

Basic Concepts

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

CONVENTION 1. The following conventions will be used throughout the book. Given sets $A, B \subseteq N$ the juxtaposition AB will denote their union $A \cup B$. The following symbols will be reserved for sets of numbers: \mathbb{R} will denote *real numbers*, \mathbb{Q} *rational numbers*, \mathbb{Z} *integers*, \mathbb{Z}^+ *non-negative integers* (including 0), \mathbb{N} *natural numbers* (that is, positive integers excluding 0). The symbol $|A|$ will be used to denote the number of elements of a finite set A , that is, its *cardinality*. The symbol $|x|$ will also denote the *absolute value* of a real number x , that is, $|x| = \max\{x, -x\}$. \diamond

2.1 Conditional independence

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

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

Moreover, let us accept two natural conventions. First, the marginal of P for $A = N$ is P itself, that is, $P^N \equiv P$. Second, a fully formal convention is that the marginal of P for $A = \emptyset$ is a probability measure on a (fixed appended)

measurable space $(X_\emptyset, \mathcal{X}_\emptyset)$ with a trivial σ -algebra $\mathcal{X}_\emptyset = \{\emptyset, X_\emptyset\}$. Observe that a measurable space of this kind only admits one probability measure P^\emptyset .

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

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

Note that no restriction concerning the mappings $A \mapsto P_{A|C}(A|x)$, $x \in X_C$ (often called the regularity requirement – see Section A.6.4, Remark A.1) is needed within this general approach. Let me emphasize that $P_{A|C}$ only depends on the marginal P^{AC} and that it is defined, for a fixed $A \in \mathcal{X}_A$, uniquely within the equivalence P^C -almost everywhere (P^C -a.e.). Observe that, owing to the convention above, if $C = \emptyset$ then the conditional probability $P_{A|C}$ coincides, in fact, with the marginal for A , that means, one has $P_{A|\emptyset} \equiv P^A$ (because a constant function can be identified with its value).

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

$$\bar{\mathcal{X}}_A = \{A \times X_{N \setminus A}; A \in \mathcal{X}_A\} = \{B \in \mathcal{X}_N; B = A \times X_{N \setminus A} \text{ for } A \subseteq X_A\}.$$

Thus, $A \subseteq B \subseteq N$ is reflected by $\bar{\mathcal{X}}_A \subseteq \bar{\mathcal{X}}_B$ and it is natural to require that the empty set \emptyset is represented by the trivial σ -algebra $\bar{\mathcal{X}}_\emptyset$ over X_N and N is represented by $\bar{\mathcal{X}}_N = \mathcal{X}_N$. Using this point of view, the marginal P^A corresponds to the restriction of P to $\bar{\mathcal{X}}_A$, and $P_{A|C}$ corresponds to the concept of conditional probability with respect to the σ -algebra $\bar{\mathcal{X}}_C$. Thus, the existence and the uniqueness of $P_{A|C}$ mentioned above follows from basic measure-theoretical facts. For details see the Appendix, Section A.6.4. \triangle

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

$$P_{AB|C}(A \times B|x) = P_{A|C}(A|x) \cdot P_{B|C}(B|x) \quad \text{for } P^C\text{-a.e. } x \in X_C. \quad (2.1)$$

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

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

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

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

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

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

However, the concept of conditional independence is not exclusively a probabilistic concept. This concept was introduced in several non-probabilistic frameworks, namely in various calculi for dealing with uncertainty in artificial intelligence – for details and overview see [133, 117, 31]. Formal properties of respective conditional independence concepts may differ in general, but an important fact is that certain basic properties of conditional independence appear to be valid in all these frameworks.

2.2 Semi-graphoid properties

Several authors independently drew attention to the above-mentioned basic formal properties of conditional independence. In modern statistics, they were first accentuated by Dawid [29], then mentioned by Mouchart and Rolin [93], and van Putten and van Shuppen [103]. Spohn [124] interpreted them in the context of philosophical logic. Finally, their significance in (probabilistic approach to) artificial intelligence was discerned and highlighted by Pearl and Paz [99]. Their terminology [100] was later widely accepted, so that researchers in artificial intelligence started to call them the *semi-graphoid properties*.

2.2.1 Formal independence models

Formally, a *conditional independence statement over N* is a statement of the form “ A is conditionally independent of B given C ” where $A, B, C \subseteq N$ are pairwise disjoint subsets of N . A statement of this kind should always be understood with respect to a certain mathematical object \mathbf{o} over N , for example, a probability measure over N . However, several other objects can occur in place of \mathbf{o} ; for example, a graph over N (see Chapter 3), a possibility

distribution over N [18, 149], a relational database over N [112] and a structural imset over N (see Section 4.4.1). The notation $A \perp\!\!\!\perp B | C [\mathbf{o}]$ will be used in those cases, but the symbol $[\mathbf{o}]$ can be omitted if it is suitable.

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

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

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

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

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

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

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

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

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

Remark 2.4. To avoid misunderstanding, the reader should be aware that the noun *model* may have any of three different meanings in this monograph. First, it can be used in its general sense in which case it is usually used without an adjective. Second, it is a part of the phrase “(formal) independence model” in which case the word *independence* indicates that one has in mind the concept introduced in this section. Note that this terminology comes from the area of artificial intelligence – see Pearl [100]. Third, it can be a part of the phrase “statistical model” in which case the adjective *statistical* indicates that one has in mind the concept mentioned in Section A.9.2, that is, a class of probability measures. Note that this terminology is often used in statistics – see Remark A.3 for more detailed explanation.

However, there is a simple reason why two different concepts are named by the same noun. The reason is that every formal independence model $\mathcal{M} \subseteq \mathcal{T}(N)$ can be understood as a statistical model \mathbb{M} , provided that a distribution framework Ψ (see Section A.9.5) is fixed. Indeed, one can put

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

Every statistical model of this kind will be called the *statistical model of CI structure*. Note that this concept generalizes the classic concept of a graphical model [157, 70]. Indeed, the reader can learn in Chapter 3 that a graph G having N as the set of nodes usually induces the class \mathbb{M}_G of Markovian measures over N , that is, a statistical model. This graphical statistical model is, however, defined by means of the formal independence model \mathcal{M}_G . Note that the class \mathbb{M}_G is often introduced in another way – see Section 8.2.1 for equivalent definitions in case of acyclic directed graphs in terms of recursive factorization and in terms of parameterization. \triangle

2.2.2 Semi-graphoids

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

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

Note that the terminology above was proposed by Pearl [100], who formulated the formal properties above in the form of inference rules, gave them special names and interpretation, and called them the *semi-graphoid axioms*. Of course, the restriction of a semi-graphoid over N to $\mathcal{T}(T)$ for non-empty $T \subseteq N$ is a semi-graphoid over T . The following fact is important.

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

Proof. This can be derived easily from Corollary A.2 proved in the Appendix (see p. 235). Indeed, having a probability measure P over N defined on a measurable space $(\mathbf{X}_N, \mathcal{X}_N)$ one can identify every subset $A \subseteq N$ with a coordinate σ -algebra $\bar{\mathcal{X}}_A \subseteq \mathcal{X}_N$ as described in Remark 2.1. Then, for a disjoint triplet $\langle A, B | C \rangle$ over N , the statement $A \perp\!\!\!\perp B | C [P]$ is equivalent to the requirement $\bar{\mathcal{X}}_A \perp\!\!\!\perp \bar{\mathcal{X}}_B | \bar{\mathcal{X}}_C [P]$ introduced in Section A.7. Having in mind that $\bar{\mathcal{X}}_{AB} = \bar{\mathcal{X}}_A \vee \bar{\mathcal{X}}_B$ for $A, B \subseteq N$ the rest follows from Corollary A.2. \square

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

Remark 2.5. The limitation to disjoint triplets in the definition of a semi-graphoid is not substantial. One can introduce an *abstract semi-graphoid* on a join semi-lattice (\mathcal{S}, \vee) as a ternary relation $\star \perp\!\!\!\perp \star | \star$ over elements A, B, C, D of \mathcal{S} satisfying

- $A \perp\!\!\!\perp B | C$ whenever $B \vee C = C$,
- $A \perp\!\!\!\perp B | C$ iff $B \perp\!\!\!\perp A | C$,
- $A \perp\!\!\!\perp B \vee D | C$ iff $[A \perp\!\!\!\perp B | D \vee C \text{ and } A \perp\!\!\!\perp D | C]$.

Taking $\mathcal{S} = \mathcal{P}(N)$ one obtains the definition of a non-disjoint semi-graphoid over N . A more complicated example is the semi-lattice of all σ -algebras $\mathcal{A} \subseteq \mathcal{X}$ in a measurable space $(\mathbf{X}, \mathcal{X})$ and the relation $\perp\!\!\!\perp$ of conditional independence for σ -algebras with respect to a probability measure on $(\mathbf{X}, \mathcal{X})$ (see Corollary A.2). Note that the above concept of an abstract semi-graphoid is essentially equivalent to the concept of a *separoid* introduced by Dawid [33], which is a mathematical structure unifying a variety of notions of “conditional independence” arising in probability, statistics, artificial intelligence, and other fields.

Let me conclude this remark by a note which indicates the obstacles that authors in mathematics meet if they want to establish new terminology. Pearl and Paz [99] decided to use the word “graphoid” to name a new concept they introduced (see p. 29 for this concept). However, it appeared that this word had already been “occupied”: it was used to name one of equivalent definitions of a matroid [155]. One of the motives which led Dawid [33] to use the word

“separoid” to name his general concept was to avoid a terminological clash. However, it appeared that this word had also been used independently by Strausz [128] to name a certain abstract binary relation between sets whose aim is to generalize geometric separation of sets in \mathbb{R}^n by hyperplanes. An interesting observation is that, by coincidence, there is a weak connection between two concepts of a separoid. For example, an undirected graph G and the relation of separation for sets of nodes in G , which is defined as in Section 3.1 but non-disjoint sets are allowed, can give an example of both separoids. The difference is that Dawid’s separoid is a ternary relation $A \perp\!\!\!\perp B | C [G]$ while a binary relation $A \perp\!\!\!\perp B | \emptyset [G]$ can serve as an example of Strausz’s separoid. \triangle

2.2.3 Elementary independence statements

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

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

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

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

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

Proof. (see also [79]) The necessity of the condition (2.2) is easily derivable using the decomposition and the weak union properties combined with the symmetry property.

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

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

Remark 2.6. Another reduction of memory demands for semi-graphoid representation follows from the symmetry property. Instead of keeping a pair of mutually symmetric statements $a \perp\!\!\!\perp b \mid K$ and $b \perp\!\!\!\perp a \mid K$ one can choose only one of them according to a suitable criterion. In particular, to represent a semi-graphoid over N with $|N| = n$ it suffices to have only $n \cdot (n - 1) \cdot 2^{n-3}$ bits. Note that the idea above is also reflected in Section 4.2.1 where just one elementary imset corresponds to a “symmetric” pair of elementary statements.

However, further reduction of the class of considered statements is not possible. The reason is as follows: every elementary triplet $\langle a, b \mid K \rangle$ over N generates a semi-graphoid over N consisting of $\langle a, b \mid K \rangle$, its symmetric image $\langle b, a \mid K \rangle$ and trivial triplets over N (cf. Lemmas 4.6 and 4.5). In fact, these are minimal non-trivial semi-graphoids over N and one has to distinguish them from other semi-graphoids over N . These observations influenced the terminology: the adjective “elementary” is used to indicate the respective disjoint triplets and independence statements. \triangle

2.2.4 Problem of axiomatic characterization

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

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

Another formal property of this sort was later derived in An et al. [3]. Consequently, a natural question occurred. Can conditional independence models arising in a discrete probabilistic setting be characterized in terms of a finite number of formal properties of this type? This question is known as the *problem of axiomatic characterization* because a result of this kind would have been a substantial step towards a syntactic description of these models in the sense of mathematical logic. Indeed, as explained in § 5 of Studený [132], then it would have been possible to construct a deductive system that is an analog of the notion of a “formal axiomatic theory” from Mendelson [92]. The considered formal properties then would have played the role of syntactic inference rules of an axiomatic theory of this sort. Unfortunately, the answer to the question above is also negative. It was shown in Studený [132] (for a more didactic proof see [144]) that, for every $n \in \mathbb{N}$, there exists a formal property

of (discrete) probabilistic conditional independence models which applies to a set of variables N with $|N| = n$ but which cannot be revealed on a set of smaller cardinality. Note that a basic tool for derivation of these properties was the multiinformation function introduced in Section 2.3.4.

On the other hand, having fixed N , a finite number of possible probabilistic conditional independence models over N suggests that they can be characterized in terms of a finite number of formal properties of semi-graphoid type. Thus, a related task is, for a small cardinality of N , to characterize them in that way. It is no problem to verify that they coincide with semi-graphoids in the case $|N| = 3$ (see Figure 5.6 for illustration). Discrete probabilistic conditional independence models over N with $|N| = 4$ were characterized in a series of papers by Matúš [84, 85, 87]; for an overview see Studený and Boček [136] where the respective formal properties of these models are explicitly formulated – one has 18300 different models of this kind and these can be characterized by more than 28 formal properties.

Remark 2.7. On the other hand, several results on relative completeness of semi-graphoid properties were achieved. In Geiger et al. [45] and independently in Matúš [82] models of “unconditional” stochastic independence (that is, submodels consisting of *unconditioned independence statements* of the form $A \perp\!\!\!\perp B \mid \emptyset$) were characterized by means of properties derivable from the semi-graphoid properties. An analogous result for the class of *saturated* or *fixed-context conditional independence statements* – that is, statements $A \perp\!\!\!\perp B \mid C$ with $ABC = N$ – was achieved independently by Geiger and Pearl [46] and by Malvestuto [77]. The result from Studený [138] can be interpreted as a specific relative-completeness result, saying that the semi-graphoid generated by a pair of conditional independence statements is always a conditional independence model induced by a discrete probability measure. Note that the problem of axiomatic characterization of CI models mentioned above differs from the problem of axiomatization (in the sense of mathematical logic) of a single CI structure over an infinite set of variables N , which was treated in Kramosil [62]. \triangle

2.3 Classes of probability measures

There is no uniformly accepted conception of the notion of a *probability distribution* in the literature. In probability theory, authors usually understand by a distribution of a (n -dimensional real) random vector an induced probability measure on the respective sample space (\mathbb{R}^n endowed with the Borel σ -algebra), that is, a **set function** on the sample (measurable) space. On the other hand, authors in artificial intelligence usually identify a distribution of a (finitely valued) random vector with a **pointwise function** on the respective (finite) sample space, ascribing probability to every configuration of values (= to every element of the sample space $\prod_{i \in N} \mathbf{X}_i$, where \mathbf{X}_i are finite sets). In

statistics, either the meaning wavers between these two basic approaches, or authors even avoid the dilemma by describing specific distributions directly by their parameters (e.g., elements of the covariance matrix of a Gaussian distribution). Therefore, no exact meaning is assigned to the phrase “probability distribution” in this book; it is used only in its general sense, mainly in vague motivational parts. Moreover, terminological distinction is made between those two above-mentioned approaches. The concept of a *probability measure* over N from Section 2.1 more likely reflects the first approach, which is more general. To relate this to the second approach one has to make an additional assumption on a probability measure P so that it can also be described by a pointwise function, called the *density* of P . Note that many authors simply make an assumption of this type implicitly without mentioning it.

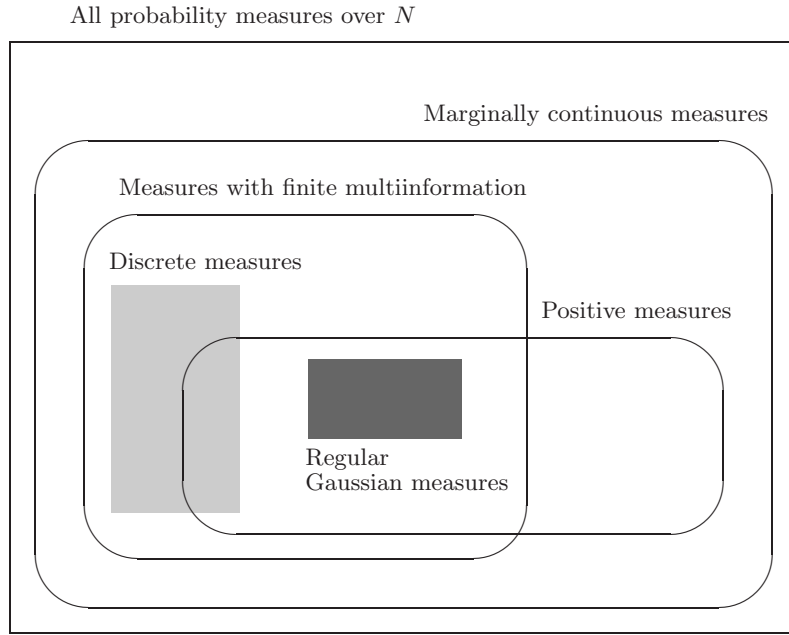


Fig. 2.1. A comparison of basic classes of probability measures over N .

In this section, basic facts about these special probability measures are recalled and several important subclasses of the class of measures having density, called “marginally continuous measures”, are introduced. One of them, the class of measures with finite multiinformation, is strongly related to the method of structural imsets described in later chapters. The information-theoretical methods are applicable to measures belonging to this class which, fortunately, involves typical measures used in practice. Inclusion relationships among introduced classes of measures are depicted in Figure 2.1.

2.3.1 Marginally continuous measures

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

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

Proof. (see also § 1.2.2 in [37]) It was shown in [130], Proposition 1, that in the case $|N| = 2$ one has $P \ll \prod_{i \in N} P^{\{i\}}$ iff there are probability measures λ_i on (X_i, \mathcal{X}_i) with $P \ll \prod_{i \in N} \lambda_i$. One can easily show that for every non-zero σ -finite measure μ_i on (X_i, \mathcal{X}_i) a probability measure λ_i on (X_i, \mathcal{X}_i) with $\mu_i \ll \lambda_i \ll \mu_i$ exists. Hence, the condition above is equivalent to the requirement for the existence of σ -finite measures μ_i with $P \ll \prod_{i \in N} \mu_i$. Finally, one can use the induction on $|N|$ to get the desired conclusion. \square

Thus, the marginal continuity of P is equivalent to the existence of a *dominating measure* μ for P , that is, the product $\mu = \prod_{i \in N} \mu_i$ of some σ -finite measures μ_i on (X_i, \mathcal{X}_i) , $i \in N$ such that $P \ll \mu$. In particular, every discrete measure over N is marginally continuous since the counting measure on X_N can serve as its dominating measure. Note that nearly all multidimensional measures used in practice are marginally continuous (see Sections 2.3.5, 2.3.6 and 4.1.3 for other examples). However, there are probability measures over N which are not marginally continuous; in particular, some singular Gaussian measures – see Example 2.3 on p. 35.

Having fixed a dominating measure μ for a marginally continuous measure P over N by a *density of P with respect to μ* will be understood (every version of) the Radon-Nikodym derivative of P with respect to μ .

Remark 2.8. Let us note without explaining details (see Remark 1 in [130]) that the assumption that a probability measure P over N is marginally continuous also implies that, for every disjoint $A, C \subseteq N$, there exists a regular version of conditional probability $P_{A|C}$ on X_A given \mathcal{X}_C in the sense of Loève [74]. The regularity of conditional probability is usually derived as a consequence of special topological assumptions on (X_i, \mathcal{X}_i) , $i \in N$ (see the Appendix, Remark A.1). Thus, the marginal continuity is a non-topological assumption implying the regularity of conditional probabilities. The concept of marginal continuity is closely related to the concept of a *dominated experiment* in Bayesian statistics – see § 1.2.2 and § 1.2.3 in the book by Florens et al. [37]. \triangle

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

CONVENTION 2. Suppose that a marginally continuous probability measure P on (X_N, \mathcal{X}_N) is given. Let us fix one-dimensional σ -finite measures which define a dominating measure μ for P . More specifically, $P \ll \mu \equiv \prod_{i \in N} \mu_i$ where μ_i is a σ -finite measure on (X_i, \mathcal{X}_i) for every $i \in N$.

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

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

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

Remark 2.9. This is to explain the way of defining marginal densities in Convention 2. First, let me emphasize that the marginal density is **not** the Radon-Nikodym derivative of respective marginals of P and μ since $\mu_A = \prod_{i \in A} \mu_i$ need not coincide with the marginal μ^A of $\mu = \prod_{i \in N} \mu_i$ unless every μ_i is a probability measure.

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

$$f^{\downarrow A}(y) = \int_{X_{N \setminus A}} f(y, z) d\mu_{N \setminus A}(z) \quad \text{for } y \in X_A.$$

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

$$f^{\downarrow \emptyset}(\star) = \int_{X_N} f(x) d\mu(x) = 1.$$

follows this line. \triangle

Lemma 2.4. Let P be a marginally continuous measure over N . Let us accept Convention 2. Given $\langle A, B | C \rangle \in \mathcal{T}(N)$ one has $A \perp\!\!\!\perp B | C [P]$ iff the following equality holds

$$f_{ABC}(x_{ABC}) \cdot f_C(x_C) = f_{AC}(x_{AC}) \cdot f_{BC}(x_{BC}) \quad \text{for } \mu\text{-a.e. } x \in X_N. \quad (2.3)$$

Proof. Note that minor omitted details of the proof (e.g. verification of equalities μ -a.e.) can be verified with the aid of basic measure-theoretical facts gathered in Section A.6.

I. First, choose and fix a density $f : \mathbf{X}_N \rightarrow [0, \infty)$ of P such that

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

and, moreover, for every disjoint $A, C \subseteq N$, one has

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

where conventions $f^{\downarrow N} = f$ and $f^{\downarrow \emptyset} \equiv 1$ are accepted. Indeed, these relationships hold μ -a.e. for every version f of $dP/d\mu$ and every version can be overdefined by 0 whenever these relationships do not hold. It is no problem to verify that $f^{\downarrow A} = dP^A/d\mu_A$ for every $\emptyset \neq A \subseteq N$.

II. Second, for every disjoint pair of sets $A, C \subseteq N$, introduce a function $h_{A|C} : \mathbf{X}_A \times \mathbf{X}_C \rightarrow [0, \infty)$ as follows:

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

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

$$(A, z) \mapsto P_{A|C}(A|z) \equiv \int_A h_{A|C}(x|z) \, d\mu_A(x) \quad \text{where } A \in \mathcal{X}_A, z \in \mathbf{X}_C,$$

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

III. Realize that (2.3) can be written as follows (see Remark 2.9):

$$f^{\downarrow ABC}(x_{ABC}) \cdot f^{\downarrow C}(x_C) = f^{\downarrow AC}(x_{AC}) \cdot f^{\downarrow BC}(x_{BC}) \quad (2.5)$$

for μ -a.e. $x \in \mathbf{X}_N$. Further, this can be rewritten in the form

$$h_{AB|C}(x_{AB}|x_C) \cdot f^{\downarrow C}(x_C) = h_{A|C}(x_A|x_C) \cdot h_{B|C}(x_B|x_C) \cdot f^{\downarrow C}(x_C) \quad (2.6)$$

for μ -a.e. $x \in \mathbf{X}_N$. Indeed, owing to (2.4), (2.5) and (2.6) are trivially valid on the set $\{x \in \mathbf{X}_N; f^{\downarrow C}(x_C) = 0\}$ while they are equivalent on its complement.

IV. The next step is to observe that (2.6) is equivalent to the requirement that $\forall A \in \mathcal{X}_A, \forall B \in \mathcal{X}_B, \forall C \in \mathcal{X}_C$ it holds

$$\begin{aligned} & \int_C \int_{A \times B} h_{AB|C}(x_{AB}|x_C) \, d\mu_{AB}(x_{AB}) \, dP^C(x_C) = \\ & = \int_C \int_A h_{A|C}(x_A|x_C) \, d\mu_A(x_A) \cdot \int_B h_{B|C}(x_B|x_C) \, d\mu_B(x_B) \, dP^C(x_C). \end{aligned}$$

Indeed, as mentioned in Section A.6.1 the equality in (2.6) is equivalent to the requirement that their integrals with respect to μ_{ABC} over all measurable rectangles $A \times B \times C$ coincide. This can be rewritten using the Fubini theorem, the Radon-Nikodym theorem and basic properties of the Lebesgue integral in the form above.

V. As explained in Step II, the last equation can be understood as follows:

$$\int_C P_{AB|C}(A \times B|z) dP^C(z) = \int_C P_{A|C}(A|z) \cdot P_{B|C}(B|z) dP^C(z). \quad (2.7)$$

Having fixed $A \in \mathcal{X}_A$ and $B \in \mathcal{X}_B$ the equality (2.7) for every $C \in \mathcal{X}_C$ is equivalent to the condition that the integrated functions are equal P^C -a.e. Hence, one can conclude that the condition (2.1) from p. 10 holds for every $A \in \mathcal{X}_A$ and $B \in \mathcal{X}_B$, that is, $A \perp\!\!\!\perp B | C [P]$. \square

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

2.3.2 Factorizable measures

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

$$f_D(x_D) = \prod_{S \in \mathcal{D}} g_S(x_S) \quad \text{for } \mu\text{-a.e. } x \in X_N, \quad (2.8)$$

where $g_S : X_S \rightarrow [0, \infty)$, $S \in \mathcal{D}$ are \mathcal{X}_S -measurable functions, called *potentials*. An equivalent formulation is that there exists a version of f_D of $dP^D/d\mu$ and potentials g_S such that (2.8) holds for every $x \in X_N$. In fact, the factorization does not depend on the choice of a dominating measure μ . One can show that the validity of (2.8) relative to a general dominating product measure $\mu = \prod_{i \in D} \mu_i$ where all μ_i are σ -finite, is equivalent to the validity of (2.8) relative to $\prod_{i \in D} P^{\{i\}}$ and with other potentials (this can be verified with the help of Lemma 2.3). Of course, the factorization after \mathcal{D} is equivalent to the factorization after \mathcal{D}^{\max} , and potentials are not unique unless $|\mathcal{D}| = 1$.

Further equivalent definition of conditional independence for marginally continuous measures is formulated in terms of factorization (see also [70], §3.1).

Lemma 2.5. Let P be a marginally continuous measure over N , μ a dominating measure for P^{ABC} and $\langle A, B|C \rangle$ a disjoint triplet over N . Then $A \perp\!\!\!\perp B|C [P]$ if and only if P factorizes after $\mathcal{D} = \{AC, BC\}$ relative to μ . More specifically, if Convention 2 is accepted then $A \perp\!\!\!\perp B|C [P]$ iff there exist an \mathcal{X}_{AC} -measurable function $g : \mathbf{X}_{AC} \rightarrow [0, \infty)$ and an \mathcal{X}_{BC} -measurable function $h : \mathbf{X}_{BC} \rightarrow [0, \infty)$ such that

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

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

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

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

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

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

Now, using the Fubini theorem and basic properties of the integral mentioned in Section A.6.1, one can derive from (2.10) by integrating

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

where the functions

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

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

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

which is equivalent to (2.3). \square

As a consequence, one can derive a certain formal property of conditional independence which was already mentioned in the discrete case (see [3, 125] and Proposition 4.1 in [81]).

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

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

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

$$f(x_{ABCD}) \cdot f(x_{AB}) \cdot f(x_\emptyset) \cdot f(x_C) \cdot f(x_D) = \\ = f(x_{ABC}) \cdot f(x_{ABD}) \cdot f(x_\emptyset) \cdot f(x_C) \cdot f(x_D) \quad \text{for } \mu\text{-a.e. } x \in X_N.$$

Now, again using Lemma 2.4, the assumptions $A \perp\!\!\!\perp B \mid \emptyset$, $A \perp\!\!\!\perp B \mid C$ and $A \perp\!\!\!\perp B \mid D$ imply that

$$f(x_{ABCD}) \cdot f(x_A) \cdot f(x_B) \cdot f(x_C) \cdot f(x_D) = \\ = f(x_{AC}) \cdot f(x_{BC}) \cdot f(x_{AD}) \cdot f(x_{BD}) \cdot f(x_\emptyset) \quad \text{for } \mu\text{-a.e. } x \in X_N.$$

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

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

Hence, by Lemma 2.5 one has $A \perp\!\!\!\perp B \mid CD$. \square

2.3.3 Multiinformation and conditional product

Let P be a marginally continuous measure over N . The *multiinformation* of P is the relative entropy $H(P \mid \prod_{i \in N} P^{\{i\}})$ of P with respect to the product of its one-dimensional marginals. It is always a value in $[0, +\infty]$ (see Lemma A.4 in Section A.6.3). A common formal convention is that the multiinformation of P is $+\infty$ in case P is not marginally continuous.

Remark 2.10. The term “multiinformation” was proposed by my PhD supervisor Albert Perez in the late 1980s. Note that miscellaneous other terms were used earlier in the literature (even by Perez himself); for example “total correlation” [154], “dependence tightness” [101] or “entaxy” [76]. The main reason for Perez’s later terminology is that the above concept directly generalizes a widely accepted information-theoretical concept of “mutual information” of two random variables; multiinformation can be applied to the case of any finite number of random variables. Indeed, it can serve as a measure of global

stochastic dependence among a finite collection of random variables (see § 4 in Studený and Vejnarová [144]). Asymptotic behavior of “empirical multiinformation”, which can be used as a statistical estimate of multiinformation on the basis of data, was examined in Studený [129]. \triangle

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

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

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

Moreover, $P^{ABC} \ll Q \ll \prod_{i \in ABC} P^{\{i\}}$ and the following equality holds true (the symbol H denotes the relative entropy introduced in Section A.6.3):

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

Proof. Note again that omitted technical details can be verified by means of basic measure-theoretical facts from Section A.6.

I. First, let us verify the uniqueness of Q . Supposing both Q^1 and Q^2 satisfy (2.12) one can observe that $(Q^1)^C = (Q^2)^C$ and $Q_{A|C}^1 \approx Q_{A|C}^2$, $Q_{B|C}^1 \approx Q_{B|C}^2$, where \approx indicates the respective equivalence of conditional probabilities (on \mathbf{X}_A resp. \mathbf{X}_B) given C mentioned in Section 2.1. Because of $A \perp\!\!\!\perp B|C [Q^i]$, $i = 1, 2$, one can derive using (2.1) that $Q_{AB|C}^1 \approx Q_{AB|C}^2$ for measurable rectangles which together with $(Q^1)^C = (Q^2)^C$ implies $Q^1 = Q^2$.

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

$$\forall x \in \mathbf{X}_N \quad \forall \text{ disjoint } A, C \subseteq N \quad f(x_C) = 0 \Rightarrow f(x_{AC}) = 0. \quad (2.14)$$

III. Let us define a function $g : \mathbf{X}_N \rightarrow [0, \infty)$ by

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

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

$$Q(D) = \int_D g(x) \, d\mu(x) \quad \text{for } D \in \mathcal{X}_N = \mathcal{X}_{ABC}.$$

IV. Under the convention $f(x_{AC})/f(x_C) \equiv 0$ in the case $f(x_C) = 0$ one can write for every $E \in \mathcal{X}_{AC}$ using the Fubini theorem, (2.14), and the Radon-Nikodym theorem:

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

Hence, $Q^{AC} = P^{AC}$ and Q is a probability measure. Replace (X_A, \mathcal{X}_A) by (X_B, \mathcal{X}_B) in the preceding consideration to obtain $Q^{BC} = P^{BC}$. The way Q has been defined implies $Q \ll \mu$ and $g = dQ/d\mu$. This form of g implies that Q is factorizable after $\{AC, BC\}$ so that $A \perp\!\!\!\perp B \mid C [Q]$ by Lemma 2.5.

V. To see $P^{ABC} \ll Q$ observe that (2.14) implies $g(x) = 0 \Rightarrow f(x) = 0$ for every $x \in X_N$, accept the convention $f(x)/g(x) \equiv 0$ in the case $g(x) = 0$, and write for every $D \in \mathcal{X}_N$ using the Radon-Nikodym theorem

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

Thus, $P \ll Q$ and $f/g = dP/dQ$.

VI. To derive (2.13) realize that it follows from the definition of g (under the convention above) that

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

Hence, of course

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

According to (A.3) and Lemma A.4 in Section A.6.3, each of the five logarithmic terms above is P -quasi-integrable and the integral is a value in $[0, \infty]$ – use the fact that $\int_{X_N} h(x_D) \, dP(x) = \int_{X_D} h(x_D) \, dP^D(x_D)$ for every $D \subseteq N$. Thus, (2.13) can be derived. \square

Remark 2.11. The measure Q satisfying (2.12) can be interpreted as a *conditional product of P^{AC} and P^{BC}* . Indeed, one can define the conditional

product for every pair of *consonant probability measures* – that is, measures sharing marginals – in this way. However, in general, some obscurities can occur. First, there exists a pair of consonant measures such that no joint measure having them as marginals exists. Second, even if joint measures of this type exist, it may happen that none of them complies with the required conditional independence statement. For both examples see Dawid and Studený [32].

Thus, the assumption of marginal continuity implies the existence of a conditional product. Note that the regularity of conditional probabilities $P_{A|C}$ or $P_{B|C}$ in the sense of Remark A.1 is a more general sufficient condition for the existence of a conditional product (see Proposition 2 in [130]). The value of $H(P^{ABC}|Q)$ in (2.13) is known in information theory as the *conditional mutual information of A and B given C (with respect to P)*. In the case of $C = \emptyset$ just the mutual information $H(P^{AB}|P^A \times P^B)$ is obtained, so that it can be viewed as a generalization of mutual information (but from a different perspective than multiinformation). Conditional mutual information is known as a good measure of stochastic dependence between A and B conditional on knowledge of C ; for an analysis in a discrete case see §3 in Studený and Vejnarová [144]. \triangle

2.3.4 Properties of multiinformation function

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

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

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

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

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

and is *supermodular*, that is, for every $\langle A, B|C \rangle \in \mathcal{T}(N)$

$$m_P(ABC) + m_P(C) - m_P(AC) - m_P(BC) \geq 0. \quad (2.16)$$

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

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

Proof. The relation (2.15) is evident. Given a set $S \subseteq N$, let us substitute $\langle A, B|C \rangle = \langle S, N \setminus S | \emptyset \rangle$ in Lemma 2.6. Equation (2.13) gives

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

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

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

where Q is the conditional product of P^{AC} and P^{BC} . Using Lemma A.4 we derive (2.16). It suffices to see $m_P(S) \leq m_P(T)$ whenever $|T \setminus S| = 1$, which follows directly from (2.16) with $\langle A, B|C \rangle = \langle S, T \setminus S | \emptyset \rangle$ and (2.15). The uniqueness of the conditional product Q mentioned in Lemma 2.6 implies that $A \perp\!\!\!\perp B | C [P]$ iff $P^{ABC} = Q$, that is, $H(P^{ABC} | Q) = 0$ by Lemma A.4. Hence (2.17) follows. \square

The class of *probability measures having finite multiinformation* is, by definition, a subclass of the class of marginally continuous measures. It will be shown in Section 4.1 that it is quite a wide class of measures, involving several classes of measures used in practice. The relation (2.17) provides a very useful equivalent definition of conditional independence for measures with finite multiinformation, namely by means of an algebraic identity. Note that just the relations (2.16) and (2.17) establish a basic method for handling conditional independence used in this monograph. Because these relations originate from information theory – the expression in (2.16) is nothing but the conditional mutual information mentioned in Remark 2.11 – I dare to call them *information-theoretical tools*. For example, all formal properties of conditional independence from Section 2.2.2 and the result mentioned at the beginning of Section 2.2.4 were derived using these tools. Corollary 2.2 also implies that the class of measures with finite multiinformation is closed under the operation of taking marginals. Note without further explanation that it is closed under the operation of conditional product as well.

The following observation appears to be useful later.

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

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

Proof. This is just a rough sketch (for technical details see Section A.6). Suppose without loss of generality $S = N$ and put $\nu = \prod_{i \in N} P^{\{i\}}$. By Lemma 2.3 one knows $P \ll \nu$. Since $P^{\{i\}} \ll \mu_i$ for every $i \in N$ choose versions of

$dP/d\nu$ and $dP^{\{i\}}/d\mu_i$ and observe that $dP/d\nu \cdot \prod_{i \in N} dP^{\{i\}}/d\mu_i$ is a version of $dP/d\mu$, defined uniquely P -a.e. (as $P \ll \nu \ll \mu$). Hence we derive

$$\ln \frac{dP}{d\nu} = \ln \frac{dP}{d\mu} - \sum_{i \in N} \ln \frac{dP^{\{i\}}}{d\mu_i} \quad \text{for } P\text{-a.e. } x \in X_N.$$

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

2.3.5 Positive measures

A marginally continuous measure P over N is *positive* if there exists a dominating measure μ for P whose density $f = dP/d\mu$ is (strictly) positive, that is, $f(x) > 0$ for μ -a.e. $x \in X_N$. Note that the positivity of a density may depend on the choice of a dominating measure. However, whenever a measure μ of this kind exists one has $\mu \ll P$. Since $P \ll \prod_{i \in N} P^{\{i\}}$ and $\prod_{i \in N} P^{\{i\}} \ll \prod_{i \in N} \mu_i \equiv \mu$ one can equivalently introduce a positive measure P over N by a simple requirement that $P \ll \prod_{i \in N} P^{\{i\}} \ll P$ and always take $\prod_{i \in N} P^{\{i\}}$ in place of μ .

A typical example is a *positive discrete measure* P on $X_N = \prod_{i \in N} X_i$ with $1 \leq |X_i| < \infty$, $i \in N$ such that $P(\{x\}) > 0$ for every $x \in X_N$ (or, more generally, only for $x \in \prod_{i \in N} Y_i$ with $Y_i = \{y \in X_i; P^{\{i\}}(\{y\}) > 0\}$). These measures play an important role in (the probabilistic approach to) artificial intelligence. Pearl [100] noticed that conditional independence models induced by these measures further satisfy a special formal property (in addition to the semi-graphoid properties), and introduced the following terminology.

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

$$\begin{aligned} 6. \text{ intersection} \quad & A \perp\!\!\!\perp B \mid DC [\mathcal{M}] \text{ and } A \perp\!\!\!\perp D \mid BC [\mathcal{M}] \\ & \text{implies } A \perp\!\!\!\perp BD \mid C [\mathcal{M}]. \end{aligned}$$

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

Proposition 2.1. Let P be a marginally continuous measure over N and sets $A, B, C, D \subseteq N$ be pairwise disjoint. If P^{BCD} is a positive measure over BCD then

$$A \perp\!\!\!\perp B \mid DC [P] \text{ and } A \perp\!\!\!\perp D \mid BC [P] \Rightarrow A \perp\!\!\!\perp BD \mid C [P].$$

Proof. (see also [70] for an alternative proof under additional restrictive assumption) This is a rough hint only. Let μ be a dominating measure for P such that $f = dP/d\mu$ is a density with $f_{BCD}(x_{BCD}) \equiv f(x_{BCD}) > 0$ for μ -a.e. $x \in X_N$ (I am again following the notational convention from the proof of

Corollary 2.1, p. 24). The assumptions $A \perp\!\!\!\perp B \mid DC \ [P]$ and $A \perp\!\!\!\perp D \mid BC \ [P]$ imply by Lemma 2.4 (one can assume $f(x_E) > 0$ for μ -a.e. $x \in \mathbf{X}_N$ whenever $E \subseteq BCD$)

$$\frac{f(x_{ACD}) \cdot f(x_{BCD})}{f(x_{CD})} = f(x_{ABCD}) = \frac{f(x_{ABC}) \cdot f(x_{BCD})}{f(x_{BC})}$$

for μ -a.e. $x \in \mathbf{X}_N$. The terms $f(x_{BCD})$ can be cancelled, so that one derives by dividing

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

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

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

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

Let us note that there are discrete probability measures whose induced conditional independence model is not a graphoid, that is, it does not satisfy the intersection property (see Example 2.3 on p. 35). On the other hand, Proposition 2.1 holds also under weaker assumptions on P^{BCD} .

2.3.6 Gaussian measures

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

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

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

where Σ^{-1} denotes the inverse of the covariance matrix Σ , called the *concentration matrix*. Its elements are sometimes considered to be alternative parameters of a regular Gaussian measure. Since the density $f_{\mathbf{e}, \Sigma}$ in (2.19) is positive, regular Gaussian measures are positive in the sense of Section 2.3.5.

On the other hand, if Σ is not regular then the respective *singular Gaussian measure* $\mathcal{N}(e, \Sigma)$ (for a detailed definition see Section A.8.3) is concentrated on an affine subspace in $\mathbb{R}^N = \mathbf{X}_N$ having the Lebesgue measure 0. Thus, singular Gaussian measures are not marginally continuous except for some rare cases (when the subspace has the form $\{\mathbf{y}\} \times \mathbf{X}_A$, $A \subset N$ for $\mathbf{y} \in \mathbf{X}_{N \setminus A}$); for illustration, see Example 2.3 on p. 35.

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

$$P_{A|C}(\star | \mathbf{z}) = \mathcal{N}(e_A + \Sigma_{A \cdot C} \cdot \Sigma_{C \cdot C}^- \cdot (\mathbf{z} - e_C), \Sigma_{A \cdot A} - \Sigma_{A \cdot C} \cdot \Sigma_{C \cdot C}^- \cdot \Sigma_{C \cdot A})$$

for every $\mathbf{z} \in \mathbf{X}_C$, where $\Sigma_{A \cdot C}$ denotes the respective submatrix of Σ and $\Sigma_{C \cdot C}^-$ denotes the generalized inverse of $\Sigma_{C \cdot C}$ (see Section A.8.1, p. 237). The point is that, for every $\mathbf{z} \in \mathbf{X}_C$, it is again a Gaussian measure whose covariance matrix $\Sigma_{A|C} = \Sigma_{A \cdot A} - \Sigma_{A \cdot C} \cdot \Sigma_{C \cdot C}^- \cdot \Sigma_{C \cdot A}$ actually does not depend on the choice of \mathbf{z} (see Section A.8.3 for further details on the conditioned Gaussian measure). Therefore, the matrix $\Sigma_{A|C}$ is called a *conditional covariance matrix*. Recall that in the case $C = \emptyset$ one has $\Sigma_{A|C} = \Sigma_{A|\emptyset} = \Sigma_{A \cdot A}$ by a convention. Elements of miscellaneous conditional covariance matrices can serve as convenient parameters of Gaussian measures – e.g. Andersson et al. [9].

An important related fact is that the expectation vector of a Gaussian measure is not significant from the point of view of conditional independence. It is implied by the following lemma that the covariance matrix alone contains all information about conditional independence structure. Therefore it is used in practice almost exclusively.

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

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

Proof. The key idea is that topological assumptions (see Remark A.1) imply the existence of a regular version of conditional probability on \mathbf{X}_{AB} given C , that is, a version $\bar{P}_{AB|C}$ such that the mapping $D \mapsto \bar{P}_{AB|C}(D | \mathbf{z})$ is a probability measure on \mathbf{X}_{AB} for every $\mathbf{z} \in \mathbf{X}_C$. Clearly, for every $A \in \mathcal{X}_A$, the mapping $\mathbf{z} \mapsto \bar{P}_{AB|C}(A \times \mathbf{X}_B | \mathbf{z})$, $\mathbf{z} \in \mathbf{X}_C$, is a version of conditional probability on \mathbf{X}_A given C ; an analogous claim is true for $B \in \mathcal{X}_B$. Thus, (2.1) can be rewritten in the form $\forall A \in \mathcal{X}_A, \forall B \in \mathcal{X}_B$,

$$\bar{P}_{AB|C}(A \times B | \mathbf{z}) = \bar{P}_{AB|C}(A \times \mathbf{X}_B | \mathbf{z}) \cdot \bar{P}_{AB|C}(\mathbf{X}_A \times B | \mathbf{z}) \quad (2.20)$$

for P^C -a.e. $\mathbf{z} \in \mathbf{X}_C$. Since all involved versions of conditional probability are probability measures for every $\mathbf{z} \in \mathbf{X}_C$, it is equivalent to the requirement that (2.20) hold for every $A \in \mathcal{Y}_A, B \in \mathcal{Y}_B$ where \mathcal{Y}_A resp. \mathcal{Y}_B are countable classes

closed under a finite intersection such that $\sigma(\mathcal{Y}_A) = \mathcal{X}_A$ resp. $\sigma(\mathcal{Y}_B) = \mathcal{X}_B$. This can be shown using Lemma A.3 since, given $B \in \mathcal{X}_B$ and $z \in \mathcal{X}_C$, the class of sets $A \in \mathcal{X}_A$ satisfying (2.20) is closed under proper set difference and monotone countable union. The classes \mathcal{Y}_A resp. \mathcal{Y}_B exist in case of Borel σ -algebras on \mathbb{R}^A resp. \mathbb{R}^B . The set of $z \in \mathcal{X}_C$ for which (2.20) holds for every $A \in \mathcal{Y}_A$ and $B \in \mathcal{Y}_B$ has P^C measure 1 (since \mathcal{Y}_A and \mathcal{Y}_B are countable). For these $z \in \mathcal{X}_C$ then (2.20) holds for every $A \in \mathcal{X}_A$ and $B \in \mathcal{X}_B$ by the above mentioned consideration. Hence,

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

However, in this special case one can suppose that $\bar{P}_{AB|C}(\star|z)$ is a Gaussian measure (see Section A.8.3) with the same covariance matrix $\Sigma_{AB|C}$ for every $z \in \mathcal{X}_C$ (while the expectation does depend on z). It is a well-known fact that – regardless of the expectation vector – one has $A \perp\!\!\!\perp B \mid \emptyset$ with respect to a Gaussian measure iff the $A \times B$ -submatrix of its covariance matrix consists of zeros; see (A.9) in Section A.8.3. \square

The previous lemma involves the following well-known criteria for elementary conditional independence statements (see also Proposition 5.2 in [70], Corollaries 6.3.3 and 6.3.4 in [157] and Exercise 3.8 in [100]).

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

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

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

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

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

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

Proof. The first part is an immediate consequence of Lemma 2.8 since we implicitly assume $\sigma_{ii} > 0$ for $i \in N$. For the last fact, first observe by elementary computation that a non-diagonal element of a regular 2×2 -matrix vanishes iff the same element vanishes in its inverse matrix. In particular,

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

The second observation is that $(\Sigma_{D|N \setminus D})^{-1} = (\Sigma^{-1})_{D \cdot D} = \Lambda_{D \cdot D}$ for every non-empty set $D \subseteq N$ (see Section A.8.1). In particular, one has $((\Sigma_{D|N \setminus D})^{-1})_{ab} = (\Lambda_{D \cdot D})_{ab} = \kappa_{ab}$ for $D = \{a, b\}$. \square

Remark 2.12. The proof of Lemma 2.8 reveals a notable difference between the Gaussian and discrete case. While in the discrete case a conditional independence statement $A \perp\!\!\!\perp B \mid C [P]$ is equivalent to the collection of requirements

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

in the Gaussian case it is equivalent to a single requirement

$$A \perp\!\!\!\perp B \mid \emptyset [P_{AB|C}(\star|z)] \quad \text{for at least one } z \in \mathbf{X}_C,$$

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

Thus, one can expect many special formal properties of conditional independence models arising in a Gaussian framework. For example, the following property of a disjoint semi-graphoid \mathcal{M} was recognized by Pearl [100] as a typical property of graphical models (see Chapter 3):

$$\begin{aligned} 7. \text{ composition} \quad & A \perp\!\!\!\perp B \mid C [\mathcal{M}] \text{ and } A \perp\!\!\!\perp D \mid C [\mathcal{M}] \\ & \text{implies } A \perp\!\!\!\perp BD \mid C [\mathcal{M}] \end{aligned}$$

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

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

$$A \perp\!\!\!\perp B \mid C [P] \text{ and } A \perp\!\!\!\perp D \mid C [P] \Rightarrow A \perp\!\!\!\perp BD \mid C [P].$$

Proof. Given a covariance matrix Σ observe that $(\Sigma_{ABD|C})_{AB \cdot AB} = \Sigma_{AB|C}$ and $(\Sigma_{ABD|C})_{AD \cdot AD} = \Sigma_{AD|C}$ (see Section A.8.1 – this holds for a general positive semi-definite matrix Σ since one can fix a pseudoinverse matrix $(\Sigma)_{\bar{C} \cdot \bar{C}}$). The premises of the rule $(\Sigma_{ABD|C})_{A \cdot B} = \mathbf{0}$ and $(\Sigma_{ABD|C})_{A \cdot D} = \mathbf{0}$ imply $(\Sigma_{ABD|C})_{A \cdot BD} = \mathbf{0}$. \square

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

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

$$a \perp\!\!\!\perp b \mid \emptyset [P] \text{ and } a \not\perp\!\!\!\perp b \mid \{c\} [P] \text{ for any distinct } a, b, c \in N.$$

Indeed, put $X_i = \{0, 1\}$ for $i \in N$ and ascribe the probability $1/4$ to all of the following configurations of values: $(0, 0, 0)$, $(0, 1, 1)$, $(1, 0, 1)$ and $(1, 1, 0)$. An example of a positive measure can be obtained by minor modification: one chooses a parameter $0 < \varepsilon < 1/8$, ascribes the probability $1/4 - \varepsilon$ to the above-mentioned configurations and ε to the remaining ones. \diamond

Another special property of Gaussian conditional independence models is the following one which was also mentioned by Pearl [100] in the context of graphical models:

$$\begin{aligned} 8. \text{ weak transitivity } \quad & A \perp\!\!\!\perp B \mid C [\mathcal{M}] \text{ and } A \perp\!\!\!\perp B \mid Cd [\mathcal{M}] \\ & \text{implies } A \perp\!\!\!\perp d \mid C [\mathcal{M}] \text{ or } d \perp\!\!\!\perp B \mid C [\mathcal{M}] \end{aligned}$$

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

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

$$A \perp\!\!\!\perp B \mid C [P] \text{ and } A \perp\!\!\!\perp B \mid Cd [P] \Rightarrow \{ A \perp\!\!\!\perp d \mid C [P] \text{ or } d \perp\!\!\!\perp B \mid C [P] \}.$$

Proof. It suffices to assume that A and B are singletons. Indeed, owing to Corollary 2.4 (and semi-graphoid properties) $A \perp\!\!\!\perp B \mid C$ is equivalent to the condition $\{a \perp\!\!\!\perp b \mid C \text{ for every } a \in A, b \in B\}$ and a similar observation can be made for the other CI statement involved in the premise. There is no pair $a \in A, b \in B$ with $\neg\{a \perp\!\!\!\perp d \mid C\}$ and $\neg\{d \perp\!\!\!\perp b \mid C\}$ because this contradicts the fact $\{a \perp\!\!\!\perp b \mid C \text{ and } a \perp\!\!\!\perp b \mid Cd\}$ implied by the premise. In other terms, either $\{\forall a \in A \ a \perp\!\!\!\perp d \mid C\}$ or $\{\forall b \in B \ d \perp\!\!\!\perp b \mid C\}$ and one can again use Corollary 2.4 to get the desired conclusion.

Lemma 2.8 allows one to reduce the general case to the case $C = \emptyset$. Indeed, one can consider $\Sigma_{N \setminus C \mid C}$ in place of the covariance matrix Σ which is also a positive semi-definite matrix (see Section A.8.1) and therefore it is a covariance matrix of a Gaussian measure over $N \setminus C$ (see Section A.8.3).

If $A = \{a\}$, $B = \{b\}$ and $C = \emptyset$ then two cases can be distinguished. If $\sigma_{ii} > 0$ for $i \in abd$ then apply Corollary 2.3 to the correlation matrix $\Gamma = (\varrho_{ij})_{i,j \in abd}$ of P^{abd} : $0 = \varrho_{ab} = \varrho_{ad} \cdot \varrho_{db}$. Hence $\varrho_{ad} = 0$ or $\varrho_{db} = 0$ which yields the desired fact. If $\sigma_{aa} = 0$ then the fact that the covariance matrix Σ is positive semi-definite implies $\det(\Sigma_{ad \cdot ad}) \geq 0$ (see Section A.8.1) which implies $\sigma_{ad} = 0$ and $a \perp\!\!\!\perp d \mid \emptyset$ by Lemma 2.8. An analogous consideration can be repeated if $\sigma_{bb} = 0$ or $\sigma_{dd} = 0$. \square

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

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

In fact, the above counterexample means that the poset of CI models induced by regular Gaussian measures over N (ordered by inclusion) is not a lattice. Note that in case $|N| = 3$ this poset coincides with the poset of DAG models (see Section 3.2) which is shown in Figure 7.4. However, if $|N| > 3$ then these posets differ – see Exercise 3.8b in [100].

An additional important fact is that every regular Gaussian measure has finite multiinformation. This follows from Lemma 2.7.

Corollary 2.6. Let P be a regular Gaussian measure with a correlation matrix Γ . Then its multiinformation has the value

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

Proof. Take the Lebesgue measure λ on $(\mathbf{X}_N, \mathcal{X}_N)$ in place of μ in Lemma 2.7. Substitution of (A.12) from Section A.8.3 into (2.18) gives

$$\begin{aligned} & -\frac{|N|}{2} \cdot \ln(2\pi) - \frac{|N|}{2} - \frac{1}{2} \cdot \ln(\det(\Sigma)) - \sum_{i \in N} \left\{ \frac{-\ln(2\pi)}{2} - \frac{1}{2} - \frac{1}{2} \cdot \ln(\sigma_{ii}) \right\} \\ & = \frac{1}{2} \sum_{i \in N} \ln \sigma_{ii} - \frac{1}{2} \cdot \ln(\det(\Sigma)) = -\frac{1}{2} \cdot \ln \frac{\det(\Sigma)}{\prod_{i \in N} \sigma_{ii}} = -\frac{1}{2} \cdot \ln(\det(\Gamma)), \end{aligned}$$

which is the fact that was needed to show. \square

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

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

$$a \perp\!\!\!\perp b | \{c\} [P] \quad \text{and} \quad a \not\perp\!\!\!\perp b | \emptyset [P] \quad \text{for any distinct } a, b, c \in N.$$

Put $P = \mathcal{N}(\mathbf{0}, \Sigma)$ where $\Sigma = (\sigma_{ij})_{i,j \in N}$ with $\sigma_{ij} = 1$ for every $i, j \in N$ and apply Corollary 2.3. It is easy to verify (see Section A.8.3) that P is concentrated on the subspace $\{(x, x, x); x \in \mathbb{R}\}$ while $P^{\{i\}} = \mathcal{N}(0, 1)$ for

every $i \in N$. Since $\prod_{i \in N} P^{\{i\}}$ is absolutely continuous with respect to the Lebesgue measure on \mathbb{R}^N , P is not marginally continuous.

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

2.3.7 Basic construction

The following lemma provides a basic method for constructing probability measures with prescribed CI structure.

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

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

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

Let us take the unifying perspective indicated in Remark 2.1: $(\mathbf{Z}_N, \mathcal{Z}_N)$ and R are fixed, and respective coordinate σ -algebras $\bar{\mathcal{X}}_A, \bar{\mathcal{Y}}_A, \bar{\mathcal{Z}}_A \subseteq \mathcal{Z}_N$ are ascribed to every $A \subseteq N$. Then P corresponds to the restriction of R to $\bar{\mathcal{X}}_N$, Q to the restriction of R to $\bar{\mathcal{Y}}_N$ and (2.22) takes the form (see Section A.7 for related concepts):

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

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

$$R(\mathbf{A}^x \cap \mathbf{A}^y \cap \mathbf{B}^x \cap \mathbf{B}^y | \bar{\mathcal{Z}}_C)(z) = R(\mathbf{A}^x \cap \mathbf{A}^y | \bar{\mathcal{Z}}_C)(z) \cdot R(\mathbf{B}^x \cap \mathbf{B}^y | \bar{\mathcal{Z}}_C)(z) \quad (2.24)$$

for R -a.e. $z \in \mathbf{Z}_N$. On the other hand, $\bar{\mathcal{X}}_A \perp\!\!\!\perp \bar{\mathcal{X}}_B | \bar{\mathcal{X}}_C [R]$ is equivalent, by a usual definition of conditional independence in terms of σ -algebras, to the requirement: $\forall \mathbf{A}^x \in \bar{\mathcal{X}}_A, \mathbf{B}^x \in \bar{\mathcal{X}}_B$

$$P(\mathbf{A}^x \cap \mathbf{B}^x | \bar{\mathcal{X}}_C)(x) = P(\mathbf{A}^x | \bar{\mathcal{X}}_C)(x) \cdot P(\mathbf{B}^x | \bar{\mathcal{X}}_C)(x) \quad (2.25)$$

for R -a.e. $z = (x, y) \in Z_N$. I write $P(\star | \bar{\mathcal{X}}_C)(x)$ instead of $R(\star | \bar{\mathcal{X}}_C)(z)$ because it is a function of x which only depends on P . Analogously, the fact $\bar{\mathcal{Y}}_A \perp\!\!\!\perp \bar{\mathcal{Y}}_B | \bar{\mathcal{Y}}_C [R]$ is equivalent to the requirement: $\forall A^y \in \bar{\mathcal{Y}}_A, B^y \in \bar{\mathcal{Y}}_B$

$$Q(A^y \cap B^y | \bar{\mathcal{Y}}_C)(y) = Q(A^y | \bar{\mathcal{Y}}_C)(y) \cdot Q(B^y | \bar{\mathcal{Y}}_C)(y) \quad (2.26)$$

for R -a.e. $z = (x, y) \in Z_N$. Now, given A^x, A^y, B^x, B^y , one can show using Lemma A.5 (see Section A.6.4) that, given a version of conditional probability $P(A^x \cap B^x | \bar{\mathcal{X}}_C)$ and a version of $Q(A^y \cap B^y | \bar{\mathcal{Y}}_C)$, their product is a version of conditional probability $R(A^x \cap A^y \cap B^x \cap B^y | \bar{\mathcal{Z}}_C)$. More specifically, the condition (W) in Lemma A.5 can be used with the class \mathcal{G} consisting of sets $C^x \cap C^y$ where $C^x \in \bar{\mathcal{X}}_C, C^y \in \bar{\mathcal{Y}}_C$, and one uses the assumption $R = P \times Q$ and the Fubini theorem. Hence, the uniqueness of conditional probability implies that

$$R(A^x \cap A^y \cap B^x \cap B^y | \bar{\mathcal{Z}}_C)(z) = P(A^x \cap B^x | \bar{\mathcal{X}}_C)(x) \cdot Q(A^y \cap B^y | \bar{\mathcal{Y}}_C)(y) \quad (2.27)$$

for R -a.e. $z = (x, y) \in Z_N$. Thus, to evidence (2.24) \Rightarrow (2.25) put $A^y = B^y = Z_N$, use (2.27) and the fact $Q(Z_N | \bar{\mathcal{Y}}_C)(y) = 1$ for R -a.e. $z = (x, y) \in Z_N$; to evidence (2.24) \Rightarrow (2.26) put $A^x = B^x = Z_N$. Conversely, (2.25), (2.26) \Rightarrow (2.24) by the repeated use of (2.27), which means that (2.23) was verified.

If both P and Q have finite multiinformation then $R^{\{i\}} = P^{\{i\}} \times Q^{\{i\}}$ are marginals of R on (Z_i, \mathcal{Z}_i) for $i \in N$ and $R \ll \prod_{i \in N} P^{\{i\}} \times \prod_{j \in N} Q^{\{j\}} = \prod_{k \in N} P^{\{k\}} \times Q^{\{k\}}$. Thus, R is a marginally continuous measure over N . Moreover, one can also apply Lemma 2.6 to R with “doubled” $N = N_x \cup N_y$ and $\langle A, B | C \rangle = \langle N_x, N_y | \emptyset \rangle$ to see that

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

Note for explanation that, in the considered case, R is the conditional product of P and Q and therefore the term $H(P^{ABC} | Q)$ in (2.13) vanishes by Lemma A.4 from Section A.6.3. In particular, the multiinformation of R is the sum of the multiinformations P and Q and, therefore, it is finite. The statement concerning discrete and positive discrete measures easily follows from the given construction. \square

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

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

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

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

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

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

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

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

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

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

Note that (2.17) in Corollary 2.2 can be used to verify the above claim. Then Lemma 2.9 can be applied repeatedly to get a discrete probability measure over N satisfying (2.28). \square

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

Corollary 2.7. If $n = |N| \geq 3$ then the number of distinct CI structures induced by discrete probability measures over N exceeds the number $2^{2^{\lfloor n/2 \rfloor}}$ where $\lfloor n/2 \rfloor$ denotes the lower integer part of $n/2$.

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

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

Similarly, in the case $l = (n+1)/2$ write

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

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

2.4 Imsets

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

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

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

Remark 2.13. The word “multiset” is taken from combinatorial theory [1] while the word “imset” is an abbreviation for **integer-valued multiset**. Later in this book certain special imsets will be used to describe probabilistic conditional independence structures (see Section 4.2.3). \triangle

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

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

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

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

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

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

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

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

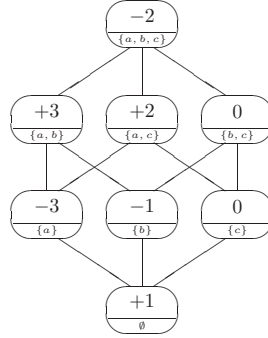


Fig. 2.2. Hasse diagram of an imset over $N = \{a, b, c\}$.

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

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

Alternatively, the second condition in the preceding line can be formulated in the form $\sum_{S \subseteq N \setminus \{j\}} u(S) = 0$ for every $j \in N$. An imset u , respectively a real function u on $\mathcal{P}(N)$, is *ℓ-standardized* if

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

and *u-standardized* if

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

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

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

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

<http://www.springer.com/978-1-85233-891-6>

Probabilistic Conditional Independence Structures

Studeny, M.

2005, XIV, 285 p. 42 illus., Hardcover

ISBN: 978-1-85233-891-6