OXFORD STATISTICAL SCIENCE SERIES • 17

Graphical Models

STEFFEN L. LAURITZEN

OXFORD SCIENCE PUBLICATIONS

Graphical Models

STEFFEN L. LAURITZEN

Department of Mathematics and Computer Science
Aalborg University

CLARENDON PRESS • OXFORD
1996

Preface

In 1976 Terry Speed invited me to Perth, Australia where he conducted a research seminar exploring relations between statistics and statistical physics. Among other things we studied the relation between the notion of interaction as used in contingency table analysis and in thermodynamics. To our delight they were formally the same and one of the most inspiring periods in my life as a researcher was initiated. In the next couple of months we worked day and night and laid essentially the foundations for the papers Darroch *et al.* (1980) and Lauritzen *et al.* (1984).

In 1979 I was invited to lecture at the Swedish summer school in statistics. Here I met Nanny Wermuth and with her as the main source of inspiration we set out to investigate possibilities for making graphical models that simultaneously dealt with discrete and continuous variables.

In 1985 David Spiegelhalter contacted me with the purpose of discussing possibilities for using graphical models in artificial intelligence.

These three meetings have had a profound influence on my life as researcher in general and in particular on the development of the material described in the present book. I am deeply indebted to the inspiration and ideas that Terry, Nanny, David, and others have provided.

Over the years I have enjoyed being part of many research groups. Here I will particularly mention three. The Danes that have been enthusiastically interested in graphical models from their early beginning: Jens Henrik Badsberg, David Edwards, Morten Frydenberg, Svend Kreiner, and a bit later also Poul Svante Eriksen. The BAIES group working on probabilistic expert systems, consisting of Robert Cowell, Phil Dawid, and David Spiegelhalter. The ODIN group in Aalborg involving far too many to mention, but Stig Andersen, Finn Jensen, Frank Jensen, Uffe Kjærulff, and Kristian Olesen have been there all the time. The enthusiasm from members of these groups has kept up my spirits.

Countless colleagues in Denmark and throughout the world have encouraged me every time I was losing hope and energy. It is plainly impossible to mention them all, but it is clear that without these the book would never have existed. It is a privilege to belong to a scientific community with so many fine people.

Heidi Andersen, David Edwards, Jinglong Wang and Dorte Sørensen have read parts of the book in some detail and I am extremely grateful for their comments and criticism. This also applies to the students at Aalborg

University who were exposed to the somewhat compact material in the spring of 1993 and survived.

Various institutions deserve thanks for hosting me while I was writing parts of the book. This includes the Statistical Laboratory in Cambridge, where I began the writing during my sabbatical leave in 1987. Chapter 6 was largely written during a wonderful stay at the institution of San Cataldo, a former nunnery beautifully situated in Southern Italy and run as a study home for Danish researchers and artists.

The Danish Research Councils and the Carlsberg Foundation have in various ways contributed financially to the research.

Oxford University Press and my colleague Frank Jensen have been of invaluable assistance with typographical matters.

Aalborg S.L.L.
December 1995

Contents

1
Introduction

1.1 Graphical models

Graphical models have their origin in several scientific areas. One of these is statistical physics. Here the ideas can be traced back to Gibbs (1902). The object of interest is a large system of particles, possibly the atoms of a gas or solid. Each particle is occupying a site where it can be in different states. The total energy of the system is composed by an external potential plus a potential due to *interaction* of groups of particles. It is usually assumed that only particles at sites close to each other interact. Sites that are close to each other are termed *neighbours* and an undirected graph is determined by the neighbour relationship. The total energy of the system is determining the behaviour of the system through the so-called Gibbsian distribution

$$p(x) = Z^{-1} \exp\{-E(x)\},$$

where $E(x)$ is the total energy of the system when this is in configuration x, and Z is a normalizing constant.

Another origin is in the subject of genetics, where the graphical models go back to Wright (1921, 1923, 1934), who founded the so-called path analysis. Here one is studying heritable properties of natural species and the graph relations are directed, with arrows moving from parent to child. Ideas of path analysis were later taken up in economics and social sciences (Wold 1954; Blalock 1971).

The third source has less obvious relation to graphs. Bartlett (1935) studied the notion of interaction in a three-way contingency table. It turns out that, apart from differences due to notational conventions, this notion of interaction is formally identical to the notion used in statistical physics (Darroch *et al.* 1980).

With this understanding, the graphical models are ready at hand for use in statistics. Their fundamental and universal applicability is due to a number of factors. Firstly, the graphs can visually represent the scientific content of a given model and facilitate communication between researcher and statistician. Secondly, the models are naturally modular so that complex problems can be described and handled by careful combination of

simple elements. Thirdly, the graphs are natural data structures for modern digital computers. Thus models can be efficiently communicated to these and the road is paved for exploiting their computational power.

Over the years the theory and methodology have developed and been extended in a multitude of directions. This is true to such an extent that it has not been possible to cover the area properly in the present book. Luckily a number of other books have been written that take care of some of the omissions in the present. Whittaker (1990*b*) focuses on the ideas behind graphical modelling and has many examples. The same holds for the book by Edwards (1995), just that here the emphasis is on models for mixed data and special attention is paid to problems that are suitable for analysis using the program MIM. Cox and Wermuth (1996) is concentrating attention on the interpretation and application of multivariate systems, in particular graphical models. Neapolitan (1990) deals with probabilistic expert systems based on graphical models, and the collection Oliver and Smith (1990) is primarily concerned with graphical models as influence diagrams, i.e. used as tools in decision analysis.

The present book is primarily concerned with the fundamental mathematical and statistical theory of graphical models. The book is mostly based on a traditional statistical approach, discussing aspects of maximum likelihood methods and significance testing in the different variety of models. It is believed that these results are basic and therefore hopefully will have interest for anyone who wants to enter the world of graphical models, independently of statistical paradigms.

The book is a research monograph and it has been written primarily for researchers and graduate students in mathematical statistics. However, I have tried to keep the mathematics at a reasonable level of abstraction and exactness. No doubt, certain sections are more difficult to read than others, but it is my hope that a relatively wide audience will be able to take advantage of the central parts of the book.

1.2 Outline of book

There are five main chapters. In Chapter 2 the basic graph theory is developed. It may not be advisable to read this chapter in detail before continuing to chapters with a more statistical content, but some familiarity with the notions is necessary to understand even the basic ideas in the subsequent developments. Some of the heavier parts can be omitted at first reading and then returned to when necessary.

Chapter 3 is concerned with conditional independence and Markov properties and these are also essential for later chapters. If desired, the reader can skip the description of directed and chain graph Markov properties at first reading, and then return to these later.

The next three chapters of the book form the core of the statistical theory. Chapter 4 is concerned with discrete variables, Chapter 5 with variables that are normally distributed, and Chapter 6 describes models for systems that contain both discrete and continuous variables. These three chapters are largely written so that they can be read independently of each other. However, as the results of Chapter 6 unify and generalize the results in Chapters 4 and 5, some readers may want to read them in the given order. On the other hand, Chapter 6 is the chapter which contains most original material that has not appeared elsewhere in the literature and others may want to attack this chapter directly.

Chapter 7 is largely a guide to the literature on topics that have not been treated in the book but constitute important and natural extensions.

Finally the book contains four appendices with some material that is used at various places and may not be sufficiently familiar to the reader. Appendix A is a collection of various little topics. Appendix B has some results from linear algebra. Appendix C treats the multivariate normal distribution in some detail, and Appendix D describes basic elements of the theory of exponential families. By having these ready at hand I hope to help readers who do not have all of this present in their minds. However, it is not advisable to learn the material in all four appendices from the given expositions alone.

2
Graphs and hypergraphs

2.1 Graphs

2.1.1 Notation and terminology

A *graph*, as we use it throughout this book, is a pair $\mathcal{G} = (V, E)$, where V is a finite set of *vertices* and the set of *edges* E is a subset of the set $V \times V$ of ordered pairs of distinct vertices. Thus our graphs are *simple*, i.e. there are no multiple edges and they have no loops.

Edges $(\alpha, \beta) \in E$ with both (α, β) and (β, α) in E are called *undirected*, whereas an edge (α, β) with its *opposite* (β, α) not in E is called *directed*.

In a large part of the book we deal with graphs where the vertices are *marked* in the sense that they are partitioned into groups. We then use the term *marked graph*. Throughout the book, vertices in marked graphs are partitioned into two types, such that the vertex set has the structure

$$V = \Delta \cup \Gamma \text{ with } \Delta \cap \Gamma = \emptyset.$$

The vertices in the set Δ represent qualitative variables and those in Γ quantitative variables. Therefore we say that the vertices in Δ are *discrete* and the vertices in Γ are *continuous*. A graph is *pure* if it has only one kind of vertex.

A basic feature of the notion of a graph is that it is a visual object. It is conveniently represented by a picture, where a *dot* is used for a discrete vertex and a circle for a continuous. Further, a *line* joining α to β represents an undirected edge, whereas an *arrow* from α pointing towards β is used for a directed edge (α, β) with $(\beta, \alpha) \notin E$. Figure 2.1 contains an illustration of these conventions. Correspondingly we use the notation $\alpha \rightarrow \beta$, $\alpha \sim \beta$ to signify that

$$(\alpha, \beta) \in E \wedge (\beta, \alpha) \notin E, \quad (\alpha, \beta) \in E \wedge (\beta, \alpha) \in E$$

and $\alpha \not\rightarrow \beta$, $\alpha \not\sim \beta$ for

$$(\alpha, \beta) \notin E, \quad (\alpha, \beta) \notin E \wedge (\beta, \alpha) \notin E.$$

Note that if $\alpha \not\sim \beta$ then there is neither an arrow nor a line between the vertices α and β.

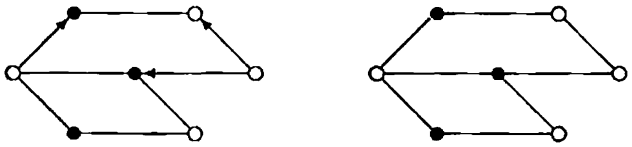


Fig. 2.1. A marked graph and its undirected version. Discrete vertices are represented by dots and continuous vertices by circles. Directed edges are represented by arrows and undirected edges by lines.

If the graph has only undirected edges it is an *undirected* graph and if all edges are directed, the graph is said to be *directed*. In an undirected graph it is often more convenient to represent the edges as unordered pairs $\{\alpha, \beta\}$.

The *undirected version* $\mathcal{G}^{\sim}$ of a graph $\mathcal{G}$ is the undirected graph obtained from $\mathcal{G}$ by substituting lines for arrows. Conversely, suppose that an irreflexive order relation $\prec$ on the vertex set V of a graph $\mathcal{G}$ is given. Then the corresponding *ordered* graph $\mathcal{G}^{\prec}$ has edges between exactly the same vertices α and β as the original graph $\mathcal{G}$, but the edge is directed if $\alpha \prec \beta$, and undirected otherwise. If $\mathcal{G}^{\prec}$ is a directed graph, we say that it is a *directed version* of $\mathcal{G}$.

If $A \subseteq V$ is a subset of the vertex set, it induces a subgraph $\mathcal{G}_A = (A, E_A)$, where the edge set $E_A = E \cap (A \times A)$ is obtained from $\mathcal{G}$ by keeping edges with both endpoints in A.

A graph is *complete* if all vertices are joined by an arrow or a line. A subset is *complete* if it induces a complete subgraph. A complete subset that is maximal (with respect to $\subseteq$) is called a *clique*.

If there is an arrow from α pointing towards β, α is said to be a *parent* of β and β a *child* of α. The set of parents of β is denoted as $\mathrm{pa}(\beta)$ and the set of children of α as $\mathrm{ch}(\alpha)$.

If there is a line between α and β, α and β are said to be *adjacent* or *neighbours*. If there is a neither a line nor an arrow between α and β, i.e. $\alpha \not\sim \beta$, then α and β are said to be *non-adjacent*. The set of neighbours of a vertex α is denoted as $\mathrm{ne}(\alpha)$.

The expressions $\mathrm{pa}(A)$, $\mathrm{ch}(A)$, and $\mathrm{ne}(A)$ denote the collection of parents, children, and neighbours of vertices in A that are not themselves elements of A:

$$\begin{aligned} \mathrm{pa}(A) &= \cup_{\alpha \in A} \mathrm{pa}(\alpha) \setminus A \\ \mathrm{ch}(A) &= \cup_{\alpha \in A} \mathrm{ch}(\alpha) \setminus A \\ \mathrm{ne}(A) &= \cup_{\alpha \in A} \mathrm{ne}(\alpha) \setminus A. \end{aligned}$$

The *boundary* $\mathrm{bd}(A)$ of a subset A of vertices is the set of vertices in $V \setminus A$ that are parents or neighbours to vertices in A. In symbols we then

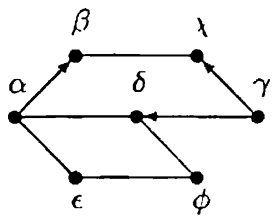


Fig. 2.2. Illustration of graph theoretic concepts. We have $\alpha \to \beta$ and also $\gamma \to \delta$ but $\delta \not\to \gamma, \alpha \not\to \lambda$ whereas, for example, $\epsilon \sim \phi$. Also $\mathrm{pa}(\chi) = \{\gamma\}$ and $\mathrm{ch}(\gamma) = \{\delta, \lambda\}$ as well as $\mathrm{bd}(\delta) = \{\alpha, \phi, \gamma\}$.

have $\mathrm{bd}(A) = \mathrm{pa}(A) \cup \mathrm{ne}(A)$. The *closure* of A is $\mathrm{cl}(A) = A \cup \mathrm{bd}(A)$. See Fig. 2.2 for further illustration.

A *path* of length n from α to β is a sequence $\alpha = \alpha_0, \ldots, \alpha_n = \beta$ of distinct vertices such that $(\alpha_{i-1}, \alpha_i) \in E$ for all $i = 1, \ldots, n$. If there is a path from α to β we say that α *leads to* β and write $\alpha \mapsto \beta$. If both $\alpha \mapsto \beta$ and $\beta \mapsto \alpha$ we say that α and β *connect* and write $\alpha \rightleftharpoons \beta$. Clearly $\rightleftharpoons$ is an equivalence relation and the corresponding equivalence classes $[\alpha]$, where

$$\beta \in [\alpha] \iff \alpha \rightleftharpoons \beta,$$

are the *connectivity components* of $\mathcal{G}$. If $\alpha \in A \subseteq V$, the symbol $[\alpha]_A$ denotes the connectivity component of α in $\mathcal{G}_A$.

A subset $C \subseteq V$ is said to be an (α, β)-*separator* if all paths from α to β intersect C. Thus, in an undirected graph, C is an (α, β)-separator if and only if

$$[\alpha]_{V \setminus C} \neq [\beta]_{V \setminus C}.$$

The subset C is said to *separate* A *from* B if it is an (α, β)-separator for every $\alpha \in A, \beta \in B$.

The vertices α such that $\alpha \mapsto \beta$ and $\beta \not\mapsto \alpha$ are the *ancestors* $\mathrm{an}(\beta)$ of β, and the *descendants* $\mathrm{de}(\alpha)$ of α are the vertices β such that $\alpha \mapsto \beta$ and $\beta \not\mapsto \alpha$. The *non-descendants* are $\mathrm{nd}(\alpha) = V \setminus (\mathrm{de}(\alpha) \cup \{\alpha\})$.

If $\mathrm{bd}(\alpha) \subseteq A$ for all $\alpha \in A$ we say that A is an *ancestral* set. In a directed graph the set A is ancestral if and only if $\mathrm{an}(\alpha) \subseteq A$ for all $\alpha \in A$. In an undirected graph, the ancestral sets are unions of connectivity components. The intersection of a collection of ancestral sets is again ancestral. Hence, for any subset A of vertices there is a smallest ancestral set containing A which is denoted by $\mathrm{An}(A)$.

A *chain* of length n from α to β is a sequence $\alpha = \alpha_0, \ldots, \alpha_n = \beta$ of distinct vertices such that $\alpha_{i-1} \to \alpha_i$ or $\alpha_i \to \alpha_{i-1}$ for all $i = 1, \ldots, n$.

An *n-cycle* is a path of length n with the modification that $\alpha = \beta$, i.e. it begins and ends in the same point. The cycle is said to be *directed* if it contains an arrow.

A *tree* is a connected, undirected graph without cycles. It has a unique path between any two vertices. A *rooted tree* is the directed acyclic graph

obtained from a tree by choosing a vertex as root and directing all edges away from this root. A *forest* is an undirected graph where all connectivity components are trees.

A class of graphs of special interest to us is the class of *chain graphs*. These are the graphs where the vertex set V can be partitioned into numbered subsets, forming a so-called *dependence chain* $V = V(1) \cup \cdots \cup V(T)$ such that all edges between vertices in the same subset are undirected and all edges between different subsets are directed, pointing from the set with lower number to the one with higher number. Such graphs are characterized by *having no directed cycles* and the connectivity components form a partitioning of the graph into *chain components*. A graph is a chain graph if and only if its connectivity components induce undirected subgraphs. The graph in Fig. 2.2 is a chain graph. Its chain components are $\{\alpha, \delta, \epsilon, \phi\}, \{\gamma\}, \{\beta, \chi\}$. The chain components are most easily found by removing all arrows before taking connectivity components. An undirected graph is a special case of a chain graph. A directed, acyclic graph is a chain graph with all chain components consisting of one vertex.

For a chain graph $\mathcal{G}$ we define its *moral graph* $\mathcal{G}^m$ as the undirected graph with the same vertex set but with α and β adjacent in $\mathcal{G}^m$ if and only if either $\alpha \to \beta$ or $\beta \to \alpha$ or if there are γ_1, γ_2 in the same chain component such that $\alpha \to \gamma_1$ and $\beta \to \gamma_2$. In the graph of Fig. 2.2, the moral graph is obtained by adding an edge between α and γ that both have children in the same chain component $\{\beta, \chi\}$, and then ignoring directions. If no edges have to be added to form the moral graph, the chain graph is said to be *perfect*. We warn the reader that the notion of a perfect graph in most graph theory literature refers to something quite different.

In the special case of a directed, acyclic graph the moral graph is obtained from the original graph by 'marrying parents' with a common child and subsequently deleting directions on all arrows.

A chain component C is said to be *terminal* if none of the vertices in C have children. A chain graph has always at least one terminal chain component. A terminal component with only one vertex is a *terminal vertex*.

3

Conditional independence and Markov properties

3.1 Conditional independence

Throughout this text a central notion is that of conditional independence of random variables, the graphs keeping track of the conditional independence relations.

Formally, if X, Y, Z are random variables with a joint distribution P, we say that X *is conditionally independent of* Y *given* Z *under* P, and write $X \perp\!\!\!\perp Y \mid Z\ [P]$, if, for any measurable set A in the sample space of X, there exists a version of the conditional probability $P(A \mid Y, Z)$ which is a function of Z alone. Usually P will be fixed and omitted from the notation. If Z is trivial we say that X *is independent of* Y, and write $X \perp\!\!\!\perp Y$. The notation is due to Dawid (1979) who studied the notion of conditional independence in a systematic fashion. Dawid (1980) gives a formal treatment.

When X, Y, and Z are discrete random variables the condition for $X \perp\!\!\!\perp Y \mid Z$ simplifies as

$$P(X = x, Y = y \mid Z = z) = P(X = x \mid Z = z)P(Y = y \mid Z = z),$$

where the equation holds for all z with $P(Z = z) > 0$. When the three variables admit a joint density with respect to a product measure μ, we have

$$X \perp\!\!\!\perp Y \mid Z \iff f_{XY \mid Z}(x, y \mid z) = f_{X \mid Z}(x \mid z) f_{Y \mid Z}(y \mid z), \tag{3.1}$$

where this equation is to hold almost surely with respect to P. If all densities are continuous, the equality in (3.1) must hold for all z with $f_Z(z) > 0$. Here it is understood that all functions on a discrete space are considered continuous functions. The condition (3.1) can be rewritten as

$$X \perp\!\!\!\perp Y \mid Z \iff f_{XYZ}(x, y, z) f_Z(z) = f_{XZ}(x, z) f_{YZ}(y, z) \tag{3.2}$$

and this equality must hold *for all values of* z when the densities are continuous.

The ternary relation $X \perp\!\!\!\perp Y \mid Z$ has the following properties, where h denotes an arbitrary measurable function on the sample space of X:

(C1) if $X \perp\!\!\!\perp Y \mid Z$ then $Y \perp\!\!\!\perp X \mid Z$;

(C2) if $X \perp\!\!\!\perp Y \mid Z$ and $U = h(X)$, then $U \perp\!\!\!\perp Y \mid Z$;

(C3) if $X \perp\!\!\!\perp Y \mid Z$ and $U = h(X)$, then $X \perp\!\!\!\perp Y \mid (Z, U)$;

(C4) if $X \perp\!\!\!\perp Y \mid Z$ and $X \perp\!\!\!\perp W \mid (Y, Z)$, then $X \perp\!\!\!\perp (W, Y) \mid Z$.

We leave the proof of these facts to the reader. Note that the converse to (C4) follows from (C2) and (C3).

If we use f as generic symbol for the probability density of the random variables corresponding to its arguments, the following statements are true:

$$X \perp\!\!\!\perp Y \mid Z \iff f(x,y,z) = f(x,z)f(y,z)/f(z) \tag{3.3}$$
$$X \perp\!\!\!\perp Y \mid Z \iff f(x \mid y,z) = f(x \mid z) \tag{3.4}$$
$$X \perp\!\!\!\perp Y \mid Z \iff f(x,z \mid y) = f(x \mid z)f(z \mid y) \tag{3.5}$$
$$X \perp\!\!\!\perp Y \mid Z \iff f(x,y,z) = h(x,z)k(y,z) \text{ for some } h,k \tag{3.6}$$
$$X \perp\!\!\!\perp Y \mid Z \iff f(x,y,z) = f(x \mid z)f(y,z). \tag{3.7}$$

The equalities above hold apart from a set of triples (x, y, z) with probability zero. If the densities are continuous functions (in particular if the state spaces are discrete), the equations hold whenever the quantitites involved are well defined, i.e. when the densities of all conditioning variables are positive. We also leave the proof of these equivalences to the reader.

Another property of the conditional independence relation is often used:

(C5) if $X \perp\!\!\!\perp Y \mid Z$ and $X \perp\!\!\!\perp Z \mid Y$ then $X \perp\!\!\!\perp (Y, Z)$.

However (C5) does not hold universally, but only under additional conditions — essentially that there be no non-trivial logical relationship between Y and Z. A trivial counterexample appears when $X = Y = Z$ with $P\{X = 1\} = P\{X = 0\} = 1/2$. We have however

Proposition 3.1 *If the joint density of all variables with respect to a product measure is positive and continuous, then the statement* (C5) *will hold true.*

Proof: Assume that the variables have a continuous density $f(x,y,z) > 0$ and that $X \perp\!\!\!\perp Y \mid Z$ as well as $X \perp\!\!\!\perp Z \mid Y$. Then (3.6) gives for all values of (x, y, z) that

$$f(x,y,z) = k(x,z)l(y,z) = g(x,y)h(y,z)$$

for suitable strictly positive functions g, h, k, l. Thus, as the density is assumed continuous, we have that for all z,

$$g(x, y) = \frac{k(x, z)l(y, z)}{h(y, z)}.$$

Choosing a fixed $z = z_0$ we have $g(x, y) = \pi(x)\rho(y)$ where $\pi(x) = k(x, z_0)$ and $\rho(y) = l(y, z_0)/h(y, z_0)$. Thus $f(x, y, z) = \pi(x)\rho(y)h(y, z)$ and hence $X \perp\!\!\!\perp (Y, Z)$ as desired. □

The proposition can be weakened to more general functions than continuous functions, but we abstain from pursuing this here.

It is illuminating to think of the properties (C1)–(C5) as purely formal expressions, with a meaning that is not necessarily tied to probability. If we interpret the symbols used for random variables as abstract symbols for pieces of knowledge obtained from, say, reading books, and further interpret the symbolic expression $X \perp\!\!\!\perp Y \mid Z$ as:

Knowing Z, reading Y is irrelevant for reading X,

the properties (C1)–(C4) translate to the following:

(I1) if, knowing Z, reading Y is irrelevant for reading X, then so is reading X for reading Y;

(I2) if, knowing Z, reading Y is irrelevant for reading the book X, then reading Y is irrelevant for reading any chapter U of X;

(I3) if, knowing Z, reading Y is irrelevant for reading the book X, it remains irrelevant after having read any chapter U of X;

(I4) if, knowing Z, reading the book Y is irrelevant for reading X and even after having also read Y, reading W is irrelevant for reading X, then reading of both Y and W is irrelevant for reading X.

Thus one can view the relations (C1)–(C4) as pure formal properties of the notion of irrelevance. The property (C5) is slightly more subtle. In a certain sense, also the symmetry (C1) is a somewhat special property of probabilistic conditional independence, rather than general irrelevance.

It is thus tempting to use the relations (C1)–(C4) as formal axioms for conditional independence or irrelevance. A *semi-graphoid* is an algebraic structure which satisfies (C1)–(C4) where X, Y, Z are disjoint subsets of a finite set and $U = h(X)$ is replaced by $U \subseteq X$ (Pearl 1988). If also (C5) holds for disjoint subsets, it is called a *graphoid*. Below we give further examples of such structures.

Example 3.2 A very important example of a model for the irrelevance axioms above is that of *graph separation* in undirected graphs. Let A, B, and C be subsets of the vertex set V of a finite undirected graph $\mathcal{G} = (V, E)$. Define

$$A \overset{\mathcal{G}}{\perp} B \,|\, C \iff C \text{ separates } A \text{ from } B \text{ in } \mathcal{G}.$$

Then it is not difficult to see that graph separation has the following properties:

(S1) if $A \overset{\mathcal{G}}{\perp} B \,|\, C$ then $B \overset{\mathcal{G}}{\perp} A \,|\, C$;

(S2) if $A \overset{\mathcal{G}}{\perp} B \,|\, C$ and U is a subset of A, then $U \overset{\mathcal{G}}{\perp} B \,|\, C$;

(S3) if $A \overset{\mathcal{G}}{\perp} B \,|\, C$ and U is a subset of B, then $A \overset{\mathcal{G}}{\perp} B \,|\, (C \cup U)$;

(S4) if $A \overset{\mathcal{G}}{\perp} B \,|\, C$ and $A \overset{\mathcal{G}}{\perp} D \,|\, (B \cup C)$, then $A \overset{\mathcal{G}}{\perp} (B \cup D) \,|\, C$.

Even the analogue of (C5) holds when all involved subsets are disjoint. Hence graph separation satisfies the graphoid axioms. □

Example 3.3 As another fundamental example, consider *geometric orthogonality* in Euclidean vector spaces. Let L, M, and N be linear subspaces of a Euclidean space V and define

$$L \perp M \,|\, N \iff (L \ominus N) \perp (M \ominus N), \tag{3.8}$$

where $L \ominus N = L \cap N^{\perp}$. If (3.8) is satisfied, then L and M are said to *meet orthogonally in* N. Again, it is not hard to see that the orthogonal meet has the following properties:

(O1) if $L \perp M \,|\, N$ then $M \perp L \,|\, N$;

(O2) if $L \perp M \,|\, N$ and U is a linear subspace of L, then $U \perp M \,|\, N$;

(O3) if $L \perp M \,|\, N$ and U is a linear subspace of M, then $L \perp M \,|\, (N + U)$;

(O4) if $L \perp M \,|\, N$ and $L \perp R \,|\, (M + N)$, then $L \perp (M + R) \,|\, N$.

The analogue of (C5) does not hold in general; for example if $M = N$ we may have

$$L \perp M \mid N \text{ and } L \perp N \mid M,$$

but if L and M are not orthogonal then it is false that $L \perp (M + N)$. □

An abstract theory for conditional independence based on graphoids and semi-graphoids has recently developed (Studený 1989, 1993; Matúš 1992*a*). It was conjectured (Pearl 1988) that the properties (C1)–(C4) were sound and complete axioms for probabilistic conditional independence. However, the completeness fails. In fact, finite axiomatization of probabilistic conditional independence is not possible (Studený 1992). See also Geiger and Pearl (1993) for a systematic study of the logical implications of conditional independence.

3.2 Markov properties

In this section we consider conditional independence in the special situation where we have a collection of random variables $(X_\alpha)_{\alpha\in V}$ taking values in probability spaces $(\mathcal{X}_\alpha)_{\alpha\in V}$. The probability spaces are either real finite-dimensional vector spaces or finite and discrete sets. For A being a subset of V we let $\mathcal{X}_A = \times_{\alpha\in A}\mathcal{X}_\alpha$ and further $\mathcal{X} = \mathcal{X}_V$. Typical elements of $\mathcal{X}_A$ are denoted as $x_A = (x_\alpha)_{\alpha\in A}$. Similarly $X_A = (X_\alpha)_{\alpha\in A}$. We then use the short notation

$$A \perp\!\!\!\perp B \mid C$$

for

$$X_A \perp\!\!\!\perp X_B \mid X_C$$

and so on. The set V is assumed to be the vertex set of a graph $\mathcal{G} = (V, E)$.

3.2.1 Markov properties on undirected graphs

Associated with an undirected graph $\mathcal{G} = (V, E)$ and a collection of random variables $(X_\alpha)_{\alpha\in V}$ as above there is a range of different Markov properties. A probability measure P on $\mathcal{X}$ is said to obey

(P) *the pairwise Markov property*, relative to $\mathcal{G}$, if for any pair (α, β) of non-adjacent vertices

$$\alpha \perp\!\!\!\perp \beta \mid V \setminus \{\alpha, \beta\};$$

(L) *the local Markov property*, relative to $\mathcal{G}$, if for any vertex $\alpha \in V$

$$\alpha \perp\!\!\!\perp V \setminus \mathrm{cl}(\alpha) \mid \mathrm{bd}(\alpha);$$

(G) *the global Markov property*, relative to $\mathcal{G}$, if for any triple (A, B, S) of disjoint subsets of V such that S separates A from B in $\mathcal{G}$

$$A \perp\!\!\!\perp B \mid S.$$

The Markov properties are related as described in the proposition below.

Proposition 3.4 *For any undirected graph $\mathcal{G}$ and any probability distribution on $\mathcal{X}$ it holds that*

$$(\mathrm{G}) \implies (\mathrm{L}) \implies (\mathrm{P}). \tag{3.9}$$

Proof: Firstly, (G) implies (L) because $\mathrm{bd}(\alpha)$ separates α from $V \setminus \mathrm{cl}(\alpha)$. Assume next that (L) holds. We have $\beta \in V \setminus \mathrm{cl}(\alpha)$ because α and β are non-adjacent. Hence

$$\mathrm{bd}(\alpha) \cup ((V \setminus \mathrm{cl}(\alpha)) \setminus \{\beta\}) = V \setminus \{\alpha, \beta\},$$

and it follows from (L) and (C3) that

$$\alpha \perp\!\!\!\perp V \setminus \mathrm{cl}(\alpha) \mid V \setminus \{\alpha, \beta\}.$$

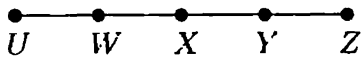
Application of (C2) then gives $\alpha \perp\!\!\!\perp \beta \mid V \setminus \{\alpha, \beta\}$ which is (P). □

It is worth noting that (3.9) only depends on the properties (C1)–(C4) of conditional independence and hence holds for any semi-graphoid. The various Markov properties are different in general, as the following examples show.

Example 3.5 Define the joint distribution of five binary random variables U, W, X, Y, Z as follows: U and Z are independent with

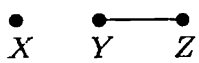
$$P(U = 1) = P(Z = 1) = P(U = 0) = P(Z = 0) = 1/2,$$

$W = U$, $Y = Z$, and $X = WY$. The joint distribution so defined is easily seen to satisfy (L) but not (G) for the graph below.



In fact, Matúš (1992*b*) shows that the global and local Markov properties coincide if and only if the dual graph $\check{\mathcal{G}}$ (defined as $\alpha \sim \beta$ if and only if $\alpha \not\perp \beta$) does not have the 4-cycle as an induced subgraph. □

Example 3.6 A simple example of a probability distribution of (X, Y, Z) that satisfies the pairwise Markov property (P) with respect to the graph



but does not satisfy the local Markov property (L) can be constructed by letting $X = Y = Z$ with $P\{X = 1\} = P\{X = 0\} = 1/2$. It can be shown (Matúš 1992*b*) that the global and pairwise Markov properties coincide if and only if the dual graph $\check{\mathcal{G}}$ does not have a subset of three vertices with two or three edges in its induced subgraph. □

If it holds for all disjoint subsets A, B, C, and D that

$$\text{if } A \perp\!\!\!\perp B \,|\, (C \cup D) \text{ and } A \perp\!\!\!\perp C \,|\, (B \cup D) \text{ then } A \perp\!\!\!\perp (B \cup C) \,|\, D, \quad (3.10)$$

then the Markov properties are all equivalent. This condition is analogous to (C5) and holds, for example, if P has a positive and continuous density with respect to a product measure μ. This is seen as in Proposition 3.1. The result is stated in the theorem below, due to Pearl and Paz (1987): see also Pearl (1988).

Theorem 3.7 (Pearl and Paz) *If a probability distribution on $\mathcal{X}$ is such that (3.10) holds for disjoint subsets A, B, C, D then*

$$(\mathrm{G}) \iff (\mathrm{L}) \iff (\mathrm{P}).$$

Proof: We need to show that (P) implies (G), so assume that S separates A from B in $\mathcal{G}$ and that (P) as well as (3.10) hold. Without loss of generality we can also assume that both A and B are non-empty. The proof is then backward induction on the number of vertices $n = |S|$ in S. If $n = |V| - 2$ then both A and B consist of one vertex and the required conditional independence follows from (P).

So assume $|S| = n < |V| - 2$ and that separation implies conditional independence for all separating sets S with more than n elements. We first assume that $V = A \cup B \cup S$, implying that at least one of A and B has more than one element, A, say. If $\alpha \in A$ then $S \cup \{\alpha\}$ separates $A \setminus \{\alpha\}$ from B and also $S \cup A \setminus \{\alpha\}$ separates α from B. Thus by the induction hypothesis

$$A \setminus \{\alpha\} \perp\!\!\!\perp B \mid S \cup \{\alpha\} \text{ and } \alpha \perp\!\!\!\perp B \mid S \cup A \setminus \{\alpha\}.$$

Now (3.10) gives $A \perp\!\!\!\perp B \mid S$.

If $A \cup B \cup S \subset V$ we choose $\alpha \in V \setminus (A \cup B \cup S)$. Then $S \cup \{\alpha\}$ separates A and B, implying $A \perp\!\!\!\perp B \mid S \cup \{\alpha\}$. Further, either $A \cup S$ separates B from $\{\alpha\}$ or $B \cup S$ separates A from $\{\alpha\}$. Assuming the former gives $\alpha \perp\!\!\!\perp B \mid A \cup S$. Using (3.10) and (C2) we derive that $A \perp\!\!\!\perp B \mid S$. The latter case is similar. □

Note that the proof only exploits (C1)–(C4) and (3.10) and therefore applies to any graphoid.

The global Markov property (G) is important because it gives a general criterion for deciding when two groups of variables A and B are conditionally independent given a third group of variables S.

As conditional independence is intimately related to factorization, so are the Markov properties. A probability measure P on $\mathcal{X}$ is said to *factorize* according to $\mathcal{G}$ if for all complete subsets $a \subseteq V$ there exist non-negative functions ψ_a that depend on x through x_a only, and there exists a product

measure $\mu = \otimes_{a \in V} \mu_a$ on $\mathcal{X}$, such that P has density f with respect to μ where f has the form

$$f(x) = \prod_{a \text{ complete}} \psi_a(x). \tag{3.11}$$

The functions ψ_a are not uniquely determined. There is arbitrariness in the choice of μ, but also groups of functions ψ_a can be multiplied together or split up in different ways. In fact one can without loss of generality assume — although this is not always practical — that only cliques appear as the sets a, i.e. that

$$f(x) = \prod_{c \in \mathcal{C}} \psi_c(x), \tag{3.12}$$

where $\mathcal{C}$ is the set of cliques of $\mathcal{G}$. If P factorizes, we say that P has property (F) and the set of such probability measures is denoted by $M_F(\mathcal{G})$. We have

Proposition 3.8 *For any undirected graph $\mathcal{G}$ and any probability distribution on $\mathcal{X}$ it holds that*

$$\text{(F)} \Longrightarrow \text{(G)} \Longrightarrow \text{(L)} \Longrightarrow \text{(P)}.$$

Proof: We only have to show that (F) implies (G) as the remaining implications are given in (3.9). Let (A, B, S) be any triple of disjoint subsets such that S separates A from B. Let $\tilde{A}$ denote the connectivity components in $\mathcal{G}_{V \setminus S}$ which contain A and let $\tilde{B} = V \setminus (\tilde{A} \cup S)$. Since A and B are separated by S, their elements are in different connectivity components of $\mathcal{G}_{V \setminus S}$ and any clique of $\mathcal{G}$ is either a subset of $\tilde{A} \cup S$ or of $\tilde{B} \cup S$. If $\mathcal{C}_A$ denotes the cliques contained in $\tilde{A} \cup S$, we thus obtain from (3.12) that

$$f(x) = \prod_{c \in \mathcal{C}} \psi_c(x) = \prod_{c \in \mathcal{C}_A} \psi_c(x) \prod_{c \in \mathcal{C} \setminus \mathcal{C}_A} \psi_c(x) = h(x_{\tilde{A} \cup S}) k(x_{\tilde{B} \cup S}).$$

Hence (3.6) gives that $\tilde{A} \perp\!\!\!\perp \tilde{B} \mid S$. Applying (C2) twice gives the desired independence. □

In the case where P has a positive and continuous density we can use the Möbius inversion lemma to show that (P) implies (F), and thus all Markov properties are equivalent. This result seems to have been discovered in various forms by a number of authors (Speed 1979) but is usually attributed to Hammersley and Clifford (1971) who proved the result in the discrete case. The condition that the density be continuous can probably be considerably relaxed (Koster 1994), whereas the positivity is essential. More precisely, we have

Theorem 3.9 (Hammersley and Clifford) *A probability distribution P with positive and continuous density f with respect to a product measure μ satisfies the pairwise Markov property with respect to an undirected graph $\mathcal{G}$ if and only if it factorizes according to $\mathcal{G}$.*

Proof: If P factorizes, it is pairwise Markov as shown in Proposition 3.8, so we just have to show that (P) implies (F).

Since the density is positive, we may take logarithms on both sides of (3.11). Hence this equation can be rewritten as

$$\log f(x) = \sum_{a:a\subseteq V} \phi_a(x), \tag{3.13}$$

where $\phi_a(x) = \log \psi_a(x)$ and $\phi_a \equiv 0$ unless a is a complete subset of V.

Assume then that P is pairwise Markov and choose a fixed but arbitrary element $x^* \in \mathcal{X}$. Define for all $a \subseteq V$

$$H_a(x) = \log f(x_a, x^*_{a^c}),$$

where $(x_a, x^*_{a^c})$ is the element y with $y_\gamma = x_\gamma$ for $\gamma \in a$ and $y_\gamma = x^*_\gamma$ for $\gamma \notin a$. Since x^* is fixed, H_a depends on x through x_a only. Let further for all $a \subseteq V$

$$\phi_a(x) = \sum_{b:b\subseteq a} (-1)^{|a\setminus b|} H_b(x).$$

From this relation it is also clear that ϕ_a depends on x through x_a only. Next we can apply Lemma A.2 (Möbius inversion) to obtain that

$$\log f(x) = H_V(x) = \sum_{a:a\subseteq V} \phi_a(x)$$

such that we have proved the theorem if we can show that $\phi_a \equiv 0$ whenever a is not a complete subset of V. So let us assume that $\alpha, \beta \in a$ and $\alpha \not\sim \beta$. Let further $c = a \setminus \{\alpha, \beta\}$. If we write H_a as short for $H_a(x)$ we have

$$\phi_a(x) = \sum_{b:b\subseteq c} (-1)^{|c\setminus b|} \left\{ H_b - H_{b\cup\{\alpha\}} - H_{b\cup\{\beta\}} + H_{b\cup\{\alpha,\beta\}} \right\}. \tag{3.14}$$

Let $d = V \setminus \{\alpha, \beta\}$. Then, by the pairwise Markov property and (3.7), we have

$$\begin{aligned} H_{b\cup\{\alpha,\beta\}}(x) - H_{b\cup\{\alpha\}}(x) &= \log \frac{f(x_b, x_\alpha, x_\beta, x^*_{d\setminus b})}{f(x_b, x_\alpha, x^*_\beta, x^*_{d\setminus b})} \\ &= \log \frac{f(x_\alpha \mid x_b, x^*_{d\setminus b}) f(x_\beta, x_b, x^*_{d\setminus b})}{f(x_\alpha \mid x_b, x^*_{d\setminus b}) f(x^*_\beta, x_b, x^*_{d\setminus b})} \end{aligned}$$

$$
\begin{aligned}
&= \log \frac{f(x_\alpha^* \mid x_b, x_{d\setminus b}^*) f(x_\beta, x_b, x_{d\setminus b}^*)}{f(x_\alpha^* \mid x_b, x_{d\setminus b}^*) f(x_\beta^*, x_b, x_{d\setminus b}^*)} \\
&= \log \frac{f(x_b, x_\alpha^*, x_\beta, x_{d\setminus b}^*)}{f(x_b, x_\alpha^*, x_\beta^*, x_{d\setminus b}^*)} \\
&= H_{b\cup\{\beta\}}(x) - H_b(x).
\end{aligned}
$$

Thus all terms in the curly brackets in (3.14) add to zero and henceforth the entire sum is zero. This completes the proof. □

The expression inside curly brackets in (3.14) is the logarithm of what is known as the *partial cross-product ratio* so that we can alternatively write

$$\phi_a(x) = \sum_{b:b\subseteq c} (-1)^{|c\setminus b|} \log \mathrm{cpr}(x_\alpha, x_\beta; x_\alpha^*, x_\beta^* \mid x_b, x_{d\setminus b}^*).$$

The pairwise Markov property ensures that all these partial cross-product ratios are equal to 1.

Example 3.10 The following example is due to Moussouris (1974) and shows that the global Markov property (G) may not imply the factorization property (F) without positivity assumptions on the density.

The example is concerned with the distribution of four binary random variables, denoted by (X_1, X_2, X_3, X_4). The following combinations are assumed to have positive probabilities, in fact each of them given a probability equal to 1/8:

$$
\begin{array}{cccc}
(0,0,0,0) & (1,0,0,0) & (1,1,0,0) & (1,1,1,0) \\
(0,0,0,1) & (0,0,1,1) & (0,1,1,1) & (1,1,1,1).
\end{array}
$$

The distribution so specified satisfies the global Markov property (G) with respect to the chordless four-cycle

with the vertices identified cyclically with the random variables. This is seen as follows.

For example, if we consider the conditional distribution of (X_1, X_3), given that $(X_2, X_4) = (0, 1)$, we find

$$P\{X_1 = 0 \mid (X_2, X_4) = (0, 1)\} = 1.$$

Since the conditional distribution of X_1 is degenerate, it is trivially independent of X_3. All other combinations of conditions on (X_2, X_4) give in a

similar way degenerate distributions for one of the remaining variables and this picture is repeated when conditioning on (X_1, X_3). Hence, we have

$$X_1 \perp\!\!\!\perp X_3 \mid (X_2, X_4) \text{ and } X_2 \perp\!\!\!\perp X_4 \mid (X_1, X_3),$$

which shows that the distribution is globally Markov with respect to the graph displayed.

But the density does not factorize. This is seen by an indirect argument. Assume the density factorizes. Then

$$0 \neq 1/8 = f(0,0,0,0) = \phi_{\{1,2\}}(0,0)\phi_{\{2,3\}}(0,0)\phi_{\{3.4\}}(0,0)\phi_{\{4,1\}}(0,0).$$

But also

$$0 = f(0,0,1,0) = \phi_{\{1,2\}}(0,0)\phi_{\{2,3\}}(0,1)\phi_{\{3,4\}}(1,0)\phi_{\{4,1\}}(0,0),$$

whereby

$$\phi_{\{2,3\}}(0,1)\phi_{\{3,4\}}(1,0) = 0.$$

Since now

$$0 \neq 1/8 = f(0,0,1,1) = \phi_{\{1,2\}}(0,0)\phi_{\{2,3\}}(0,1)\phi_{\{3,4\}}(1,1)\phi_{\{4,1\}}(1,0),$$

this implies $\phi_{\{2,3\}}(0,1) \neq 0$ and hence $\phi_{\{3,4\}}(1,0) = 0$, which contradicts that

$$0 \neq 1/8 = f(1,1,1,0) = \phi_{\{1,2\}}(1,1)\phi_{\{2,3\}}(1,1)\phi_{\{3,4\}}(1,0)\phi_{\{4,1\}}(0,1).$$

Hence the density cannot factorize. □

In general none of the Markov properties are preserved under weak limits. This is because weak convergence of joint distributions does not imply convergence of conditional distributions. This fact is illustrated in the following example.

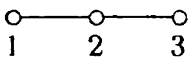
Example 3.11 Let $Y = (Y_1, Y_2, Y_3)^\top$ be a trivariate normal random variable with mean zero and covariance matrix

$$\Sigma = \begin{pmatrix} 1 & \frac{1}{\sqrt{n}} & \frac{1}{2} \\ \frac{1}{\sqrt{n}} & \frac{2}{n} & \frac{1}{\sqrt{n}} \\ \frac{1}{2} & \frac{1}{\sqrt{n}} & 1 \end{pmatrix}.$$

Using Proposition C.5 the conditional distribution of $(Y_1, Y_3)^\top$ given Y_2 is bivariate normal with covariance matrix

$$\Sigma_{13|2} = \begin{pmatrix} 1 & \frac{1}{2} \\ \frac{1}{2} & 1 \end{pmatrix} - \begin{pmatrix} \frac{1}{\sqrt{n}} \\ \frac{1}{\sqrt{n}} \end{pmatrix} \left(\frac{n}{2}\right) \left(\frac{1}{\sqrt{n}}, \frac{1}{\sqrt{n}}\right) = \begin{pmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{2} \end{pmatrix}$$

and hence $Y_1 \perp\!\!\!\perp Y_3 \mid Y_2$, which means that Y satisfies the global Markov property on the graph



As we shall see later in Proposition 5.2, this can alternatively be seen from inspection of the inverse covariance matrix

$$K = \Sigma^{-1} = \begin{pmatrix} 2 & -\sqrt{n} & 0 \\ -\sqrt{n} & \frac{3n}{2} & -\sqrt{n} \\ 0 & -\sqrt{n} & 2 \end{pmatrix}.$$

The distribution of Y' converges weakly to the normal distribution with covariance matrix

$$\bar{\Sigma} = \begin{pmatrix} 1 & 0 & \frac{1}{2} \\ 0 & 0 & 0 \\ \frac{1}{2} & 0 & 1 \end{pmatrix}$$

which is not Markov on the graph considered, since Y_2 is degenerate and

$$\bar{\Sigma}_{13|2} = \bar{\Sigma}_{13} = \begin{pmatrix} 1 & \frac{1}{2} \\ \frac{1}{2} & 1 \end{pmatrix}.$$

Hence the Markov property is not in general stable under weak limits. □

There are, however, important exceptions where the Markov property is preserved under limits and the case of discrete sample space is one.

Proposition 3.12 *Assume that a sequence of probability distributions P_n on a discrete sample space converges to P. If P_n all satisfy any of the Markov properties* (P), (L), *or* (G), *then P satisfies the same Markov property.*

Proof: The density p of the limiting distribution is given as $p = \lim p_n$, where p_n is the density of P_n. Let A, B, C be disjoint and assume that $A \perp\!\!\!\perp B \mid C$ for all P_n. If $p(i_C) > 0$ we get

$$\begin{aligned} p(i_A, i_B, i_C) = \lim_{n\to\infty} p_n(i_A, i_B, i_C) &= \lim_{n\to\infty} \frac{p_n(i_A, i_C)p_n(i_B, i_C)}{p_n(i_C)} \\ &= \frac{p(i_A, i_C)p(i_B, i_C)}{p(i_C)}. \end{aligned}$$

whereby $A \perp\!\!\!\perp B \mid C$ also in the limit. Thus we have shown that all conditional independences are preserved, as desired. □

Observe that the property (F) is not preserved under weak limits, even in the discrete case, as the following example shows.

Example 3.13 Consider four binary random variables (X_1, X_2, X_3, X_4) and let

$$f_n(x_1, x_2, x_3, x_4) = \frac{n^{x_1x_2+x_2x_3+x_3x_4-x_1x_4-x_2-x_3+1}}{8+8n}.$$

Direct calculations show that, as $n \to \infty$, f_n converges to the distribution in Example 3.10. Hence we deduce that (F) is not stable under limits. □

It holds however, in the discrete case, that any probability distribution which satisfies (F) can be obtained as a limit of positive probabilities that factorize. This is true because if p factorizes we get

$$p(i) = \prod_a \psi_a(i) = \lim_{\epsilon \to 0} \frac{\prod_a \{\psi_a(i) + \epsilon\}}{\sum_j \prod_a \{\psi_a(j) + \epsilon\}} = \lim_{\epsilon \to 0} p_\epsilon(i),$$

where p_ϵ clearly is positive and factorizes.

In the discrete case we refer to limits of positive Markov probabilities as *extended Markov* probabilities and denote these by $M_E(\mathcal{G})$. The argument above shows that $M_F(\mathcal{G}) \subseteq M_E(\mathcal{G})$ but the inclusion is strict for a general graph $\mathcal{G}$.

Distributions in $M_E(\mathcal{G})$ are identified through their clique marginals when their state space is finite. This fact is a special case of the following lemma (when $\mathcal{A}$ is taken to be the set of cliques of the graph $\mathcal{G}$).

Lemma 3.14 *Let $\mathcal{A}$ be a set of subsets of a finite set Δ and P and Q be probability mesures on the product space $\mathcal{I} = \times_{\delta \in \Delta} \mathcal{I}_\delta$ where, for each δ, $\mathcal{I}_\delta$ is a finite set. If P and Q have identical marginals to the sets $a \in \mathcal{A}$ and both are limits of measures whose densities with respect to a product measure μ factorize over $\mathcal{A}$, then $P = Q$.*

Proof: It is no restriction to assume that μ is the counting measure on $\mathcal{I}$ such that we have

$$p(i) = \lim_{n \to \infty} \prod_{a \in \mathcal{A}} \phi_a^n(i_a), \quad q(i) = \lim_{n \to \infty} \prod_{a \in \mathcal{A}} \psi_a^n(i_a).$$

Then we get

$$\begin{aligned}
\sum_i p(i) \log q(i) &= \sum_i p(i) \log \left\{ \lim_{n \to \infty} \prod_{a \in \mathcal{A}} \psi_a^n(i_a) \right\} \\
&= \lim_{n \to \infty} \sum_i p(i) \sum_{a \in \mathcal{A}} \log \psi_a^n(i_a) \\
&= \lim_{n \to \infty} \sum_{a \in \mathcal{A}} \sum_i p(i) \log \psi_a^n(i_a) \\
&= \lim_{n \to \infty} \sum_{a \in \mathcal{A}} \sum_i q(i) \log \psi_a^n(i_a)
\end{aligned}$$

$$= \lim_{n\to\infty} \sum_i q(i) \sum_{a\in\mathcal{A}} \log \psi_a^n(i_a)$$
$$= \sum_i q(i) \log q(i),$$

where we have exploited that $\sum_i p(i) \log \psi_a^n(i_a)$ depends on P through its a-marginal only and therefore is equal to $\sum_i q(i) \log \psi_a^n(i_a)$.

Interchanging the role of p and q in the above calculations yields

$$\sum_i q(i) \log p(i) = \sum_i p(i) \log p(i).$$

Combining these identities with the information inequality (A.4) gives

$$\begin{aligned} \sum_i p(i) \log p(i) &= \sum_i q(i) \log p(i) \le \sum_i q(i) \log q(i) \\ &= \sum_i p(i) \log q(i) \le \sum_i p(i) \log p(i). \end{aligned}$$

Since (A.4) is strict unless $p(i) \equiv q(i)$, we must have $P = Q$ and the lemma is proved. □

Lemma 3.14 still holds in the finite case for a general μ and a general system of subsets $\mathcal{A}$. For countable state spaces and general μ it does not hold, i.e. we may have two different probability distributions P and Q that both have factorizing densities (H. G. Kellerer, personal communication). This fact seems to contradict a remark in Csiszár (1975).

If $\mathcal{A}$ is the set of cliques of a decomposable graph, the lemma holds under quite general circumstances (Kellerer 1964*a*, 1964*b*). It is not known whether the conclusion of the lemma holds for factorizing probability distributions in the case of a product measure μ and countable (or more general) state spaces. The critical point in the proof is the interchange of the two sums.

The following example is due to Matúš and Studený (1995) and shows that not all global Markov probabilities can be obtained as limits of factorizing distributions.

Example 3.15 Let $\mathcal{G}$ be the four-cycle as in Example 3.10, but let all four variables have three possible values a, b, and c. Let P be the distribution with each of the following nine states having probability equal to 1/9:

$$\begin{array}{lll} (a,a,a,a) & (b,a,b,c) & (c,a,c,b) \\ (a,b,b,b) & (b,b,c,a) & (c,b,a,c) \\ (a,c,c,c) & (b,c,a,b) & (c,c,b,a) \end{array}$$

This distribution is globally Markov with respect to $\mathcal{G}$ because the conditional distribution of (X_1, X_3) given (X_2, X_4) is always degenerate and this is also true for the conditional distribution of (X_2, X_4) given (X_1, X_3). It cannot be obtained as a limit of factorizing distributions, as the argument below shows.

All pairs (X_i, X_{i+1}) are marginally independent in P (where we have let $X_5 = X_1$). Hence P has the same clique marginals as the distribution Q defined by $X_1, \ldots, X_4$ being mutually independent and uniformly distributed on the space $\{a, b, c\}$. Clearly $Q \in M_E(\mathcal{G})$. By Lemma 3.14 there is only one element in $M_E(\mathcal{G})$ with these marginals so we must have $P \notin M_E(\mathcal{G})$. □

We introduce the notation $M^+(\mathcal{G})$ for the positive Markov probabilities, $M_P(\mathcal{G})$ for the pairwise, $M_L(\mathcal{G})$ for the local, $M_G(\mathcal{G})$ for the global, and recall that we use $M_F(\mathcal{G})$ for those that factorize. We can then summarize the previous considerations as follows. In the general case we have the chain of inclusions

$$M^+(\mathcal{G}) \subset M_F(\mathcal{G}) \subseteq M_G(\mathcal{G}) \subseteq M_L(\mathcal{G}) \subseteq M_P(\mathcal{G}). \tag{3.15}$$

In the discrete case we have further that $M_G(\mathcal{G})$, $M_L(\mathcal{G})$, and $M_P(\mathcal{G})$ are all closed under limits and

$$M_F(\mathcal{G}) \subseteq M_E(\mathcal{G}) \subseteq M_G(\mathcal{G}). \tag{3.16}$$

All inclusions in (3.15) and (3.16) are strict for a general graph $\mathcal{G}$ as shown through various examples.

When (A, B, S) form a weak decomposition of $\mathcal{G}$ the Markov properties decompose accordingly. This is expressed formally in three propositions below.

Proposition 3.16 *Assume that (A, B, S) decompose $\mathcal{G} = (V, E)$ weakly. Then a probability distribution P factorizes with respect to $\mathcal{G}$ if and only if both its marginal distributions $P_{A \cup S}$ and $P_{B \cup S}$ factorize with respect to $\mathcal{G}_{A \cup S}$ and $\mathcal{G}_{B \cup S}$ respectively and the densities f satisfy*

$$f(x) f_S(x_S) = f_{A \cup S}(x_{A \cup S}) f_{B \cup S}(x_{B \cup S}). \tag{3.17}$$

Proof: Suppose that p factorizes with respect to $\mathcal{G}$ such that

$$f(x) = \prod_{c \in C} \psi_c(x).$$

Since (A, B, S) decomposes $\mathcal{G}$, all cliques are either subsets of $A \cup S$ or of $B \cup S$. Let $\mathcal{A}$ denote the cliques that are subsets of $A \cup S$ and $\mathcal{B}$ those that are subsets of $B \cup S$. Then

$$f(x) = \prod_{c \in \mathcal{A}} \psi_c(x) \prod_{c \in \mathcal{B} \setminus \mathcal{A}} \psi_c(x) = h(x_{A \cup S}) k(x_{B \cup S}).$$

By direct integration we find

$$f_{A\cup S}(x_{A\cup S}) = h(x_{A\cup S})\bar{k}(x_S)$$

where

$$\bar{k}(x_S) = \int k(x_{B\cup S})\mu_B(dx_B),$$

and similarly with the other marginals. This gives (3.17) as well as the factorizations of both marginal densities.

Conversely, assume that (3.17) holds and that $f_{A\cup S}$ and $f_{B\cup S}$ factorize. Then let

$$\psi_S(x_S) = \begin{cases} \frac{1}{f_S(x_S)} & \text{if } f_S(x_S) \neq 0 \\ 0 & \text{otherwise.} \end{cases}$$

Since f_S is obtained from integration of f, the latter must also be almost everywhere zero when f_S is. Hence

$$\tilde{f}(x) = f_{A\cup S}(x_{A\cup S})f_{B\cup S}(x_{B\cup S})\psi_S(x_S)$$

is a density for P and P factorizes. □

The analogous result is also true for the global Markov property (G), as we shall now show.

Proposition 3.17 *Assume that (A, B, S) decompose $\mathcal{G} = (V, E)$ weakly. Then a probability distribution P is globally Markov with respect to $\mathcal{G}$ if and only if both marginal distributions $P_{A\cup S}$ and $P_{B\cup S}$ are globally Markov with respect to $\mathcal{G}_{A\cup S}$ and $\mathcal{G}_{B\cup S}$ respectively and the densities satisfy (3.17).*

Proof: Suppose that a probability distribution P satisfies the global Markov property with respect to $\mathcal{G}$. Then (3.17) follows because S separates A from B. To show that $P_{A\cup S}$ is globally Markov with respect to $\mathcal{G}_{A\cup S}$ we show that for any triple $(\tilde{A}, \tilde{B}, \tilde{S})$ of disjoint subsets of $A \cup S$ with $\tilde{S}$ separating $\tilde{A}$ from $\tilde{B}$ in $\mathcal{G}_{A\cup S}$, $\tilde{S}$ separates $\tilde{A}$ from $\tilde{B}$ in $\mathcal{G}$. So let $(\tilde{A}, \tilde{B}, \tilde{S})$ be such a triple and let $\alpha \in \tilde{A}$ and $\beta \in \tilde{B}$. Because S is complete, at least one of them, say α, is in A. If there were a path from α to β avoiding $(A \cup S) \setminus \{\alpha, \beta\}$ it would go via B, contradicting that A is separated from B by S.

Next, assume that $P_{A\cup S}$ and $P_{B\cup S}$ both satisfy the global Markov property and the joint density f factorizes as in (3.17). Let U, W and C be disjoint subsets of V such that C separates U from W and assume first that $V = U \cup W \cup C$.

Since S is complete and C separates U from W, either $U \cap S$ or $W \cap S$ must be empty; we assume the former. The entire set of variables is then partitioned into eight disjoint subsets

$$U \cap A,\ C \cap A,\ W \cap A,\ C \cap S,\ W \cap S,\ U \cap B,\ C \cap B,\ W \cap B$$

and we denote the corresponding subsets of random variables by $X_1, \ldots, X_8$; see the diagram below.

	U	C	W
A	X_1	X_2	X_3
S		X_4	X_5
B	X_6	X_7	X_8

It must hold that $C \cap (A \cup S)$ separates $U \cap A$ from $W \cap (A \cup S)$ in $\mathcal{G}_{A \cup S}$, for otherwise C would not separate U from W as assumed. Hence, from the global Markov property of $P_{A \cup S}$ and (3.6) we must have the factorization

$$f_{A \cup S}(x_{A \cup S}) = h_1(x_1, x_2, x_4) k_1(x_2, x_3, x_4, x_5).$$

By symmetry we further find

$$f_{B \cup S}(x_{B \cup S}) = h_2(x_4, x_6, x_7) k_2(x_4, x_5, x_7, x_8).$$

Combined with (3.17) this gives

$$f(x) f_S(x_S) = h_1(x_1, x_2, x_4) k_1(x_2, x_3, x_4, x_5) h_2(x_4, x_6, x_7) k_2(x_4, x_5, x_7, x_8).$$

Using that $x_S = (x_4, x_5)$ and letting $h = h_1 h_2$ and $k = k_1 k_2 / f_S$ yields

$$f(x) = h(x_1, x_2, x_4, x_6, x_7) k(x_2, x_3, x_4, x_5, x_7, x_8) = h(x_{U \cup C}) k(x_{W \cup C}),$$

which shows that $U \perp\!\!\!\perp W \mid C$.

If $U \cup W \cup C$ is not the entire set of variables V we proceed as in the proof of Proposition 3.8. Details are omitted. □

In the case of discrete sample spaces we have the analogous result for extended Markov probabilities.

Proposition 3.18 *Assume that (A, B, S) decompose $\mathcal{G} = (V, E)$ weakly and the sample space is discrete. Then a probability distribution P is extended Markov with respect to $\mathcal{G}$ if and only if both marginal distributions $P_{A \cup S}$ and $P_{B \cup S}$ are extended Markov with respect to $\mathcal{G}_{A \cup S}$ and $\mathcal{G}_{B \cup S}$ respectively and the densities satisfy*

$$p(i) p(i_S) = p(i_{A \cup S}) p(i_{B \cup S}). \tag{3.18}$$

Proof: Assume that P is extended Markov. Then

$$p(i) = \lim_{n \to \infty} p_n(i),$$

where p_n are positive and factorize. By Proposition 3.16 we therefore have

$$p_n(i) p_n(i_S) = p_n(i_{A \cup S}) p_n(i_{B \cup S}),$$

where the factors on the right-hand side factorize on $\mathcal{G}_{A\cup S}$ and $\mathcal{G}_{B\cup S}$ respectively. Hence their limits are extended Markov. Letting n tend to infinity in the equation yields thus the desired factorization.

Suppose conversely that (3.18) holds with the factors being extended Markov. Then

$$p(i_{A\cup S}) = \lim_{n\to\infty} q_n(i_{A\cup S}), \quad p(i_{B\cup S}) = \lim_{n\to\infty} r_n(i_{B\cup S}), \tag{3.19}$$

where q_n and r_n factorize appropriately. Let now

$$\pi_n(i) = \frac{q_n(i_{A\cup S})}{q_n(i_S)} r_n(i_{B\cup S}). \tag{3.20}$$

Then π_n defines a probability distribution that factorizes. We need to show that

$$p(i) = \lim_{n\to\infty} \pi_n(i).$$

This is trivial if $p(i_S) \neq 0$. If $p(i_S) = 0$ then both $p(i) = 0$ and $p(i_{B\cup S}) = 0$. Hence, by (3.19), the second factor in (3.20) converges to zero. The first factor in (3.20) is bounded above by one and therefore the product converges to zero as desired. □

As we have seen, the various Markov properties are different for general, undirected graphs, unless the densities are all positive. However, in the special case of decomposable graphs some Markov properties coincide.

Proposition 3.19 *Let $\mathcal{G}$ be weakly decomposable. Then*

$$M_F(\mathcal{G}) = M_G(\mathcal{G}).$$

Proof: The proof is by induction on $|\mathcal{C}|$, the number of cliques in the graph $\mathcal{G}$.

The result is trivial for $|\mathcal{C}| = 1$, so let us assume the equation to be true for all graphs with at most n cliques and let $\mathcal{G}$ have $n+1$ cliques.

We only have to show that $M_G(\mathcal{G}) \subseteq M_F(\mathcal{G})$. Assume that P is globally Markov with respect to $\mathcal{G}$ and let (A, B, S) be a proper weak decomposition of $\mathcal{G}$. By Proposition 3.17, both marginals $P_{A\cup S}$ and $P_{B\cup S}$ are globally Markov and (3.17) holds. By the induction assumption the marginal distributions both factorize. The full factorization of P is now obtained from (3.17) by dividing with f_S. □

And a direct consequence of this and (3.16) is the corollary below.

Corollary 3.20 *Assume that $\mathcal{G}$ is weakly decomposable and the state space is discrete. Then*

$$M_F(\mathcal{G}) = M_E(\mathcal{G}) = M_G(\mathcal{G}).$$

It is an easy consequence of Example 3.10 that the converse to Proposition 3.19 holds in the sense that if the graph is not decomposable, it has a chordless cycle and one can by analogy construct a distribution which does not factorize but is globally Markov.

The global Markov property is the strongest of the Markov properties in the sense that the associated list of conditional independence statements strictly contains the statements associated with the other properties. Moreover, it cannot be further strengthened. For example it holds (Frydenberg 1990*b*) that if all state spaces are binary, i.e. $\mathcal{X}_\alpha = \{1, -1\}$, then

$$A \perp\!\!\!\perp B \mid S \text{ for all } P \in M_F(G) \iff S \text{ separates } A \text{ from } B. \tag{3.21}$$

In other words, if A and B are not separated by S then there is a factorizing distribution that makes them conditionally dependent. Geiger and Pearl (1993) conjecture that to any undirected graph $\mathcal{G}$ and fixed state space $\mathcal{X}$ one can find a single $P \in M_F(G)$ such that for this P it holds that

$$A \perp\!\!\!\perp B \mid S \iff S \text{ separates } A \text{ from } B.$$

This is clearly a stronger statement but no proof is known.

3.2.2 Markov properties on directed acyclic graphs

Before we proceed to the case of a general chain graph we consider the same setup as in the previous subsection, only now the graph $\mathcal{G}$ is assumed to be directed and acyclic. The Markov property on a directed acyclic graph was first studied systematically in Kiiveri *et al.* (1984) but see also for example Pearl and Verma (1987), Verma and Pearl (1990*a*), J.Q. Smith (1989), Geiger and Pearl (1990), Lauritzen *et al.* (1990) and other references given below.

We say that a probability distribution P admits a *recursive factorization* according to $\mathcal{G}$, if there exist non-negative functions, henceforth referred to as *kernels*, $k^\alpha(\cdot,\cdot), \alpha \in V$ defined on $\mathcal{X}_\alpha \times \mathcal{X}_{\mathrm{pa}(\alpha)}$, such that

$$\int k^\alpha(y_\alpha, x_{\mathrm{pa}(\alpha)})\mu_\alpha(dy_\alpha) = 1$$

and P has density f with respect to μ, where

$$f(x) = \prod_{\alpha \in V} k^\alpha(x_\alpha, x_{\mathrm{pa}(\alpha)}).$$

We then also say that P has property (DF). It is an easy induction argument to show that if P admits a recursive factorization as above, then the kernels $k^\alpha(\cdot, x_{\mathrm{pa}(\alpha)})$ are densities for the conditional distribution of X_α, given $X_{\mathrm{pa}(\alpha)} = x_{\mathrm{pa}(\alpha)}$. Also it is immediate that if we form the undirected moral graph $\mathcal{G}^m$ (marrying parents and deleting directions) such as described towards the end of Section 2.1.1, we have

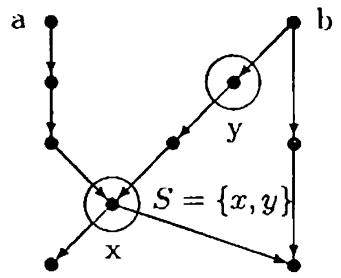


Fig. 3.1. The directed global Markov property. Is $a \perp\!\!\!\perp b \mid S$?

Lemma 3.21 *If P admits a recursive factorization according to the directed, acyclic graph $\mathcal{G}$, it factorizes according to the moral graph $\mathcal{G}^m$ and obeys therefore the global Markov property relative to $\mathcal{G}^m$.*

Proof: The factorization follows from the fact that, by construction, the sets $\{\alpha\} \cup \mathrm{pa}(\alpha)$ are complete in $\mathcal{G}^m$ and we can therefore let $\psi_{\{\alpha\}\cup\mathrm{pa}(\alpha)} = k^\alpha$. The remaining part of the statement follows from the fact that (F) implies (G) in the undirected case; see Proposition 3.8. □

It clearly also holds that

Proposition 3.22 *If P admits a recursive factorization according to the directed, acyclic graph $\mathcal{G}$ and A is an ancestral set, then the marginal distribution P_A admits a recursive factorization according to $\mathcal{G}_A$.*

From this it directly follows that

Corollary 3.23 *Let P factorize recursively according to $\mathcal{G}$. Then*

$$A \perp\!\!\!\perp B \mid S$$

whenever A and B are separated by S in $(\mathcal{G}_{\mathrm{An}(A\cup B\cup S)})^m$, the moral graph of the smallest ancestral set containing $A \cup B \cup S$.

The property in Corollary 3.23 will be referred to as the *directed global Markov property* (DG). The directed global Markov property has the same role as the global Markov property has in the case of an undirected graph, in the sense that it gives the sharpest possible rule for reading conditional independence relations off the directed graph. The procedure is illustrated in the following example.

Example 3.24 Consider a directed Markov field on the graph in Fig. 3.1 and the problem of deciding whether $a \perp\!\!\!\perp b \mid S$. The moral graph of the smallest ancestral set containing all the variables involved is shown in Fig. 3.2. It is immediate that S separates a from b in this graph, implying $a \perp\!\!\!\perp b \mid S$. □

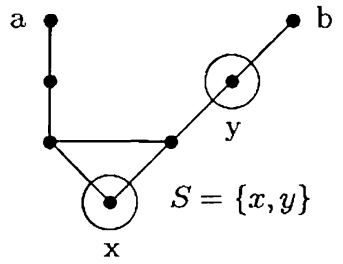


Fig. 3.2. The moral graph of the smallest ancestral set in the graph of Fig. 3.1 containing $\{a\} \cup \{b\} \cup S$. Clearly S separates a from b in this graph, implying $a \perp\!\!\!\perp b \mid S$.

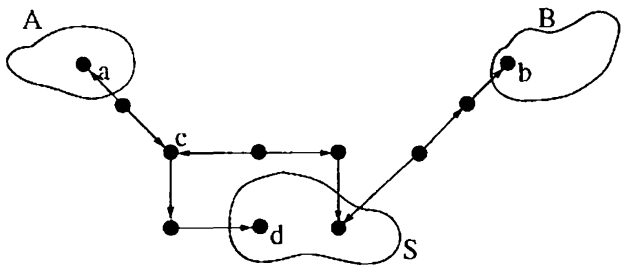


Fig. 3.3. Example of an active chain from A to B. The path from c to d is not part of the chain, but indicates that c must have descendants in S.

An alternative formulation of the directed global Markov property was given by Pearl (1986*a*, 1986*b*) with a full formal treatment in Verma and Pearl (1990*a*, 1990*b*). A chain π from a to b in a directed, acyclic graph $\mathcal{G}$ is said to be *blocked* by S, if it contains a vertex $\gamma \in \pi$ such that either

- $\gamma \in S$ and arrows of π do not meet head-to-head at γ, or
- $\gamma \notin S$ nor has γ any descendants in S, and arrows of π do meet head-to-head at γ.

A chain that is not blocked by S is said to be *active*. Two subsets A and B are now said to be *d-separated* by S if all chains from A to B are blocked by S. We then have

Proposition 3.25 *Let A, B and S be disjoint subsets of a directed, acyclic graph $\mathcal{G}$. Then S d-separates A from B if and only if S separates A from B in $(\mathcal{G}_{\mathrm{An}(A \cup B \cup S)})^m$.*

Proof: Suppose S does not d-separate A from B. Then there is an active chain from A to B such as, for example, indicated in Fig. 3.3. All vertices in this chain must lie within $\mathrm{An}(A \cup B \cup S)$. This follows because if the arrows

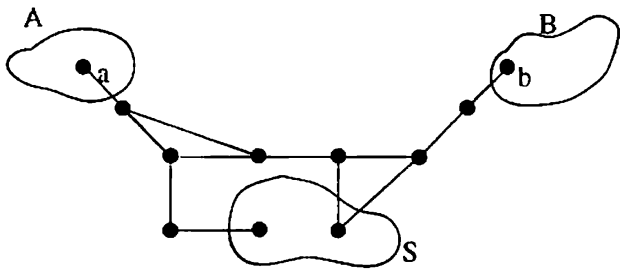


Fig. 3.4. The moral graph corresponding to the active chain in $\mathcal{G}$.

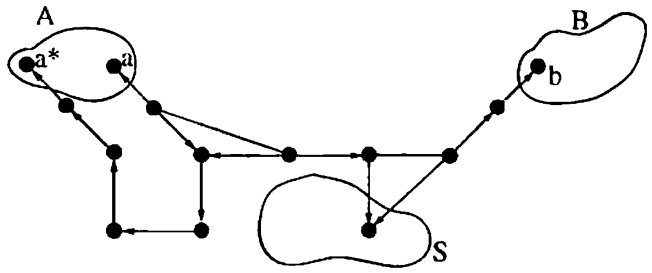


Fig. 3.5. The chain in the graph $(\mathcal{G}_{\mathrm{An}(A\cup B\cup S)})^m$ makes it possible to construct an active chain in $\mathcal{G}$ from A to B.

meet head-to-head at some vertex γ, either $\gamma \in S$ or γ has descendants in S. And if not, either of the subpaths away from γ either meets another arrow, in which case γ has descendants in S, or leads all the way to A or B. Each of these head-to-head meetings will give rise to a marriage in the moral graph such as illustrated in Fig. 3.4, thereby creating a chain from A to B in $(\mathcal{G}_{\mathrm{An}(A\cup B\cup S)})^m$, circumventing S.

Suppose conversely that A is not separated from B in $(\mathcal{G}_{\mathrm{An}(A\cup B\cup S)})^m$. Then there is a chain in this graph that circumvents S. The chain has pieces that correspond to edges in the original graph and pieces that correspond to marriages. Each marriage is a consequence of a meeting of arrows head-to-head at some vertex γ. If γ is in S or it has descendants in S, the meeting does not block the chain. If not, γ must have descendants in A or B, since the ancestral set was smallest. In the latter case, a new chain can be created with one head-to-head meeting fewer, using the line of descent, such as illustrated in Fig. 3.5. Continuing this substitution process eventually leads to an active chain from A to B and the proof is complete. □

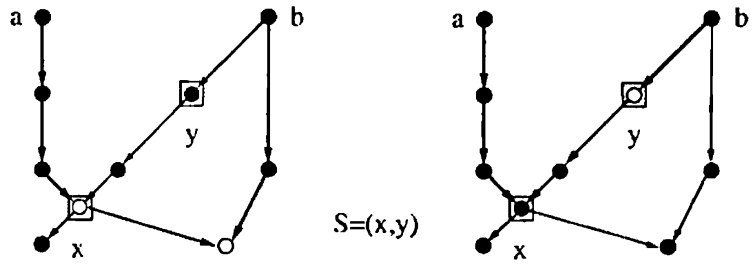


Fig. 3.6. Illustration of Pearl's separation criterion. There are two chains from a to b, drawn with bold lines. Both are blocked, but different vertices γ, indicated with open circles, play the role of blocking vertices.

We illustrate the concept of d-separation by applying it to the query of Example 3.24. As Fig. 3.6 indicates, all chains between a and b are blocked by S, whereby the global Markov property gives that $a \perp\!\!\!\perp b \mid S$.

Geiger and Pearl (1990) show in their Theorem 5 that the criterion of d-separation cannot be improved, in the sense that for any given directed acyclic graph $\mathcal{G}$, one can find state spaces $\mathcal{X}_\alpha, \alpha \in V$ and a probability P such that

$$A \perp\!\!\!\perp B \mid S \iff S \text{ d-separates } A \text{ from } B. \tag{3.22}$$

An argument analogous to that given in Frydenberg (1990*b*) shows that the analogue of (3.21) holds for d-separation in directed, acyclic graphs. Geiger and Pearl (1990) show that in the case of real sample spaces, a Gaussian distribution satisfying (3.22) exists and conjecture the similar result to be true also for the case where the state spaces all have two points, i.e. $\mathcal{X}_\alpha = \{1, -1\}$.

Also the directed case has analogues of the pairwise and local Markov properties. We say that P obeys the *directed pairwise Markov property* (DP) if for any pair (α, β) of non-adjacent vertices with $\beta \in \mathrm{nd}(\alpha)$,

$$\alpha \perp\!\!\!\perp \beta \mid \mathrm{nd}(\alpha) \setminus \{\beta\}.$$

Similarly P obeys the *directed local Markov property* (DL) if any variable is conditionally independent of its non-descendants, given its parents:

$$\alpha \perp\!\!\!\perp \mathrm{nd}(\alpha) \mid \mathrm{pa}(\alpha).$$

As $\mathrm{pa}(\alpha) \subseteq \mathrm{nd}(\alpha)$, it follows from properties (C2) and (C3) of conditional independence that the directed local Markov property (DL) implies the pairwise (DP). The converse is not true in general, as the following example shows.

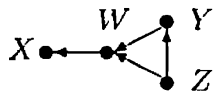


Fig. 3.7. The local and pairwise directed Markov properties are not equivalent in general. This is illustrate by the directed acyclic graph above.

Example 3.26 Let $X = Y = Z$ and W be independent of X with

$$P\{X = 1\} = P\{X = 0\} = P\{W = 1\} = P\{W = 0\} = 1/2.$$

This distribution is pairwise Markov with respect to the graph in Fig. 3.7. However, it is not locally Markov, as this would imply $X \perp\!\!\!\perp (Y, Z) \mid W$, which clearly is not the case. A consequence of Theorem 3.27 is that the density in this example does not admit a recursive factorization. □

In contrast to the undirected case we have that the remaining three properties (DF), (DL) and (DG) are equivalent, just assuming the existence of the density f.

Theorem 3.27 *Let $\mathcal{G}$ be a directed, acyclic graph. For a probability distribution P on $\mathcal{X}$ which has density with respect to a product measure μ, the following conditions are equivalent:*

(DF) *P admits a recursive factorization according to $\mathcal{G}$;*

(DG) *P obeys the directed global Markov property, relative to $\mathcal{G}$;*

(DL) *P obeys the directed local Markov property, relative to $\mathcal{G}$.*

Proof: That (DF) implies (DG) is Corollary 3.23. That (DG) implies (DL) follows by observing that $\{\alpha\} \cup \mathrm{nd}(\alpha)$ is an ancestral set and that $\mathrm{pa}(\alpha)$ obviously separates $\{\alpha\}$ from $\mathrm{nd}(\alpha) \setminus \mathrm{pa}(\alpha)$ in $(\mathcal{G}_{\{\alpha\} \cup \mathrm{nd}(\alpha)})^m$. The final implication is shown by induction on the number of vertices $|V|$ of $\mathcal{G}$. Let α_0 be a terminal vertex of $\mathcal{G}$. Then we can let k^{α_0} be the conditional density of X_{α_0}, given $X_{V \setminus \{\alpha_0\}}$, which by (DL) can be chosen to depend on $x_{\mathrm{pa}(\alpha_0)}$ only. The marginal distribution of $X_{V \setminus \{\alpha_0\}}$ trivially obeys the directed local Markov property and admits a factorization by the inductive assumption. Combining this factorization with k^{α_0} yields the factorization for P. This completes the proof. □

In fact (DL) and (DG) are equivalent, even without assuming the existence of a density with respect to a product measure (Lauritzen *et al.* 1990).

If the probability distribution P satisfies (3.10), for example if the density is positive, then it is not difficult to see that (DP) implies (DL) and all directed Markov properties are equivalent.

Since the three conditions in Theorem 3.27 are all equivalent, we choose to speak of a *directed Markov distribution* as one where any of the conditions (DF), (DL) or (DG) is satisfied. The set of such distributions is denoted by $M(\mathcal{G})$ where $\mathcal{G}$ is a directed and acyclic graph.

It follows as in Proposition 3.12 that in the case of a discrete sample space, the directed Markov properties are preserved under weak limits. So there is no need for a special term or symbol for extended Markov probabilities on a directed acyclic graph.

In the particular case when the directed acyclic graph $\mathcal{G}$ is perfect (see Section 2.1.3) the directed Markov property on $\mathcal{G}$ and the factorization Markov property on its undirected version $\mathcal{G}^\sim$ coincide. This is contained in the following

Proposition 3.28 *Let $\mathcal{G}$ be a perfect directed acyclic graph and $\mathcal{G}^\sim$ its undirected version. Then P admits a recursive factorization with respect to $\mathcal{G}$ if and only if it factorizes according to $\mathcal{G}^\sim$.*

Proof: That the graph is perfect means that $\mathrm{pa}(\alpha)$ is complete for all $\alpha \in V$. Hence $\mathcal{G}^m = \mathcal{G}^\sim$. From Lemma 3.21 it then follows that any $P \in M(\mathcal{G})$ also factorizes with respect to $\mathcal{G}^\sim$.

The reverse inclusion is established by induction on the number of vertices $|V|$ of $\mathcal{G}$. For $|V| = 1$ there is nothing to show. For $|V| = n + 1$ let $P \in M_F(\mathcal{G}^\sim)$ and find a terminal vertex $\alpha \in V$. This vertex has $\mathrm{pa}_\mathcal{G}(\alpha) = \mathrm{bd}_{\mathcal{G}^\sim}(\alpha)$ and, since $\mathcal{G}$ is perfect, this set is complete in both graphs as well. Hence $(V \setminus \{\alpha\}, \{\alpha\}, \mathrm{bd}(\alpha))$ is a weak decomposition of $\mathcal{G}^\sim$ and Proposition 3.16 gives the factorization

$$f(x) = f(x_{V\setminus\{\alpha\}}) f(x_{\mathrm{cl}(\alpha)}) / f(x_{\mathrm{bd}(\alpha)}),$$

where the first factor factorizes according to $\mathcal{G}^\sim_{V\setminus\{\alpha\}}$. Using the induction assumption on this factor gives the full recursive factorization of P. □

In the discrete case Proposition 3.18 implies that the above result can be strengthened to comprise the extended Markov property.

Proposition 3.29 *Let $\mathcal{G}$ be a perfect directed acyclic graph and $\mathcal{G}^\sim$ its undirected version. If all vertices are discrete and hence the total sample space, then*

$$M(\mathcal{G}) = M_F(\mathcal{G}^\sim) = M_E(\mathcal{G}^\sim).$$

In particular, any extended Markov probability factorizes.

Proof: The proof that $M(\mathcal{G}) = M_E(\mathcal{G}^\sim)$ is completely analogous to the proof of the preceding proposition. Since $M(\mathcal{G}) \subseteq M_F(\mathcal{G}^\sim) \subseteq M_E(\mathcal{G}^\sim)$ the result follows. □

3.2.3 Markov properties on chain graphs

In the present section we deal with the Markov properties for a general chain graph $\mathcal{G} = (V, E)$ thereby unifying the directed and undirected cases. A detailed study of the Markov property on chain graphs can be found in Frydenberg (1990*a*). Here we give the most basic results and definitions.

The factorization in the case of a chain graph is more complex and involves two parts. Denote the set of chain components of the graph by $\mathcal{T}$.

Then we first assume a factorization of the density as in the directed acyclic case:

$$f(x) = \prod_{\tau \in \mathcal{T}} f(x_\tau \mid x_{\mathrm{pa}(\tau)}). \tag{3.23}$$

But there is one more assumption. To explain this assumption, let τ_* denote the undirected graph that has $\tau \cup \mathrm{pa}(\tau)$ as nodes and undirected edges between a pair (α, β) if either both of these are in $\mathrm{pa}(\tau)$ or there is an edge, directed or undirected, between them in the chain graph $\mathcal{G}$. Thus for chain components that are singletons, τ_* is complete. This is the case for all chain components if $\mathcal{G}$ is a directed acyclic graph. In an undirected graph, τ_* is just the subgraph induced by the connected component τ.

The second requirement in the chain graph factorization is then that the factors in (3.23) can be further factorized as

$$f(x_\tau \mid x_{\mathrm{pa}(\tau)}) = \prod_a \phi_a(x), \tag{3.24}$$

where a varies over all subsets of $\tau \cup \mathrm{pa}(\tau)$ that are complete in τ_*, and $\phi_a(x)$ as usual represent functions that depend on x through x_a only.

This factorization clearly unifies the directed and undirected cases. A probability distribution P on $\mathcal{X}$ that has a density f with respect to a product measure which satisfies (3.23) and (3.24) is said to *factorize* according to $\mathcal{G}$, and we also say that P has property (FC).

The chain graph factorization has several equivalent formulations. To describe these, let the vertex set be partitioned in a dependence chain as $V = V(1) \cup \cdots \cup V(T)$ such that each of the sets $V(t)$ only has lines between vertices, and arrows point from vertices in sets with lower number to those with higher number. The set of *concurrent* variables relative to this partitioning is next defined to be the set $C(t) = V(1) \cup \cdots \cup V(t)$. Let $\mathcal{G}^*(t)$ be the undirected graph with vertex set $C(t)$ and α adjacent to β in $\mathcal{G}^*(t)$ if either $(\alpha, \beta) \in E$ or $(\beta, \alpha) \in E$ or if $\{\alpha, \beta\} \subseteq C(t-1)$, i.e. $C(t-1)$

is made complete in $\mathcal{G}^*(t)$ by adding all missing edges between these and directions on existing edges are ignored. Let further $B(t) = \mathrm{pa}\{V(t)\}$ and $\mathcal{G}_*(t) = \mathcal{G}^*(t)_{V(t)\cup B(t)}$. Then we have the following equivalent formulations of the chain graph factorization. There are obviously others, for example described through conditioning. We abstain from listing them all.

Proposition 3.30 *For a probability distribution P on $\mathcal{X}$ that admits a density with respect to a product measure, the following conditions are equivalent:*

(i) *P factorizes according to $\mathcal{G}$;*

(ii) *P admits a density that factorizes as*

$$f(x) = \prod_{\tau\in\mathcal{T}} \frac{f\left(x_{\tau\cup\mathrm{pa}(\tau)}\right)}{f\left(x_{\mathrm{pa}(\tau)}\right)} \tag{3.25}$$

and each of the denumerators factorizes on the graph τ_*;

(iii) *for any dependence chain, P admits a density that factorizes as*

$$f(x) = \prod_{t=1}^{T} \frac{f\left(x_{C(t)}\right)}{f\left(x_{C(t-1)}\right)} \tag{3.26}$$

and each of the denumerators factorizes on the graph $\mathcal{G}^*(t)$;

(iv) *for any dependence chain, P admits a density that factorizes as*

$$f(x) = \prod_{t=1}^{T} \frac{f\left(x_{V(t)\cup B(t)}\right)}{f\left(x_{B(t)}\right)} \tag{3.27}$$

and each of the denumerators factorizes on the graph $\mathcal{G}_*(t)$;

(v) *for some dependence chain, P admits a density that factorizes as in either of* (iii) *or* (iv).

Proof: We leave the details of this to the reader. □

Also in the case of chain graphs, there are pairwise, local and global Markov properties, in an even greater variety. We say that a probability P satisfies

(PB) the *pairwise block-recursive Markov property* relative to the dependence chain $V(1),\ldots,V(T)$, if for any pair (α,β) of non-adjacent vertices we have

$$\alpha \perp\!\!\!\perp \beta \mid C(t^*) \setminus \{\alpha,\beta\},$$

where t^* is the smallest t that has $\alpha,\beta \in C(t)$;

(PC) *the pairwise chain Markov property*, relative to $\mathcal{G}$, if for any pair (α, β) of non-adjacent vertices with $\beta \in \mathrm{nd}(\alpha)$

$$\alpha \perp\!\!\!\perp \beta \mid \mathrm{nd}(\alpha) \setminus \{\beta\};$$

(LC) *the local chain Markov property*, relative to $\mathcal{G}$, if for any vertex $\alpha \in V$

$$\alpha \perp\!\!\!\perp \mathrm{nd}(\alpha) \mid \mathrm{bd}(\alpha);$$

(GC) *the global chain Markov property*, relative to $\mathcal{G}$, if for any triple (A, B, S) of disjoint subsets of V such that S separates A from B in $(\mathcal{G}_{\mathrm{An}(A \cup B \cup S)})^m$, the moral graph of the smallest ancestral set containing $A \cup B \cup S$, we have

$$A \perp\!\!\!\perp B \mid S.$$

In general, the property (PB) depends on the particular dependence chain.

Recently Bouckaert and Studený (1995) have given a separation criterion which is analogous to d-separation and equivalent to the global chain Markov property (GC). They also show that any conditional independence statement that is derivable from the local chain Markov property and the properties (C1)–(C5) is represented in the global chain Markov property GC). Further, they show that probability distributions exist that satisfy the local chain Markov property and violate all conditional independence statements that are not of the form given by (GC). In this sense, the global chain Markov property is therefore the strongest possible.

Note that the Markov properties unify the corresponding properties for the directed and undirected cases. For in the undirected case we have that $\mathrm{nd}(\alpha) = V \setminus \{\alpha\}$ and $\mathcal{G} = (\mathcal{G}_{\mathrm{An}(A \cup B \cup S)})^m$. And in the directed case $\mathrm{bd}(\alpha) = \mathrm{pa}(\alpha)$.

Generally all these Markov properties are therefore different without additional assumptions, just as in the undirected case. In the remaining part of the section we assume that all probability measures have positive densities, implying that all five of the basic properties of conditional independence (C1)–(C5) hold. As we shall see, this implies that all Markov properties are equivalent. We first need a few lemmas.

Lemma 3.31 *If $V(1), \ldots, V(T)$ is a dependence chain for $\mathcal{G}$ and and P satisfies the pairwise block-recursive Markov property relative to $\mathcal{G}$, then $P_{C(T-1)}$ satisfies the pairwise block-recursive Markov property relative to $\mathcal{G}^*(T-1)$.*

Proof: This is trivial, as the pairwise block-recursive Markov property is defined to be identical to the combination of the pairwise undirected

Markov properties for the marginal distributions $P_{C(t)}$ of the concurrent variables $C(t)$ on the graphs $\mathcal{G}^*(t)$, for $t = 1, \ldots, T$. □

In the case of the p . vise chain graph Markov property, the corresponding result needs an extra condition.

Lemma 3.32 *If C is a terminal chain component in $\mathcal{G}$ and P satisfies* (3.10) *and the pairwise chain Markov property relative to $\mathcal{G}$, then $P_{V\setminus C}$ satisfies the pairwise chain Markov property relative to $\mathcal{G}_{V\setminus C}$.*

Proof: Let $\alpha, \beta \in V \setminus C$ be non-adjacent in $\mathcal{G}$ and assume $\beta \in \mathrm{nd}(\alpha)$. We have to show that $\alpha \perp\!\!\!\perp \beta \mid \big(\mathrm{nd}(\alpha) \setminus (C \cup \{\beta\})\big)$, since $\mathrm{nd}(\alpha) \setminus C$ is the set of non-descendants of α in $\mathcal{G}_{V\setminus C}$. If $C \cap \mathrm{nd}(\alpha) = \emptyset$ this is a direct consequence of the fact that (PC) holds for $\mathcal{G}$. Else we must have $C \subseteq \mathrm{nd}(\alpha)$. Since no pair δ, γ with $\delta \in \mathrm{de}(\alpha) \cup \{\alpha\}$ and $\gamma \in C$ can be adjacent and $\mathrm{nd}(\gamma) = V \setminus \{\gamma\}$ for all such γ, the pairwise chain Markov property implies that $\delta \perp\!\!\!\perp \gamma \mid V \setminus \{\delta, \gamma\}$ for any such pair. Repeated use of (3.10) yields

$$\mathrm{de}(\alpha) \cup \{\alpha\} \perp\!\!\!\perp C \mid \mathrm{nd}(\alpha) \setminus C$$

and thus by (C2),

$$\alpha \perp\!\!\!\perp C \mid \mathrm{nd}(\alpha) \setminus C. \tag{3.28}$$

The pairwise property gives directly that

$$\alpha \perp\!\!\!\perp \beta \mid \mathrm{nd}(\alpha) \setminus \{\beta\}. \tag{3.29}$$

Thus (3.10) used on (3.28) and (3.29) yields

$$\alpha \perp\!\!\!\perp (C \cup \{\beta\}) \mid \mathrm{nd}(\alpha) \setminus (C \cup \{\beta\})$$

and the result follows from (C2). □

Lemma 3.33 *If P satisfies the pairwise chain Markov property* (PC) *with respect to $\mathcal{G}$ as well as* (3.10)*, it satisfies the pairwise undirected Markov property* (P) *with respect to $\mathcal{G}^m$.*

Proof: We use induction on the number of chain components. If there is only one chain component in $\mathcal{G}$, it is undirected and connected, $\mathcal{G}^m = \mathcal{G}$ and there is nothing to show.

Assume then the lemma to hold for all graphs with n or fewer chain components and let $\mathcal{G}$ have $n+1$ components. Let α, β be non-adjacent in $\mathcal{G}^m$. Without loss of generality we assume that $\beta \in \mathrm{nd}(\alpha)$. If $V \setminus \{\alpha\} = \mathrm{nd}(\alpha)$, the conditional independence $\alpha \perp\!\!\!\perp \beta \mid V \setminus \{\alpha, \beta\}$ follows directly. Else there must be a terminal chain component $C \subseteq \big(V \setminus (\{\alpha\} \cup \mathrm{nd}(\alpha))\big)$.

Because α and β are non-adjacent in $\mathcal{G}^m$ they will also be in $(\mathcal{G}_{V\setminus C})^m$. The induction assumption together with Lemma 3.32 gives that

$$\alpha \perp\!\!\!\perp \beta \mid V \setminus (C \cup \{\alpha, \beta\}). \tag{3.30}$$

Also at least one of α and β, say the former, cannot be in $\mathrm{bd}(C)$, since otherwise they would be adjacent in the moral graph. Thus

$$\alpha \perp\!\!\!\perp \gamma \mid V \setminus \{\alpha, \gamma\} \text{ for all } \gamma \in C. \tag{3.31}$$

Repeated use of property (3.10) of conditional independence on (3.31) yields

$$\alpha \perp\!\!\!\perp C \mid V \setminus (C \cup \{\alpha\}). \tag{3.32}$$

We can now use (C4) on (3.30) and (3.32) to obtain that

$$\alpha \perp\!\!\!\perp \{\beta\} \cup C \mid V \setminus (C \cup \{\alpha, \beta\}).$$

From (C3) and (C2) it follows that $\alpha \perp\!\!\!\perp \beta \mid V \setminus \{\alpha, \beta\}$, which was to be shown. □

Clearly, from arguments analogous to the directed and undirected cases, we have in general that

$$\text{(FC)} \Longrightarrow \text{(GC)} \Longrightarrow \text{(PC)} \Longrightarrow \text{(PB) for any dependence chain,}$$

but if we assume (3.10), all Markov properties are equivalent.

Theorem 3.34 *Assume that P is such that* (3.10) *holds for disjoint subsets of V; then*

$$\text{(GC)} \Longleftrightarrow \text{(LC)} \Longleftrightarrow \text{(PC)} \Longleftrightarrow \text{(PB) for some dependence chain.}$$

Proof: We must show that the pairwise block-recursive property for any dependence chain implies the global chain Markov property (GC). We argue first that (PB) implies (PC) and then that (PC) implies (GC).

The first part is shown by using induction on the number of chain components of $\mathcal{G}$. If there is only one chain component the statement is trivially true. Assume that (PB) implies (PC) for all chain graphs with at most n chain components and let $\mathcal{G}$ have $n+1$ chain components. Assume then that P satisfies (PB) with respect to a given dependence chain and let α and β be non-adjacent vertices.

If $t^* < T$, marginalization to $C(t^*)$ and the inductive assumption will give that

$$\alpha \perp\!\!\!\perp \beta \mid \mathrm{nd}(\alpha) \setminus \{\beta\}.$$

If $t^* = T$ and $V(T)$ has only one chain component, the same conclusion can be drawn because then $\mathrm{nd}(\alpha) = C(T) \setminus \{\alpha\}$.

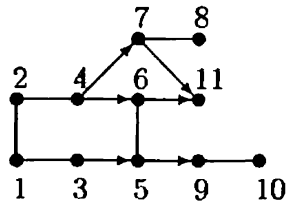


Fig. 3.8. A chain graph. The chain components are {1,2,3,4}, {5,6}, {7,8}, {9,10}, {11}. Is $3 \perp\!\!\!\perp 8 \mid \{2,5\}$? Is $3 \perp\!\!\!\perp 8 \mid 2$?

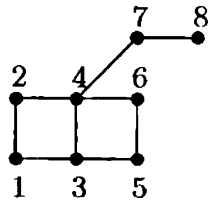


Fig. 3.9. The moral graph of the smallest ancestral set in the graph of Fig. 3.8 containing {2,3,5,8}. A connection between 3 and 4 has been introduced since these both have children in the same chain component $\{5,6\}$. We cannot conclude $3 \perp\!\!\!\perp 8 \mid \{2,5\}$.

If there is more than one chain component in $V(T)$, Theorem 3.7 yields that P is globally Markov with respect to $\mathcal{G}^*(T)$ and hence the same conclusion holds. Thus P satisfies (PC).

Assume next that P satisfies (PC). If R is any ancestral set in $\mathcal{G}$, it can be obtained from V by stepwise removal of terminal chain components. By Lemma 3.32 each removal preserves the pairwise property. Combining with Lemma 3.33 we find that P_R is pairwise Markov with respect to $(\mathcal{G}_R)^m$. Theorem 3.7 yields that P_R is globally Markov relative to the same graph. Thus, the result follows by letting $R = \mathrm{An}(A \cup B \cup S)$. □

Example 3.35 As an illustration of the global chain Markov property, consider the graph in Fig. 3.8 and the question of deciding whether it holds that $3 \perp\!\!\!\perp 8 \mid \{2,5\}$. The smallest ancestral set containing these variables is the set {1,2,3,4,5,6,7,8}. The moral graph of this adds an edge between 3 and 4, because these both have children in the chain component {5,6}. Thus the graph in Fig. 3.9 appears. Since there is a path between 3 and 8 circumventing 2 and 5 in this graph, we cannot conclude that $3 \perp\!\!\!\perp 8 \mid \{2,5\}$.

If we instead consider the question whether $3 \perp\!\!\!\perp 8 \mid 2$, the smallest ancestral set becomes {1,2,3,4,7,8}, no edge has to be added between 3 and 4 and Fig. 3.10 reveals that $3 \perp\!\!\!\perp 8 \mid 2$. □

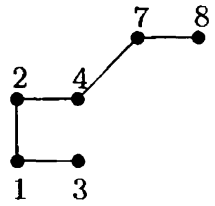


Fig. 3.10. The moral graph of the smallest ancestral set in the graph of Fig. 3.8 containing {2,3,8}. We conclude that $3 \perp\!\!\!\perp 8 \mid 2$.

We cannot expect factorization results to be more general for chain graphs than for undirected graphs, since the chain graphs contain these as special cases. But there is a result analogous to Theorem 3.9.

Theorem 3.36 *A probability distribution with strictly positive and continuous density f satisfies the pairwise block-recursive Markov property with respect to $\mathcal{G}$ if and only if it factorizes according to $\mathcal{G}$.*

Proof: That any chain graph Markov density factorizes as in (3.26) is immediate. That the denumerators also factorize appropriately is seen as follows. Since the density is assumed positive and continuous, (3.10) holds and the different Markov properties are equivalent. From the pairwise block-recursive Markov property it follows that any two variables α and β that are not adjacent in $\mathcal{G}^*(t)$ are conditionally independent given the remaining concurrent variables $C(t)$. But since $V(t) \perp\!\!\!\perp C(t-1) \mid B(t)$ we have that

$$\mathcal{L}\left(X_\alpha, X_\beta \mid X_{C(t)\setminus\{\alpha,\beta\}}\right) = \mathcal{L}\left(X_\alpha, X_\beta \mid X_{(V(t)\cup B(t))\setminus\{\alpha,\beta\}}\right)$$

and hence the marginal distribution of the variables in $V(t) \cup B(t)$ is pairwise Markov with respect to the undirected graph $\mathcal{G}^*(t)$. Hence the densities factorize by Theorem 3.9.

Conversely, assume a factorization given with the stated properties and let α, β be non-adjacent in $\mathcal{G}$. Let t^* be the smallest t with $\alpha, \beta \in C(t)$. Clearly, if the density factorizes, so does the marginal density of $C(t)$ for all t and hence we can assume that $t^* = T$. If α and β are not both in $V(T) \cup B(T)$, the expression (3.27) directly gives that α and β are independent given the remaining variables in $C(t^*) = C(T) = V$. Else if $\alpha, \beta \in V(T) \cup B(T)$ they are not both in $B(T)$ since $t^* = T$. Therefore they are also non-adjacent in $\mathcal{G}^*(T)$. Hence $f(x_{V(T)\cup B(T)})$ is a product of functions, one of which does not depend on β and one of which does not depend on α. Since this is the only factor in (3.27) containing both variables, the same is true for the full joint density. Hence α and β are conditionally independent given the remaining variables, and the pairwise block-recursive Markov property has been established. □

If P is not strictly positive but extended Markov, i.e. P is the limit of positive chain graph Markov distributions, there is a modified version of Theorem 3.36:

Corollary 3.37 *A probability distribution on a discrete sample space satisfies the extended Markov property with respect to a chain graph $\mathcal{G}$ if and only if it factorizes as*

$$p(i) = \prod_{t=1}^{T} \frac{p\left(i_{V(t)\cup B(t)}\right)}{p\left(i_{B(t)}\right)}, \tag{3.33}$$

where $0/0 = 0$ and each of the terms in the denumerator are extended Markov with respect to the graph $\mathcal{G}^*(t)$.

Proof: This follows as in the undirected case from Theorem 3.36 by taking limits. □

Also this factorization can be written in a number of equivalent ways.

3.3 Notes

There are several aspects of graphs and conditional independence that have not been covered here. Cox and Wermuth (1993) discuss alternative ways of encoding conditional independence assumptions through graphs. In Andersson and Perlman (1993), conditional independence models for the multivariate normal distribution are determined through certain distributive lattices of subspaces. The precise relation between such conditional independence restrictions and those given by graphs is discussed in Andersson *et al.* (1995*b*). Basically, conditional independence restrictions given by lattices correspond to those given by directed acyclic graphs that are transitive, i.e. they satisfy that $\alpha \mapsto \beta$ implies $\alpha \rightarrow \beta$.

An important question is related to the notion of *Markov equivalence*. Two graphs are called Markov equivalent if they induce the same conditional independence restrictions. Conditions for two chain graphs to be Markov equivalent were given by Frydenberg (1990*a*) and further results related to this notion are given in Andersson *et al.* (1995*a*, 1996).

Recently, Koster (1996) has studied Markov properties of *reciprocal graphs*. These generalize chain graphs but permit directed cycles and appear therefore to be natural objects for studying systems with feedback, such as encountered in connection with structural equation models (Goldberger and Duncan 1973; Jöreskog 1977). Paz and Geva (1996) study *annotated graphs* that are even more general.

Another recent development is the systematic use of directed acyclic graphs for the interpretation, conjecture and discovery of causal relations

(Spirtes *et al.* 1993; Pearl 1995; Shafer 1996). This development is a natural extension of earlier work by Wright (1921), Wold (1954, 1960), Blalock (1971), and others.