1.10 Imsets and geometric views 32
1.10.1 The concept of a structural imset 32
1.10.2 Imsets for statistical learning 33
1.11 CI inference 33
Acknowledgements 34

Bibliography 34

The aim of the chapter

In this chapter, the concept of *conditional independence* (CI) is recalled and an overview of both former and recent results on the description of CI structures is given. The traditional graphical models, namely those ascribed to *undirected graphs* (UGs) and *directed acyclic graphs* (DAGs), can be interpreted as special cases of statistical models of a CI structure. Therefore, an overview of Markov properties for these two basic types of graphs is also given. Markov properties for more general graphs are discussed in Chapter 2.

1.1 Introduction: historical overview and an example

In this section, some earlier results on CI are recalled and an example is given to informally illustrate the concept of probabilistic CI.

1.1.1 Stochastic conditional independence

Loève [27] already defined the concept of CI in terms of σ-algebras in his book on probability theory in the 1950s. Phil Dawid [12] was probably the first statistician who explicitly formulated certain basic formal properties of stochastic CI. He observed that several statistical concepts, e.g., the one of a sufficient statistic, can equivalently be defined in terms of generalized CI and this observation allows one to derive many results in an elegant way with the aid of those formal properties. These basic formal properties of stochastic CI were later independently formulated in the context of philosophical logic by Spohn [50], who was interested in the interpretation of the concept of CI and its relation to causality. The same properties, this time formulated in terms of σ-algebras, were also explored by Mouchart and Rolin [39]. The author of this chapter was told that the conditional independence symbol $\perp\!\!\!\perp$ was proposed by Dawid and Mouchart in their discussion in the late 1970s.

The significance of the CI concept for probabilistic reasoning was later recognized by Pearl and Paz [43], who observed that the above-mentioned basic formal properties of CI are also valid for certain ternary separation relations induced by undirected graphs. This led them to an idea of describing such formal ternary relations by graphs and introducing an abstract concept of a *semi-graphoid*. The even more abstract concept of a *separoid* was later suggested by Dawid [13]. Pearl and Paz [43] also raised a conjecture that semi-graphoids coincide with probabilistic CI structures, which was refuted by myself in [52] using some tools of information theory.

A lot of effort and time has been devoted to the problem of characterizing all possible CI structures induced by four discrete random variables. The final solution to that problem was achieved by Matúš [36, 33, 34]; the number of these structures is 18478 [66] and they are decomposed into 1098 types.

1.1.2 Graphs and local computation method

The idea to use graphs whose nodes correspond to random variables in order to describe CI structures had appeared in statistics earlier than Pearl and Paz suggested this approach in the context of computer science. One can distinguish between two basic traditional trends,

namely, using undirected and directed (acyclic) graphs. Note that statistical models described by such graphs can be understood as the models of (special) CI structures.

Undirected graphs (UGs) appeared in the 1970s in statistical physics as tools to describe relations among discrete random variables. Moussouris [40] introduced several Markov properties relative to an UG for distributions with positive density and showed their equivalence with a factorization condition. Darroch, Lauritzen, and Speed [10] realized that UGs can be used to describe statistical models arising in the theory of contingency tables, so they introduced a special class of (undirected) graphical models and interpreted them in terms of CI. At the same time, the use of UGs was considered in the area of multivariate statistical analysis. Dempster [15] introduced covariance selection models for continuous real random variables, which were interpreted in terms of CI by Wermuth [67].

In the 1980s, directed acyclic graphs (DAGs) found their applications in the decision-making theory in connection with influence diagrams. Smith [49] used the above-mentioned formal properties of CI to easily show the correctness of some operations with influence diagrams. Pearl's book [42] on probabilistic reasoning had a substantial impact on promotion of graphical methods in artificial intelligence; in the book, he defined a directional separation criterion (d-separation) for DAGs and pinpointed the role of CI.

The theoretical breakthrough leading to (graphical) probabilistic expert systems was the *local computation method*. Lauritzen and Spiegelhalter [26] offered a methodology to perform efficient computations of conditional probabilities for (discrete) measures which are Markovian with respect to a DAG.

1.1.3 Conditional independence in other areas

Probability theory and statistics are not the only fields in which the concept of CI was introduced and examined. An analogous concept of *embedded multivalued dependency* (EMVD) was studied in the 1970s in the theory of relational databases. Sagiv and Walecka [44] showed that there is no finite axiomatic characterization of EMVD structures. Shenoy [48] observed that one can introduce the concept of CI within various calculi for dealing with knowledge and uncertainty in artificial intelligence (AI), including Spohn's theory of natural (ordinal) conditional functions, Zadeh's possibility theory and the Dempster-Shafer theory of evidence.

This motivated several papers devoted to formal properties of CI within various uncertainty calculi in AI. For example, Vejnarová [63] studied the properties of CI in the frame of possibility theory and it was shown in [54] that there is no finite axiomatic characterization of the CI structures arising in the context of natural conditional functions. Various concepts of conditional irrelevance have also been introduced and their formal properties were examined in the theory of *imprecise probabilities*; let us mention the concept of epistemic irrelevance introduced by Cozman and Walley [9].

1.1.4 Geometric approach and methods of modern algebra

The observation that graphs cannot describe all possible discrete stochastic CI structures led me to proposing a linear-algebraic method of their description in [57]. In this approach, certain vectors whose components are integers and correspond to subsets of the set of variables, called (structural) *imsets*, are used to describe the CI structures. The approach allows one to apply geometric methods of combinatorial optimization to learning graphical models and to approaching the CI implication problem. Hemmecke et al. [21] answered two of the open problems related to the method of imsets and disproved a geometric conjecture from [57] about the cone corresponding to the (structural) imsets.

The application of methods of modern algebra and (polyhedral) geometry to problems

arising in mathematical statistics has recently led to establishing a new field of *algebraic statistics*. Drton, Sturmfels and Sullivant [16] in their book on this topic devoted one chapter to advanced algebraic tools for describing statistical models of CI structure. The topic of probabilistic CI thus naturally became one of the topics of interest in that area.

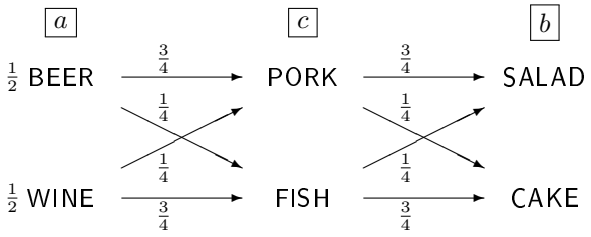
1.1.5 A motivational example

This section contains a small story to explain the intuitive sense of CI. The section can be skipped without affecting the flow of the chapter.

Imagine that organizers of a conference entitled *Probabilistic Graphical Models*, to be held in September 2018 in Prague, have the task to organize a lunch for the participants in a student cafeteria during a lunch break. Because of time limitations, they give the participants a limited choice of drinks and dishes. Cruel organizers intentionally decide to ignore human rights of vegetarians and teetotalers; thus, 3 items are to be served consecutively:

a ... a drink (exclusive choice is either BEER or white WINE),
c ... the main course (the choice is PORK or FISH),
b ... a dessert (the choice is SALAD or CAKE).

The participants are asked to decide about their main courses c after obtaining their drinks a. This can substantially influence their decisions: a well-known fact is that white wine pairs with fish while beer is the best fit with a traditional Czech dish which consists of roasted pork, sauerkraut and dumplings. Thus, a typical participant already drinking wine chooses fish, while only a minority of wine-drinkers take pork. Analogously, a typical beer-drinker goes for pork and a minority of them take fish. An analogous decision problem for participants occurs when they finish their main courses. Since the pork with sauerkraut is fat, a typical pork-eater decides to compensate that by choosing a light salad; on the other hand, fish has low calories, which makes a majority of fish-eaters decide for a sweet cake. Assume for simplicity that the proportion of non-typical participants in each group is $\frac{1}{4}$ and that drinking preferences are equal. This leads to the following scheme, allowing us to compute the overall probabilities:



It is clear from the description of the situation that (the decision about) the last event b dominantly depends on (the previous decision about) the event c; in fact, it is independent of what the results of the event a was. This description characterizes the situation when *events a and b are conditionally independent given the values of the event c*, which is conventionally denoted by $a \perp\!\!\!\perp b \,|\, c$.

This example also illustrates the intuitive difference between conditional and unconditional independence of events. One can observe higher correlation between beer and salad,

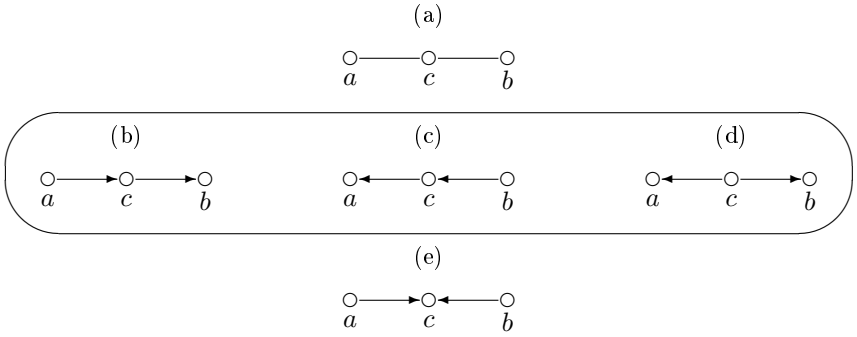


FIGURE 1.1: Examples of graphs for a description of a CI structure.

respectively between wine and cake:

$$
\begin{array}{lll}
(\mathsf{BEER},\mathsf{SALAD}) & \left.\begin{array}{lll}(\mathsf{BEER},\mathsf{PORK},\mathsf{SALAD}) & \to & \frac{1}{2}\cdot\frac{3}{4}\cdot\frac{3}{4}=\frac{9}{32}\\ (\mathsf{BEER},\mathsf{FISH},\mathsf{SALAD}) & \to & \frac{1}{2}\cdot\frac{1}{4}\cdot\frac{1}{4}=\frac{1}{32}\end{array}\right\} & \to \ \frac{10}{32},\\
(\mathsf{BEER},\mathsf{CAKE}) & \left.\begin{array}{lll}(\mathsf{BEER},\mathsf{PORK},\mathsf{CAKE}) & \to & \frac{1}{2}\cdot\frac{3}{4}\cdot\frac{1}{4}=\frac{3}{32}\\ (\mathsf{BEER},\mathsf{FISH},\mathsf{CAKE}) & \to & \frac{1}{2}\cdot\frac{1}{4}\cdot\frac{3}{4}=\frac{3}{32}\end{array}\right\} & \to \ \frac{6}{32},\\
(\mathsf{WINE},\mathsf{SALAD}) & \left.\begin{array}{lll}(\mathsf{WINE},\mathsf{PORK},\mathsf{SALAD}) & \to & \frac{1}{2}\cdot\frac{1}{4}\cdot\frac{3}{4}=\frac{3}{32}\\ (\mathsf{WINE},\mathsf{FISH},\mathsf{SALAD}) & \to & \frac{1}{2}\cdot\frac{3}{4}\cdot\frac{1}{4}=\frac{3}{32}\end{array}\right\} & \to \ \frac{6}{32},\\
(\mathsf{WINE},\mathsf{CAKE}) & \left.\begin{array}{lll}(\mathsf{WINE},\mathsf{PORK},\mathsf{CAKE}) & \to & \frac{1}{2}\cdot\frac{1}{4}\cdot\frac{1}{4}=\frac{1}{32}\\ (\mathsf{WINE},\mathsf{FISH},\mathsf{CAKE}) & \to & \frac{1}{2}\cdot\frac{3}{4}\cdot\frac{3}{4}=\frac{9}{32}\end{array}\right\} & \to \ \frac{10}{32}.
\end{array}
$$

Note that situations when a confounding variable c exists for correlated variables a and b often occurs in connection with the so-called *Simpson's paradox*.

Graphs can be used to depict such conditional in/dependence relations among (random) variables; for example, the undirected graph in Figure 1.1(a) is traditionally used to depict the above-described situation. Directed edges (= arrows) can then be used to give some additional information or interpretation. Thus, the directed graph in Figure 1.1(b), which brings the same CI information as that in Figure 1.1(a), can be used to show the time direction. Nevertheless, the role of variables a and b is exchangeable in the sense that the probability distribution remains the same if one swaps a and b. Indeed, one can consider an absurd situation when the cruel organizers decide the next day to reverse the order of courses in order to torture the participants with the thirst. The point is that the probability distribution will be the same but the graph in Figure 1.1(c) then better reflects the additional information about the time direction. The third option is that one decides to use arrows to pinpoint the central role of the variable c, which leads to the graph shown in Figure 1.1(d). All the directed graphs in the oval within Figure 1.1 give the same CI information (about the underlying distribution); in this case we say they are *independence* or *Markov equivalent*.

The directed graph in Figure 1.1(e) is, however, interpreted in another way. This graph is traditionally used to describe the situation when variables a and b are unconditionally independent but conditionally dependent given c; in notation, $a \perp\!\!\!\perp b$ and $a \not\perp\!\!\!\perp b \,|\, c$. Let us give another silly example of when such a probability distribution can occur. Imagine that a poor man has the last two coins to be spent this evening and he has to decide whether to buy some food or a bottle of beer he likes a lot. Thus, he has to decide exclusively between the hunger and the thirst; nevertheless, he slightly tends to avoid the hunger. Therefore, he

decides to toss the coins and if two heads occur then he buys the beer, otherwise the food. Then the result a of tossing the first coin will be independent of the result b of tossing the second coin while the event c whether he buys food or beer depends on the joint configuration of values of a and b. In this case, a and b are *not* conditionally independent given the values of c. Some authors then say that variables a and b are *marginally independent* but not conditionally independent given (the values of) variable c.

The latter example also allows us to explain the difference between two traditional interpretations of graphical models. In this chapter we deal with the basic *CI interpretation* when the graph in Figure 1.1(e) is understood solely as a record of CI information about the distribution. Other authors have given an extended *causal interpretation* of graphs, where arrows are used to encode expected causal relationships among variables. Thus, the graph in Figure 1.1(e) can be understood as a pictorial representation of the fact that the variable c causally/functionally depends on (combination of) variables a and b. Note, however, that causal interpretation of graphs is based on very specific assumptions on data generation mechanism. The causal interpretation is discussed in Part IV of this handbook.

1.2 Notation and elementary concepts

In this section, notation is introduced and elementary notions are recalled. Throughout the chapter N is a finite non-empty index set whose elements correspond to random *variables* (and to nodes of graphs in graphical context). The symbol $\mathcal{P}(N) := \{ A : A \subseteq N \}$ will denote the *power set* of N.

1.2.1 Discrete probability measures

This section mainly deals with the discrete case and does not require any special previous knowledge from the reader.

Definition 1.2.1. A *discrete probability measure over* N is defined as follows:

(i) For each $i \in N$ a non-empty finite set X_i is given, which is the *individual sample space* for the variable i. This defines a *joint sample space*, which is the Cartesian product $\mathsf{X}_N := \prod_{i \in N} \mathsf{X}_i$.

(ii) A probability measure P on X_N is given; it is determined by its *density*, which is a function $p : \mathsf{X}_N \to [0, 1]$ such that $\sum_{x \in \mathsf{X}_N} p(x) = 1$. Then $P(\mathbb{A}) = \sum_{x \in \mathbb{A}} p(x)$ for any $\mathbb{A} \subseteq \mathsf{X}_N$.

A general *probability measure over* N is defined analogously, but instead of a finite set X_i, a measurable space $(\mathsf{X}_i, \mathcal{X}_i)$ is assumed for any $i \in N$. The joint sample space is endowed with the product σ-algebra $\bigotimes_{i \in N} \mathcal{X}_i$. Some measures on $(\mathsf{X}_N, \bigotimes_{i \in N} \mathcal{X}_i)$ cannot be determined by densities in the general case.

Given $A \subseteq N$, any list of elements $[x_i]_{i \in A}$ such that $x_i \in \mathsf{X}_i$ for $i \in A$ will be called a *configuration* for A. The set X_A of configurations for A is then the *sample space for* A. Given disjoint $A, B \subseteq N$, we will use concatenation AB as a shorthand for (disjoint) union $A \cup B$. Given disjoint configurations $a \in \mathsf{X}_A$ and $b \in \mathsf{X}_B$, the symbol $[a, b]$ will denote their *join*, i.e., the joint list. If the joint configuration is an argument of a function, say of a density $p : \mathsf{X}_{AB} \to \mathbb{R}$, then the brackets will be omitted and we will write $p(a, b)$ instead of $p([a, b])$; similarly in the case of the join of three or more disjoint configurations.

In the case of $A \subseteq B$ and $b \in \mathsf{X}_B$, the symbol b_A will denote the *restriction* of the configuration b for A, that is, the restricted list. The mapping from X_B to X_A ascribing b_A to $b \in \mathsf{X}_B$ is the corresponding marginal *projection*. In particular, the symbol $b_\emptyset$ will denote the *empty configuration*, that is, the empty list of elements.

Given $i \in N$, the symbol i will often be used as an abbreviation for the singleton $\{i\}$. In particular, if $i \in A \subseteq N$ and $a \in \mathsf{X}_A$ then the symbol a_i will be a simplified notation for the marginal configuration $a_{\{i\}}$; of course, it is nothing but the i-th component of the configuration a.

Given disjoint $A, B \subseteq N$ and configuration sets $\mathbb{A} \subseteq \mathsf{X}_A$, $\mathbb{B} \subseteq \mathsf{X}_B$, we introduce $\mathbb{A} \times \mathbb{B} := \{[a, b] : a \in \mathbb{A} \,\&\, b \in \mathbb{B}\}$. Note that $\mathbb{A} \times \mathbb{B}$ is typically the Cartesian product but if $A = \emptyset$ and $\mathbb{A} \neq \emptyset$, that is, if $\mathbb{A} = \{a_\emptyset\}$ only contains the empty configuration, then one has $\mathbb{A} \times \mathbb{B} = \mathbb{B}$; analogously in the case of $B = \emptyset \neq \mathbb{B}$.

Definition 1.2.2. Given $A \subseteq N$ and a probability measure P over N, the *marginal measure for* A is the measure P_A over A defined by the relation

$$P_A(\mathbb{A}) := P(\{x \in \mathsf{X}_N : x_A \in \mathbb{A}\}) \quad \text{for } \mathbb{A} \subseteq \mathsf{X}_A \quad (\mathbb{A} \in \textstyle\bigotimes_{i \in A} \mathcal{X}_i \text{ in general}).$$

In the discrete case, the *marginal density for* A is the density of P_A; it is given by the formula

$$p_A(a) = P(\{x \in \mathsf{X}_N : x_A = a\}) = \sum_{c \in \mathsf{X}_{N \setminus A}} p(a, c) \quad \text{for } a \in \mathsf{X}_A,$$

where p is the (joint) density of the probability measure P.

Note that a simple *vanishing principle* for marginal densities will be tacitly used in § 1.3.1: if $x \in \mathsf{X}_N$, $C \subseteq B \subseteq N$ then $p_C(x_C) = 0$ implies $p_B(x_B) = 0$. The next elementary concept in the discrete case is that of a conditional probability, where the conditioning objects are (marginal) configurations.

Definition 1.2.3. Given disjoint sets $A, C \subseteq N$ of variables and a discrete probability measure P over N, the *conditional probability on* X_A *given* C is a (partial) function of two arguments denoted by $P_{A|C}(*|*)$, where the asterisks stand for the respective arguments. Specifically,

$$P_{A|C}(\mathbb{A}|c) \;:=\; \frac{P_{AC}(\mathbb{A} \times \{c\})}{P_C(\{c\})} \equiv \frac{P_{AC}(\mathbb{A} \times \{c\})}{p_C(c)}$$
$$\text{where } \mathbb{A} \subseteq \mathsf{X}_A \text{ and } c \in \mathsf{X}_C \text{ with } p_C(c) > 0.$$

The *conditional density for* A *given* C is also a (partial) function, in this case both arguments are the respective marginal configurations:

$$p_{A|C}(a|c) := \frac{p_{AC}(a, c)}{p_C(c)} \equiv P_{A|C}(\{a\}|c) \quad \text{for } a \in \mathsf{X}_A,\ c \in \mathsf{X}_C \text{ with } p_C(c) > 0.$$

Observe that the marginal measure can be viewed as a special case of the conditional probability where the conditioning configuration is empty, that is, $C = \emptyset$. Another observation is that, for any *positive configuration*, that is, $c \in \mathsf{X}_C$ with $p_C(c) > 0$, the function $\mathbb{A} \subseteq \mathsf{X}_A \mapsto P_{A|C}(\mathbb{A}|c)$ is a probability measure over A. It is clear that $P_{A|C}(*|*)$ only depends on the marginal P_{AC}.

In the computer science community, the conditional density is sometimes called a *conditional probability table*. Let us emphasize that the ratio defining the conditional density is not defined for conditioning on *zero configurations* $c \in \mathsf{X}_C$ with $p_C(c) = 0$, an important detail which is, unfortunately, omitted or even ignored in some machine learning (text)books.

Note that the assumption that the density is *strictly positive*, that is, $p(x) > 0$ for all $x \in \mathsf{X}_N$, is too restrictive in the area of probabilistic expert systems because it does not allow for modeling functional dependencies between random variables.

In the discrete case, one does not need to extend the conditional probability to zero configurations in order to define the notion of CI; however, in the general case, one has to consider different versions of conditional probability, which makes the general definition of CI more technical (see § 1.3.2).

1.2.2 Continuous distributions

In this section, which can be skipped by beginners, we assume that the reader is familiar with the standard notions of measure theory. The meaning of the term *probability distribution* encountered in the literature depends on the field in which it is actually encountered. In probability theory, it usually means a (general) probability measure, while in statistics its meaning is typically restricted to measures given by densities, and in computer science it is often identified with the concept of a density function.

In statistics, one typically works with real continuous distributions and these are defined through densities. There is a quite wide class of probability measures for which the concept of density (function) makes good sense.

Definition 1.2.4. A probability measure over N is *marginally continuous* if it is absolutely continuous with respect to the product of its one-dimensional marginals, that is, if

$$(\bigotimes_{i\in N} P_i)(\mathbb{A}) = 0 \quad \text{implies} \quad P(\mathbb{A}) = 0 \qquad \text{for any } \mathbb{A} \in \bigotimes_{i\in N} \mathcal{X}_i,$$

in notation $P \ll \bigotimes_{i\in N} P_i$, where the symbol $\otimes$ is used to denote both the product of (probability) measures and the product of σ-algebras.

An equivalent definition of a marginally continuous measure is that there exists a (dominating) system of σ-finite measures μ^i on $(\mathsf{X}_i, \mathcal{X}_i)$ for $i \in N$ such that $P \ll \bigotimes_{i\in N} \mu^i$ (see [57, Lemma 2.3]). It is easy to verify that every discrete probability measure over N is marginally continuous: the dominating system of measures is the system of counting measures, that is, $\mu^i(\mathbb{A}) = |\mathbb{A}|$ for any $i \in N$ and $\mathbb{A} \subseteq \mathsf{X}_i$. Another standard example is a *regular Gaussian measure* over N; in this case, for any $i \in N$, $\mathsf{X}_i = \mathbb{R}$ is the set of real numbers endowed with the Borel σ-algebra and μ^i is the Lebesgue measure.

Having fixed individual sample spaces and a dominating system of σ-finite measures, every marginally continuous measure P can be introduced through its *joint density* f, which is the Radon-Nikodym derivative of P with respect to $\mu := \bigotimes_{i\in N} \mu^i$. For all $A \subseteq N$, we put $\mathcal{X}_A := \bigotimes_{i\in A} \mathcal{X}_i$ and accept a convention that $\mathcal{X}_\emptyset := \{\emptyset, \mathsf{X}_\emptyset\}$ is the only (trivial) σ-algebra on $\mathsf{X}_\emptyset$.

The *marginal density for* $A \subseteq N$ is then defined as the Radon-Nikodym derivative f_A of the marginal P_A with respect to $\mu^A := \bigotimes_{i\in A} \mu^i$, where $\mu^\emptyset$ is the only probability measure on $(\mathsf{X}_\emptyset, \mathcal{X}_\emptyset)$ by a convention. Recall that it is an $\mathcal{X}_A$-measurable function satisfying $P_A(\mathbb{A}) = \int_{x\in\mathbb{A}} f_A(x)\, \mathrm{d}\,\mu^A(x)$ for any $\mathbb{A} \in \mathcal{X}_A$. The marginal density f_A can be understood as a function on the joint sample space X_N depending only on the marginal configuration x_A. The joint and marginal densities are determined uniquely in the sense of μ-everywhere.

1.3 The concept of conditional independence

In this section, several equivalent definitions of probabilistic CI in the discrete case are presented; the general case is discussed in the end of the section.

1.3.1 Conditional independence in the discrete case

In this section, a number of equivalent definitions of probabilistic CI are given and illustrated by two examples. Our attention will intentionally be restricted to the discrete case in order to keep the text accessible to beginners.

The following symmetric definition of CI was chosen as the basic one because it is analogous to the definition of stochastic independence, which is characterized by the requirement that the joint distribution is the product of marginal ones.

Definition 1.3.1. Let $A, B, C \subseteq N$ be pairwise disjoint sets of variables and P a discrete probability measure over N. We say that *A and B are conditionally independent given C with respect to P* and write $A \perp\!\!\!\perp B \,|\, C\ [P]$ if

$$\forall\, \mathbb{A} \subseteq \mathsf{X}_A \quad \forall\, \mathbb{B} \subseteq \mathsf{X}_B \quad \forall\, c \in \mathsf{X}_C \ \text{ such that } \ p_C(c) > 0$$
$$P_{AB|C}(\mathbb{A} \times \mathbb{B}|c) = P_{A|C}(\mathbb{A}|c) \cdot P_{B|C}(\mathbb{B}|c). \tag{1.1}$$

It follows from the definition that the validity of $A \perp\!\!\!\perp B \,|\, C\ [P]$ only depends on the marginal measure P_{ABC}. Clearly, a modified formulation of (1.1) is that, for each positive configuration $c \in \mathsf{X}_C$, the conditional probability $P_{AB|C}(*|c)$ is the product of some measures over A and B. The condition (1.1) has a natural interpretation of *conditional irrelevance*: once the value $c \in \mathsf{X}_C$ for C is known, the variables in A and B do not influence each other, i.e., the occurrence of a value $b \in \mathsf{X}_B$ does not influence the probability of occurrence of $a \in \mathsf{X}_A$, and vice versa. Also, (1.1) can be extended to a general case, as explained in § 1.3.2. On the other hand, (1.1) is not suitable for verification.

Fortunately, there are elegant equivalent conditions given in terms of densities. Specifically, given pairwise disjoint $A, B, C \subseteq N$ and a discrete probability measure P over N, the CI statement $A \perp\!\!\!\perp B \,|\, C\ [P]$ has the following equivalent formulation in terms of *marginal densities*:

$$\forall\, x \in \mathsf{X}_{ABC} \qquad p_C(x_C) \cdot p_{ABC}(x) = p_{AC}(x_{AC}) \cdot p_{BC}(x_{BC})\,, \tag{1.2}$$

which easily implies a seemingly weaker condition

$$\forall\, x \in \mathsf{X}_{ABC} \text{ with } p_{ABC}(x) > 0 \quad p_{ABC}(x) = \frac{p_{AC}(x_{AC}) \cdot p_{BC}(x_{BC})}{p_C(x_C)}\,. \tag{1.3}$$

Using the vanishing principle, the reader can easily see that (1.1)⇒(1.2)⇒(1.3); the implication (1.3)⇒(1.1) follows from the next fact.

Observation 1.3.1. There exists a probability measure $\bar{P}$ on X_{ABC} such that

$$\bar{P}_{AC} = P_{AC}, \quad \bar{P}_{BC} = P_{BC}, \text{ and } \ A \perp\!\!\!\perp B \,|\, C\ [\bar{P}].$$

The measure $\bar{P}$ is uniquely determined and satisfies $P_{ABC} \ll \bar{P}$.

Proof. We define the value $\bar{p}(x)$ of the density of $\bar{P}$ by the formula on the right hand side of (1.3) for $x \in \mathsf{X}_{ABC}$ with $p_C(x_C) > 0$ and $\bar{p}(x) = 0$ in the case of $p_C(x_C) = 0$. The remaining statements are left to the reader as an exercise. □

Observation 1.3.1 even holds for any pair of discrete probability measures Q on X_{AC} and R on X_{BC} satisfying $Q_C = R_C$ in place of P_{AC} and P_{BC}. The measure $\bar{P}$ can then be called the *conditional product of Q and R*; this result implies that, for any such *consonant* pair of measures Q and R, a distribution P over ABC exists having them as marginals, namely $\bar{P}$.

To verify (1.3)⇒(1.1), we use the construction from the proof of Observation 1.3.1 and apply (1.3) to see that $\bar{p}(x) = p_{ABC}(x)$ in the case of $p_{ABC}(x) > 0$. Then we realize that the values of both $\bar{p}$ and p_{ABC} sum up to 1 to extend the equality $\bar{p}(x) = p_{ABC}(x)$ to the case of $p_{ABC}(x) = 0$.

Another CI characterization in terms of marginal densities appeared in [40]; it can be interpreted as a *cross-exchange condition* for configurations:

$$\forall\, a, \bar{a} \in \mathsf{X}_A,\ \ \forall\, b, \bar{b} \in \mathsf{X}_B,\ \ \forall\, c \in \mathsf{X}_C \ \text{ one has}$$
$$p_{ABC}(a,b,c) \cdot p_{ABC}(\bar{a},\bar{b},c) = p_{ABC}(a,\bar{b},c) \cdot p_{ABC}(\bar{a},b,c)\,. \tag{1.4}$$

To verify (1.2)⇒(1.4), we distinguish between the cases of $p_C(c) = 0$, when (1.4) is evident, and $p_C(c) > 0$. In the latter case, derive (1.4) whose sides are both multiplied by $p_C(c) \cdot p_C(c)$ from equalities (1.2) applied to $x = [a,b,c]$, $x = [\bar{a},\bar{b},c]$, $x = [\bar{a},b,c]$, and $x = [a,\bar{b},c]$. The implication (1.4)⇒(1.2) can be shown by summing over $\bar{a}$ and $\bar{b}$ in (1.4). The condition (1.4) is particularly easy to verify in the binary case, when $|\mathsf{X}_i| = 2$ for all $i \in N$.

Example 1.3.2. To illustrate the application of the above equivalent definitions of (discrete probabilistic) CI, let us take $N = \{a,b,c\}$, $\mathsf{X}_i = \{0,1\}$ and introduce a binary probability measure P on X_N by its density p as follows:

$$p(0,0,0) = p(0,1,1) = p(1,0,1) = p(1,1,0) = \frac{1}{8} + \varepsilon\,,$$
$$p(0,0,1) = p(0,1,0) = p(1,0,0) = p(1,1,1) = \frac{1}{8} - \varepsilon\,,$$

for some $0 \le \varepsilon \le \frac{1}{8}$. We have $p_{ab}(0,0) = p_{ab}(0,1) = p_{ab}(1,0) = p_{ab}(1,1) = \frac{1}{4}$. No matter what the parameter ε is, the cross-exchange condition (1.4) holds:

$$p_{ab}(0,0) \cdot p_{ab}(1,1) = \frac{1}{4} \cdot \frac{1}{4} = \frac{1}{16} = \frac{1}{4} \cdot \frac{1}{4} = p_{ab}(0,1) \cdot p_{ab}(1,0)\,,$$

which means $a \perp\!\!\!\perp b \,|\, \emptyset\ [P]$ or, in a briefly record, $a \perp\!\!\!\perp b\ [P]$. One can also test that using the condition (1.2): since one has $p_a(0) = p_a(1) = \frac{1}{2}$ and $p_b(0) = p_b(1) = \frac{1}{2}$, using the fact that $p_\emptyset(x_\emptyset) = 1$ for any $x \in \mathsf{X}_{ab}$ we have

$$p_\emptyset(x_\emptyset) \cdot p_{ab}(x_{ab}) = 1 \cdot \frac{1}{4} = \frac{1}{4} = \frac{1}{2} \cdot \frac{1}{2} = p_a(x_a) \cdot p_b(x_b)\,,$$

again showing $a \perp\!\!\!\perp b \,|\, \emptyset\ [P]$. On the other hand, for all $0 < \varepsilon \le \frac{1}{8}$, the cross-exchange condition (1.4) does not hold with $c = 0$:

$$p(0,0,0) \cdot p(1,1,0) = (\frac{1}{8} + \varepsilon)^2 \neq (\frac{1}{8} - \varepsilon)^2 = p(0,1,0) \cdot p(1,0,0)\,,$$

which means $a \not\perp\!\!\!\perp b \,|\, c\ [P]$. The density p is strictly positive except $\varepsilon = \frac{1}{8}$, which is one of the classic examples of interesting zero-admitting densities. Recall a traditional tale from probabilistic reasoning on how such a distribution can occur. Imagine that variables a and b describe the result of a simultaneous (independent) toss of two fair coins, with outcomes 0 and 1, and a witness rings a bell whenever different outcomes occur. The variable c has the value 1 if the bell rings, otherwise it has the value 0.

Factorization and other equivalent definitions

An elegant characterization of a CI statement is in terms of *factorization*:

$$\exists\, f : \mathsf{X}_{AC} \to \mathbb{R},\ \exists\, g : \mathsf{X}_{BC} \to \mathbb{R} \ \text{ such that}$$
$$\forall\, x \in \mathsf{X}_{ABC} \quad p_{ABC}(x) = f(x_{AC}) \cdot g(x_{BC})\,, \tag{1.5}$$

where the functions f and g are called *potentials*. To show (1.2)⇒(1.5), put $f = p_{AC}$ and $g(z) = \frac{p_{BC}(z)}{p_C(z_C)}$ in the case of $p_C(z_C) > 0$ and $g(z) = 0$ otherwise. To show (1.5)⇒(1.2) introduce marginal potentials $f_C(c) = \sum_{a \in \mathsf{X}_A} f(a,c)$, $g_C(c) = \sum_{b \in \mathsf{X}_B} g(b,c)$ for $c \in \mathsf{X}_C$ and observe by summing in (1.5) that $p_{AC} = f \cdot g_C$, $p_{BC} = f_C \cdot g$ and $p_C = f_C \cdot g_C$. Then substitute these equalities and (1.5) to both sides of (1.2). In comparison with the condition (1.3), the factorization condition (1.5) does not require the potentials to be expressed in terms of marginal densities, which makes (1.5) more suitable for verification.

The concept of CI is often introduced in terms of *conditional densities*. An elegant symmetric definition of CI in these terms is the following one:

$$\forall\, x \in \mathsf{X}_{ABC} \ \text{ such that } p_C(x_C) > 0, \text{ one has}$$
$$p_{AB|C}(x_{AB}|x_C) = p_{A|C}(x_A|x_C) \cdot p_{B|C}(x_B|x_C)\,. \tag{1.6}$$

To see it is equivalent to the previous conditions observe that (1.2)⇒(1.6)⇒(1.3). Nevertheless, the most popular definition in the terms of conditional densities is the next asymmetric one, which basically says that the conditional distribution $P_{A|BC}$ *does not depend on the variables in* B:

$$\forall\, x \in \mathsf{X}_{ABC} \text{ with } p_{BC}(x_{BC}) > 0 \quad p_{A|BC}(x_A|x_{BC}) = p_{A|C}(x_A|x_C)\,. \tag{1.7}$$

One can easily show that (1.2)⇒(1.7)⇒(1.3). The interpretation of the condition (1.7), which is common in the theory of Markov processes, is that the *future* A depends on the *past* B only through the *present* C. Of course, there are lots of modifications of this condition, for example that $p_{A|BC}(*|*)$ only depends on AC, but these modifications are omitted in this chapter.

Example 1.3.3. To illustrate the application of the factorization property, take another example of a discrete distribution with zero-admitting density. Put again $N = \{a,b,c\}$, $\mathsf{X}_i = \{0,1\}$ and introduce a probability measure P on X_N by its density p as follows:

$$p(0,0,0) = p(1,1,1) = \frac{1}{2}, \quad p(x) = 0 \text{ for remaining configurations } x \in \mathsf{X}_N.$$

To verify that $a \perp\!\!\!\perp b \,|\, c\ [P]$ holds using the condition (1.5), introduce functions $f : \mathsf{X}_{ac} \to \mathbb{R}$ and $g : \mathsf{X}_{bc} \to \mathbb{R}$ as follows:

$$f(0,0) = f(1,1) = \frac{1}{2}, \qquad f(0,1) = f(1,0) = 0\,,$$
$$g(0,0) = g(1,1) = 1, \qquad g(0,1) = g(1,0) = 0\,.$$

For $x \in \mathsf{X}_N$, one has $f(x_{ac}) \cdot g(x_{bc}) \neq 0$ iff either $x = (0,0,0)$ or $x = (1,1,1)$ and the value is $\frac{1}{2}$ then. Thus, (1.5) holds and, by symmetry argument, we observe that $i \perp\!\!\!\perp j \,|\, k\ [P]$ is true for any choice of distinct $i,j,k \in N$. On the other hand, one has $p_{ab}(0,0) = p_{ab}(1,1) = \frac{1}{2}$ and $p_{ab}(0,1) = p_{ab}(1,0) = 0$, which allows one to observe, using the cross-exchange condition (1.4), that $a \not\perp\!\!\!\perp b \,|\, \emptyset\ [P]$; hence, by symmetry, $i \not\perp\!\!\!\perp j \,|\, \emptyset\ [P]$ for any distinct $i,j \in N$.

1.3.2 More general CI concepts

This section, to be skipped by beginners, assumes that the reader is familiar with deeper notions of measure theory. Its aim is to explain how probabilistic CI is defined in terms of σ-algebras and how this abstract definition is reduced to the cases of general and marginally continuous probability measures.

A crucial concept is that of *conditional probability*, where the conditioning object is a σ-algebra. Let $\boldsymbol{P}$ be a probability measure on a measurable space $(\mathsf{X}, \mathcal{X})$, $\mathcal{C} \subseteq \mathcal{X}$ a σ-algebra and $\tilde{\mathbb{A}} \in \mathcal{X}$ an event. A version of *conditional probability* of $\tilde{\mathbb{A}}$ given $\mathcal{C}$ (= conditioned by $\mathcal{C}$) is any $\mathcal{C}$-measurable function $h : \mathsf{X} \to [0,1]$, denoted by $\boldsymbol{P}[\tilde{\mathbb{A}}|\mathcal{C}]$, such that

$$\forall\, \tilde{\mathbb{C}} \in \mathcal{C} \qquad \boldsymbol{P}(\tilde{\mathbb{A}} \cap \tilde{\mathbb{C}}) = \int_{\tilde{\mathbb{C}}} h(x)\ \mathrm{d}\boldsymbol{P}(x) \equiv \int_{\tilde{\mathbb{C}}} \boldsymbol{P}[\tilde{\mathbb{A}}|\mathcal{C}](x)\ \mathrm{d}\boldsymbol{P}(x)\,. \tag{1.8}$$

It follows from the Radon-Nikodym theorem that such a function h exists and is unique in the sense of $\boldsymbol{P}_{\mathcal{C}}$-everywhere equality, where $\boldsymbol{P}_{\mathcal{C}}$ denotes the restriction of $\boldsymbol{P}$ to the measurable space $(\mathsf{X}, \mathcal{C})$. One can introduce the concept of CI for σ-algebras as follows: given σ-algebras $\mathcal{A}, \mathcal{B}, \mathcal{C} \subseteq \mathcal{X}$, we say that $\mathcal{A}$ and $\mathcal{B}$ are *conditionally independent given* $\mathcal{C}$ and write $\mathcal{A} \perp\!\!\!\perp \mathcal{B} \,|\, \mathcal{C}$ if

$$\begin{aligned} &\forall\, \tilde{\mathbb{A}} \in \mathcal{A}\ \ \forall\, \tilde{\mathbb{B}} \in \mathcal{B} \\ &\qquad \boldsymbol{P}[\tilde{\mathbb{A}} \cap \tilde{\mathbb{B}}|\mathcal{C}](x) = \boldsymbol{P}[\tilde{\mathbb{A}}|\mathcal{C}](x) \cdot \boldsymbol{P}[\tilde{\mathbb{B}}|\mathcal{C}](x) \quad \text{for } \boldsymbol{P}_{\mathcal{C}}\text{-a.e } x \in \mathsf{X}\,. \end{aligned} \tag{1.9}$$

Note that the validity of (1.9) does not depend on the choice of particular versions of conditional probabilities; its equivalent formulation is the condition

$$\forall\, \tilde{\mathbb{A}} \in \mathcal{A} \quad \text{there exists } \mathcal{C}\text{-measurable version of } \boldsymbol{P}[\tilde{\mathbb{A}}|\mathcal{B} \vee \mathcal{C}]\,,$$

where $\mathcal{B} \vee \mathcal{C}$ is the σ-algebra generated by $\mathcal{B} \cup \mathcal{C}$; see [57, Lemma A.6]. This condition can be interpreted as an analogue of the discrete condition (1.7).

Let us now describe how the CI definition (1.9) works in the case of a (general) *probability measure P over N* mentioned in Definition 1.2.1. In this case, we put $(\mathsf{X}, \mathcal{X}) := (\mathsf{X}_N, \bigotimes_{i \in N} \mathcal{X}_i)$, $\boldsymbol{P} := P$. Recall from § 1.2.2 that, for $A \subseteq N$, $\mathcal{X}_A \equiv \bigotimes_{i \in A} \mathcal{X}_i$ denotes the product σ-algebra on X_A, with $\mathcal{X}_\emptyset \equiv \{\emptyset, \mathsf{X}_\emptyset\}$. It can be ascribed the respective *coordinate σ-algebra* $\mathcal{A} := \{\mathbb{A} \times \mathsf{X}_{N \setminus A} : \mathbb{A} \in \mathcal{X}_A\}$ of subsets of $\mathsf{X} = \mathsf{X}_N$; one then has $\mathcal{A} \subseteq \mathcal{X}$.

Given disjoint $A, C \subseteq N$, let $\mathcal{C}$ denote the coordinate σ-algebra for $\mathcal{X}_C$. Any event $\mathbb{A} \in \mathcal{X}_A$ can be ascribed its cylindrical extension $\tilde{\mathbb{A}} := \mathbb{A} \times \mathsf{X}_{N \setminus A}$; the conditional probability $x \in \mathsf{X}_N \mapsto \boldsymbol{P}[\tilde{\mathbb{A}}|\mathcal{C}](x)$ then depends on x_C and can be identified with an $\mathcal{X}_C$-measurable function on X_C, to be denoted by $c \in \mathsf{X}_C \mapsto P_{A|C}(\mathbb{A}|c)$. Thus, (1.8) allows one to introduce the concept of *conditional probability on X_A given C* as a function $P_{A|C} : \mathcal{X}_A \times \mathsf{X}_C \to [0,1]$ of two arguments such that, for any $\mathbb{A} \in \mathcal{X}_A$, the function $c \in \mathsf{X}_C \mapsto P_{A|C}(\mathbb{A}|c)$ is $\mathcal{X}_C$-measurable and satisfies

$$P_{AC}(\mathbb{A} \times \mathbb{C}) = \int_{\mathbb{C}} P_{A|C}(\mathbb{A}|c)\ \mathrm{d}P_C(c) \quad \text{for any } \mathbb{C} \in \mathcal{X}_C\,.$$

Observe that this is a natural generalization of the concept from Definition 1.2.3. Given pairwise disjoint $A, B, C \subseteq N$, the condition $\mathcal{A} \perp\!\!\!\perp \mathcal{B} \,|\, \mathcal{C}$ from (1.9) turns into the requirement

$$\begin{aligned} &\forall\, \mathbb{A} \in \mathcal{X}_A\ \ \forall\, \mathbb{B} \in \mathcal{X}_B \\ &\qquad P_{AB|C}(\mathbb{A} \times \mathbb{B}|c) = P_{A|C}(\mathbb{A}|c) \cdot P_{B|C}(\mathbb{B}|c) \quad \text{for } P_C\text{-a.e. } c \in \mathsf{X}_C \end{aligned}$$

which directly generalizes (1.1) and can be considered as a definition of the CI statement $A \perp\!\!\!\perp B \,|\, C\ [P]$ in the case of a (general) measure P over N.

In the case of a *marginally continuous* measure P over N (see § 1.2.2) one can introduce CI in terms of marginal densities. Specifically, it was shown in [57, Lemma 2.4] that, provided a dominating system of measures μ^i on $(\mathsf{X}_i, \mathcal{X}_i)$, $i \in N$, is fixed, one has $A \perp\!\!\!\perp B \,|\, C \; [P]$ for pairwise disjoint $A, B, C \subseteq N$ iff

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

where f_D, $D \subseteq N$, denotes the marginal density for D. This condition generalizes (1.2) and one can also generalize the other equivalent conditions from § 1.3.1 in terms of densities. For example, (1.5) takes the form: there exist $\mathcal{X}_{AC}$-measurable $h : \mathsf{X}_{AC} \to \mathbb{R}$ and $\mathcal{X}_{BC}$-measurable $g : \mathsf{X}_{BC} \to \mathbb{R}$ such that

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

1.4 Basic properties of conditional independence

In this section, we introduce (probabilistic) CI structures and recall their basic formal properties. We also relate formal CI models to classic statistical models.

1.4.1 Conditional independence structure

A *disjoint triplet over* N is an ordered triplet $A, B, C \subseteq N$ of pairwise disjoint subsets of N. Notation $\langle A, B|C\rangle$ will be used to indicate the intended interpretation of such a triplet as a formal statement that the variables in A are independent of/dependent on the variables in B conditionally the variables in C. The system of all disjoint triplets over N will be denoted by $\mathcal{T}(N)$.

A *formal independence model over* N is a subset $\mathcal{M}$ of $\mathcal{T}(N)$, whose elements are interpreted as independence statements. We write $A \perp\!\!\!\perp B \,|\, C \; [\mathcal{M}]$ to indicate that $\langle A, B|C\rangle \in \mathcal{M}$ is interpreted as an independence statement and $A \not\perp\!\!\!\perp B \,|\, C$ if $\langle A, B|C\rangle$ is interpreted as a dependence statement.

The *conditional independence structure* induced by a probability measure P over N is a formal independence model (over N) composed of those triplets which represent valid CI statements with respect to P:

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

Not every formal independence model is a CI structure. The next proposition presents basic formal properties of CI structures.

Observation 1.4.1. Let P be a probability measure over N. Then one has for (pairwise disjoint) $A, B, C, D \subseteq N$:

(i) $\emptyset \perp\!\!\!\perp B \,|\, C \; [P]$,

(ii) $A \perp\!\!\!\perp B \,|\, C \; [P] \;\Leftrightarrow\; B \perp\!\!\!\perp A \,|\, C \; [P]$,

(iii) $A \perp\!\!\!\perp BD \,|\, C \; [P] \;\Leftrightarrow\; \{\, A \perp\!\!\!\perp D \,|\, C \; [P] \;\;\&\;\; A \perp\!\!\!\perp B \,|\, DC \; [P] \,\}$.

Moreover, if P has a strictly positive density then

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

Recall that a discrete measure P on X_N has a (strictly) positive density if $p(x) > 0$ for all $x \in \mathsf{X}_N$. In the general case (see § 1.2.2) a measure P over N has a positive density if it is marginally continuous and a dominating system μ^i, $i \in N$, of σ-finite measures exists such that $\mu \equiv \bigotimes_{i \in N} \mu^i \ll P$.

The proof of Observation 1.4.1 given below is intentionally restricted to the discrete case so that beginners can understand it fully. A reader familiar with calculus of densities can modify the proof to cover the marginally continuous case (see § 1.2.2), provided that he/she is familiar with peculiarities of the almost-everywhere equality of densities. Nevertheless, in a general case of CI in terms of σ-algebras (see § 1.3.2), deeper measure-theoretical considerations are needed to derive the result; see [57, § A.7].

Proof. We assume the discrete case throughout the proof. To verify (i), we use (1.1) and realize that in the case of $A = \emptyset$ one has either $\mathbb{A} = \emptyset = \mathbb{A} \times \mathbb{B}$ or $\{\, \mathbb{A} \neq \emptyset \;\&\; \mathbb{A} \times \mathbb{B} = \mathbb{B} \,\}$. The condition (ii) is evident. To verify (iii), we combine (1.2) and (1.3). For the implication $A \perp\!\!\!\perp BD \,|\, C \;\Rightarrow\; A \perp\!\!\!\perp D \,|\, C$, we use (1.2): the summation over B-configurations in $p_C \cdot p_{ABDC} = p_{AC} \cdot p_{BDC}$ gives $p_C \cdot p_{ADC} = p_{AC} \cdot p_{DC}$. As concerns $A \perp\!\!\!\perp BD \,|\, C \Rightarrow A \perp\!\!\!\perp B \,|\, DC$, we multiply the above-derived equalities (the latter with swapped sides) to get

$$p_C \cdot p_{ABDC} \cdot p_{AC} \cdot p_{DC} = p_{AC} \cdot p_{BDC} \cdot p_C \cdot p_{ADC} \,.$$

Because canceling is possible here for positive $ABDC$-configurations, one gets

$$\forall\, p_{ABDC} > 0 \quad p_{ABDC} \cdot p_{DC} = p_{ADC} \cdot p_{BDC} \,,$$

which is, by (1.3), $A \perp\!\!\!\perp B \,|\, DC$. The proof of the converse, that is, the implication $\{\, A \perp\!\!\!\perp D \,|\, C \;\&\; A \perp\!\!\!\perp B \,|\, DC \,\} \;\Rightarrow A \perp\!\!\!\perp BD \,|\, C$, is analogous.

To verify $\{\, A \perp\!\!\!\perp B \,|\, DC \;\&\; A \perp\!\!\!\perp D \,|\, BC \,\} \;\Rightarrow A \perp\!\!\!\perp BD \,|\, C$ in (iv), we use (1.3) for both CI statements and get by canceling (because of $p_{BDC} > 0$):

$$\frac{p_{ADC} \cdot p_{BDC}}{p_{DC}} \;=\; p_{ABDC} \;=\; \frac{p_{ABC} \cdot p_{BDC}}{p_{BC}} \quad\Rightarrow\quad \frac{p_{ADC}}{p_{DC}} = \frac{p_{ABC}}{p_{BC}} \,.$$

Choose and fix a configuration $b \in \mathsf{X}_B$ and write

$$\forall\, [a,d,c] \in \mathsf{X}_{ADC} \qquad p_{A|DC}(a|d,c) = \frac{p_{ADC}(a,d,c)}{p_{DC}(d,c)} = \frac{p_{ABC}(a,b,c)}{p_{BC}(b,c)} \,,$$

which means that $p_{A|DC}$ does not depend on $d \in \mathsf{X}_D$. By condition (1.7), one has $A \perp\!\!\!\perp D \,|\, C\ [P]$. By (iii), this together with $A \perp\!\!\!\perp B \,|\, DC\ [P]$ implies $A \perp\!\!\!\perp BD \,|\, C\ [P]$. □

Note that the property in Observation 1.4.1(iv), called *intersection*, need not be valid for a discrete distribution P which is not strictly positive. Indeed, Example 1.3.3 shows that one can have $i \perp\!\!\!\perp j \,|\, k\ [P]$ for all distinct $i, j, k \in N$, while $i \not\perp\!\!\!\perp j \,|\, \emptyset\ [P]$; the latter implies $i \not\perp\!\!\!\perp \{j,k\} \,|\, \emptyset\ [P]$. Chapter 3 in this book analyzes the validity of the intersection property in more details.

Side-remark about relational databases

Formal independence models satisfying the conditions (i)-(iii) from Observation 1.4.1 occur also outside of statistics. There is one area in computer science where a concept analogous to the concept of probabilistic CI has been studied. It is the theory of *relational databases*, which is approximately 15 years older than probabilistic reasoning. The problem there is how to efficiently organize data in large data banks [7]. Researchers in this area became interested in special concepts of functional and multi-valued dependencies (in databases),

which allowed them to reduce the memory demands, and tried to axiomatize them [3]. Moreover, there is a concept of *embedded multivalued dependency* (EMVD) [44], which is completely analogous to probabilistic CI and exhibits similar formal properties.

In this theory, the elements of N are called *attributes*, and each attribute $i \in N$ is ascribed a finite (individual) sample space X_i of possible values. A *relational database over* N is simply a set of configurations over N.

One can introduce natural operations with relational databases, some of which were already mentioned in § 1.2.1. Given $A \subseteq B \subseteq N$ and a relational database $\mathbb{D} \subseteq \mathsf{X}_B$ over B, the *projection* of $\mathbb{D}$ onto A is a relational database over A defined by $\mathbb{D}_A := \{b_A : b \in \mathbb{D}\}$. The second important operation is that of a combination, which is an analogue of the operation of conditional product for discrete probability measures from Observation 1.3.1. Specifically, given a disjoint triplet $\langle A, B|C\rangle$ over N and databases $\mathbb{D}^1 \subseteq \mathsf{X}_{AC}$, $\mathbb{D}^2 \subseteq \mathsf{X}_{BC}$ its *combination* is a relational database over ABC defined as follows:

$$\mathbb{D}^1 \bowtie \mathbb{D}^2 \;:=\; \{\, [a,b,c] \in \mathsf{X}_{ABC} \;:\; [a,c] \in \mathbb{D}^1 \;\&\; [b,c] \in \mathbb{D}^2 \,\}\,.$$

There is an analogy of the CI concept: given $\langle A, B|C\rangle \in \mathcal{T}(N)$ and a database $\mathbb{D}$ over N, we say that an *embedded multivalued dependency* (EMVD) statement $A \perp\!\!\!\perp B \,|\, C$ $[\mathbb{D}]$ holds if $\mathbb{D}_{ABC} = \mathbb{D}_{AC} \bowtie \mathbb{D}_{BC}$, in words, if the projection of $\mathbb{D}$ onto ABC is the combination of its projections onto AC and BC.

We leave it to the reader to verify that the formal independence model induced by $\mathbb{D}$ satisfies the conditions (i)-(iii) from Observation 1.4.1.

1.4.2 Statistical model of a CI structure

The aim of this section is to explain that formal independence models can be interpreted as common statistical models. Recall that by a (mathematical) *statistical model* is meant a class of probability measures $\mathbb{M}$ on a prescribed sample space, which is a measurable space $(\mathsf{X}, \mathcal{X})$. In multivariate statistical analysis, one usually has a *joint sample space* $(\mathsf{X}_N, \mathcal{X}_N)$ in the place of $(\mathsf{X}, \mathcal{X})$.

Typically, a statistical model $\mathbb{M}$ is a parameterized class of measures and all of them are absolutely continuous with respect to some given σ-finite measure μ on $(\mathsf{X}, \mathcal{X})$, which is a product measure $\mu = \bigotimes_{i\in N} \mu^i$ in the case of $(\mathsf{X}_N, \mathcal{X}_N)$. Each probability measure in $\mathbb{M}$ is then determined by its density with respect to μ and, quite often, they are assumed to be mutually absolutely continuous. The parameters usually belong to a convex subset $\Theta \subseteq \mathbb{R}^n$ for some $n \geq 1$.

Assume that a *distribution framework* is specified, that is, a collection Ψ of probability measures on the sample space is determined from which the probability measures in $\mathbb{M}$ should be chosen. For example, in the discrete case, Ψ could be the class of all measures with positive density, while in the continuous case with $\mathsf{X}_i = \mathbb{R}$ for $i \in N$, one can have the class of regular Gaussian distributions on $\mathbb{R}^N$ in the place of Ψ. Then, every formal independence model $\mathcal{M} \subseteq \mathcal{T}(N)$ over N can be ascribed a class of probability measures

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

which can be called the *statistical model of CI structure* given by $\mathcal{M}$.

This concept generalizes the classic concept of a *graphical model* [68, 24]. Indeed, the reader can learn in § 1.7.1 that every UG G over N induces the class $\mathbb{M}_G$ of Markovian measures over N through a formal independence model $\mathcal{M}_G$ induced by G. In general, statistical models of CI structures are very complicated; however, graphical models provide a subclass of nice models.

1.5 Semi-graphoids, graphoids, and separoids

The notions discussed in this section have been inspired by the research on stochastic CI, but they rather belong to the area of discrete mathematics. Pearl and Paz [43] introduced the following concept in 1987.

Definition 1.5.1. A *disjoint semi-graphoid over* N is a formal independence model $\mathcal{M}$ over N satisfying the following conditions/axioms:

$\emptyset \perp\!\!\!\perp B \,|\, C \; [\mathcal{M}]$ — triviality,
$A \perp\!\!\!\perp B \,|\, C \; [\mathcal{M}] \;\Rightarrow\; B \perp\!\!\!\perp A \,|\, C \; [\mathcal{M}]$ — symmetry,
$A \perp\!\!\!\perp BD \,|\, C \; [\mathcal{M}] \;\Rightarrow\; A \perp\!\!\!\perp B \,|\, DC \; [\mathcal{M}]$ — weak union,
$A \perp\!\!\!\perp BD \,|\, C \; [\mathcal{M}] \;\Rightarrow\; A \perp\!\!\!\perp D \,|\, C \; [\mathcal{M}]$ — decomposition,
$A \perp\!\!\!\perp D \,|\, C \; [\mathcal{M}] \;\&\; A \perp\!\!\!\perp B \,|\, DC \; [\mathcal{M}] \;\Rightarrow\; A \perp\!\!\!\perp BD \,|\, C \; [\mathcal{M}]$ — contraction.

A disjoint semi-graphoid $\mathcal{M}$ will be called a *graphoid* (over N) if it satisfies

$A \perp\!\!\!\perp B \,|\, DC \; [\mathcal{M}] \;\&\; A \perp\!\!\!\perp D \,|\, BC \; [\mathcal{M}] \;\Rightarrow\; A \perp\!\!\!\perp BD \,|\, C \; [\mathcal{M}]$ — intersection.

Given $\mathcal{M} \subseteq \mathcal{T}(N)$, its *semi-graphoid closure* is the smallest semi-graphoid over N containing $\mathcal{M}$. The *graphoid closure of* $\mathcal{M}$ can be introduced analogously.

Semi/graphoid closures are well defined because set intersection of semi/graphoids over N is a semi/graphoid over N. The CI implications in Definition 1.5.1 are nothing else but detailed conditions from Observation 1.4.1, which basically says that every probabilistic CI structure is a disjoint semi-graphoid, and even a graphoid if the distribution has a positive density.

There are areas different from probability theory in which semi-graphoids have occurred. We have seen in the side-remark from § 1.4.1 that every relational database can be ascribed a disjoint semi-graphoid. The undirected separation criterion from § 1.7.1 allows one to ascribe a graphoid to every UG over N and the same holds for the directional separation criterion from § 1.8.1. Let us give three more examples; their verification is left to the reader.

A class of subsets: take $\mathcal{T} \subseteq \mathcal{P}(N) \equiv \{A \,:\, A \subseteq N\}$ and define

$$A \perp\!\!\!\perp B \,|\, C \; [\mathcal{T}] \;:=\; \forall\, T \in \mathcal{T} \quad T \subseteq ABC \;\Rightarrow\; [\,T \subseteq AC \text{ or } T \subseteq BC\,].$$

A natural conditional function: given a finite joint sample space X_N,
this is a function $\kappa : \mathsf{X}_N \to \mathbb{Z}$ such that $\min\{\,\kappa(x) \,:\, x \in \mathsf{X}_N\} = 0$.
Introduce a marginal (function) for any $A \subseteq N$ by the formula:
$\kappa_A(y) := \min\{\,\kappa(y,z) \,:\, z \in \mathsf{X}_{N\setminus A}\,\}$ for any $y \in \mathsf{X}_A$. Define

$$\begin{aligned} A \perp\!\!\!\perp B \,|\, C \; [\kappa] \;&:=\; \forall\, x \in \mathsf{X}_N \\ &\kappa_C(x_C) + \kappa_{ABC}(x_{ABC}) = \kappa_{AC}(x_{AC}) + \kappa_{BC}(x_{BC})\,. \end{aligned}$$

Note that this is a concept taken over from [51].

A supermodular function: this is a set function $m : \mathcal{P}(N) \to \mathbb{R}$ such that $m(D \cup E) + m(D \cap E) \geq m(D) + m(E)$ for all $D, E \subseteq N$. Define

$$A \perp\!\!\!\perp B \,|\, C \; [m] \;:=\; m(C) + m(ABC) = m(AC) + m(BC)\,.$$

Note that semi-graphoids defined in this way coincide with *structural semi-graphoids* mentioned in § 1.10.1.

Some authors do not regard the restriction to disjoint triplets over N as necessary and consider a *general semi-graphoid* over N, which is a set of ordered triplets $A \perp\!\!\!\perp B \,|\, C$ of (not necessarily disjoint) subsets of N, which satisfies the following three conditions:

- $B \subseteq C \;\Rightarrow\; A \perp\!\!\!\perp B \,|\, C$,
- $A \perp\!\!\!\perp B \,|\, C \;\Leftrightarrow\; B \perp\!\!\!\perp A \,|\, C$,
- $A \perp\!\!\!\perp B \cup D \,|\, C \;\Leftrightarrow\; \{\, A \perp\!\!\!\perp D \,|\, C \;\&\; A \perp\!\!\!\perp B \,|\, D \cup C \,\}$.

A general semi-graphoid is induced by a discrete probability measure P over N through the condition (1.2) where non-disjoint triplets are allowed. Then $A \perp\!\!\!\perp A \,|\, C\ [P]$ means that $\forall\, p_{AC} > 0$ one has $p_{AC} = p_C$, which corresponds to *functional dependency of A on C*; note that an axiomatic characterization of probabilistic functional dependency structures was given by Matúš [30]. Thus, general semi-graphoids are broader than disjoint semi-graphoids because they involve functional dependency relation modeling.

Dawid took an even more general point of view and introduced an abstract concept of a separoid; below we describe a simplification of his definition [13].

Definition 1.5.2. Let $\mathbb{S}$ be a joint semi-lattice, that is, a partially ordered set in which every two elements a, b have a supremum (= a join), denoted by $a \vee b$. A set of ordered triplets $a \perp\!\!\!\perp b \,|\, c$ of elements of $\mathbb{S}$ will be called a *separoid* if

- $b \vee c = c \;\Rightarrow\; a \perp\!\!\!\perp b \,|\, c$,
- $a \perp\!\!\!\perp b \,|\, c \;\Leftrightarrow\; b \perp\!\!\!\perp a \,|\, c$,
- $a \perp\!\!\!\perp b \vee d \,|\, c \;\Leftrightarrow\; \{\, a \perp\!\!\!\perp d \,|\, c \;\;\&\;\; a \perp\!\!\!\perp b \,|\, d \vee c \,\}$.

Of course, every general semi-graphoid over N is a separoid on the lattice $(\mathcal{P}(N), \subseteq)$. Another prominent example requires for the reader to be familiar with measure theory: given a probability measure $\boldsymbol{P}$ on a measurable space $(\mathsf{X}, \mathcal{X})$, let $\mathbb{S}$ be the set of all σ-algebras contained in $\mathcal{X}$, ordered by inclusion. Then the ternary relation $\mathcal{A} \perp\!\!\!\perp \mathcal{B} | \mathcal{C}$ introduced in § 1.3.2 is a separoid.

1.5.1 Elementary and dominant triplets

To represent a (disjoint) semi-graphoid over N in the memory of a computer, one does not need all $|\mathcal{T}(N)| = 4^{|N|}$ bits.

Definition 1.5.3. A disjoint triplet $\langle A, B|C\rangle$ over N will be called *trivial* if either $A = \emptyset$ or $B = \emptyset$; it will be called *elementary* if $|A| = 1 = |B|$. The system of elementary triplets over N will be denoted by $\mathcal{T}_\epsilon(N)$.

Clearly, the trivial triplets can always be excluded from considerations because they are contained in any semi-graphoid. On the other hand, the elementary triplets are substantial because of the following fact.

Observation 1.5.1. Let $\mathcal{M}$ be a disjoint semi-graphoid over N. Then, for every disjoint triplet $\langle A, B|C\rangle \in \mathcal{T}(N)$, one has $A \perp\!\!\!\perp B \,|\, C\ [\mathcal{M}]$ iff

$$\forall\, i \in A \quad \forall\, j \in B \quad \forall\, K \text{ with } C \subseteq K \subseteq ABC \setminus \{i, j\} \qquad i \perp\!\!\!\perp j \,|\, K\ [\mathcal{M}]\,. \tag{1.10}$$

In particular, for two disjoint semi-graphoids $\mathcal{M}^1$ and $\mathcal{M}^2$ over N, one has $\mathcal{M}^1 \subseteq \mathcal{M}^2$ iff $\mathcal{M}^1 \cap \mathcal{T}_\epsilon(N) \subseteq \mathcal{M}^2 \cap \mathcal{T}_\epsilon(N)$, which implies that any semi-graphoid $\mathcal{M}$ is uniquely determined by its elementary trace $\mathcal{M} \cap \mathcal{T}_\epsilon(N)$.

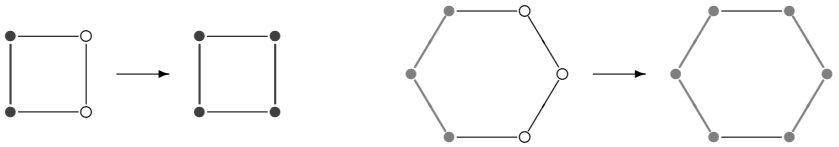


FIGURE 1.2: Illustration of the square and hexagon axioms.

Proof. The necessity of (1.10) can be easily derived using the decomposition and weak union properties combined with the symmetry property. For the converse implication, suppose that $\langle A, B|C\rangle$ is not trivial and use induction on $|AB|$; the instance $|AB| = 2$ is evident. Supposing $|AB| > 2$ either A or B is not a singleton. Owing to the symmetry property, one can – without the loss of generality – consider $|B| \geq 2$, choose $b \in B$ and put $B' = B \setminus \{b\}$. By the induction assumption, the condition (1.10) implies both $A \perp\!\!\!\perp B' \,|\, C$ $[\mathcal{M}]$ and $A \perp\!\!\!\perp b \,|\, B'C$ $[\mathcal{M}]$. Thus, the contraction property gives $A \perp\!\!\!\perp B \,|\, C$ $[\mathcal{M}]$. □

One can also easily show that $\mathcal{N} \subseteq \mathcal{T}_\epsilon(N)$ is a trace of a semi-graphoid iff the *symmetry* condition $i \perp\!\!\!\perp j \,|\, K$ $[\mathcal{N}] \Leftrightarrow j \perp\!\!\!\perp i \,|\, K$ $[\mathcal{N}]$ and the *exchange* property $i \perp\!\!\!\perp j \,|\, kL$ $[\mathcal{N}]$ & $i \perp\!\!\!\perp k \,|\, L$ $[\mathcal{N}] \Leftrightarrow i \perp\!\!\!\perp k \,|\, jL$ $[\mathcal{N}]$ & $i \perp\!\!\!\perp j \,|\, L$ $[\mathcal{N}]$ hold. Thus, the semi-graphoid closure can be described in terms of elementary triplets. Since $|\mathcal{T}_\epsilon(N)| = |N| \cdot (|N| - 1) \cdot 2^{|N|-2}$ it is enough to have $\binom{|N|}{2} \cdot 2^{|N|-2}$ bits to represent a semi-graphoid over N.

Matúš [35] was interested in the intricacy of the semi-graphoid inference between elementary CI statements and showed that the length of the derivation sequence can be exponential in $|N|$. Nonetheless, there is an alternative way to represent semi-graphoids in the memory of a computer.

Definition 1.5.4. We say that $\langle A, B|C\rangle \in \mathcal{T}(N)$ *dominates* $\langle A', B'|C'\rangle \in \mathcal{T}(N)$ if $A' \subseteq A$, $B' \subseteq B$ and $C \subseteq C' \subseteq ABC$. The triplets in a semi-graphoid which are maximal with respect to this partial order on $\mathcal{T}(N)$ are called *dominant*.

If one restricts oneself to non-trivial triplets then elementary triplets in a (fixed) semi-graphoid $\mathcal{M}$ are minimal with respect to the dominance ordering; thus, the dominant and elementary triplets are somehow opposite to each other. An alternative way to represent a semi-graphoid in the memory of a computer is by a list of its non-trivial (symmetrized) dominant triplets.

One can also implement the semi-graphoid and graphoid closures in these terms, as shown by Baioletti, Busanello and Vantaggi [2]. Dominant triplets were also employed as an useful tool in [55] to show that the semi-graphoid closure of two disjoint triplets over N is always a probabilistic CI structure. This fact can be interpreted as a result on *relative completeness* of semi-graphoid implications for probabilistic CI inference if the input list has at most 2 items (see § 1.11). Semi/graphoids over a fixed set N can also be classified according to their *semi/graphoid complexity*, by which we mean the minimal cardinality of a semi/graphoid generator [56].

For readers familiar with (advanced) polyhedral geometry, we mention two interesting equivalent geometric definitions/interpretations of the concept of a semi-graphoid, which were offered by Morton and his co-authors [37, 38]. Both equivalent geometric definitions come from the semi-graphoid description in terms of elementary triplets.

The first equivalent definition is related to a special polytope, called a *permutohedron*, which was previously introduced by Shouté in 1911 [46]. The idea is that all permutations over a set $N = \{1, 2, \ldots, n\} \equiv [n]$ are interpreted as vectors in $\mathbb{R}^N$ and their convex

hull is taken. There is a certain standard way to label one-dimensional faces (= geometric edges) of this polytope by elementary triplets over N. Thus, $\mathcal{N} \subseteq \mathcal{T}_\epsilon(N)$ is identified with a set of geometric edges of the permutohedron. The two above-mentioned conditions on $\mathcal{N}$ characterizing a semi-graphoid then have an elegant geometric interpretation. Every two-dimensional face of the permutohedron is either a square or a regular hexagon. The symmetry condition can be interpreted as a *square axiom* requiring that if a geometric edge of a square belongs to $\mathcal{N}$ then so does the opposite edge. The exchange property corresponds to a *hexagon axiom*, which says that if a pair of touching edges of a hexagon belongs to $\mathcal{N}$ then the same holds for the pair of edges opposite to them in the hexagon; see Figure 1.2.

The second equivalent definition is in terms of (complete) *polyhedral fans*, which are certain collections of polyhedral cones covering $\mathbb{R}^N$. There is a prominent polyhedral fan induced by a special equivalence of vectors in $\mathbb{R}^N$, where $u, v \in \mathbb{R}^N$ are equivalent if $\forall\, i, j \in N$ one has $u_i \leq u_j \Leftrightarrow v_i \leq v_j$. That fan is called the S_n*-fan* (for $n = |N|$) by Morton [38] or *braid arrangement* by other authors. Semi-graphoids are then in a one-to-one correspondence with polyhedral fans which coarsen the prominent S_n-fan.

1.6 Elementary graphical concepts

In this section we introduce basic graphical concepts to be used in the following three sections. Recall from § 1.2 that N is a generic symbol for a finite non-empty index set whose elements correspond to random variables and occur as nodes of graphs in a graphical context.

By a graph *over* N we will understand a graph which has N as the set of *nodes*. Graphs considered in this chapter have no multiple edges; and there are two possible types of their edges.

Undirected edges are unordered pairs of distinct nodes, that is, two-element subsets of N. We will write $i - j$ to denote an undirected edge between nodes i and j from N; a pictorial representation is analogous. An *undirected graph* (UG) is a graph whose edges are all undirected; if $i - j$ in an undirected graph G then we say that i and j are *neighbors* in G. The symbol $\mathrm{ne}_G(i) := \{j \in N : i - j \text{ in } G\}$ will denote the set of all neighbors of $i \in N$ in G. A set of nodes $A \subseteq N$ is *complete* in an UG G if $i - j$ in G is true for all distinct $i, j \in A$. Maximal complete sets in G with respect to the set-inclusion ordering are called *cliques* of G. An UG G over N is *complete* if N is complete in G.

Directed edges, also called *arrows*, are ordered pairs of distinct nodes. We will write $i \to j$ to denote an arrow from node i to node j in N; similarly in figures. A *directed graph* is a graph whose edges are all arrows. If $i \to j$ in a directed graph G then we say that i is a *parent* of j in G or, dually, that j is a *child* of i. The symbol $\mathrm{pa}_G(j) := \{i \in N : i \to j \text{ in } G\}$ will denote the set of all parents of $j \in N$ in G.

Given a graph G over N (either directed or undirected) and a non-empty set of nodes $T \subseteq N$, the *induced subgraph* of G for T, denoted by G_T, is a graph over T with just those edges in G which run between nodes of T.

A *walk* in a graph G over N (either directed or undirected) is a sequence of nodes $i_1, \ldots, i_k$, $k \geq 1$, such that each consecutive pair of nodes in the sequence is adjacent by an edge in the graph G. The end-nodes of the walk are i_1 and i_k; if $k \geq 3$ then the remaining nodes i_ℓ, $1 < \ell < k$, are *internal nodes*. The number of edges in the walk, that is, $k - 1$, is called the *length* of the walk. A walk in G is called a *path* if $i_1, \ldots, i_k$ are distinct; it is called a *cycle* if $k \geq 4$, $i_1 = i_k$ and $i_1, \ldots, i_{k-1}$ are distinct. In the case of a directed graph G, a path or a cycle is called *directed* if $i_\ell \to i_{\ell+1}$ for $\ell = 1, \ldots, k-1$.

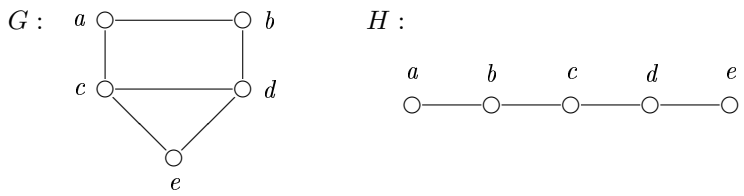


FIGURE 1.3: Two examples of undirected graphs.

A directed graph G is called *acyclic* if it contains no directed cycle. Directed graphs that are acyclic are conventionally called *directed acyclic graphs* (DAGs). A well-known equivalent characterization of a DAG is that it is such a directed graph G which admits an enumeration of nodes $i_1, \ldots, i_{|N|}$ which is *consonant* with the direction of arrows: that is, if $i_\ell \to i_k$ in G then $\ell < k$.

An important concept is that of a *chordal* (undirected) graph. It is an UG G such that each cycle in G of the length at least 4 has a *chord*, that is, an edge between nodes in the cycle which is not an edge forming the cycle. A well-known equivalent definition of a chordal graph G is that the cliques of G can be ordered into a sequence $C_1, \ldots, C_m$, $m \geq 1$, satisfying the *running intersection property*: $\forall\, k \geq 2$ exists $\ell < k$ such that $C_k \cap (\bigcup_{r<k} C_r) \subseteq C_\ell$.

1.7 Markov properties for undirected graphs

This section contains some theoretical results concerning undirected graphical models, called *Markov networks* in the context of probabilistic reasoning [42].

1.7.1 Global Markov property for an UG

Given an undirected graph G over N and a disjoint triplet $\langle A, B|C\rangle \in \mathcal{T}(N)$, we say that A and B are *separated* by C in G and write $A \perp\!\!\!\perp B \,|\, C\ [G]$ if every walk in G from a node in A to a node in B contains a node in C. Of course, this is equivalent to an identical condition with paths in place of walks. Another formulation is that after the removal of the set of nodes in C (including the edges leading to those nodes) there is no path between A and B; that is, no connected component of the induced graph $G_{N\setminus C}$ meets both A and B.

To illustrate the concept of (undirected graphical) separation, consider the graphs G and H in Figure 1.3. Clearly $A = \{a\}$ and $B = \{e\}$ are not separated by $C = \{c\}$ in G because of the path $a - b - d - e$, which avoids $C = \{c\}$. But they are separated by $C = \{c, d\}$, which means that $a \perp\!\!\!\perp e \,|\, cd\ [G]$. One can also easily observe that $a \perp\!\!\!\perp e \,|\, cd\ [H]$.

Every undirected graph G over N induces a formal independence model over N by means of the *undirected separation* criterion

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

which appears to be a (disjoint) graphoid. A probability measure P over N with $\mathcal{M}_G \subseteq \mathcal{M}_P$ is then called *Markovian* with respect to G; in an alternative terminology, P satisfies the *global Markov property* relative to G:

(G) if A and B are separated by C in G then $A \perp\!\!\!\perp B \,|\, C\ [P]$.

The (statistical) *undirected graphical model* $\mathbb{M}_G$ then consists of Markovian distributions with respect to G. As explained in § 1.4.2, the class $\mathbb{M}_G$ can be interpreted as the statistical model of the CI structure given by $\mathcal{M}_G$.

A probability measure P over N is called *perfectly Markovian* with respect to G if $\mathcal{M}_G = \mathcal{M}_P$. The existence of a discrete perfectly Markovian measure with respect to any given UG G was shown by Geiger and Pearl in [19, Theorem 11]. In particular, $\mathcal{M}_G$ is indeed a probabilistic CI structure for any UG G and the statistical model $\mathbb{M}_G$ is non-empty (in the case of non-degenerate sample spaces X_i, $i \in N$). Another related result says that formal independence models induced by UGs can be described in an axiomatic way, that is, they are characterized in terms of finitely many implications [43].

1.7.2 Local and pairwise Markov properties for an UG

Verification whether a probability measure over N is Markovian with respect to an UG over N can be difficult because of the number of CI statements to be tested, which may be very high. Nevertheless, in the case of a measure with a (strictly) positive density certain reasonable sufficient conditions exist.

We say that a probability measure P over N satisfies the *local/pairwise Markov property* relative to G if

(L) for all $i \in N$ $\quad i \perp\!\!\!\perp N \setminus (i \cup \mathrm{ne}_G(i)) \,|\, \mathrm{ne}_G(i)$ $[P]$,

(P) for all distinct $i, j \in N$ with $\neg(i - j \text{ in } G)$ $\quad i \perp\!\!\!\perp j \,|\, N \setminus \{i, j\}$ $[P]$.

It is easy to verify, using Observation 1.4.1(iii), that (G)⇒(L)⇒(P); however, examples are available in which (P)⇏(L)⇏(G) for discrete distributions [24].

Observation 1.7.1. Assume that a probability measure P over N has a strictly positive density. Then one has (G)⇔(L)⇔(P) for P.

Proof. The key fact is the property specified in Observation 1.4.1(iv), which implies that the CI structure induced by G is a graphoid. Thus, it is enough to show that the graphoid closure of the set of triplets of the form $\langle i, j | N \setminus \{i, j\}\rangle$ for non-edges $i, j \in N$, $\neg(i - j \text{ in } G)$, contains the whole formal independence model $\mathcal{M}_G$. This observation is left to the reader as an exercise. □

Of course, it is clear from the presented proof that P need not necessarily have a strictly positive density; it is sufficient for the CI structure induced by P to be a graphoid. There are weaker conditions [45] which ensure the validity of that assertion. To illustrate the above-mentioned concepts, let us again consider the UGs in Figure 1.3. The reader can easily check that the following lists of independencies are the respective requirements.

(L) for G:	(P) for G:	(L) for H:	(P) for H:
$a \perp\!\!\!\perp de \,\|\, bc$	$a \perp\!\!\!\perp d \,\|\, bce$	$a \perp\!\!\!\perp cde \,\|\, b$	$a \perp\!\!\!\perp c \,\|\, bde$
$b \perp\!\!\!\perp ce \,\|\, ad$	$a \perp\!\!\!\perp e \,\|\, bcd$	$b \perp\!\!\!\perp de \,\|\, ac$	$a \perp\!\!\!\perp d \,\|\, bce$
$c \perp\!\!\!\perp b \,\|\, ade$	$b \perp\!\!\!\perp c \,\|\, ade$	$c \perp\!\!\!\perp ae \,\|\, bd$	$a \perp\!\!\!\perp e \,\|\, bcd$
$d \perp\!\!\!\perp a \,\|\, bce$	$b \perp\!\!\!\perp e \,\|\, acd$	$d \perp\!\!\!\perp ab \,\|\, ce$	$b \perp\!\!\!\perp d \,\|\, ace$
$e \perp\!\!\!\perp ab \,\|\, cd$		$e \perp\!\!\!\perp abc \,\|\, d$	$b \perp\!\!\!\perp e \,\|\, acd$
			$c \perp\!\!\!\perp e \,\|\, abd$

Thus, by Observation 1.7.1, to check whether a probability measure P with a strictly positive density satisfies the global Markov property relative to G, it is enough to verify that four CI statements (P) for G are valid with respect to P. Analogously, as concerns H, five CI statements in (L) for H are enough to verify the global Markov property relative to H.

Note that the undirected *separation criterion* from § 1.7.1 was a result of a certain development in the theory of Markov fields, which stemmed from statistical physics. The authors, who had developed this theory in the 1970s, restricted their attention to strictly positive discrete probability distributions. Several types of Markov conditions were proposed in [40]: the original pairwise Markov property was strengthened to the local and global versions. The reader can ask whether one can possibly even strengthen the global Markov property. Note that it follows from the result on the existence of a perfectly Markovian positive discrete measure [19] that the global Markov property cannot be strengthened. Moreover, it also occurs to be the strongest possible Markov property within the framework of regular Gaussian measures.

1.7.3 Factorization property for an UG

There is another sufficient condition for the global Markov property, which does not demand for the distribution to have a positive density. Specifically, we say that a marginally continuous measure P over N *is factorized* according to an UG G over N if a dominating system of σ-finite measures μ^i, $i \in N$, exists such that, for the respective joint density f, one has

(F) there exists potentials $\psi_C : \mathsf{X}_C \to [0, \infty)$, $C \in \mathcal{C}_G$, with

$$f(x) = \prod_{C \in \mathcal{C}_G} \psi_C(x_C) \qquad \text{for } \mu\text{-a.e. } x \in \mathsf{X}_N \,,$$

where $\mathcal{C}_G$ denotes the collection of cliques of G.

Note that one always has (F)$\Rightarrow$(G); this observation can be derived from repeated application of the fact that the factorization condition (1.5) is an equivalent definition of CI; see [25, Proposition 1]. On the other hand, examples of discrete measures showing (G)$\not\Rightarrow$(F) exist [31]. Nevertheless, the conditions are quite often equivalent. The following result, whose proof is omitted, is known as the *Hammersley-Clifford theorem*, see [24, Theorem 3.9]. It is a very useful observation as discussed in Chapter 3 of this book.

Observation 1.7.2. Assume that a probability measure P over N has a strictly positive density. Then one has (F)$\Leftrightarrow$(G) for P.

1.8 Markov properties for directed graphs

This section deals with directed acyclic graphical models, called *Bayesian networks* in the context of probabilistic reasoning [42].

1.8.1 Directional separation criteria

In the directed case, several separation criteria are available to decide whether a disjoint triplet is represented in a graph; however, these apparently different criteria are equivalent with each other. They are described in this subsection, throughout which we assume that G is a directed graph over N; indeed, to introduce the criteria it is not substantial whether G is acyclic or not.

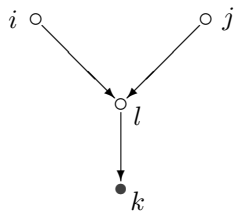


FIGURE 1.4: A simple example of a DAG.

Straightforward criterion in terms of walks

Let us start with a straightforward separation criterion for walks, which is the simplest one. Let $\rho : i_1, \ldots, i_k$, $k \geq 1$, be a walk in G. We say that a node i_ℓ in ρ occurs as a *collider* in ρ if it is an internal node in ρ and $i_{\ell-1} \to i_\ell \leftarrow i_{\ell+1}$ in G. Other occurrences of nodes in ρ, including its end-nodes, are called *non-colliders*. We say that ρ is *interrupted* by a set of nodes $C \subseteq N$ if

either a node exists which occurs as a *non-collider* in ρ and *belongs to* C,

or a node exists which occurs as a *collider* in ρ and *is outside of* C.

A walk in G which is not interrupted by a set $C \subseteq N$ will be called *free* for C, or just briefly *C-free*. Thus, $\rho : i_1, \ldots, i_k$, $k \geq 1$, is C-free provided one has, for all i_ℓ, $1 \leq \ell \leq k$:

- if i_ℓ is a non-collider node occurrence in ρ then $i_\ell \notin C$,
- if i_ℓ is a collider node occurrence in ρ then $i_\ell \in C$.

Given $\langle A, B|C\rangle \in \mathcal{T}(N)$, we say that A and B are *directionally separated* by C in G if every *walk* in G from a node in A to a node in B is interrupted by C and write $A \perp\!\!\!\perp B \,|\, C \; [G]$ then.

Thus, the interrupting condition for non-colliders is the same as in the undirected case (see § 1.7.1), while the condition for colliders is completely converse. It also follows from the definition that if a walk has a node with occurrences of both a collider and a non-collider then it must be interrupted by any $C \subseteq N$.

Note that, when testing $A \perp\!\!\!\perp B \,|\, C \; [G]$, one has to consider all walks from A to B, not just paths. For example, the only path from i do j in the graph in Figure 1.4 is $i \to l \leftarrow j$ and this path is interrupted by the set $C = \{k\}$. Nevertheless, a walk $i \to l \to k \leftarrow l \leftarrow j$ exists in the graph which is C-free.

A natural question arises whether the walk-based criterion is decidable. Below we describe a propagation algorithm which, for given disjoint sets of nodes A and C, finds the set $\bar{A}$ of nodes to which a C-free walk exists from a node in A. Thus, if B is disjoint with $\bar{A} \cup C$, then directional separation $A \perp\!\!\!\perp B \,|\, C \; [G]$ holds, otherwise it does not. The algorithm can be viewed as a kind of modification of the *Bayes-ball* algorithm [47] by Shachter.

Input: Directed graph G over N; $A, C \subseteq N$ disjoint sets of nodes.
Auxiliary sets of nodes: $U, V, W \subseteq N$.
Put $U := A$, $V := \emptyset$, $W := \emptyset$.
Apply exhaustively the following three propagation rules:

(i) $w \in U \cup W$, $w \leftarrow u$, $u \notin C \quad \Rightarrow \quad u \in U$,

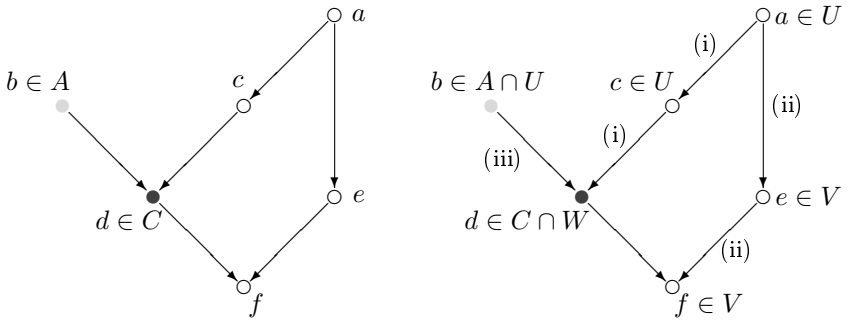


FIGURE 1.5: Illustration of the propagation algorithm for directed graphs.

(ii) $u \in U \cup V$, $u \to v$, $v \notin C \quad \Rightarrow \quad v \in V$,

(iii) $u \in U \cup V$, $u \to w$, $w \in C \quad \Rightarrow \quad w \in W$.

Output: Put $\bar{A} := U \cup V$ when the algorithm terminates.

We leave to the reader to verify that the output of the algorithm is indeed the set $\bar{A}$. This follows from the interpretation of the auxiliary sets of nodes:

- U is the set of nodes u in N such that either $u \in A$ or there exists a C-free walk from A to u which ends by an arrow pointing *out of* u,
- V is the set of nodes v in N such that there exists a C-free walk from A to v which ends by an arrow pointing *into* v,
- W is the set of nodes w in C such that there exists a C-free walk from A to some $u \in \mathrm{pa}_G(w)$.

An example of application of the algorithm is in Figure 1.5: here we start with $A = \{b\}$ and $C = \{d\}$ and by consecutive application of (iii), (i) and (ii) get $U = \{a, b, c\}$, $V = \{e, f\}$ and $W = \{d\}$. Thus, $\bar{A} = U \cup V = \{a, b, c, e, f\}$.

D-separation criterion

Another option to solve the verification problem is to modify the criterion so that only paths are considered. This leads to a traditional directional criterion, often abbreviated as *d-separation criterion*, which was promoted by Pearl and his coauthors [42]. To formulate this criterion, one needs an additional graphical concept: if there exists a directed path in G from node $i \in N$ to node $j \in N$ then we say that i is an *ancestor* of j in G, or, dually, that j is a *descendant* of i in G. Note that any node is its own descendant.

Given a path $\rho : i_1, \ldots, i_k$, $k \geq 1$, in G and $C \subseteq N$ we say that ρ is *active* for C, briefly C-*active* if, for all i_ℓ, $1 \leq \ell \leq k$:

- if i_ℓ is a non-collider in ρ then $i_\ell \notin C$,
- if i_ℓ is a collider in ρ then i_ℓ has a descendant in C.

A path which is not active with respect to C is *blocked by* C. Finally, $\langle A, B|C\rangle \in \mathcal{T}(N)$ is represented in G according to the *d-separation criterion*, if every *path* in G from A to B is blocked by C.

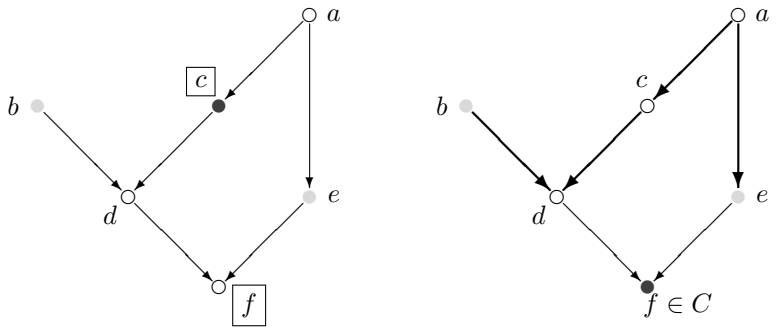


FIGURE 1.6: Illustration of the d-separation criterion for directed graphs.

Note that, in the case of a path, any node occurs at most once in ρ and must be either a collider or a non-collider. The concept of a C-active path only slightly differs from that of a C-free walk: there is a weaker requirement in the case of collider nodes. To illustrate the application of d-separation criterion, consider the graph in Figure 1.6. If we test $\langle b, e|c\rangle$ then we can observe that any path from b to e either goes through a non-collider $c \in C$ and is blocked there or it goes through a collider f and is blocked there (because f is *not* an ancestor of c). In order to see that $\langle b, e|f\rangle$ is not represented, let us consider the path $b \to d \leftarrow c \leftarrow a \to e$, which is C-active because the only collider d is an ancestor of $f \in C$.

Moralization criterion

The *moralization criterion*, promoted by Lauritzen and his co-authors [24, 8], is not straightforward in the sense that the graph is modified during the test. This criterion is based on transformation of the directed graph into a certain UG and then using the undirected separation criterion. It has three steps:

1. one removes some nodes and gets an induced subgraph of the original graph, which is relevant for the tested triplet,

2. this induced subgraph is transformed to a certain UG over the same set of nodes, which is – for certain reasons – called the *moral graph*,

3. finally the undirected separation criterion from § 1.7.1 is applied to the tested triplet and the moral graph.

Because of the first step, the moral graphs assigned to different tested triplets may be different. To formulate the criterion, one also needs additional graphical concepts. Specifically, an *immorality* in G is an induced subgraph of G of the form $i \to k \leftarrow j$. The *moral graph* of a directed graph G over N is an UG G^{mor} over the same set of nodes N such that $i - j$ in G^{mor} if

either $[i, j]$ is an edge in the original graph G,

or there exists an immorality in G of the form $i \to k \leftarrow j$.

We say that a triplet $\langle A, B|C\rangle \in \mathcal{T}(N)$ is represented in G according to the *moralization criterion* if A a B are separated by C in the undirected graph $H = (G_{\mathrm{an}_G(ABC)})^{mor}$, where the symbol $\mathrm{an}_G(ABC)$ denotes the set of *ancestors* of nodes in ABC (see the text about d-separation).

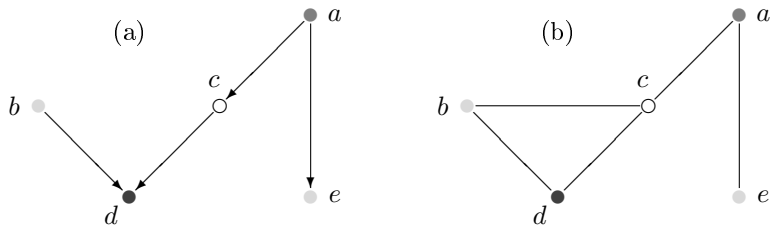


FIGURE 1.7: Illustration of the moralization criterion.

To illustrate the application of the moralization criterion, consider the graph in Figure 1.6. To test $\langle b, e|ad\rangle$, we first transform the graph into the induced subgraph for the set of ancestors of nodes in $\{a, b, d, e\}$ in Figure 1.7(a) and then to the moral graph in Figure 1.7(b). We observe that every path from b to e in the moral graph goes through a. Thus, $\langle b, e|ad\rangle$ is represented in the graph according to the moralization criterion. Testing $\langle b, e|d\rangle$ leads to the same moral graph in Figure 1.7(b); but, this time, a path $b - c - a - e$ exists which avoids the set $C = \{d\}$. This implies that $\langle b, e|d\rangle$ is not represented.

The third option

There is another criterion based on transformation of the graph, in which some edges are removed instead of added. This criterion was suggested by Massey [29] and also independently by Darwiche in his book [11]. To test a triplet $\langle A, B|C\rangle \in \mathcal{T}(N)$ the following steps are made:

1. the induced subgraph for the ancestors of nodes in ABC is constructed (this is identical with the first step of the moralization criterion),
2. the subgraph is pruned by the removal of arrows outgoing from C,
3. if there is no path between A and B in the resulting directed graph then the triplet is represented according to this criterion.

To illustrate the criterion, let us again consider the graph in Figure 1.6 and test $\langle b, e|d\rangle$. The first step leads to the graph in Figure 1.7(a), and no arrow is removed from that graph in the second step. Since there is a path between b and e in it, the triplet is not represented. On the other hand, when $\langle b, e|ad\rangle$ is tested, the graph in Figure 1.7(a) is pruned in the second step by the removal of arrows $a \to c$ and $a \to e$. Thus, e is an isolated node in the resulting graph and the triplet is represented according to this special criterion.

Equivalence of directional criteria

The equivalence of d-separation and moralization criteria (in the case of a DAG) was shown in [24, Proposition 3.25]. The equivalence of the last criterion with d-separation was proved in [11, Theorem 4.1].

Nonetheless, all the criteria are mutually equivalent even in the case of a general directed graph. Thus, the following fact is left to the reader as an exercise. A hint is that one shows, for any of the three criteria, a directed graph G over N and $\langle A, B|C\rangle \in \mathcal{T}(N)$, that the triplet is *not represented* in G according to the respective criterion iff there exists a C-free walk between A and B, that is, $A \not\perp\!\!\!\perp B \,|\, C\ [G]$.

Observation 1.8.1. Let G be a directed graph over N and $\langle A, B|C\rangle \in \mathcal{T}(N)$. Then $A \perp$

$\perp B \,|\, C\ [G]$ iff $\langle A, B|C\rangle \in \mathcal{T}(N)$ is represented in G according to any of the three above-mentioned path-based criteria.

1.8.2 Global Markov property for a DAG

Every directed acyclic graph G over N induces a formal independence model over N through the *directional separation* criterion

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

which is a disjoint graphoid. A probability measure P over N with $\mathcal{M}_G \subseteq \mathcal{M}_P$ is called *Markovian* with respect to G and we also say that P satisfies the *directed global Markov property* relative to G:

(DG) if A and B are directionally separated by C in G then $A \perp\!\!\!\perp B \,|\, C\ [P]$.

The statistical *directed graphical model* $\mathbb{M}_G$ consists of all Markovian measures with respect to G. The class $\mathbb{M}_G$ can be interpreted as the statistical model of the CI structure given by $\mathcal{M}_G$ (see § 1.4.2).

A probability measure P over N is called *perfectly Markovian* with respect to a DAG G if $\mathcal{M}_G = \mathcal{M}_P$. The existence of a perfectly Markovian measure with respect to any given DAG was shown by Geiger and Pearl [18].

Note that formal independence models induced by DAGs cannot be described completely in an axiomatic way. The reason is that these models are not closed under marginalization operation; see [57, Remark 3.5].

1.8.3 Local Markov property for a DAG

In the directed case, several variations of both local and pairwise Markov properties exist. One can distinguish between ordered versions, when an enumeration of nodes consonant with the direction of arrows is given, and the Markov property is relative to it on the one hand, and unordered versions on the other hand; see [8, § 5.3]. In this section, a basic unordered version of the local Markov property is presented.

An auxiliary graphical concept is needed to formulate this property. Recall from § 1.8.1 that a node j is a *descendant* of a node i in G if a directed path exists in G from i to j; denote the set of all descendants of node $i \in N$ in G by $\mathrm{ds}_G(i)$. Note that $i \in \mathrm{ds}_G(i)$.

A probability measure P over N satisfies a *directed local Markov property* relative to a DAG G over G if

(DL) for all $i \in N \qquad i \perp\!\!\!\perp N \setminus (\mathrm{ds}_G(i) \cup \mathrm{pa}_G(i)) \,|\, \mathrm{pa}_G(i)\ [P]$.

Observation 1.8.2. For any probability measure P over N, (DG)⇔(DL).

Proof. Given any enumeration $i_1, \ldots, i_{|N|}$ of nodes which is consonant with the direction of arrows G, it was shown in [64] that $\mathcal{M}_G$ is the semi-graphoid closure of the list of triplets of the form $\langle i_\ell, \{i_1, \ldots, i_{\ell-1}\} \setminus \mathrm{pa}_G(i_\ell) \,|\, \mathrm{pa}_G(i_\ell)\rangle$, $\ell = 2, \ldots, |N|$. Hence, $\mathcal{M}_G$ can be shown to be the semi-graphoid closure of the set of triplets of the form $\langle i, N \setminus (\mathrm{ds}_G \cup \mathrm{pa}_G(i)) \,|\, \mathrm{pa}_G(i)\rangle$; use Observation 1.4.1. □

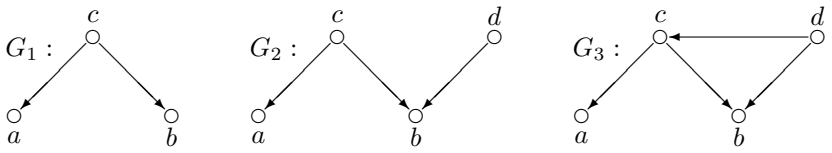


FIGURE 1.8: Illustration of the concept of a legally reversible arrow.

1.8.4 Factorization property for a DAG

Recursive factorization is a necessary and sufficient condition for a marginally continuous measure to be Markovian with respect to a *directed acyclic graph*. In the case of a discrete measure P over N it has the form

$$\textbf{(DF)} \qquad p(x) = \prod_{i \in N} p_{i|\mathrm{pa}_G(i)}(x_i | x_{\mathrm{pa}_G(i)}) \qquad \text{for every } x \in \mathsf{X}_N,$$

where a convention is accepted that $p_{A|C}(a|c) = 0$ whenever $p_C(c) = 0$ for $a \in \mathsf{X}_A$, $c \in \mathsf{X}_C$, $A, C \subseteq N$ disjoint.

The definition in the case of a marginally continuous measure is analogous, but one has to correctly introduce the conditional densities and the equation in (DF) is meant in the μ-a.e. sense, where μ is a dominating joint product measure. One can show that (DF)$\Leftrightarrow$(DG) then; see [25, Theorem 1].

Since the statistical model $\mathbb{M}_G$ for a DAG G coincides with the class of recursively factorizable distributions, there is a natural *parameterization* of this class in the discrete case; the elementary parameters are interpreted as (the values of) conditional probabilities [57, Lemma 8.1].

1.8.5 Markov equivalence for DAGs

We say that two DAGs G and H over N are *Markov equivalent* if they define the same statistical model, that is, $\mathbb{M}_G = \mathbb{M}_H$ (see § 1.8.2); note that this concept depends on the considered distribution framework Ψ (see § 1.4.2).

Analogously, two DAGs G and H over N are *independence equivalent* if they induce the same formal independence model: $\mathcal{M}_G = \mathcal{M}_H$; this notion, however, does not depend on the considered distribution framework. Clearly, independence equivalence implies Markov equivalence and the converse is also true provided that the distribution framework Ψ is non-degenerate [57, § 6.1]. For example, in the discrete case, non-degeneracy means that, for any $i \in N$, the individual sample space X_i has at least two elements. Thus, these two concepts of equivalence for DAGs typically coincide.

There could be different DAGs which are independence equivalent and a natural task is to characterize them graphically. A classic characterization of this kind was mentioned by Verma and Pearl [65]. One crucial concept here is that of *underlying undirected graph* of a DAG G, called alternatively a *skeleton* by some authors [1]: it is an UG over N in which an edge between a and b exists if either $a \to b$ in G or $a \leftarrow b$ in G. The second substantial concept is that of *immorality*: recall from § 1.8.1 that it is an induced subgraph of G of the form $i \to k \leftarrow j$. The classic graphical characterization says that two DAGs are independence equivalent iff they have the same skeleton and immoralities.

Nevertheless, there is also an indirect *transformational characterization* of equivalent DAGs proposed by Chickering [6], which often appears to be useful. It reveals an elementary graphical operation preserving independence equivalence of graphs. More specifically, given

a DAG G over N, we say that an arrow $a \to b$ in G is *legally reversible* if the graph H obtained from G by replacing $a \to b$ by $a \leftarrow b$ in H (and keeping remaining arrows untouched) is also acyclic, and, moreover, independence equivalent to G. We say then H is obtained from G by *legal arrow reversal*. The following fact is true.

Observation 1.8.3. Given a DAG G over N with an arrow $a \to b$, it is legally reversible iff $\mathrm{pa}_G(b) = \mathrm{pa}_G(a) \cup \{a\}$.

The proof can be found in [6, Lemma 1]. Note, however, that Chickering's terminology is different: he talks about a *covered edge* if the condition from Observation 1.8.3 holds. To illustrate the concept of a legally reversible arrow consider the DAGs in Figure 1.8. Both arrows in the graph G_1 are legally reversible because the parent set for c is empty and the other two nodes have c as the only parent node. If one modifies the graph by adding an arrow $d \to b$ (and the node d) then one gets the graph G_2 in which the arrow $c \to b$ will *not* be legally reversible. This can again be changed by adding a further arrow $d \to c$: then the arrow $c \to b$ will become legally reversible (in G_3).

The next observation, shown in [23, Lemma 4.2], gathers both graphical characterizations. Condition (B) is the classic direct characterization, while condition (C) is a transformational characterization.

Observation 1.8.4. The following three conditions are equivalent for any two given DAGs G and H over N:

(A) G and H are independence equivalent (that is, $\mathcal{M}_G = \mathcal{M}_H$),

(B) G and H have the same skeleton and immoralities,

(C) there exists a sequence $G = G_1, \ldots, G_m = H$, $m \geq 1$, of graphs over N, such that G_{i+1} is obtained from G_i by a legal arrow reversal for all $i = 1, \ldots, m-1$. (The graphs must be DAGs then.)

1.9 Remarks on chordal graphs

The class of *chordal undirected graphs* (see § 1.6 for the definition) plays a central role in graphical models. These graphs have widely been studied in graph-theoretical literature and a plenty of equivalent definitions/characterizations have been introduced; see, for example, [24, § 2.1].

A common alternative name is a decomposable graph, which is related to an equivalent definition of a chordal graph in terms of decompositions. Specifically, a non-trivial *decomposition* of an UG G over N is defined by a pair of sets $S, T \subseteq N$ such that

- $S \cup T = N$, $S \setminus T \neq \emptyset \neq T \setminus S$,
- $S \cap T$ is a complete set in G (see § 1.6),
- $S \setminus T \perp\!\!\!\perp T \setminus S \,|\, S \cap T$ $[G]$ (see § 1.7.1).

Then G is decomposed into its induced subgraphs G_S and G_T. An UG G is chordal iff it is *decomposable*, which means that it is either complete or can be non-trivially decomposed into decomposable graphs (over smaller sets of nodes); see [24, Proposition 2.5].

The statistical models ascribed to decomposable graphs exhibit elegant properties. For example, an explicit closed form expression for the maximum likelihood estimate exists; see

[24, § 4.4.2]. There is an analogous formula for the joint density of (any globally) Markovian measure with respect to a chordal UG in terms of the marginal densities for its cliques; see [57, § 3.4.1].

Another related equivalent definition is the existence of a *junction tree* of its cliques; see [8, Theorem 4.6]. Junction trees then form a mathematical basis for miscellaneous effective computational methods which originate from the *local computation method* [26]. Some of the chapters in Part II of the handbook discuss the computation methods.

An interesting fact illustrating the mathematical beauty of these graphs is as follows: a formal independence model $\mathcal{M}$ is induced by a chordal graph (by undirected separation) iff it is a model induced by a certain UG (by undirected separation) and by a certain DAG (by directional separation). There is a finite axiomatization of such formal independence models found by de Campos [14].

1.10 Imsets and geometric views

In this section we mention the method of structural imsets, which offers a geometric point of view on the (description of) CI structures.

1.10.1 The concept of a structural imset

Although graphs offer an elegant and intuitive interpretation of some CI structures, they are not able to describe all possible probabilistic CI structures. This motivates a proposal for a non-graphical method of their description by means of vectors, whose components are integers indexed by subsets of N; such vectors are called *imsets*.

A starting point is the concept of an *elementary imset* from [57, § 4.2.1], which is a vector in $\mathbb{R}^{\mathcal{P}(N)}$ encoding the elementary CI statement $i \perp\!\!\!\perp j \,|\, K$ corresponding to $\langle i, j|K\rangle \in \mathcal{T}_\epsilon(N)$ (see § 1.5.1). Specifically, we put

$$u_{\langle i,j|K\rangle} \;:=\; \delta_{ijK} + \delta_K - \delta_{iK} - \delta_{jK},$$

where $\delta_A \in \mathbb{R}^{\mathcal{P}(N)}$ denotes the zero-one vector identifier of a set $A \subseteq N$.

One can consider the cone $\mathcal{S}(N)$ in $\mathbb{R}^{\mathcal{P}(N)}$ of non-negative linear combinations of elementary imsets over N. *Structural imsets*, used to describe CI structures, can equivalently be introduced as vectors in $\mathcal{S}(N) \cap \mathbb{Z}^{\mathcal{P}(N)}$ [22]. There was an open problem whether every structural imset is also a *combinatorial imset*, that is, a combination of elementary imsets with non-negative integer coefficients. This is indeed true if $|N| \leq 4$ but Hemmecke et al. [21] gave an example of a structural imset over N with $|N| = 5$ which is not a combinatorial imset.

The next step is to ascribe a formal independence model over N to any structural imset u over N. There is a certain linear-algebraic criterion to decide, for each $\langle A, B|C\rangle \in \mathcal{T}(N)$, whether $A \perp\!\!\!\perp B \,|\, C\ [u]$ holds; this criterion is omitted in this chapter and can be found in [57, § 4.4.1]. The criterion can be viewed as an analogue of separation criteria used in graphical description of CI structures. The formal independence models

$$\mathcal{M}_u \;=\; \{\, \langle A, B|C\rangle \in \mathcal{T}(N) \;:\; A \perp\!\!\!\perp B \,|\, C\ [u] \,\} \qquad \text{for } u \in \mathcal{S}(N) \cap \mathbb{Z}^{\mathcal{P}(N)}$$

appear to be semi-graphoids, called *structural semi-graphoids*. Every such semi-graphoid is, in fact, induced by a combinatorial imset, which means that one can limit oneself to combinatorial imsets. Following the analogy with graphical models, one can introduce, for

any structural imset u, the corresponding statistical model $\mathbb{M}_u$ of *Markovian distributions* P with respect to u satisfying $\mathcal{M}_u \subseteq \mathcal{M}_P$. Moreover, it was shown [57, Theorem 4.1] that, for marginally continuous measure P over N the Markov property with respect to a structural imset u is equivalent to a certain factorization property, which generalizes the recursive factorization for DAGs mentioned in § 1.8.4.

The crucial result concerning structural imsets is that, for any probability measure P over N with *finite multiinfomation*, that is, with finite relative entropy of P with respect to $\bigotimes_{i\in N} P_i$, the CI structure induced by P is a structural semi-graphoid [57, Theorem 5.2]. In other words, any such distribution is *perfectly Markovian* with respect to some combinatorial imset u, which means that $\mathcal{M}_u = \mathcal{M}_P$. Note that all discrete measures and all regular Gaussian measures over N have finite multiinformation values.

Structural semi-graphoids also coincide with semi-graphoids ascribed to supermodular functions mentioned in § 1.5. A remark, which may interest a reader familiar with advanced polyhedral geometry, is that one can extend the observation according to which semi-graphoids correspond to polyhedral fans coarsening the S_n-fan (see § 1.5.1). Morton [37] also mentioned that a semi-graphoid is structural iff the corresponding polyhedral fan is a normal fan of a polytope.

1.10.2 Imsets for statistical learning

Imsets can also be applied in the context of learning Bayesian network (BN) structure. There is a certain standard translation of a DAG G over N into a combinatorial imset u_G, called the *standard imset* (for G), which has the property that the usual criteria for learning BN structure become affine functions (= sums of linear functions with constants) of the standard imset [61]. Thus, the learning task can be transformed into a *linear programming* problem; a mathematical task is then to characterize the domain in the form of finitely many linear inequalities.

It is sometimes advantageous in combinatorial optimization to work with zero-one vectors. Therefore, standard imsets were transformed by an affine invertible self-transformation of $\mathbb{Z}^{\mathcal{P}(N)}$ into *characteristic imsets*, which are zero-one vectors with an elegant graphical interpretation [20], and these vectors were applied to learning the BN structure by tools of integer linear programming [60]. This approach seems to be particularly suitable for learning decomposable models [59], in which case there is hope that the corresponding polytope will be completely characterized by linear inequalities.

1.11 CI inference

This section is concerned with the following task: given an input list $\mathcal{L}$ of CI statements over N, characterize its probabilistic *CI closure*, which is the smallest CI structure containing $\mathcal{L}$. A traditional aim is to obtain the CI closure by applying interpretable formal CI implications, analogous to the semi-graphoid inference rules from Definition 1.5.1. Although there is no finite set of inference rules characterizing probabilistic CI inference [53], one can find such an axiomatic characterization in some special instances. The semi-graphoid implications are sufficient in the case of $|\mathcal{L}| = 2$ [55] or if $\mathcal{L}$ consists of special CI statements, such as the marginal CI statements $A \perp\!\!\!\perp B \,|\, \emptyset$ [17, 32] or saturated CI statements $A \perp\!\!\!\perp B \,|\, C$ with $ABC = N$ [28, 19].

Matúš [34] characterized the CI closure for discrete measures for $|N| = 4$; in this case 24 formal properties are enough [58]. Several methods to derive implications among CI

statements can be used. The method of structural imsets [57, § 6.2] provides a sufficient condition for probabilistic CI implication; the respective linear-algebraic criterion can be tested using a computer [5]. The most efficient methods for computer testing of that linear-algebraic condition seem to be linear programming ones [4, 41]. On the other hand, there are linear-algebraic tools to derive CI implications based on different principles [62]. On top of that, advanced methods of modern algebra can be used to derive CI implications; Chapter 3 gives more details on this topic.

Acknowledgements

I am indebted to Fero Matúš for his cooperation on the topic of CI. Our work has been supported from Grant Project GAČR n. 16-12010S. I would also like to express my thanks to the reviewers, whose comments helped me improve the quality of presentation. I am also grateful to Antonín Otáhal and Cheri Dohnal for correcting my English.

Bibliography

[1] S. A. Andersson, D. Madigan, and M. D. Perlman. A characterization of Markov equivalence classes for acyclic digraphs. *Ann. Statist.*, 25(2):505–541, 1997.

[2] M. Baioletti, G. Busanello, and B. Vantaggi. Conditional independence structure and its closure: inferential rules and algorithms. *Internat. J. Approx. Reason.*, 50(7):1097–1114, 2009.

[3] C. Beeri, R. Fagin, and J. H. Howard. A complete axiomatization for functional and multivalued dependencies in database relations. In *Proceedings of the 1977 ACM SIGMOD International Conference on Management of Data*, pages 47–61. ACM, 1977.

[4] R. Bouckaert, R. Hemmecke, S. Lindner, and M. Studený. Efficient algorithms for conditional independence inference. *J. Mach. Learn. Res.*, 11:3453–3479, 2010.

[5] R. R. Bouckaert and M. Studený. Racing algorithms for conditional independence inference. *Internat. J. Approx. Reason.*, 45(2):386–401, 2007.

[6] D. M. Chickering. A transformational characterization of equivalent Bayesian network structures. In *Uncertainty in Artificial Intelligence 11*, pages 87–98. Morgan Kaufmann, San Francisco, 1995.

[7] E. F. Codd. A relational model of data for large shared data banks. *Communications of the ACM*, 13:377–387, 1970.

[8] R. G. Cowell, A. P. Dawid, S. L. Lauritzen, and D. J. Spiegelhalter. *Probabilistic Networks and Expert Systems*. Springer, New York, 1999.

[9] F. G. Cozman and P. Walley. Graphoid properties of epistemic irrelevance and independence. *Ann. Math. Artif. Intell.*, 45(1/2):173–195, 2005.

[10] J. N. Darroch, S. L. Lauritzen, and T. P. Speed. Markov fields and log-linear interaction models for contingency tables. *Ann. Statist.*, 8(3):522–539, 1980.

[11] A. Darwiche. *Modeling and Reasoning with Bayesian Networks*. Cambridge University Press, New York, 2009.

[12] A. P. Dawid. Conditional independence in statistical theory. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 41:1–31, 1979.

[13] A. P. Dawid. Separoids: a mathematical framework for conditional independence and irrelevance. *Ann. Math. Artif. Intell.*, 31(1/4):335–372, 2001.

[14] L. M. de Campos. Characterization of decomposable dependency models. *J. Artif. Intell. Res.*, 5:289–300, 1996.

[15] A. P. Dempster. Covariance selection. *Biometrics*, 28:157–175, 1972.

[16] M. Drton, B. Sturmfels, and S. Sullivant. *Lectures on Algebraic Statistics*. Birkhäuser, 2009.

[17] D. Geiger, A. Paz, and J. Pearl. Axioms and algorithms for inferences involving probabilistic independence. *Inform. and Comput.*, 91(1):128–141, 1991.

[18] D. Geiger and J. Pearl. On the logic of causal models. In *Uncertainty in Artificial Intelligence 4*, pages 3–14. North-Holland, Amsterdam, 1990.

[19] D. Geiger and J. Pearl. Logical and algorithmic properties of conditional independence and graphical models. *Ann. Statist.*, 21(4):2001–2021, 1993.

[20] R. Hemmecke, S. Lindner, and M. Studený. Characteristic imsets for learning Bayesian network structure. *Internat. J. Approx. Reason.*, 53:1336–1349, 2012.

[21] R. Hemmecke, J. Morton, A. Shiu, B. Sturmfels, and O. Wienand. Three counterexamples on semi-graphoids. *Combin. Probab. Comput.*, 17:239–257, 2008.

[22] T. Kashimura, T. Sei, A. Takemura, and K. Tanaka. Cones of elementary imsets and supermodular functions: a review and some new results. In *Proceedings of 2nd CREST-SBM International Conference*, pages 357–363. World Scientific, 2012.

[23] T. Kočka, R. R. Bouckaert, and M. Studený. On characterizing inclusion of Bayesian networks. In *Uncertainty in Artificial Intelligence 17*, pages 261–268. Morgan Kaufmann, San Francisco, 2001.

[24] S. L. Lauritzen. *Graphical Models*. Clarendon Press, Oxford, 1996.

[25] S. L. Lauritzen, A. P. Dawid, B. N. Larsen, and H.-G. Leimer. Independence properties of directed Markov fields. *Networks*, 20(5):491–505, 1990.

[26] S. L. Lauritzen and D. J. Spiegelhalter. Local computation with probabilities on graphical structures and their application to expert systems. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 50(2):157–224, 1988.

[27] M. Loève. *Probability Theory, Foundations, Random Processes*. D. van Nostrand, Toronto, 1955.

[28] F. M. Malvestuto. A unique formal system for binary decomposition of database relations, probability distributions and graphs. *Inform. Sci.*, 59:21–52, 1992.

[29] J. L. Massey. Causal interpretation of random variables (in Russian). *Problemy Peredachi Informatsii*, 32(1):112–116, 1996.

[30] F. Matúš. Abstact functional dependency structures. *Theoret. Comput. Sci.*, 81:117–126, 1991.

[31] F. Matúš. On equivalence of Markov properties over undirected graphs. *J. Appl. Probab.*, 29(3):745–749, 1992.

[32] F. Matúš. Stochastic independence, algebraic independence and abstract connectedness. *Theoret. Comput. Sci. A*, 134(2):445–471, 1994.

[33] F. Matúš. Conditional independences among four random variables II. *Combin. Probab. Comput.*, 4(4):407–417, 1995.

[34] F. Matúš. Conditional independences among four random variables III., final conclusion. *Combin. Probab. Comput.*, 8(3):269–276, 1999.

[35] F. Matúš. Lengths of semigraphoid inferences. *Ann. Math. Artif. Intell.*, 35:287–294, 2002.

[36] F. Matúš and M. Studený. Conditional independences among four random variables I. *Combin. Probab. Comput.*, 4(4):269–278, 1995.

[37] J. Morton. *Geometry of conditional independence.* PhD thesis, University of California Berkeley, 2007.

[38] J. Morton, L. Pachter, A. Shiu, B. Sturmfels, and O. Wienand. Convex rank tests and semigraphoids. *SIAM J. Discrete Math.*, 23(3):1117–1134, 2009.

[39] M. Mouchart and J.-M. Rolin. A note on conditional independence with statistical applications. *Statistica*, 44(4):557–584, 1984.

[40] J. Moussouris. Gibbs and Markov properties over undirected graphs. *J. Stat. Phys.*, 10(1):11–31, 1974.

[41] M. Niepert, M. Gyssens, B. Sayrafi, and D. van Gucht. On the conditional independence implication problem: a lattice-theoretic approach. *Artificial Intelligence*, 202:29–51, 2013.

[42] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.* Morgan Kaufmann, San Mateo, 1988.

[43] J. Pearl and A. Paz. Graphoids, graph-based logic for reasoning about relevance relations. In *Advances in Artificial Intelligence II*, pages 357–363. North-Holland, Amsterdam, 1987.

[44] Y. Sagiv and S. F. Walecka. Subset dependencies and completeness result for a subclass of embedded multivalued dependencies. *J. ACM*, 29(1):103–117, 1982.

[45] E. San Martín, M. Mouchart, and J.-M. Rolin. Ignorable common information, null sets and Basu's first theorem. *Sankhyā*, 67:674–697, 2005.

[46] P. H. Schouté. Analytic treatment of the polytopes regularly derived from regular polytopes. *Verhandelingen der Koninklijke Akademie van Wetenschappen te Amsterdam*, 11(3):370–381, 1911.

[47] R. D. Shachter. Bayes-ball, the rational pastime (for determining irrelevance and requisite information in belief networks and influence diagrams). In *Uncertainty in Artificial Intelligence 14*, pages 480–487. Morgan Kaufmann, San Francisco, 1998.

[48] P. P. Shenoy. Conditional independence in valuation-based systems. *Internat. J. Approx. Reason.*, 10(3):203–234, 1994.

[49] J. Q. Smith. Influence diagrams for statistical modelling. *Ann. Statist.*, 17(2):654–672, 1989.

[50] W. Spohn. Stochastic independence, causal independence and shieldability. *J. Philos. Logic*, 9(1):73–99, 1980.

[51] W. Spohn. Ordinal conditional functions: a dynamic theory of epistemic states. In *Causation in Decision, Belief Change, and Statistics II.*, pages 105–134. Kluwer, Dordrecht, 1988.

[52] M. Studený. Multiinformation and the problem of characterization of conditional independence relations. *Probl. Control Inform.*, 18:3–16, 1989.

[53] 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*, pages 377–396. Kluwer, Dordrecht, 1992.

[54] M. Studený. Conditional independence and natural conditional functions. *Internat. J. Approx. Reason.*, 12(1):43–68, 1995.

[55] M. Studený. Semigraphoids and structures of probabilistic conditional independence. *Ann. Math. Artif. Intell.*, 21(1):71–98, 1997.

[56] M. Studený. Complexity of structural models. In *Proceedings of the joint session of 6th Prague Symposium on Asymptotic Statistics and 13th Prague Conference*, pages 523–528. Union of Czech Mathematicians and Physicists, 1998.

[57] M. Studený. *Probabilistic Conditional Independence Structures*. Springer, London, 2005.

[58] M. Studený and P. Boček. CI-models arising among 4 random variables. In *Proceedings of WUPES'94, September 11-15, 1994, Czech Republic*, pages 268–282, 1994.

[59] M. Studený and J. Cussens. Towards using the chordal graph polytope in learning decomposable models. *Internat. J. Approx. Reason.*, 88:259–281, 2017.

[60] M. Studený and D. Haws. Learning Bayesian network structure: towards the essential graph by integer linear programming tools. *Internat. J. Approx. Reason.*, 55:1043–1071, 2014.

[61] M. Studený, J. Vomlel, and R. Hemmecke. A geometric view on learning Bayesian network structures. *Internat. J. Approx. Reason.*, 51:578–586, 2010.

[62] K. Tanaka, M. Studený, A. Takemura, and T. Sei. A linear-algebraic tool for conditional independence inference. *J. Algebr. Stat.*, 6(2):150–167, 2015.

[63] J. Vejnarová. Conditional independence in possibility theory. *Internat. J. Uncertain. Fuzziness Knowledge-Based Systems*, 12:253–269, 2000.

[64] T. Verma and J. Pearl. Causal networks, semantics and expressiveness. In *Uncertainty in Artificial Intelligence 4*, pages 69–76. North-Holland, Amsterdam, 1990.

[65] T. Verma and J. Pearl. Equivalence and synthesis of causal models. In *Uncertainty in Artificial Intelligence 6*, pages 220–227. Elsevier, Amsterdam, 1991.

[66] P. Šimeček. *Independence models (in Czech).* PhD thesis, Charles University, 2007.

[67] N. Wermuth. Analogies between multiplicative models for contingency tables and covariance selection. *Biometrics*, 32:95–108, 1976.

[68] J. Whittaker. *Graphical Models in Applied Multivariate Statistics*. John Wiley, Chichester, 1990.