Graphical Models

STEFFEN L. LAURITZEN

Department of Mathematics and Computer Science
Aalborg University

CLARENDON PRESS • OXFORD
1996

Contents

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

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

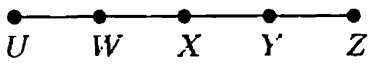
Application of (C2) then gives $\alpha \perp\!\!\!\perp \beta \mid V \setminus \{\alpha, \beta\}$ which is (P). □

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

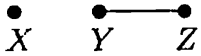
$$P(U = 1) = P(Z = 1) = P(U = 0) = P(Z = 0) = 1/2,$$

$W = U$, $Y = Z$, 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 $\alpha \tilde{\sim} \beta$ if and only if $\alpha \not\sim \beta$) 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 $\mu = \otimes_{\alpha \in V} \mu_\alpha$ 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*

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

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

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

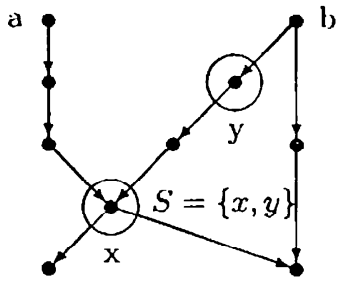


Fig. 3.1. The directed global Markov property. Is $a \perp\!\!\!\perp b \mid S$?

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 $a \perp\!\!\!\perp b \mid S$. □

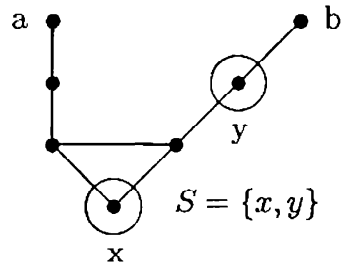


Fig. 3.2. The moral graph of the smallest ancestral set in the graph of Fig. 3.1 containing $\{a\} \cup \{b\} \cup S$. Clearly S separates a from b in this graph, implying $a \perp\!\!\!\perp b \mid S$.

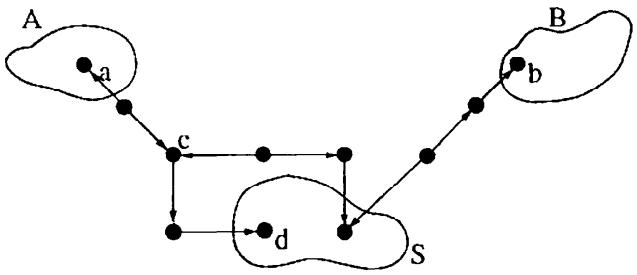


Fig. 3.3. Example of an active chain from A to B. The path from c to d is not part of the chain, but indicates that c must have descendants in S.

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

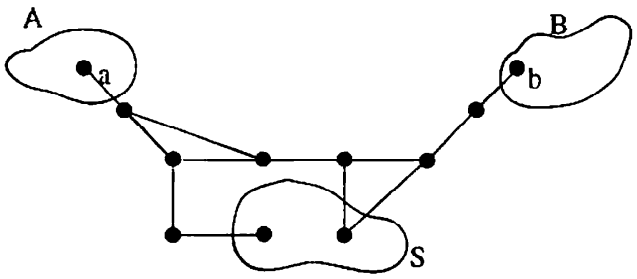


Fig. 3.4. The moral graph corresponding to the active chain in $\mathcal{G}$.

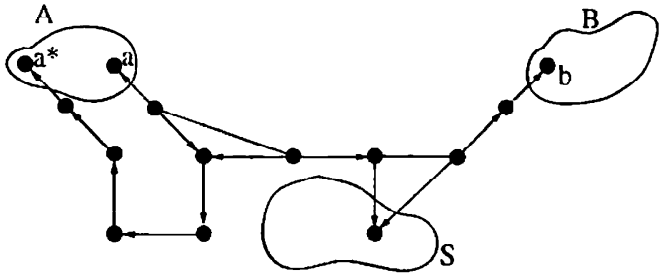


Fig. 3.5. The chain in the graph $(\mathcal{G}_{\mathrm{An}(A\cup B\cup S)})^m$ makes it possible to construct an active chain in $\mathcal{G}$ from A to B.

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. □

P and Q leads to minimum discriminant information estimation (Kullback 1959; Ireland *et al.* 1969). See also Brown (1986) and E.S. Christensen (1989).

A.3 Möbius inversion

An important combinatorial trick is contained in the following

Lemma A.2 (Möbius inversion) *Let Ψ and Φ be functions defined on the set of all subsets of a finite set V, taking values in an Abelian group. Then the following two statements are equivalent:*

(1) *for all* $a \subseteq V : \Psi(a) = \sum_{b:b\subseteq a} \Phi(b)$;

(2) *for all* $a \subseteq V : \Phi(a) = \sum_{b:b\subseteq a} (-1)^{|a\setminus b|}\Psi(b)$.

Proof: We show (2) $\Longrightarrow$ (1):

$$\begin{aligned}\sum_{b:b\subseteq a} \Phi(b) &= \sum_{b:b\subseteq a}\sum_{c:c\subseteq b} (-1)^{|b\setminus c|}\Psi(c) \\ &= \sum_{c:c\subseteq a} \Psi(c)\left\{\sum_{b:c\subseteq b\subseteq a} (-1)^{|b\setminus c|}\right\} \\ &= \sum_{c:c\subseteq a} \Psi(c)\left\{\sum_{h:h\subseteq a\setminus c} (-1)^{|h|}\right\}.\end{aligned}$$

The latter sum is equal to zero unless $a \setminus c = \emptyset$, i.e. if $c = a$, because any finite, non-empty set has the same number of subsets of even as of odd cardinality. The proof of (1) $\Longrightarrow$ (2) is performed analogously. □

The Abelian group referred to in the lemma can be the real numbers, but often also just the additive group of a real vector space V, the vector space of linear maps on V or the vector space of symmetric $k \times k$ matrices etc. More general versions of the lemma exist that relate to general lattices rather than the lattice of subsets of a set; see for example Aigner (1979).

A.4 Iterative partial maximization

For computation of maximum likelihood estimates we shall rely on procedures involving iterative partial maximization in the sense that the likelihood function is maximized over different sections in the parameter space. This is then repeated cyclically.

We consider a continuous real-valued function L on a compact set Θ, and assume that the value $\hat{\theta}$ that maximizes L is uniquely determined.