1

Conditional Independence Concept and Markov Properties for Basic Graphs

Milan Studený

Institute of Information Theory and Automation of the CAS

CONTENTS

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 classic graphical models, namely those ascribed to *undirected graphs* (UGs) and *directed acyclic graphs* (DAGs), can be interpreted as special cases of (statistical) models of CI structure. Therefore, an overview of Markov properties for these two basic types of graphs is also given.

1.1 Introduction: history overview

In this section some of earlier results on CI are recalled.

1.1.1 Stochastic conditional independence

Already in the 1950s, Loève [20] in his book on probability theory defined the concept of CI is terms od σ-algebras. Phil Dawid [7] 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 sufficient statistics, can equivalently be defined in terms of generalized CI and this observation allows one to derive many results on them in an elegant way, using the formal properties. These basic formal properties of stochastic CI were later independently formulated in the context of philosophical logic by Spohn [41], 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 [31]. The author of this chapter was told that the conditional independence symbol $\perp\!\!\!\perp$ was proposed by Dawid and Mouchart during their joint discussion in the end of the 1970s.

The significance of the concept of CI for probabilistic reasoning was later recognized by Pearl and Paz [35], who observed that the above basic formal properties of CI are also valid for certain ternary separation relations induced by undirected graphs. This led them to the idea describe such formal ternary relations by graphs and introduced an abstract concept of a *semi-graphoid*. Even more abstract concept of a *separoid* was later suggested by Dawid [8]. Pearl and Paz [35] also raised a conjecture that semi-graphoids coincide with probabilistic CI structures, which has been refuted by myself in [43] using some tools of information theory.

A lot of effort and time was devoted to the problem to characterize all possible CI structures induced by four discrete random variables. The final solution to that problem was achieved by Matúš [29, 26, 27]; the number of these structures is 18478 [55] and they decompose into 1098 types.

1.1.2 Graphs and local computation method

However, 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 that in the context of computer science. One can distinguish 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) occurred in the 1970s in statistical physics as tools to describe relations among discrete random variables. Moussouris [32] introduced several Markov properties with respect to an UG for distributions with positive density and showed their equivalence with the factorization condition. Darroch, Lauritzen, and Speed [6] realized that UGs can be used to describe statistical models arising in the theory of contingency tables, introduced a special class of (undirected) graphical models and interpreted them in terms of CI. Parallel occurrence of UGs was in the area of multivariate statistical analysis. Dempster [9] introduced covariance models for continuous real random variables, which were interpreted in terms of CI by Wermuth [56].

In the 1980s, directed acyclic graphs (DAGs) found their application in the decision making theory in connection with influence diagrams. Smith [40] used above formal properties of CI to show easily the correctness of some operations with influence diagrams. Substantial impact on propagation of graphical methods in artificial intelligence had Pearl's book [34] on probabilistic reasoning in which 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 [19] offered a methodology to perform efficiently computation of conditional measures for (discrete) distributions which are Markovian with respect a DAG.

1.1.3 Conditional independence in other areas

Nevertheless, the probability theory and statistics is not the only field in which the concept of CI was introduced and examined. An analogous concept of *embedded multivalued dependency* (EMVD) was studied in the 1970s in theory of relational databases. Sagiv and Walecka [36] showed that there is no finite axiomatic characterization of EMVD structures. Shenoy [39] 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 ordinal conditional functions, Zadeh's possibility theory and the Dempster-Shafer theory of evidence.

This motivated several papers devoted to formal properties of CI in various uncertainty calculi in AI. For example, Vejnarová [53] studied the properties of CI in the frame of possibility theory and it was shown in [44] that there is no finite axiomatization of 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 [5].

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 a proposal of a linear-algebraic method of their description in [47]. 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 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. [15] answered two of the open problems related to the method of imsets and disproved a geometric conjecture from [47] about the cone corresponding to (structural) imsets.

The application of methods of modern algebra and (polyhedral) geometry to problems arising in mathematical statistics lead recently to establishing a new field of *algebraic statistics*. Drton, Sturmfels and Sullivant [10] in their book on this topic devoted one chapter to advanced algebraic tools to describe statistical models of CI structure. Thus, the theme of probabilistic CI became naturally one of the topics of interest in that area.

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 needs no special previous reader's knowledge.

Definition 1 A *discrete probability measure over N* is defined as follows:

(i) For every $i \in N$ 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 $\mathsf{X}_N := \prod_{i\in N} \mathsf{X}_i$.

(iii) 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 named 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 *joint*, that is 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 brackets will be omitted and we will write $p(a, b)$ instead of $p([a, b])$; similarly in case of the joint of three or more disjoint configurations.

In case $A \subseteq B$ 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 $b_\emptyset$ is 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 $\mathbb{A} = \{a_\emptyset\}$ consists of the empty configuration, then one has $\mathbb{A} \times \mathbb{B} = \mathbb{B}$; analogously in case $B = \emptyset \neq \mathbb{B}$.

Definition 2 Given $A \subseteq N$ and a probability measure P over N, the *marginal measure for A* is a 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 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 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 $*$ is a substitute for the respective arguments. Specifically,

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

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 of the marginal P_{AC}.

In computer science community, the conditional density is sometimes named a *conditional probability table*. Let us emphasize that the ratio defining the conditional density is not defined for conditioning *zero configuration* $c\in\mathsf{X}_C$ with $p_C(c)=0$, which important detail 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 any $x\in\mathsf{X}_N$, is too restrictive in the area of probabilistic expert systems because it does not allow to model 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 standard notions of measure theory. The meaning of the term of a *probability distribution* in the literature depends on the field. 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 quite wide class of probability measures for which the concept of density (function) has sense.

Definition 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

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

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 [47, 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 which 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 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 any $A \subseteq N$, 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 sense μ-everywhere.

1.2.3 Graphical concepts

By a graph *over* N we will understand a graph which has the set N as the set of *nodes*. Graphs considered in this chapter have no multiple edges and two possible types of 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; the pictorial representation in figures is analogous. An *undirected graph* (UG) is a graph in which every present edge is 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 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 for every distinct $i, j \in A$. Maximal complete sets in G with respect to set inclusion are named *cliques* of G.

Directed edges, also named *arrows*, are ordered pairs of distinct nodes. We will write $i \to j$ to denote an arrow from a node i to a node j in N; similarly in figures. A *directed graph* is a graph in which every present edge is an arrow. 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 parents of $j \in N$ in G.

A *route* 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 every consecutive pair of nodes in the sequence is adjacent by an edge in the graph G. The end-nodes of the route 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 route, that is, $k-1$, is called the *length* of the route. A route 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 case of a directed graph G, a path or a cycle is called *directed* if $i_\ell \rightarrow i_{\ell+1}$ for $\ell = 1, \ldots, k-1$.

A directed graph G is called *acyclic* if it has no directed cycle. Directed graph that are acyclic are conventionally named *directed acyclic graphs* (DAGs). A well-known equivalent characterization of an DAG is that it is 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 $i_\ell \rightarrow i_k$ in G then $\ell < k$.

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 discrete case

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

Definition 5 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

$$\forall\, \mathbb{A} \subseteq \mathsf{X}_A \quad \forall\, \mathbb{B} \subseteq \mathsf{X}_B \quad \forall\, c \in \mathsf{X}_C \quad \text{such that} \quad p_C(c) > 0$$

$$P_{AB|C}(\mathbb{A} \times \mathbb{B}|c) = P_{A|C}(\mathbb{A}|c) \cdot P_{B|C}(\mathbb{B}|c). \tag{1.1}$$

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 every 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 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 conversely. 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 in term 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 next 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 1 We define the value $\bar{p}(x)$ of the density of $\bar{P}$ by the formula on the RHS of (1.3) for $x \in \mathsf{X}_{ABC}$ with $p_C(x_C) > 0$ and $\bar{p}(x) = 0$ in case $p_C(x_C) = 0$. 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* and the 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) use the construction in the proof of Observation 1.3.1 and apply (1.3) to see that $\bar{p}(x) = p_{ABC}(x)$ in case $p_{ABC}(x) > 0$. Then realize that the values of both $\bar{p}$ and p_{ABC} sum to 1 to extend the equality $\bar{p}(x) = p_{ABC}(x)$ to the case $p_{ABC}(x) = 0$.

Further CI characterization in terms of marginal densities appeared in [32]; 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) distinguish the cases $p_C(c) = 0$, when (1.4) is evident, and $p_C(c) > 0$. In the latter case derive (1.4) whose both sides are 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 summation 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 any $i \in N$.

An elegant characterization of a CI statement is in term 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 case $p_C(z_C) > 0$ 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:

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

To see it is equivalent to the previous conditions observe (1.2)⇒(1.6)⇒(1.3). However, the most popular definition in 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:

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

One can easily show (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 does depend 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 depend on AC, but these modifications are omitted in this chapter.

1.3.2 More general CI concepts

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

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 $h : X \to [0,1]$, denoted by $\boldsymbol{P}\,[\tilde{\mathbb{A}}|\mathcal{C}]$, such that

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

The existence of such function h and its uniqueness in sense $\boldsymbol{P}_{\mathcal{C}}$-everywhere follows from the Radon-Nikodym theorem, where $\boldsymbol{P}_{\mathcal{C}}$ denotes the restriction of $\boldsymbol{P}$ to the measurable space $(\mathsf{X}, \mathcal{C})$. One can introduce the concept 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 versions of conditional probabilities and 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 [47, Lemma A.6]. This condition can be interpreted as an analogue of the discrete condition (1.7).

Let us describe how the CI definition (1.9) works in case of a (general) *probability measure* P *over* N mentioned in Definition 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 cylindric 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 3. Given pairwise disjoint $A,B,C\subseteq N$, the condition $\mathcal{A}\perp\!\!\!\perp\mathcal{B}\,|\,\mathcal{C}$ from (1.9) then 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 case of a (general) measure P over N.

In 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 [47, Lemma 2.4] that, provided a dominanting 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 in/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:

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

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

Proof 2 The arguments valid in the discrete case are only given, but the result holds in general. To verify (i) use (1.1) and realize that in case $A = \emptyset$ 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) combine (1.2) and (1.3). For the implication $A \perp\!\!\!\perp BD \,|\, C \;\Rightarrow\; A \perp\!\!\!\perp D \,|\, C$ 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 equalities (the latter with exchanged sides) to get $p_C \cdot p_{ABDC} \cdot p_{AC} \cdot p_{DC} = p_C \cdot p_{ADC} \cdot p_{AC} \cdot p_{BDC}$. 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 $\{\, A \perp\!\!\!\perp D \,|\, C \;\&\; A \perp\!\!\!\perp B \,|\, DC \,\} \;\Rightarrow A \perp\!\!\!\perp BD \,|\, C$ 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 the 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 Observation1.4.1(iv) need not be valid for a discrete distribution which not strictly positive. For example, consider $|N| = 3$, $\mathsf{X}_i = \{0,1\}$ for $i \in N$ and density p such that $p(0,0,0) = \frac{1}{2} = p(1,1,1)$ and $p(x) = 0$ for remaining configurations $x \in \mathsf{X}_N$. Then one has $i \perp\!\!\!\perp j \,|\, k\ [P]$ while $i \not\perp\!\!\!\perp j \,|\, \emptyset\ [P]$, which implies $i \not\perp\!\!\!\perp \{j,k\} \,|\, \emptyset\ [P]$.

Example: relational databases

However, formal independence models satisfying the conditions (i)-(iii) from Observation 1.4.1 occur also beyond statistics. For example, analogous formal models were studied in the theory of *relational databases*.

In that area, the elements of N are called *attributes*, and every 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.

On 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 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 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

This 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 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 $(\mathsf{X}, \mathcal{X})$, being a product measure $\mu = \bigotimes_{i\in N} \mu^i$ in case of $(\mathsf{X}_N, \mathcal{X}_N)$. Each probability measure in $\mathbb{M}$ is then determined by its density with respect of 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 place of Ψ. Then, every formal independence model $\mathcal{M} \subseteq \mathcal{T}(N)$ over N can be ascribed a statistical model

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

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

Note that this concept generalizes the classic concept of a *graphical model* [57, 17]. Indeed, the reader can learn in § 1.6 that every UG G over N induces the class $\mathbb{M}_G$ of Markovian measures over N, which statistical model can also be defined using a formal independence model induced by G.

1.5 Semi-graphoids, graphoids, and separoids

The notions discussed in this section have been inspired by the research on stochastic CI, but they more belong to the area of discrete mathematics. Pearl and Paz [35] introduced in 1987 the following concept.

Definition 6 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 $\mathcal{M}$. Analogously, the *graphoid closure of* $\mathcal{M}$ can be introduced.

Semi/graphoid closures are correctly defined because every set intersection of semi/graphoids over N is a semi/graphoid over N. The CI implications in Definition 6 are nothing 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 positive density.

There are other areas than the probability theory in which semi-graphoids have occurred. We have seen in the example from § 1.4.1 that every relational database can be ascribed a disjoint semi-graphoid. An undirected separation criterion from § 1.6.1 allows one to ascribe a graphoid to every UG over N. 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

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

An ordinal 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

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

Note that this is a concept taken over from [42].

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 any $D, E \subseteq N$. Define

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

Note that semi-graphoids defined in this way coincide with *structural semi-graphoids* mentioned in § 1.8.1.

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 [23]. Thus, general semi-graphoids are broader than disjoint semi-graphoids because they involve functional dependency relations modeling.

Dawid took even more general point of view and introduced an abstract concept of a separoid; the following is a simplification of his definition [8].

Definition 7 Let $\mathbb{S}$ be a joint semi-lattice, that is, a partially ordered set in which every two elements a, b have a supremum (= joint), denoted by $a \vee b$. A set of ordered triplets $a \perp\!\!\!\perp b \,|\, c$ of elements of $\mathbb{S}$ will be named a *separoid* if

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

Of course, every general semi-graphoid over N is a separoid on the lattice $(\mathcal{P}(N), \subseteq)$. Another prominent example requires the reader being 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 8 A disjoint triplet $\langle A, B|C\rangle$ over N will be named *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 and 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 $\mathcal{M}^1$ a $\mathcal{M}^2$ 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 $\mathcal{M} \cap \mathcal{T}_\epsilon(N)$.

Proof 3 The necessity of (1.10) is easily derivable 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 case $|AB| = 2$ is evident. Supposing $|AB| > 2$ either A or B is not a singleton. Owing to the symmetry property one can consider without the loss of generality $|B| \geq 2$, choose $b \in B$ and put $B' = B \setminus \{b\}$. By the induction assumption, (1.10) implies both $A \perp\!\!\!\perp B' \,|\, C\ [\mathcal{M}]$ and $A \perp\!\!\!\perp b \,|\, B'C\ [\mathcal{M}]$. Hence, by application of the contraction property $A \perp\!\!\!\perp B \,|\, C\ [\mathcal{M}]$ is derived.

One can also show easily 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 [28] was interested in the intricacy of the semi-graphoid implication between elementary CI statements and showed that the length of the derivation sequence can be exponential in $|N|$. However, there is an alternative way to represent semi-graphoids in the memory of a computer.

Definition 9 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, 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 the 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 [1]. Dominant triplets were also an useful tool in [45] to show that the semi-graphoid closure of two disjoint triplets over N is always a probabilistic CI structure. This 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 § 1.9). Semi/graphoids over a fixed set N can also be classified according to their *semi/graphoid complexity*, by which is meant the minimal cardinality of a semi/graphoid generator [46].

For a reader 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 [30]. They both 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 had already been introduced by Shouté in 1911 [37]. 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 then interpreted as a *square axiom* requiring that if a geometric edge of a square belongs to $\mathcal{N}$ then the opposite edge does so. 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 opposite edges in the hexagon.

The second equivalent definition is in terms of (complete) *polyhedral fans*, which are certain collections of polyhedral cones partitioning $\mathbb{R}^N$. 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 or *braid arrangement* by other authors. Then semi-graphoids are in one-to-one correspondence with polyhedral fans which coarsen the prominent S_n-fan.

1.6 Markov properties for undirected graphs

This section contains some theoretical results concerning undirected graphical models, named *Markov networks* in the context of probabilistic reasoning [34].

1.6.1 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 route in G from a node in A to a node in B contains a node in C. Of course, this is equivalent to the same condition with paths in place of routes. 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.

Thus, 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; an alternative terminology is that P satisfied 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 [13, 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 case non-trivial sample spaces X_i, $i \in N$). Another related result is that formal independence models induced by UGs can be described in an axiomatic way, that is, they are characterized in terms of finitely many CI implications [35].

1.6.2 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 the number of CI statements to be tested may be very high. Nevertheless, in case of a measure with a (strictly) positive density reasonable sufficient conditions exist.

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

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

(P) for any distinct $i,j \in N$ with $\neg(i \,—\, j)$ in $G \qquad i \perp\!\!\!\perp j \,|\, N \setminus \{i,j\} \; [P]$.

It is easy to verify using Observation 1.4.1(iii) that (G)⇒(L)⇒(P); however, examples that (P)⇏(L)⇏(G) for discrete distributions are available [17].

Observation 1.6.1 Assume that a probability measure P over N has strictly positive density. Then one has (G)⇔(L)⇔(P) for P.

Proof 4 The key fact is the property 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 $i, j \in N$, $\neg(i \;\text{---}\; j \text{ in } G)$, contains the whose formal independence model $\mathcal{M}_G$. This observation is left to the reader as an exercise.

Note that the undirected *separation criterion* from § 1.6.1 was a result of some evolution in theory of Markov fields, which stemed from statistical physics. The authors who had developed this theory in the 1970s restricted their attention to positive discrete probability distributions. Several types of Markov conditions were proposed in [32]: the original pairwise Markov property was strengthened to the local and global one. 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 [13] 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.

1.6.3 Factorization property for an UG

There is another sufficient condition for the global Markov property, which does not demand the distribution to have a positive density. Specifically, we say that a marginally continuous measure P over N *factorizes* 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 $\psi_C : X_C \to [0, \infty)$, $C \in \mathcal{C}_G$, 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)⇒(G), which observation can be derived from repeated application of the fact that the factorization condition (1.5) is an equivalent definition of CI; see [18, Proposition 1]. On the other hand, examples of discrete measures showing (G)⇏(F) exist [24]. Nevertheless, the conditions are quite often equivalent. The following result, whose proof is omitted, is known as the *Hammersley-Clifford theorem*, see [17, Theorem 3.9]. It is very useful observation as discussed in chapter 3 of this book.

Observation 1.6.2 Assume that a probability measure P over N has strictly positive density. Then one has (F)⇔(G) for P.

1.7 Markov properties for directed graphs

This section deals with directed acyclic graphical models, named *Bayesian networks* in the context of probabilistic reasoning [34].

1.7.1 Directional separation criterion

In the directed case, there are different but equivalent separation criteria to decide whether a disjoint triplet is represented in a graph. In this chapter, only a straightforward directional separation criterion for routes is presented, which is probably the simplest one. In this subsection we assume that G is a directed graph over N; it is not substantial whether G is acyclic or not.

Let $\rho : i_1, \ldots, i_k$, $k \geq 1$, be a route 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 $i_{\ell-1} \rightarrow i_\ell \leftarrow i_{\ell+1}$ in G. Other occurrences of nodes in ρ, including its end-nodes, are named *non-colliders.* We say that ρ is *blocked* 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* C.

Thus, the blocking condition for non-colliders is the same as in the undirected case (see § 1.6.1), while the condition for colliders is completely converse. It also follows from the definition that if a route has a node with both collider and non-collider occurrences then it must be blocked by any $C \subseteq N$. A route in G which is not blocked by a set $C \subseteq N$ will be called C*-free.*

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 *route* in G from a node in A to a node in B is blocked by C and write $A \perp\!\!\!\perp B \,|\, C \; [G]$ then. Note that one has to consider all routes from A to B, not just paths. For example, in a graph over $N = \{i, j, k, l\}$ with arrows $i \rightarrow l$, $l \rightarrow k$ and $j \rightarrow l$ the only path from i do j is $i \rightarrow l \leftarrow j$, which is blocked by the set $C = \{k\}$. However, a route $i \rightarrow l \rightarrow k \leftarrow l \leftarrow j$ exists in the graph which is C-free.

Since the criterion is formulated in terms of routes a natural question arises whether it is decidable. Indeed, there exists a propagation *Bayes-ball* algorithm [38] which, for given disjoint sets of nodes A and C, finds the set $\bar{A}$ of nodes to which a C-free route exists from a node in A. Thus, if B is disjoint with $\bar{A}$, then directional separation holds, otherwise not.

The directional separation criterion is close to the *d-separation criterion*, which was proposed by Pearl [34]. An equivalent *moralization criterion* was suggested by Lauritzen and his co-authors [17]; it is based on transformation of the directed graph to a certain UG and using undirected separation. The equivalence of these criteria (in case of a DAG) was shown in [17, Proposition 3.25]. There are other criteria, for example Massey [22] offered another criterion based on a (different) transformation of the graph into an UG.

1.7.2 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 satisfied 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 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 [12].

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 [47, Remark 3.5].

1.7.3 Local Markov property for a DAG

In the directed case several variations of both local and pairwise Markov properties exist. One can distinguish ordered versions, when an enumeration of nodes consonant with the direction of arrows is given and the Markov property is relative to it, and unordered versions; see [4, § 5.3]. In this section, a basic unordered version of the local Markov property is presented.

To formulate it an additional graphical concept is needed. If there exists a directed path in G from a node $i \in N$ to a 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. The set of descendants of a node $i \in N$ in G will be denoted by $\mathrm{ds}_G(i)$.

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

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

Observation 1.7.1 For any probability measure P over N, (DG)⇔(DL).

Proof 5 Given any enumeration $i_1, \ldots, i_{|N|}$ of nodes which is consonant with the direction of arrows G, it was shown in [54] 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.

1.7.4 Factorization property for a DAG

Recursive factorization condition is a necessary and sufficient condition for a marginally continuous measure being Markovian with respect to a *directed acyclic graph.* In 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 is analogous in case of a marginally continuous measure, but one has to introduce correctly conditional densities and the equation in (DF) is meant in μ-a.e. sense, where μ is a dominating joint product measure. One can show that (DF)$\Leftrightarrow$(DG) then; see [18, 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 [47, Lemma 8.1].

1.8 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.

1.8.1 The concept of a structural imset

Although graphs offer an elegant and intuitive interpretation of (some of) CI structures, they are not able to describe all possible probabilistic CI structures. This motivated 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 [47, § 4.2.1], which is a vector in $\mathbb{R}^{\mathcal{P}(N)}$ encoding an 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 $\mathbb{R}^{\mathcal{P}(N)}$ generated by (all) elementary imsets over N. *Structural imsets*, used to describe CI structures, can equivalently be introduced as vectors in $\mathcal{S}(N) \cap \mathbb{Z}^{\mathcal{P}(N)}$ [16]. 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 true if $|N| \leq 4$ but Hemmecke et al. [15] gave an example of a structural imset over N with $|N| = 5$ which is not a combinatorial one.

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 every $\langle A, B|C\rangle \in \mathcal{T}(N)$, whether $A \perp\!\!\!\perp B \,|\, C\; [u]$ holds; the criterion is omitted in this chapter and can be found in [47, § 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 [47, Theorem 4.1] that, for marginal 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 § 1.7.4.

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 [47, Theorem 5.2]. In other words, any such distribution is *perfectly Markovian* with respect to some combinatorial imset u, which means $\mathcal{M}_u = \mathcal{M}_P$. Note that any discrete measure and any regular Gaussian measure over N has finite multiinformation.

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 that semi-graphoids correspond to polyhedral fans coarsening the S_n-fan (see § 1.5.1). Morton [30] also showed that a semi-graphoid is structural iff the corresponding polyhedral fan is a normal fan of a polytope.

1.8.2 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 usual criteria for learning BN structure become affine functions (= sums of linear functions with constants) of the standard imset [51]. Thus, the learning task can be transformed into a *linear programming* (LP) 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 a linear invertible self-transformation of $\mathbb{Z}^{\mathcal{P}(N)}$ into *characteristic imsets*, which are zero-one vectors with elegant graphical interpretation [14], and these vectors were applied to learning BN structure by tools of integer linear programming [50]. This approach seems to be particularly suitable for learning decomposable models [49], in which case there is hope that the corresponding polytope will be characterized completely by linear inequalities.

1.9 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 as the result of application of interpretable formal CI implications, analogous to the semi-graphoid inference rules from Definition 6. Although there is no finite set of inference rules characterizing CI inference [43] one can find such an axiomatic characterization in some special cases. The semi-graphoid implications are enough in case $|\mathcal{L}| = 2$ [45] or if $\mathcal{L}$ consists of special CI statements, like the marginal CI statements $A \perp\!\!\!\perp B \,|\, \emptyset$ [11, 25] or saturated CI statements $A \perp\!\!\!\perp B \,|\, C$ with $ABC = N$ [21, 13].

Matúš [27] characterized the CI closure for discrete measures if $|N| = 4$; in this case 24 formal properties are enough [48]. Several methods to derive implications among CI statements can be used. The method of structural imsets [47, § 6.2] provides a sufficient condition for probabilistic CI implication; the respective linear-algebraic criterion can be tested using a computer [3]. The most efficient methods for computer testing of that sufficient condition seems to be linear programming methods [2, 33]. On the other hand, there are linear-algebraic tools to derive CI implications based on different principles [52]. On the top of that, advanced methods of modern algebra can be used to derive CI implications; chapter 3 gives more details on this topic.

Acknowledgement

I am indebted to Fero Matúš for his cooperation on the topic of CI. Our work is supported from the grant project GAČR n. 16-12010S.

Bibliography

[1] 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.

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

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

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

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

[6] 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.

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

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

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

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

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

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

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

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

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

[16] 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.

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

[18] 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.

[19] 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.

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

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

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

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

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

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

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

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

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

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

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

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

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

[33] 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.

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

[35] 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.

[36] Y. Sagiv and S. F. Walecka. Subset dependencies and completeness result for a subclass of embedded multivalued dependencies. *Journal of Association for Computing Machinery*, 29(1):103–117, 1982.

[37] 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.

[38] 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.

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

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

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

[42] 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.

[43] 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.

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

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

[46] 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.

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

[48] 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.

[49] M. Studený and J. Cussens. The chordal graph polytope for learning decomposable models. In *JMLR Workshops and Conference Proceedings*, volume 52, pages 499–510, 2016.

[50] 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.

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

[52] 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.

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

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

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

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

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