Steffen L. Lauritzen

Elements of Graphical Models

Lectures from the XXXVIth International Probability Summer School in Saint-Flour, France, 2006

September 4, 2011

Springer

Your dedication goes here

Preface

Here come the golden words

place(s),
month year

First name Surname
First name Surname

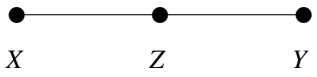
Contents

Chapter 1
Introduction

Conditional independence

The notion of conditional independence is fundamental for graphical models.

For three random variables X, Y and Z we denote this as $X \perp\!\!\!\perp Y \,|\, Z$ and graphically as

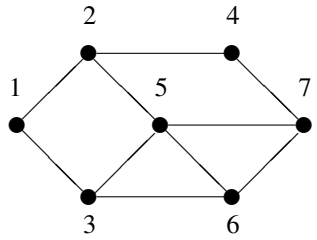


If the random variables have density w.r.t. a product measure μ, the conditional independence is reflected in the relation

$$f(x,y,z)f(z) = f(x,z)f(y,z),$$

where f is a generic symbol for the densities involved.

Graphical models



For several variables, complex systems of conditional independence can be described by undirected graphs.

Then a set of variables A is conditionally independent of set B, given the values of a set of variables C if C separates A from B.

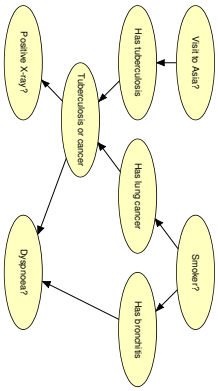
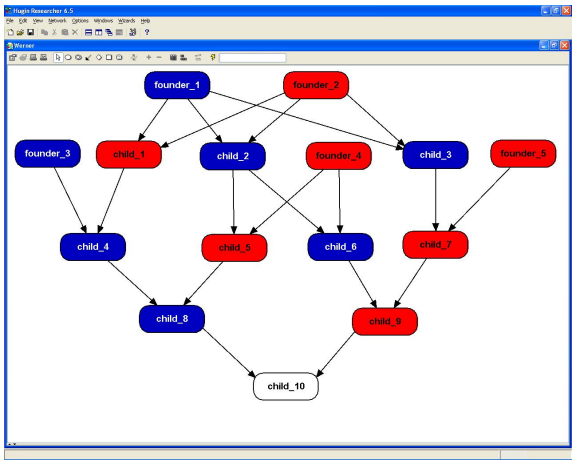


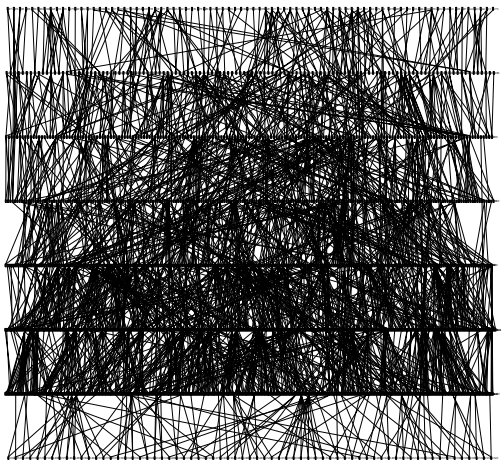
Fig. 1.1 An example of a directed graphical model describing the relationship between risk factors, lung diseases and symptoms. This model was used by Lauritzen and Spiegelhalter (1988) to illustrate important concepts in probabilistic expert systems.

A pedigree



Graphical model for a pedigree from study of Werner's syndrome. Each node is itself a graphical model.

A highly complex pedigree



Family relationship of 1641 members of Greenland Eskimo population.

Chapter 2
Markov Properties

2.1 Conditional Independence

2.1.1 Basic properties

Graphical models are based on the the notion of conditional independence:

Definition 2.1 (Conditional Independence). Random variables X and Y are *conditionally independent* given the random variable Z if the conditional distribution of X given Y and Z is the same as the conditional distribution of X given Z alone, i.e.

$$\mathscr{L}(X \,|\, Y,Z) = \mathscr{L}(X \,|\, Z). \tag{2.1}$$

We then write $X \perp\!\!\!\perp Y \,|\, Z$ or $X \perp\!\!\!\perp_P Y \,|\, Z$ if we want to emphasize that this depends on a specific probability measure P. Alternatively, conditional independence can be formulated in terms of σ-algebras:

Definition 2.2 (Conditional independence of σ-algebras). The σ-algebras $\mathscr{A}$ and $\mathscr{B}$ are *conditionally independent* given the σ-algebra $\mathscr{C}$ if the conditional expectation satisfies

$$\mathbf{E}(1_{A\cap B} \,|\, \mathscr{C}) = \mathbf{E}(1_A \,|\, \mathscr{C})\mathbf{E}(1_B \,|\, \mathscr{C}) \text{ for all } A \in \mathscr{A},\, B \in \mathscr{B}. \tag{2.2}$$

We then write $\mathscr{A} \perp\!\!\!\perp \mathscr{B} \,|\, \mathscr{C}$ and we have $X \perp\!\!\!\perp Y \,|\, Z \iff \sigma(X) \perp\!\!\!\perp \sigma(Y) \,|\, \sigma(Z)$, i.e. if the corresponding σ-algebras are independent.

It is not difficult to show that if the joint distribution of (X,Y,Z) has density with the respect to a product measure, conditional independence is equivalent to the factorizations

$$X \perp\!\!\!\perp Y \,|\, Z \iff f(x,y,z)f(z) = f(x,z)f(y,z) \tag{2.3}$$

$$\iff \exists a,b : f(x,y,z) = a(x,z)b(y,z). \tag{2.4}$$

Similarly, one can show that for random variables X, Y, Z, and W it holds

(C1) if $X \perp\!\!\!\perp Y \,|\, Z$ then $Y \perp\!\!\!\perp X \,|\, Z$;
(C2) if $X \perp\!\!\!\perp Y \,|\, Z$ and $U = g(Y)$, then $X \perp\!\!\!\perp U \,|\, Z$;
(C3) if $X \perp\!\!\!\perp Y \,|\, Z$ and $U = g(Y)$, then $X \perp\!\!\!\perp Y \,|\, (Z,U)$;
(C4) if $X \perp\!\!\!\perp Y \,|\, Z$ and $X \perp\!\!\!\perp W \,|\, (Y,Z)$, then $X \perp\!\!\!\perp (Y,W) \,|\, Z$;

If the joint distribution of the random variables have a density w.r.t. a product measure which is strictly positive, it further holds that

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

Without additional conditions on the joint distribution, (C5) does not hold, but positivity of the density is *not necessary* for (C5). For example, in the case where W is constant it is enough that $f(y,z) > 0$ for all (y,z) or $f(x,z) > 0$ for all (x,z). In the discrete and finite case it is sufficient that the bipartite graphs $\mathscr{G}_+ = (\mathscr{Y} \cup \mathscr{Z}, E_+)$ defined by

$$y \sim_+ z \iff f(y,z) > 0,$$

are all connected, or alternatively if the same condition is satisfied with X replacing Y.

Conditional independence can be seen as encoding irrelevance in a fundamental way. If we give $A \perp\!\!\!\perp B \,|\, C$ the interpretation: *Knowing C, A is irrelevant for learning B,* the properties (C1)–(C4) translate to:

(I1) If, knowing C, learning A is irrelevant for learning B, then B is irrelevant for learning A;
(I2) If, knowing C, learning A is irrelevant for learning B, then A is irrelevant for learning any part D of B;
(I3) If, knowing C, learning A is irrelevant for learning B, it remains irrelevant having learnt any part D of B;
(I4) If, knowing C, learning A is irrelevant for learning B and, having also learnt A, D remains irrelevant for learning B, then both of A and D are irrelevant for learning B.

The property (C5) does not have immediate intuitive appeal for general irrelevance. Also the symmetry (C1) is a special property of probabilistic conditional independence, rather than of general irrelevance, so (I1) does not have the same immediate appeal as the others.

2.1.2 General conditional independence

The general interpretation of conditional independence suggests the usefulness of an abstract study of algebraic structures satisfying these. So consider the set of subsets of a finite set and a ternary relation $\perp_\sigma$ among those.

Definition 2.3 (Graphoid). The relation $\perp_\sigma$ is said to be a *graphoid* if for all disjoint subsets A, B, C, and D of V:

(S1) if $A \perp_\sigma B \,|\, C$ then $B \perp_\sigma A \,|\, C$;
(S2) if $A \perp_\sigma B \,|\, C$ and $D \subseteq B$, then $A \perp_\sigma D \,|\, C$;
(S3) if $A \perp_\sigma B \,|\, C$ and $D \subseteq B$, then $A \perp_\sigma B \,|\, (C \cup D)$;
(S4) if $A \perp_\sigma B \,|\, C$ and $A \perp_\sigma D \,|\, (B \cup C)$, then $A \perp_\sigma (B \cup D) \,|\, C$;
(S5) if $A \perp_\sigma B \,|\, (C \cup D)$ and $A \perp_\sigma C \,|\, (B \cup D)$ then $A \perp_\sigma (B \cup C) \,|\, D$.

The relation is a called *a semigraphoid* if only (S1)–(S4) holds.

The properties (S1)–(S4) are known as the *semigraphoid axioms* and similarly (S1)–(S5) as the *graphoid axioms*. They originate in a slightly different form with Dawid (1979, 1980). It was conjectured by Pearl (1988) that they could be used as complete axioms for probabilistic conditional independence but this has been shown to be false; in fact, there is no finite axiom system which is complete for conditional independence (Studený 1992).

It is possible to consider more general (semi)graphoid relations, defined on other lattices than the lattice of subsets of a set, for example on the lattice of sub-σ-algebras of a σ-algebra. A completely general discussion of conditional independence structures can be based on the notion of *imsets* (Studený 1993).

In the following we shall give several examples of graphoids and semigraphoids and more examples will appear later in the notes.

Let V be a finite set and $X = (X_v, v \in V)$ random variables taking values in $\mathscr{X}_v$. For $A \subseteq V$ we let $X_A = (X_v, v \in A)$ and similarly $x_A = (x_v, v \in A) \in \mathscr{X}_A = \times_{v \in A} \mathscr{X}_v$. If we abbreviate as

$$A \perp\!\!\!\perp B \,|\, S \iff X_A \perp\!\!\!\perp X_B \,|\, X_S,$$

the basic properties of conditional independence imply that *the relation $\perp\!\!\!\perp$ on subsets of V is a semigraphoid* and *if $f(x) > 0$ for all x, the relation $\perp\!\!\!\perp$ is also a graphoid.* This is a *probabilistic semigraphoid*. Not all (semi)graphoids $\perp_\sigma$ are probabilistically representable in the sense that there is a joint distribution so that

$$A \perp_\sigma B \,|\, S \iff X_A \perp\!\!\!\perp X_B \,|\, X_S;$$

see Studený (1993) for further discussion of this point.

Second order conditional independence

Sets of random variables A and B are *partially uncorrelated* for fixed C if their residuals after *linear* regression on X_C are uncorrelated:

$$\mathrm{Cov}\{X_A - \mathbf{E}^*(X_A \,|\, X_C), X_B - \mathbf{E}^*(X_B \,|\, X_C)\} = 0,$$

in other words, if the partial correlations $\rho_{AB \cdot C}$ are equal to zero. If this holds we write $A \perp_2 B \,|\, C$. *The relation $\perp_2$ satisfies the semigraphoid axioms (S1) -(S4)*, and the graphoid axioms if there is no non-trivial linear relation between the variables in V.

Separation in undirected graphs

Let $\mathscr{G} = (V,E)$ be a finite and simple undirected graph (no self-loops, no multiple edges). For subsets A,B,S of V, let $A \perp_{\mathscr{G}} B \,|\, S$ denote that S separates A from B in $\mathscr{G}$, i.e. that all paths from A to B intersect S. It then holds that *the relation $\perp_{\mathscr{G}}$ on subsets of V is a graphoid.* Indeed, this is the reason for choosing this name for such separation relations.

Geometric orthogonality

As another fundamental example, consider geometric orthogonality in Euclidean vector spaces or Hilbert spaces. Let L, M, and N be linear subspaces of a Hilbert space H and define

$$L \perp M \,|\, N \iff (L \ominus N) \perp (M \ominus N),$$

where $L \ominus N = L \cap N^{\perp}$. If this condition is satisfied, L and M are said to *meet orthogonally in N*. This relation has 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 (S5) 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 in general it is false that $L \perp (M+N)$. Thus $\perp$ *is a semigraphoid relation* on the lattice of closed subspaces of a Hilbert space.

Variation independence

Let $\mathscr{U} \subseteq \mathscr{X} = \times_{v \in V} \mathscr{X}_v$ and define for $S \subseteq V$ and $u_S^* \in \mathscr{X}_S \cap \mathscr{U}$ the *S-section* $\mathscr{U}^{u_S^*}$ of $\mathscr{U}$ as

$$\mathscr{U}^{u_S^*} = \{u_{V \setminus S} : u_S = u_S^*, u \in \mathscr{U}\}.$$

Define further the conditional independence relation $\ddagger_{\mathscr{U}}$ as

$$A \ddagger_{\mathscr{U}} B \,|\, S \iff \forall u_S^* : \mathscr{U}^{u_S^*} = \{\mathscr{U}^{u_S^*}\}_A \times \{\mathscr{U}^{u_S^*}\}_B$$

i.e. if and only if the S-sections all have the form of a product space. *The relation $\ddagger_{\mathscr{U}}$ satisfies the semigraphoid axioms.* Note in particular that $A \ddagger_{\mathscr{U}} B \,|\, S$ holds if $\mathscr{U}$ is the support of a probability measure satisfying $A \perp\!\!\!\perp B \,|\, S$.

2.2 Markov Properties for Undirected Graphs

Graphs can be used to generate conditional independence structures in the form of Markov properties, typically described through the separation properties of the graph. Here we consider a simple undirected graph $\mathscr{G} = (V,E)$ and a conditional independence relation $\perp_\sigma$ on the subsets of V which we assume satisfies the semigraphoid axioms. The Markov properties associated with an undirected graph $\mathscr{G}$ are known as pairwise, local and global, to be detailed in the following.

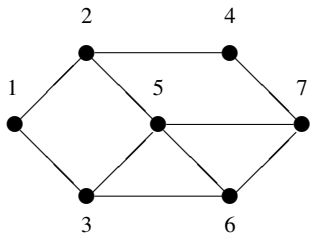


Fig. 2.1 Undirected graph used to illustrate the different Markov properties

The pairwise Markov property

The semigraphoid relation $\perp_\sigma$ satisfies *the pairwise Markov property* w.r.t. $\mathscr{G}$ if non-adjacent vertices are conditionally independent given the remaining, i.e.

$$\alpha \not\sim \beta \Rightarrow \alpha \perp_\sigma \beta \,|\, V \setminus \{\alpha,\beta\}.$$

For example, in Fig. 2.1 the pairwise Markov property states that

$$1 \perp_\sigma 5 \,|\, \{2,3,4,6,7\} \text{ and } 4 \perp_\sigma 6 \,|\, \{1,2,3,5,7\}.$$

If the relation $\perp_\sigma$ satisfies the pairwise Markov property, we also write that $\perp_\sigma$ *satisfies* (P).

The local Markov property

The semigraphoid relation $\perp_\sigma$ satisfies *the local Markov property* w.r.t. $\mathscr{G}$ if every variable is conditionally independent of the remaining, given its neighbours

$$\forall \alpha \in V : \alpha \perp_\sigma V \setminus \mathrm{cl}(\alpha) \,|\, \mathrm{bd}(\alpha).$$

For example, if $\perp_\sigma$ satisfies the local Markov property w.r.t. the graph in Fig. 2.1 it holds that $5 \perp_\sigma \{1,4\} \,|\, \{2,3,6,7\}$ and $7 \perp_\sigma \{1,2,3\} \,|\, \{4,5,6\}$. If the relation $\perp_\sigma$ satisfies the local Markov property, we also write that $\perp_\sigma$ *satisfies* (L).

The global Markov property

The semigraphoid relation $\perp_\sigma$ satisfies *the global Markov property* w.r.t. $\mathscr{G}$ if any two sets which are separated by a third are conditionally independent given the separating set

$$A \perp_{\mathscr{G}} B \,|\, S \Rightarrow A \perp_\sigma B \,|\, S.$$

To identify conditional independence relations in the graph of Fig. 2.1 one should look for separating sets, such as $\{2,3\}$, $\{4,5,6\}$, or $\{2,5,6\}$. For example, it follows that $1 \perp_\sigma 7 \,|\, \{4,5,6\}$ and $2 \perp_\sigma 6 \,|\, \{3,4,5\}$. If the relation $\perp_\sigma$ satisfies the global Markov property, we also write that $\perp_\sigma$ *satisfies* (G).

Structural relations among Markov properties

The various Markov properties are related, but different in general:

Theorem 2.1. *For any semigraphoid relation $\perp_\sigma$ it holds that*

$$\text{(G)} \Rightarrow \text{(L)} \Rightarrow \text{(P)}.$$

If $\perp_\sigma$ satisfies graphoid axioms it further holds that

$$\text{(P)} \Rightarrow \text{(G)}$$

so that in the graphoid case

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

The latter holds in particular for $\perp\!\!\!\perp$, when $f(x) > 0$, so that for probability distributions with positive densities, all the Markov properties coincide.

Proof. Since this result is so fundamental and the proof illustrates the use of graphoid axioms very well, we give the full argument here, following Lauritzen (1996).

(G) implies (L):

This holds because $\mathrm{bd}(\alpha)$ separates α from $V \setminus \mathrm{cl}(\alpha)$.

(L) implies (P):

Assume (L). Then $\beta \in V \setminus \mathrm{cl}(\alpha)$ because $\alpha \not\sim \beta$. Thus

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

Hence by (L) and (S3) we get that

$$\alpha \perp_\sigma (V \setminus \mathrm{cl}(\alpha)) \,|\, V \setminus \{\alpha, \beta\}.$$

(S2) then gives $\alpha \perp_\sigma \beta \,|\, V \setminus \{\alpha, \beta\}$ which is (P).

(P) implies (G) for graphoids:

The proof uses reverse induction to establish this for a general undirected graph.

Before we proceed to give this proof, due to Pearl and Paz (1987), it is helpful to note that the graphoid condition (S5):

$$A \perp_\sigma B \,|\, (C \cup D) \text{ and } A \perp_\sigma C \,|\, (B \cup D) \Rightarrow A \perp_\sigma (B \cup C) \,|\, D$$

exactly expresses that the pairwise Markov property (P) implies the global Markov property (G) on the graph in Fig. 2.2.

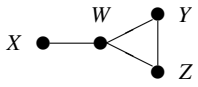


Fig. 2.2 The graphoid condition (S5) expresses that the pairwise Markov property (P) implies the global Markov property (G) on this particular graph.

Assume (P) and $A \perp_{\mathscr{G}} B \,|\, S$. We must show $A \perp_\sigma B \,|\, S$. Without loss of generality we assume that A and B are non-empty. The proof is reverse induction on $n = |S|$.

If $n = |V| - 2$ then A and B are singletons and (P) yields $A \perp_\sigma B \,|\, S$ directly.

Assume next that $|S| = n < |V| - 2$ and the conclusion has been established for $|S| > n$. Consider first the case $V = A \cup B \cup S$. Then either A or B has at least two elements, say A. If $\alpha \in A$ then $B \perp_{\mathscr{G}} (A \setminus \{\alpha\}) \,|\, (S \cup \{\alpha\})$ and also $\alpha \perp_{\mathscr{G}} B \,|\, (S \cup A \setminus \{\alpha\})$ (as $\perp_{\mathscr{G}}$ is a semi-graphoid). Thus by the induction hypothesis

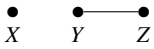
$$(A \setminus \{\alpha\}) \perp_\sigma B \,|\, (S \cup \{\alpha\}) \text{ and } \{\alpha\} \perp_\sigma B \,|\, (S \cup A \setminus \{\alpha\}).$$

Now (S5) gives $A \perp_\sigma B \mid S$.

For $A \cup B \cup S \subset V$ we choose $\alpha \in V \setminus (A \cup B \cup S)$. Then $A \perp_{\mathscr{G}} B \,|\, (S \cup \{\alpha\})$ and hence the induction hypothesis yields $A \perp_\sigma B \,|\, (S \cup \{\alpha\})$. Further, either $A \cup S$ separates B from $\{\alpha\}$ or $B \cup S$ separates A from $\{\alpha\}$. Assuming the former gives $\alpha \perp_\sigma B \,|\, A \cup S$. Using (S5) we get $(A \cup \{\alpha\}) \perp_\sigma B \,|\, S$ and from (S2) we derive that $A \perp_\sigma B \mid S$. The latter case is similar. □

The Markov properties are genuinely different in general if the graphoid axioms are not satisfied, as demonstrated by the examples below.

Example 2.1 (Pairwise Markov but not local Markov). Let $X = Y = Z$ with $P\{X = 1\} = P\{X = 0\} = 1/2$. This distribution satisfies (P) but not (L) with respect to the graph below.



The pairwise Markov property says that $X \perp\!\!\!\perp Y \,|\, Z$ and $X \perp\!\!\!\perp Z \,|\, Y$, which both are satisfied. However, we have that $\mathrm{bd}(X) = \emptyset$ so (L) would imply $X \perp\!\!\!\perp (Y, Z)$ which is false.

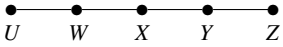
It can be shown that (L) $\iff$ (P) *if and only if* $\check{\mathscr{G}}$ *has no induced subgraph* $\check{\mathscr{G}}_A = (A, \check{E}_A)$ *with* $|A| = 3$ *and* $|\check{E}_A| \in \{2, 3\}$ (Matúš 1992).

Here the dual graph $\check{\mathscr{G}}$ is defined by $\alpha \check{\sim} \beta$ if and only if $\alpha \not\sim \beta$, i.e. has edges exactly where $\mathscr{G}$ does not.

Example 2.2 (Local Markov but not global Markov). Let U and Z be independent with

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

$W = U$, $Y = Z$, and $X = WY$. This satisfies (L) but not (G) w.r.t. the graph below.



The local Markov property follows because all variables depend deterministically on their neighbours. But the global Markov property fails; for example it is false that $W \perp\!\!\!\perp Y \,|\, X$.

It can be shown that (G) $\iff$ (L) *if and only if the dual graph* $\check{\mathscr{G}}$ *does not have the 4-cycle as an induced subgraph* (Matúš 1992).

Factorization and Markov properties

For $a \subseteq V$, $\psi_a(x)$ is a function depending on x_a only, i.e.

$$x_a = y_a \Rightarrow \psi_a(x) = \psi_a(y).$$

We can then write $\psi_a(x) = \psi_a(x_a)$ without ambiguity.

The distribution of X *factorizes w.r.t.* $\mathscr{G}$ or satisfies (F) if its density f w.r.t. product measure on $\mathscr{X}$ has the form

$$f(x) = \prod_{a \in \mathscr{A}} \psi_a(x),$$

where $\mathscr{A}$ are *complete* subsets of $\mathscr{G}$ or, equivalently, if

$$f(x) = \prod_{c \in \mathscr{C}} \tilde{\psi}_c(x),$$

where $\mathscr{C}$ are the *cliques* of $\mathscr{G}$.

Example 2.3. The cliques of the graph in Fig. 2.1 are the maximal complete subsets $\{1,2\}$, $\{1,3\}$, $\{2,4\}$, $\{2,5\}$, $\{3,5,6\}$, $\{4,7\}$, and $\{5,6,7\}$ and a complete set is any subset of these sets, for example $\{2\}$ or $\{5,7\}$. The graph corresponds to a factorization as

$$\begin{aligned} f(x) = {} & \psi_{12}(x_1,x_2)\psi_{13}(x_1,x_3)\psi_{24}(x_2,x_4)\psi_{25}(x_2,x_5) \\ & \times \psi_{356}(x_3,x_5,x_6)\psi_{47}(x_4,x_7)\psi_{567}(x_5,x_6,x_7). \end{aligned}$$

Consider a distribution with density w.r.t. a product measure and let (G), (L) and (P) denote Markov properties w.r.t. the semigraphoid relation $\perp\!\!\!\perp$.

Theorem 2.2. *It holds that*

$$\text{(F)} \Rightarrow \text{(G)}$$

and if the density is strictly positive it further holds that (P) $\Rightarrow$ (F)*, such that then all the Markov properties coincide:*

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

Proof. See Lauritzen (1996, pp. 35–36).

Without the positivity restriction (G) and (F) are genuinely different, as illustrated in the example below, due to Moussouris (1974).

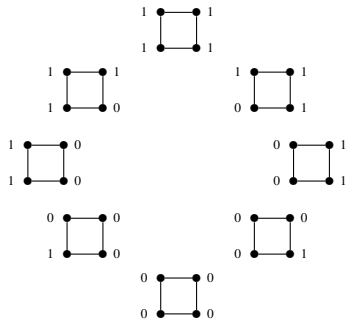


Fig. 2.3 The distribution which is uniform on these 8 configurations satisfies (G) w.r.t. the 4-cycle. Yet it does not factorize with respect to this graph.

Example 2.4 (Global but not factorizing). Consider the uniform distribution on the 8 configurations displayed in Fig. 2.3. Conditioning on opposite corners renders one corner deterministic and therefore the global Markov property is satisfied.

However, the density does not factorize. To see this we assume the density factorizes. Then e.g.

$$0 \neq 1/8 = f(0,0,0,0) = \psi_{12}(0,0)\psi_{23}(0,0)\psi_{34}(0,0)\psi_{41}(0,0)$$

so these factors are all positive. Continuing for all possible 8 configurations yields that all factors $\psi_a(x)$ are strictly positive, since all four possible configurations are possible for every clique.

But this contradicts the fact that only 8 out of 16 possible configurations have positive probability. □

In fact, we shall see later that (F) $\iff$ (G) *if and only if $\mathscr{G}$ is chordal*, i.e. does not have an n-cycle as an induced subgraph with $n \geq 4$.

Instability of conditional independence under weak limits

Consider a sequence $P_n, n = 1,2,\ldots$ of probability measures on $\mathscr{X}$ and assume that $A \perp\!\!\!\perp_{P_n} B \,|\, C$. If $P_n \to P$ weakly, it does *not* hold in general that $A \perp\!\!\!\perp_P B \,|\, C$. A simple counterexample is as follows: Consider $X = (X_1, X_2, X_3) \sim \mathscr{N}_3(0, \Sigma_n)$ with

$$\Sigma_n = \begin{pmatrix} 1 & \frac{1}{\sqrt{n}} & \frac{1}{2} \\ \frac{1}{\sqrt{n}} & \frac{2}{n} & \frac{1}{\sqrt{n}} \\ \frac{1}{2} & \frac{1}{\sqrt{n}} & 1 \end{pmatrix} \to \begin{pmatrix} 1 & 0 & \frac{1}{2} \\ 0 & 0 & 0 \\ \frac{1}{2} & 0 & 1 \end{pmatrix}$$

so in the limit it is not true that $1 \perp\!\!\!\perp_P 3 \,|\, 2$. The concentration matrix K_n is

$$K_n = \Sigma_n^{-1} = \begin{pmatrix} 2 & -\sqrt{n} & 0 \\ -\sqrt{n} & \frac{3n}{2} & -\sqrt{n} \\ 0 & -\sqrt{n} & 2 \end{pmatrix}$$

so for all n it holds that $1 \perp\!\!\!\perp_{P_n} 3 \,|\, 2$. The critical feature is that K_n does not converge, hence the densities do not converge.

Stability of conditional independence under limits

If $\mathscr{X}$ is discrete and finite and $P_n \to P$ pointwise, conditional independence is preserved: This follows from the fact that

$$X \perp\!\!\!\perp_{P_n} Y \,|\, Z \iff f_n(x,y,z) f_n(z) = f_n(x,z) f_n(y,z)$$

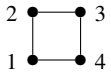
and this relation is clearly stable under pointwise limits. Hence (G), (L) and (P) *are closed under pointwise limits in the discrete case*.

In general, *conditional independence is preserved if $P_n \to P$ in total variation* (A. Klenke, personal communication, St Flour 2006).

Example 2.5 (Instability of factorization under limits). Even in the discrete case, (F) *is not closed under pointwise limits in general.* Consider four binary variables X_1, X_2, X_3, X_4 with joint distribution

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

This factorizes w.r.t. the graph below.



It holds that $f_n(x) = n/(8+8n)$ for each of the configurations below

$$(0,0,0,0)\ (1,0,0,0)\ (1,1,0,0)\ (1,1,1,0)$$
$$(0,0,0,1)\ (0,0,1,1)\ (0,1,1,1)\ (1,1,1,1),$$

whereas $f_n(x) = 1/(8+8n)$ for the remaining 8 configurations. Thus, when $n \to \infty$ the density f_n converges to $f(x) = 1/8$ for each of the configurations above and $f(x) = 0$ otherwise, i.e. to the distribution in Example 2.4 which is globally Markov but does not factorize.

Markov faithfulness

A distribution P is said to be *Markov faithful* to a graph $\mathscr{G}$ if it holds that

$$A \perp_{\mathscr{G}} B \,|\, S \iff A \perp\!\!\!\perp_P B \,|\, S.$$

It can be shown by a dimensional argument that *if* $|\mathscr{X}_v| \geq 2$ *for all* $v \in V$*, then there is a distribution* P *which is Markov faithful to* $\mathscr{G}$. For a Markov faithful P, the graphoids $\perp_{\mathscr{G}}$ and $\perp\!\!\!\perp_P$ are isomorphic.

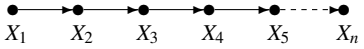
In fact, in the discrete and finite case, *the set of Markov distributions which are not faithful to a given graph is a Lebesgue null-set in the set of Markov distributions.* No formal proof seems to be published, but Meek (1995) gives a proof for the case of directed acyclic graphs and indicates how this can be extended to undirected graphs.

2.3 Markov Properties for Directed Acyclic Graphs

A *directed acyclic graph* $\mathscr{D}$ over a finite set V is a simple graph with all edges directed and no directed cycles in the sense that that following arrows in the graph, it is impossible to return to any point.

Graphical models based on DAGs have proved fundamental and useful in a wealth of interesting applications, including expert systems, genetics, complex biomedical statistics, causal analysis, and machine learning, see for example Fig. 1.1 and other examples in Chapter 1.

The directed Markov properties are straight-forward generalizations of the notion of a Markov chain with $X_{i+1} \perp\!\!\!\perp \{X_1, \ldots, X_{i-1}\} \,|\, X_i$ for $i = 3, \ldots, n$:



Local directed Markov property

A semigraphoid relation $\perp_\sigma$ satisfies *the local Markov property* (L) w.r.t. a directed acyclic graph $\mathscr{D}$ if all variables are conditionally independent of its non-descendants given its parents.

$$\forall \alpha \in V : \alpha \perp_\sigma \{\mathrm{nd}(\alpha) \setminus \mathrm{pa}(\alpha)\} \,|\, \mathrm{pa}(\alpha).$$

Here $\mathrm{nd}(\alpha)$ denotes the *non-descendants* of α.

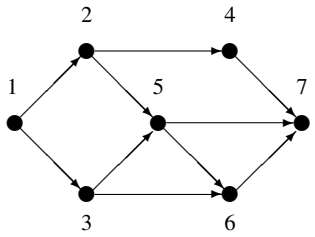


Fig. 2.4 A directed, acyclic graph

The local Markov property for the DAG in Fig. 2.4 yields, for example, that $4 \perp_\sigma \{1,3,5,6\} \,|\, 2$, $5 \perp_\sigma \{1,4\} \,|\, \{2,3\}$, and $3 \perp_\sigma \{2,4\} \,|\, 1$.

Ordered Markov property

Suppose the vertices V of a DAG $\mathscr{D}$ are *well-ordered* in the sense that they are linearly ordered in a way which is compatible with $\mathscr{D}$, i.e. so that

$$\alpha \in \mathrm{pa}(\beta) \Rightarrow \alpha < \beta.$$

We then say that the semigraphoid relation $\perp_\sigma$ satisfies the *ordered Markov property* (O) w.r.t. a well-ordered DAG $\mathscr{D}$ if

$$\forall \alpha \in V : \alpha \perp_\sigma \{\mathrm{pr}(\alpha) \setminus \mathrm{pa}(\alpha)\} \,|\, \mathrm{pa}(\alpha).$$

Here $\mathrm{pr}(\alpha)$ are the *predecessors* of α, i.e. those which are before α in the well-ordering.

The numbering in Fig. 2.4 corresponds to a well-ordering. The ordered Markov property says for example that $4 \perp_\sigma \{1,3\} \,|\, 2$, $5 \perp_\sigma \{1,4\} \,|\, \{2,3\}$, and $3 \perp_\sigma \{2\} \,|\, 1$.

Separation in DAGs

The global Markov property for directed acyclic graphs is expressed in terms of a type of separation which is somewhat involved compared to the undirected case.

A *trail* τ from α to β is a sequence $v_1, v_2, \ldots, v_n$ of edges with $\alpha = v_1$, $\beta = v_n$ and all consecutive vertices being adjacent. A trail τ in $\mathscr{D}$ is *blocked* by a set S if it contains a vertex $\gamma \in \tau$ such that

- either $\gamma \in S$ and edges of τ do not meet head-to-head at γ, or
- γ and all its descendants are not in S, and edges of τ meet head-to-head at γ.

A trail that is not blocked is *active*. Two subsets A and B of vertices are *d-separated* by S if all trails from A to B are blocked by S. We write $A \perp_{\mathscr{D}} B \,|\, S$.

In the DAG of Fig. 2.4 we have, for example, that for $S = \{5\}$, the trail $(4,2,5,3,6)$ is *active*, whereas the trails $(4,2,5,6)$ and $(4,7,6)$ are *blocked*. For $S = \{3,5\}$ all these trails are blocked. Hence it holds that $4 \perp_{\mathscr{D}} 6 \,|\, 3,5$, but it is *not* true that $4 \perp_{\mathscr{D}} 6 \,|\, 5$ nor that $4 \perp_{\mathscr{D}} 6$.

Global directed Markov property

A semigraphoid relation $\perp_\sigma$ satisfies the *global Markov property* (G) w.r.t. a directed acyclic graph $\mathscr{D}$ if

$$A \perp_{\mathscr{D}} B \,|\, S \Rightarrow A \perp_\sigma B \,|\, S.$$

In Fig. 2.4 the global Markov property thus entails that $4 \perp\!\!\!\perp 6 \,|\, 3,5$ and $2 \perp\!\!\!\perp 3 \,|\, 1$.

Equivalence of Markov properties

In the directed case the relationship between the alternative Markov properties is much simpler than in the undirected case.

Proposition 2.1. *It holds for any directed acyclic graph $\mathscr{D}$ and any semigraphoid relation $\perp_\sigma$ that all directed Markov properties are equivalent:*

$$\text{(G)} \iff \text{(L)} \iff \text{(O)}.$$

We omit the proof of this fact and refer to Lauritzen et al (1990) for details. **FiXme Fatal: give the proof here?**

There is also a pairwise property (P), but it is less natural than in the undirected case and it is weaker than the others, see Lauritzen (1996, page 51).

Factorisation with respect to a DAG

A probability distribution P over $\mathscr{X} = \mathscr{X}_V$ *factorizes* over a DAG $\mathscr{D}$ if its density f w.r.t. some product measure μ has the form

$$(\mathrm{F}): \quad f(x) = \prod_{v \in V} k_v(x_v \,|\, x_{\mathrm{pa}(v)})$$

where $k_v \geq 0$ and $\int_{\mathscr{X}_v} k_v(x_v \,|\, x_{\mathrm{pa}(v)})\, \mu_v(dx_v) = 1$. It can be easily shown by induction that (F) *is equivalent to* (F*), where

$$(\mathrm{F}^*): \quad f(x) = \prod_{v \in V} f(x_v \,|\, x_{\mathrm{pa}(v)}),$$

i.e. it follows from (F) that k_v *in fact are conditional densities.* The graph in Fig. 2.4 thus corresponds to the factorization

$$\begin{aligned} f(x) = {} & f(x_1) f(x_2 \,|\, x_1) f(x_3 \,|\, x_1) f(x_4 \,|\, x_2) \\ & \times f(x_5 \,|\, x_2, x_3) f(x_6 \,|\, x_3, x_5) f(x_7 \,|\, x_4, x_5, x_6). \end{aligned}$$

Markov properties and factorization

Assume that the probability distribution P has a density w.r.t. some product measure on $\mathscr{X}$. It is then always true that (F) holds if and only if $\perp\!\!\!\perp_P$ satisfies (G), so all directed Markov properties are equivalent to the factorization property!

$$(\mathrm{F}) \iff (\mathrm{G}) \iff (\mathrm{L}) \iff (\mathrm{O}). \tag{2.5}$$

FiXme Fatal: give the proof here?

Ancestral marginals

The directed Markov properties are closed under marginalization to *ancestral subsets,* i.e. sets which contain the parents of all its vertices

$$\alpha \in A \Rightarrow \mathrm{pa}(\alpha) \in A.$$

Proposition 2.2. *If P factorizes w.r.t. $\mathscr{D}$ and $A \subseteq V$ is ancestral, it factorizes w.r.t. $\mathscr{D}_A$.*

Proof. Induction after $|V|$, using that if A is ancestral and $A \neq V$, there is a terminal vertex α_0 with $\alpha_0 \notin A$. Hence the A-marginal can be obtained by first marginalizing to $V' = V \setminus \{\alpha_0\}$ and subsequently marginalizing to A from V' which has one vertex less than V. □

Moralization and undirected factorizations

The *moral graph* $\mathscr{D}^m$ of a DAG $\mathscr{D}$ is obtained by adding undirected edges between unmarried parents and subsequently dropping directions, as illustrated in Fig. 2.5.

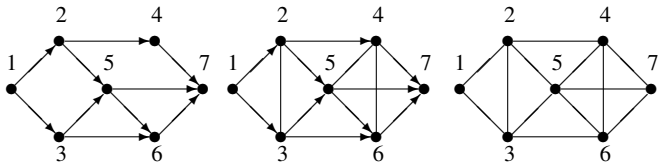


Fig. 2.5 Illustration of the moralization process. Undirected edges are added to parents with a common child. Directions on edges are subsequently dropped.

Markov properties of directed and undirected graphs are different in general. However, there are obvious important connections between directed and undirected factorizations. We have for example the following

Proposition 2.3. *If P factorizes w.r.t. $\mathscr{D}$, it factorizes w.r.t. the moralized graph $\mathscr{D}^m$.*

Proof. This is seen directly from the factorization:

$$f(x) = \prod_{v \in V} f(x_v \,|\, x_{\mathrm{pa}(v)}) = \prod_{v \in V} \psi_{\{v\} \cup \mathrm{pa}(v)}(x),$$

since $\{v\} \cup \mathrm{pa}(v)$ are all complete in $\mathscr{D}^m$. □

Hence if P satisfies any of the directed Markov properties w.r.t. $\mathscr{D}$, it satisfies all Markov properties for $\mathscr{D}^m$.

Perfect DAGs

The *skeleton* $\sigma(\mathscr{D})$ of a DAG is the undirected graph obtained from $\mathscr{D}$ by ignoring directions.

A DAG $\mathscr{D}$ is *perfect* if all parents are married or, in other words if $\sigma(\mathscr{D}) = \mathscr{D}^m$. It follows directly from Proposition 2.3 that the directed and undirected properties are identical for a perfect DAG $\mathscr{D}$:

Corollary 2.1. *P factorizes w.r.t a perfect DAG $\mathscr{D}$ if and only if it factorizes w.r.t. its skeleton $\sigma(\mathscr{D})$.*

Not that a *rooted tree* with arrows pointing away from the root is a perfect DAG. Thus for such a rooted tree the directed and undirected Markov properties are the same.

In particular this yields the well-known fact that *any Markov chain is also a Markov field.*

We shall later see that *an undirected graph $\mathcal{G}$ can be oriented to form a perfect DAG if and only if $\mathcal{G}$ is chordal.*

Alternative equivalent separation criterion

The criterion of d-separation can be difficult to verify in some cases, although efficient algorithms to settle d-separation queries exists. For example, Geiger et al (1990) describe an algorithm with worst case complexity $O(|E|)$ for finding all vertices α which satisfy $\alpha \perp_{\mathcal{D}} B \,|\, S$ for fixed sets B and S.

Algorithms for settling such queries can also be based on the following alternative separation criterion given by Lauritzen et al (1990) which is based on Propositions 2.2 and 2.3. For a query involving three sets A, B, S we perform the following

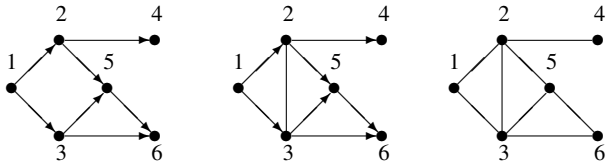


Fig. 2.6 To settle the query "$4 \perp_m 6 \,|\, 3,5$?" we first form the subgraph induced by all ancestors of vertices involved. The moralization adds an undirected edge between 2 and 3 with common child 5 and drops directions. Since $\{3,5\}$ separates 4 from 6 in the resulting graph, we conclude that $4 \perp_m 6 \,|\, 3,5$.

operations:

1. Reduce to subgraph induced by ancestral set $\mathcal{D}_{\mathrm{An}(A \cup B \cup S)}$ of $A \cup B \cup S$;
2. Moralize to form $(\mathcal{D}_{\mathrm{An}(A \cup B \cup S)})^m$;
3. Say that *S m-separates A from B* and write $A \perp_m B \mid S$ if and only if S separates A from B in this undirected graph.

The procedure is illustrated in Fig. 2.6. It now follows directly from Propositions 2.2 and 2.3 that

Corollary 2.2. *If P factorizes w.r.t. $\mathcal{D}$ it holds that*

$$A \perp_m B \,|\, S \Rightarrow A \perp\!\!\!\perp B \,|\, S.$$

Proof. This holds because then P factorizes w.r.t. $\mathcal{D}^m_{\mathrm{An}(A \cup B \cup S)}$ and hence satisfies (G) for this graph. □

Indeed the concepts of m-separation and d-separation are equivalent:

Proposition 2.4. *$A \perp_m B \mid S$ if and only if $A \perp_{\mathcal{D}} B \mid S$.*

Proof. This is Proposition 3.25 of Lauritzen (1996). □

Note however that Richardson (2003) has pointed out that the proof given in Lauritzen et al (1990) and Lauritzen (1996) needs to allow self-intersecting paths to be correct. **FiXme Fatal: give the correct proof here**

It holds for any DAG $\mathscr{D}$ that $\perp_{\mathscr{D}}$ (and hence $\perp_m$) *satisfies graphoid axioms* (Verma and Pearl 1990).

To show this is true, it is sometimes easy to use $\perp_m$, sometimes $\perp_{\mathscr{D}}$. For example, (S2) is trivial for $\perp_{\mathscr{D}}$, whereas (S5) is trivial for $\perp_m$. So, equivalence of $\perp_{\mathscr{D}}$ and $\perp_m$ can be very useful.

Faithfulness

As in the undirected case, a distribution P is said to be *Markov faithful* for a DAG $\mathscr{D}$ if it holds that

$$A \perp_{\mathscr{D}} B \,|\, S \iff A \perp\!\!\!\perp_P B \,|\, S.$$

For a Markov faithful P, the graphoids $\perp_{\mathscr{D}}$ and $\perp\!\!\!\perp_P$ are isomorphic.

If $|\mathscr{X}_v| \geq 2$ for all $v \in V$, then there is a distribution P which is Markov faithful for $\mathscr{D}$, and it holds further that *the set of directed Markov distributions which are not faithful is a Lebesgue null-set* in the set of directed Markov distributions (Meek 1995), confirming in particular that the criterion of d-separation is indeed the strongest possible.

Markov equivalence

Two DAGS $\mathscr{D}$ and $\mathscr{D}'$ are said to be *Markov equivalent* if the separation relations $\perp_{\mathscr{D}}$ and $\perp_{\mathscr{D}'}$ are identical. Markov equivalence between DAGs is easy to identify, as shown by Frydenberg (1990a) and Verma and Pearl (1990).

Fig. 2.7 The two DAGs to the left are Markov equivalent whereas those to the right are not. Although those to the right have the same skeleton they do not share the same unmarried parents.

Proposition 2.5. *Two directed acyclic graphs $\mathscr{D}$ and $\mathscr{D}'$ are Markov equivalent if and only if $\mathscr{D}$ and $\mathscr{D}'$ have the same skeleton and the same unmarried parents.*

The use of this result is illustrated in Fig. 2.7.

A DAG $\mathscr{D}$ is *Markov equivalent* to an undirected $\mathscr{G}$ if the separation relations $\perp_{\mathscr{D}}$ and $\perp_{\mathscr{G}}$ are identical.

This happens if and only if $\mathscr{D}$ is perfect and $\mathscr{G} = \sigma(\mathscr{D})$. So the graphs below are all equivalent

but not equivalent to the directed acyclic graph below

2.4 Summary

We conclude by a summary of the most important definitions and facts given in the present chapter.

Markov properties for undirected graphs

(P) *Pairwise Markov property:* $\alpha \not\sim \beta \Rightarrow \alpha \perp\!\!\!\perp \beta \,|\, V \setminus \{\alpha, \beta\}$;
(L) *Local Markov property:* $\alpha \perp\!\!\!\perp V \setminus \mathrm{cl}(\alpha) \,|\, \mathrm{bd}(\alpha)$;
(G) *Global Markov property:* $A \perp_{\mathscr{G}} B \,|\, S \Rightarrow A \perp\!\!\!\perp B \,|\, S$;
(F) *Factorization property:* $f(x) = \prod_{a \in \mathscr{A}} \psi_a(x)$, $\mathscr{A}$ being complete subsets of V.

It then holds that

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

If $f(x) > 0$ even

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

Markov properties for directed acyclic graphs

(O) *Ordered Markov property:* $\alpha \perp\!\!\!\perp \{\mathrm{pr}(\alpha) \setminus \mathrm{pa}(\alpha)\} \,|\, \mathrm{pa}(\alpha)$;
(L) *Local Markov property:* $\alpha \perp\!\!\!\perp \{\mathrm{nd}(\alpha) \setminus \mathrm{pa}(\alpha)\} \,|\, \mathrm{pa}(\alpha)$;
(G) *Global Markov property:* $A \perp_{\mathscr{D}} B \,|\, S \Rightarrow A \perp\!\!\!\perp B \,|\, S$.
(F) *Factorization property:* $f(x) = \prod_{v \in V} f(x_v \,|\, x_{\mathrm{pa}(v)})$.

It then *always* holds that

$$\text{(F)} \iff \text{(G)} \iff \text{(L)} \iff \text{(O)}.$$

Relation between Markov properties on different graphs

If P is directed Markov w.r.t. $\mathscr{D}$ *then P factorizes w.r.t.* $\mathscr{D}^m$.

A DAG $\mathscr{D}$ is *perfect* if skeleton $\mathscr{G} = \sigma(\mathscr{D}) = \mathscr{D}^m$, implying that *directed and undirected separation properties are identical,* i.e. $A \perp_{\mathscr{G}} B \,|\, S \iff A \perp_{\mathscr{D}} B \,|\, S$.

An undirected graph $\mathscr{G}$ *is the skeleton of a perfect DAG* $\mathscr{D}$*, i.e.* $\mathscr{G} = \sigma(\mathscr{D}) = \mathscr{D}^m$*, if and only if* $\mathscr{G}$ *is chordal.*

Two DAGs $\mathscr{D}$ and $\mathscr{D}'$ are *Markov equivalent,* i.e. $A \perp_{\mathscr{D}} B \,|\, S \iff A \perp_{\mathscr{D}'} B \,|\, S$, *if and only if* $\sigma(\mathscr{D}) = \sigma(\mathscr{D}')$ *and* $\mathscr{D}$ *and* $\mathscr{D}'$ *have the same unmarried parents.*

Chapter 3
Graph Decompositions and Algorithms

One important feature of graphical models is modularity; probabilistic information in complex stochastic systems is distributed over smaller modules exploiting conditional independence relations. The present chapter is specifically concerned with such aspects.

3.1 Graph Decompositions and Markov Properties

Definition 3.1 (Graph decomposition). A partitioning of V into a triple (A,B,S) of subsets of V forms a *decomposition* of an undirected graph $\mathscr{G}$ if both of the following holds:

(i) $A \perp_{\mathscr{G}} B \,|\, S$;
(ii) S is complete.

The decomposition is *proper* if $A \neq \emptyset$ and $B \neq \emptyset$ and the *components* of $\mathscr{G}$ are the induced subgraphs $\mathscr{G}_{A\cup S}$ and $\mathscr{G}_{B\cup S}$. A graph is said to be *prime* if no proper decomposition exists. Examples of prime graphs and graph decompositions are given in Fig. 3.1 and Fig. 3.2. *Any finite undirected graph can be recursively decomposed into its uniquely defined prime components* (Wagner 1937; Tarjan 1985; Diestel 1987, 1990), as illustrated in Fig. 3.3.

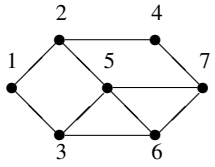


Fig. 3.1 An example of a prime graph. This graph has no complete separators.

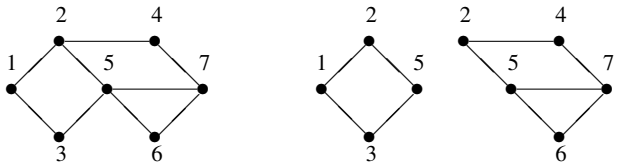


Fig. 3.2 Decomposition with $A = \{1,3\}$, $B = \{4,6,7\}$ and $S = \{2,5\}$.

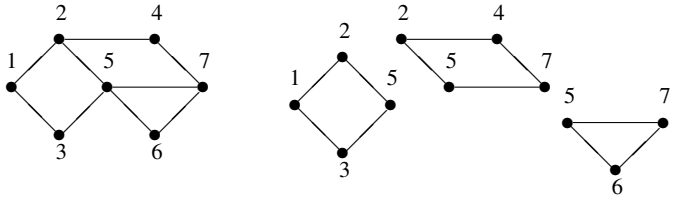


Fig. 3.3 Recursive decomposition of a graph into its unique prime components.

Definition 3.2 (Decomposable graph). A graph is said to be *decomposable* if its prime components are cliques.

It would make more sense to say that such a graph is fully decomposable and reserve the term decomposable for a graph that is not prime. However, this has not been the tradition in the statistical literature.

Decomposition of Markov properties

Graph decompositions are important because they correspond to decomposition and thus modularization of the Markov properties, as captured in the following result.

Proposition 3.1. *Let (A,B,S) be a decomposition of $\mathcal{G}$. Then P factorizes w.r.t. $\mathcal{G}$ if and only if both of the following hold:*

(i) $P_{A\cup S}$ and $P_{B\cup S}$ factorize w.r.t. $\mathcal{G}_{A\cup S}$ and $\mathcal{G}_{B\cup S}$;
(ii) $f(x)f_S(x_S) = f_{A\cup S}(x_{A\cup S})f_{B\cup S}(x_{B\cup S})$.

Proof. This is Proposition 3.16 of (Lauritzen 1996). □

Recursive decomposition of a decomposable graph yields:

$$f(x)\prod_{S\in\mathcal{S}} f_S(x_S)^{\nu(S)} = \prod_{C\in\mathcal{C}} f_C(x_C).$$

Here $\mathcal{S}$ is the set of *complete separators* occurring in the decomposition process and $\nu(S)$ the number of times a given S appears. More generally, if $\mathcal{Q}$ denotes the prime components of $\mathcal{G}$ we have:

$$f(x)\prod_{S\in\mathscr{S}} f_S(x_S)^{\nu(S)} = \prod_{Q\in\mathscr{Q}} f_Q(x_Q). \tag{3.1}$$

Combinatorial consequences

If in (3.1) we let $\mathscr{X}_v = \{0,1\}$ and f be uniform, i.e. $f(x) = 2^{-|V|}$, this yields

$$2^{-|V|}\prod_{S\in\mathscr{S}} 2^{-|S|\nu(S)} = \prod_{Q\in\mathscr{Q}} 2^{-|Q|}$$

and hence we must have

$$\sum_{Q\in\mathscr{Q}} |Q| = \sum_{S\in\mathscr{S}} |S|\nu(S) + |V|. \tag{3.2}$$

Similarly the right and left hand sides of (3.1) must have the same number of factors as every decomposition yields an extra factor on both sides of the equation and hence it holds that

$$|\mathscr{Q}| = \sum_{S\in\mathscr{S}} \nu(S) + 1.$$

These identities were also derived in a slightly different form in Lauritzen et al (1984).

Example 3.1. An undirected tree τ is decomposable with prime components being the edges and the separators equal to the vertices, with the multiplicity $\nu(v)$ being one less than the *degree* of the vertex

$$\nu(v) = \deg(v) - 1 = |\operatorname{bd}(v)|,$$

and the combinatorial identities above reduce to

$$2|E| = \sum_{v\in V} \deg(v), \quad |E| = \sum_{v\in V} (\deg(v)+1) + 1 = |V| + 1 + \sum_{v\in V} \deg(v)$$

implying in particular

$$|E| = |V| + 1.$$

3.2 Chordal Graphs and Junction Trees

Properties associated with decomposability

A numbering $V = \{1,\ldots,|V|\}$ of the vertices of an undirected graph is said to be *perfect* if the induced oriented graph is a perfect DAG or, equivalently, if

$$\forall j = 2,\ldots,|V| : \operatorname{bd}(j) \cap \{1,\ldots,j-1\} \text{ is complete in } \mathscr{G}.$$

An undirected graph $\mathcal{G}$ is said to be *chordal* if it has no n-cycle with $n \geq 4$ as an induced subgraph. Chordal graphs are also known as *rigid circuit* graphs (Dirac 1961) or *triangulated* graphs (Berge 1973). A set S is called an (α,β)-*separator* if $\alpha \perp_{\mathcal{G}} \beta \,|\, S$. Chordal graphs can now be characterized as follows.

Proposition 3.2. *The following conditions are equivalent for any undirected graph* $\mathcal{G}$.

(i) $\mathcal{G}$ *is chordal;*
(ii) $\mathcal{G}$ *is decomposable;*
(iii) $\mathcal{G}$ *admits a perfect numbering;*
(iv) *All minimal* (α,β)*-separators are complete.*

Two important lemmas will be used in the following and will therefore be quoted here. They are simple reformulations of Lemma 2.19 and Lemma 2.21 of Lauritzen (1996) and the reader is referred to this text for formal proofs:

Lemma 3.1. *Let* $\mathcal{G} = (V,E)$ *be a chordal graph and* $\mathcal{G}' = (V,E')$ *a subgraph with exactly one less edge. Then* $\mathcal{G}'$ *is chordal if and only the endpoints of this edge are contained in exactly one clique of* $\mathcal{G}$.

The next lemma ensures that any chordal subgraph of a chordal graph can be obtained from the larger by removing a single edge at a time without violating chordality.

Lemma 3.2. *Let* $\mathcal{G} = (V,E)$ *be a chordal graph and* $\mathcal{G}' = (V,E')$ *a chordal subgraph with k less edges. Then there exists a sequence of chordal graphs* $\mathcal{G}' = \mathcal{G}_0 \subset \mathcal{G}_1 \subset \cdots \subset \mathcal{G}_k = \mathcal{G}$ *which all differ by a single edge.*

Identifying chordal graphs

There are several algorithms for identifying chordal graphs. Here is a greedy algorithm for checking chordality based on the fact that chordal graphs are those that admit perfect numberings:

Algorithm 3.1. `Greedy algorithm for checking chordality of a graph and identifying a perfect numbering:`

1. `Look for a vertex` v^* `with` $\mathrm{bd}(v^*)$ `complete.`
   If no such vertex exists, the graph is not chordal.
2. `Form the subgraph` $\mathcal{G}_{V\setminus v^*}$ `and let` $v^* = |V|$;
3. `Repeat the process under 1;`
4. *If the algorithm continues until only one vertex is left, the graph is chordal and the numbering is perfect.*

The worst-case complexity of this algorithm is $O(|V|^2)$ as $|V| - k$ vertices must be queried to find the vertex to be numbered as $|VZ| - k$. The algorithm is illustrated in Fig. 3.4 and Fig. 3.5.

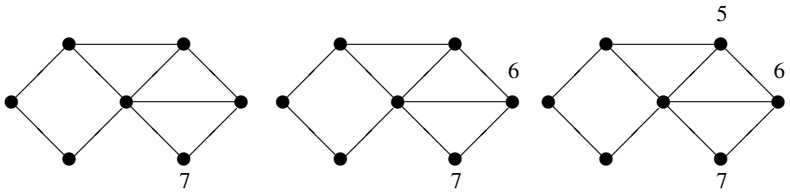


Fig. 3.4 The greedy algorithm at work. This graph is *not* chordal, as there is no candidate for number 4.

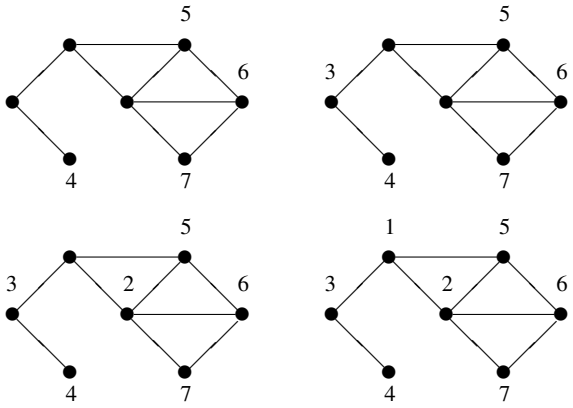


Fig. 3.5 The greedy algorithm at work. Initially the algorithm proceeds as in Fig. 3.4. This graph is chordal and the numbering obtained is a perfect numbering.

Maximum cardinality search

This simple algorithm is due to Tarjan and Yannakakis (1984) and has complexity $O(|V|+|E|)$. It checks chordality of the graph and generates a perfect numbering if the graph is chordal. In addition, as we shall see in a moment, the cliques of the chordal graph can be identified as the algorithm runs.

Algorithm 3.2 (Maximum Cardinality Search). Checking chordality of a chordal graph and identifying a perfect numbering:

1. Choose $v_0 \in V$ arbitrary and let $v_0 = 1$;
2. When vertices $\{1,2,\ldots,j\}$ have been identified, choose $v = j+1$ among $V \setminus \{1,2,\ldots,j\}$ with highest cardinality of its numbered neighbours;
3. *If* $\mathrm{bd}(j+1) \cap \{1,2,\ldots,j\}$ *is not complete,* $\mathscr{G}$ is not chordal;
4. Repeat from 2;
5. *If the algorithm continues until only one vertex is left, the graph is chordal and the numbering is perfect.*

The algorithm is illustrated in Fig. 3.7 and Fig. 3.6.

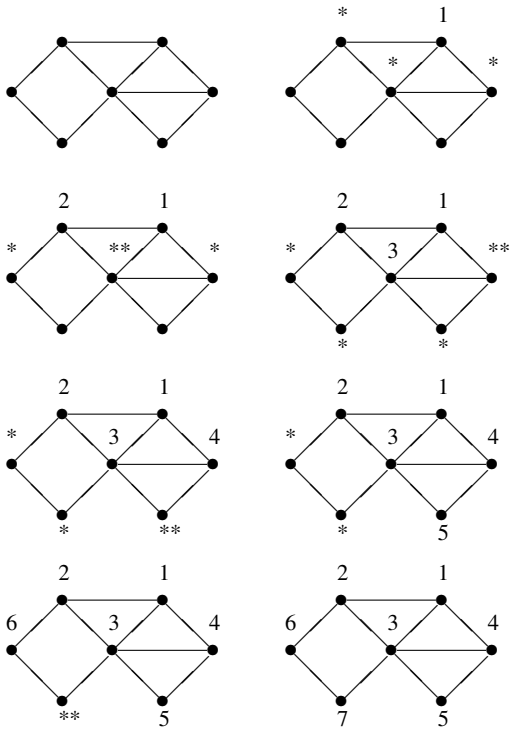


Fig. 3.6 Maximum Cardinality Search at work. When a vertex is numbered, a counter for each of its unnumbered neighbours is increased with one, marked here with the symbol $*$. The counters keep track of the numbered neighbours of any vertex and are used to identify the next vertex to be numbered. This graph is *not* chordal as discovered at the last step because 7 does not have a complete boundary.

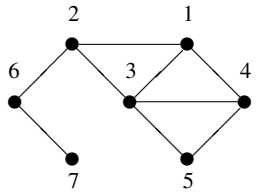


Fig. 3.7 MCS numbering for a chordal graph. The algorithm runs essentially as in the non-chordal case.

Finding the cliques

Finding the cliques of a general graph is an NP-complete problem. But the cliques of a chordal graph can be found in a simple fashion from a MCS numbering $V = \{1,\dots,|V|\}$. More precisely we let

$$S_\lambda = \mathrm{bd}(\lambda) \cap \{1,\dots,\lambda-1\}$$

and $\pi_\lambda = |B_\lambda|$. Say that λ is a *ladder vertex* if $\lambda = |V|$ or if $\pi_{\lambda+1} < \pi_\lambda + 1$ and let Λ be the set of ladder vertices.

It then holds that the cliques of $\mathscr{G}$ are $C_\lambda = \{\lambda\} \cup B_\lambda, \lambda \in \Lambda$. For a proof of this assertion see e.g. Cowell et al (1999, page 56).

Example 3.2. For the MCS ordering in Fig. 3.7 we find $\pi_\lambda = (0,1,2,2,2,1,1)$ yielding the ladder nodes $\{3,4,5,6,7\}$ and the corresponding cliques

$$\mathscr{C} = \{\{1,2,3\},\{1,3,4\},\{3,4,5\},\{2,6\},\{6,7\}\}.$$

Junction tree

Let $\mathscr{A}$ be a collection of finite subsets of a set V. A *junction tree* $\mathscr{T}$ of sets in $\mathscr{A}$ is an undirected tree with $\mathscr{A}$ as a vertex set, satisfying the *junction tree property: If $A, B \in \mathscr{A}$ and C is on the unique path in $\mathscr{T}$ between A and B, then $A \cap B \subset C$.*

If the sets in $\mathscr{A}$ are pairwise incomparable, *they can be arranged in a junction tree if and only if $\mathscr{A} = \mathscr{C}$ where $\mathscr{C}$ are the cliques of a chordal graph.*

The junction tree can be constructed directly from the MCS ordering $C_\lambda, \lambda \in \Lambda$. More precisely, since

$$B_\lambda = \mathrm{bd}(\lambda) \cap \{1,\dots,\lambda-1\}$$

is complete for all $\lambda \in \Lambda$ it holds that

$$C_\lambda \cap (\cup_{\lambda' < \lambda} C_{\lambda'}) = C_\lambda \cap C_{\lambda^*} = S_\lambda$$

for some $\lambda^* < \lambda$. A junction tree is now easily constructed by attaching C_λ to any C_{λ^*} satisfying the above. Although λ^* may not be uniquely determined, S_λ is. Indeed, the sets S_λ are the minimal complete separators and *the numbers $\nu(S)$ are* $\nu(S) = |\{\lambda \in \Lambda : S_\lambda = S\}|$. Junction trees can be constructed in many other ways as well (Jensen and Jensen 1994). **FiXme Fatal: make figure to illustrate**

3.3 Probability Propagation and Junction Tree Algorithms

Junction trees of prime components

In general, the *prime components* of any undirected graph can be arranged in a junction tree in a similar way using an algorithm of Tarjan (1985), see also Leimer (1993).

Then *every pair of neighbours* (C,D) *in the junction tree represents a decomposition of* $\mathscr{G}$ into $\mathscr{G}_{\tilde{C}}$ and $\mathscr{G}_{\tilde{D}}$, where $\tilde{C}$ is the set of vertices in cliques connected to C but separated from D in the junction tree, and similarly with $\tilde{D}$.

Tarjan's algorithm is based on first numbering the vertices by a slightly more sophisticated algorithm (Rose et al 1976) known as *Lexicographic Search* (LEX) which runs in $O(|V|^2)$ time.

Markov properties of junction tree

The factorization property of an undirected graph can be seen as an 'outer' factorization over the junction tree into prime components, combined with 'inner' or 'local' factorizations on each prime component. More precisely, if we let $Q \in \mathscr{Q}$ be the prime components of a graph $\mathscr{G}$, arranged in a junction tree $\mathscr{T}$ and use that any graph decomposition also yields a decomposition of the Markov properties we have the following.

Proposition 3.3. *The distribution of* $X = (X_v, v \in V)$ *factorizes w.r.t.* $\mathscr{G}$ *if and only if the distribution of* $X_Q, Q \in \mathscr{Q}$ *factorizes w.r.t.* $\mathscr{T}$ *and each of* X_Q *factorizes w.r.t.* $\mathscr{G}_Q$.

If $\mathscr{G}$ *is decomposable,* $X = (X_v, v \in V)$ *factorizes w.r.t.* $\mathscr{G}$ *if and only if* $X_C, C \in \mathscr{C}$ *factorizes w.r.t.* $\mathscr{T}$.

3.4 Local Computation

Local computation algorithms similar to probability propagation have been developed independently in a number of areas with a variety of purposes. This includes, for example:

- Kalman filter and smoother (Thiele 1880; Kalman and Bucy 1961);
- Solving sparse linear equations (Parter 1961);
- Decoding digital signals (Viterbi 1967; Bahl et al 1974);
- Estimation in hidden Markov models (Baum 1972);
- Peeling in pedigrees (Elston and Stewart 1971; Cannings et al 1976);
- Belief function evaluation (Kong 1986; Shenoy and Shafer 1986);
- Probability propagation (Pearl 1986; Lauritzen and Spiegelhalter 1988; Jensen et al 1990);

- Optimizing decisions (Jensen et al 1994; Lauritzen and Nilsson 2001).

All algorithms are using, explicitly or implicitly, a *graph decomposition* and *a junction tree* or similar to make the computations.

An abstract perspective

Before we describe the local computation algorithms for probability propagation in detail, it is helpful to look at things from an abstract perspective.

We consider a *large* finite set V and a collection $\mathscr{C}$ of *small* subsets of V. Our elementary objects $\phi_C, C \in \mathscr{C}$ are *valuations* with *domain* C. These can be *combined* as

$$\phi_A \otimes \phi_B$$

to form more complex valuations with domain $A \cup B$. The combination operation $\otimes$ is assumed to *commutative* and *associative*:

$$\phi_A \otimes \phi_B = \phi_A \otimes \phi_B, \quad (\phi_A \otimes \phi_B) \otimes \phi_C = \phi_A \otimes (\phi_B \otimes \phi_C). \tag{3.3}$$

Valuations can be *marginalised:* For $A \subset V$, $\phi^{\downarrow A}$ denotes the *A-marginal* of ϕ. $\phi^{\downarrow A}$ has domain A. The marginalisation is assumed to satisfy *consonance:*

$$\phi^{\downarrow(A\cap B)} = \left(\phi^{\downarrow B}\right)^{\downarrow A} \tag{3.4}$$

and *distributivity:*

$$(\phi \otimes \phi_C)^{\downarrow B} = \left(\phi^{\downarrow B}\right) \otimes \phi_C \text{ if } C \subseteq B. \tag{3.5}$$

The conditions (3.3), (3.4) and (3.5) are known as the *Shenoy–Shafer axioms* after Shenoy and Shafer (1990) who first studied local computation in an abstract perspective. The specific algorithms described here only work when the semigroup of valuations is also *separative*, i.e. satisfies

$$\phi_A \otimes \phi_B = \phi_A \otimes \phi_A = \phi_B \otimes \phi_B \Rightarrow \phi_A = \phi_B,$$

which implies that division of valuations can be partially defined (Lauritzen and Jensen 1997).

Computational challenge

The computational challenge is to calculate marginals $\psi_A = \phi^{\downarrow A}$ of a *joint valuation*

$$\phi = \otimes_{C\in\mathscr{C}} \phi_C$$

with domain $V = \cup_{C\in\mathscr{C}} C$.

We are interested in cases where the direct computation of $\phi^{\downarrow A}$ is impossible if V is large. Hence we wish to calculate $\phi^{\downarrow A}$ using only *local* operations, i.e. operating on factors ψ_B with domain $B \subseteq C$ for some $C \in \mathscr{C}$, taking advantage of the fact that C are rather small.

Typically there also a *second purpose* of the calculation. Let us consider some examples.

Example 3.3 (Probability propagation). We consider a factorizing density on $\mathscr{X} = \times_{v \in V} \mathscr{X}_v$ with V and $\mathscr{X}_v$ finite:

$$p(x) = \prod_{C \in \mathscr{C}} \phi_C(x).$$

The *potentials* $\phi_C(x)$ depend on $x_C = (x_v, v \in C)$ only. The basic task to calculate a *marginal* (likelihood)

$$p(x^*E) = p^{\downarrow E}(x_E^*) = \sum_{y_{V \setminus E}} p(x_E^*, y_{V \setminus E})$$

for $E \subseteq V$ and fixed x_E^*, but the sum has too many terms. A second purpose is to calculate the *predictive probabilities* $p(x_v \,|\, x_E^*) = p(x_v, x_E^*)/p(x_E^*)$ for $v \in V$.

Example 3.4 (Sparse linear equations). Here valuations ϕ_C are *equation systems* involving variables with labels C. The combination operation $\phi_A \otimes \phi_B$ *concatenates* equation systems. The marginal $\phi_B^{\downarrow A}$ *eliminates* variables in $B \setminus A$, resulting in an equation system involving only variables in A. The marginal $\phi^{\downarrow A}$ of the joint valuation thus reduces the system of equations to a smaller one. A second computation finds a *solution* of the equation system.

Example 3.5 (Constraint satisfaction). Here the valuations ϕ_C represent *constraints* involving variables in C; the combination $\phi_A \otimes \phi_B$ *concatenates* the constraints; the marginal $\phi_B^{\downarrow A}$ finds *implied constraints*. The second computation identifies *jointly feasible configurations.* If represented by indicator functions, $\otimes$ is ordinary product and $\phi^{\downarrow E}(x_E^*) = \oplus_{y_{V \setminus E}} \phi(x_E^*, y_{V \setminus E})$, where $1 \oplus 1 = 1 \oplus 0 = 0 \oplus 1 = 1$ and $0 \oplus 0 = 0$.

Computational structure

Algorithms all implicitly or explicitly arrange the collection of sets $\mathscr{C}$ in a *junction tree* $\mathscr{T}$. Thus *the algorithms work if and only if $\mathscr{C}$ are cliques of chordal graph $\mathscr{G}$.*

If this is not so from the outset, a *triangulation* is used to construct a chordal graph $\mathscr{G}'$ with $E \subseteq E'$. This triangulation can be made in different ways with different computational complexities resulting. Typically, what must be controlled is *the maximal clique size,* i.e. the cardinality of the largest $C \in \mathscr{C}$. Optimizing this step is known to be NP-complete (Yannakakis 1981), but there are several heuristic algorithms which find good triangulations. In fact, there are algorithms, which in most cases run at reasonable computational speed and are guaranteed to return an

optimal triangulation. Such an algorithm has been implemented in version 6 of the commercially available software HUGIN (Andersen et al 1989). This algorithm is based on work of Shoiket and Geiger (1997), Berry et al (2000), and Bouchitté and Todinca (2001), and it is described in Jensen (2002).

Clearly, in a probabilistic perspective, *if P factorizes w.r.t. $\mathscr{G}$ it factorizes w.r.t. $\mathscr{G}'$*. Henceforth we assume such a triangulation has been made so we work with a chordal graph $\mathscr{G}$.

Setting up the structure

In many applications P is initially factorizing over a *directed acyclic graph* $\mathscr{D}$. The computational structure is then set up in several steps:

1. *Moralization:* Constructing $\mathscr{D}^m$, exploiting that if P factorizes on $\mathscr{D}$, it factorizes over $\mathscr{D}^m$.
2. *Triangulation:* Adding edges to find chordal graph $\mathscr{G}$ with $\mathscr{D}^m \subseteq \mathscr{G}$ as mentioned above;
3. *Constructing junction tree:*
4. *Initialization:* Assigning potential functions ϕ_C to cliques.

Basic computation

The basic computation now involves following steps

1. *Incorporating observations:* If $X_E = x_E^*$ is observed, we modify potentials as

$$\phi_C(x_C) \leftarrow \phi_C(x) \prod_{e \in E \cap C} \delta(x_e^*, x_e),$$

   with $\delta(u,v) = 1$ if $u = v$ and else $\delta(u,v) = 0$. Then:

$$p(x \,|\, X_E = x_E^*) = \frac{\prod_{C \in \mathscr{C}} \phi_C(x_C)}{p(x_E^*)}.$$

2. Marginals $p(x_E^*)$ and $p(x_C \,|\, x_E^*)$ are then calculated by a local *message passing* algorithm, to be described in further detail below.

Assigning potentials

Between any two cliques C and D which are neighbours in the junction tree their intersection $S = C \cap D$ is one of the minimal separators appearing in the decomposition sequence. We now explicitly represent these separators in the junction tree and also assign potentials to them, initially $\phi_S \equiv 1$ for all $S \in \mathscr{S}$, where $\mathscr{S}$ is the set of separators. We also let

$$\kappa(x) = \frac{\prod_{C\in\mathscr{C}} \phi_C(x_C)}{\prod_{S\in\mathscr{S}} \phi_S(x_S)}, \tag{3.6}$$

and now it holds that $p(x|x_E^*) = \kappa(x)/p(x_E^*)$. *The expression (3.6) will be invariant* under the message passing.

Marginalization

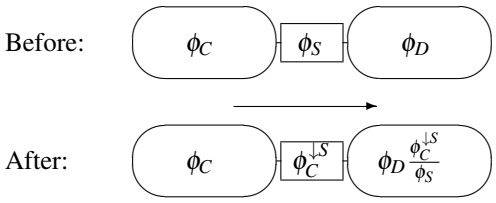
The *A-marginal* of a potential ϕ_B for $A \subseteq B$ is

$$\phi_B^{\downarrow A}(x) = \sum_{y_B : y_A = x_A} \phi_B(y)$$

If ϕ_B depends on x through x_B only and $B \subseteq V$ is 'small', marginal can be computed easily. The marginalisation clearly satisfies consonance (3.4) and distributivity (3.5).

Messages

When *C sends message* to *D*, the following happens:



Note that this computation is *local*, involving only variables within the pair of cliques. The expression in (3.6) is *invariant under the message passing* since $\phi_C\phi_D/\phi_S$ is:

$$\frac{\phi_C\,\phi_D\,\frac{\phi_C^{\downarrow S}}{\phi_S}}{\phi_C^{\downarrow S}} = \frac{\phi_C\phi_D}{\phi_S}.$$

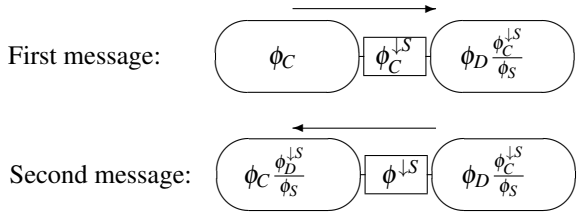
After the message has been sent, D contains the D-marginal of $\phi_C\phi_D/\phi_S$. To see this, we calculate

$$\left(\frac{\phi_C\phi_D}{\phi_S}\right)^{\downarrow D} = \frac{\phi_D}{\phi_S}\phi_C^{\downarrow D} = \frac{\phi_D}{\phi_S}\phi_C^{\downarrow S},$$

where we have used distributivity and consonance.

Second message

Before we proceed to discuss the case of a general junction tree, we shall investigate what happens when *D returns message* to *C*:



Now all sets contain the relevant marginal of $\phi = \phi_C\phi_D/\phi_S$*, including the separator.* This is seen as follows. The separator contains

$$\phi^{\downarrow S} = \left(\frac{\phi_C\phi_D}{\phi_S}\right)^{\downarrow S} = (\phi^{\downarrow D})^{\downarrow S} = \left(\phi_D\frac{\phi_C^{\downarrow S}}{\phi_S}\right)^{\downarrow S} = \frac{\phi_C^{\downarrow S}\phi_D^{\downarrow S}}{\phi_S}.$$

The clique *C* contains

$$\phi_C\frac{\phi^{\downarrow S}}{\phi_C^{\downarrow S}} = \frac{\phi_C}{\phi_S}\phi_D^{\downarrow S} = \phi^{\downarrow C}$$

since, as before

$$\left(\frac{\phi_C\phi_D}{\phi_S}\right)^{\downarrow C} = \frac{\phi_D}{\phi_S}\phi_C^{\downarrow D} = \frac{\phi_C}{\phi_S}\phi_D^{\downarrow S}.$$

Note that now *further messages between C and D are neutral.* Nothing will change if a message is repeated.

Message passing schedule

To describe the message passing algorithm fully we need to arrange for a scheduling of messages to be delivered. As we have seen above, it never harms to send a message, since the expression (3.6) is invariant under the operation. However, for computational efficiency it is desirable to send messages in such a way that redundant messages are avoided. The schedule to be described here is used in HUGIN and has two phases:

COLLINFO*:*

In this first phase, messages are sent from leaves towards arbitrarily chosen root *R*. It then holds that *after* COLLINFO*, the root potential satisfies* $\phi_R(x_R) = p(x_R, x_E^*)$.

DISTINFO*:*

In the second phase messages are sent from the root R towards the leaves of the junction tree. *After* COLLINFO *and subsequent* DISTINFO*, it holds that*

$$\phi_B(x_B) = p(x_B, x_E^*) \text{ for all } B \in \mathscr{C} \cup \mathscr{S}. \tag{3.7}$$

Hence $p(x_E^*) = \sum_{x_S} \phi_S(x_S)$ for any $S \in \mathscr{S}$ and $p(x_v \,|\, x_E^*)$ can readily be computed from any ϕ_S with $v \in S$.

Alternative scheduling of messages

Another efficient way of scheduling the messages is via *local control*. We then allow clique to send a message if and only if it has already received message from all other of its neighbours. Such messages are *live.* Using this protocol, there will be one clique who first receives messages from all its neighbours. This is effectively the root R in COLLINFO and DISTINFO. *Exactly two live messages along every branch are needed* to ensure that (3.7) holds.

Maximization

Another interesting task is to find the configuration with maximum probability, also known as the MAP. To solve this, we simply replace the standard sum-marginal with *max-marginal:*

$$\phi_B^{\downarrow A}(x) = \max_{y_B : y_A = x_A} \phi_B(y).$$

This marginalization also satisfies consonance and distributivity, and hence the same message passing schemes as above will apply. After COLLINFO and subsequent DISTINFO, the potentials satisfy

$$\phi_B(x_B) = \max_{y_{V \setminus (B \cup E)}} p(x_B, x_E^*, y_{V \setminus (B \cup E)}) = p(x_B, x_E^*, \hat{x}_{V \setminus (B \cup E)}) \text{ for all } B \in \mathscr{C} \cup \mathscr{S},$$

where $\hat{x}$ is the most probable configuration. Hence

$$\max p(y_{V \setminus E}, x_E^*) p(\hat{x}_{V \setminus E}, x_E^*) = \max_{x_S} \phi_S(x_S) \text{ for any } S \in \mathscr{S}$$

and the most probable configuration can now readily be identified (Cowell et al 1999, page 98). Viterbi's decoding algorithm for Hidden Markov Models (Viterbi 1967) is effectively a special instance of max-propagation.

It is also possible to find the k most probable configurations by a local computation algorithm (Nilsson 1998).

Since (3.6) remains invariant, *one can switch freely between max- and sum-propagation* without reloading original potentials.

Random propagation

Another variant of the message passing scheme picks a random configuration with distribution $p(x|x_E^*)$. Recall that after COLLINFO, the root potential is $\phi_R(x) \propto p(x_R|x_E)$. We then modify DISTINFO as follows:

1. Pick random configuration $\check{x}_R$ from ϕ_R;
2. Send message to neighbours C as $\check{x}_{R\cap C} = \check{x}_S$ where $S = C \cap R$ is the separator;
3. Continue by picking $\check{x}_C$ according to $\phi_C(x_{C\setminus S}, \check{x}_S)$ and send message further away from root.

When the sampling stops at the leaves of the junction tree, a configuration $\check{x}$ has been generated from $p(x|x_E^*)$.

There is an abundance of variants of the basic propagation algorithm; see Cowell et al (1999) for many of these.

3.5 Summary

Graph decompositions

A partitioning (A,B,S) of V forms a *decomposition* if S is complete and $A \perp_{\mathscr{G}} B|S$. A graph is *prime* if it has no proper decomposition exists. The *prime components* of a graph are the prime induced subgraphs and *any finite undirected graph can be recursively decomposed into its prime components.*

Chordal graphs

A graph is *chordal* if it has no induced cycles of length greater than three. The following are equivalent for any undirected graph $\mathscr{G}$.

(i) *$\mathscr{G}$ is chordal;*
(ii) *$\mathscr{G}$ is decomposable;*
(iii) *All prime components of $\mathscr{G}$ are cliques;*
(iv) *$\mathscr{G}$ admits a perfect numbering;*
(v) *Every minimal (α,β)-separator are complete.*

Trees are chordal graphs and thus decomposable. The prime components are the branches.

Maximum Cardinality Search (MCS) (Tarjan and Yannakakis 1984) *identifies whether a graph is chordal or not.* If a graph $\mathscr{G}$ is chordal, MCS *yields a perfect numbering* of the vertices. In addition it *finds the cliques* of $\mathscr{G}$:

Junction tree

A *junction tree* $\mathscr{T}$ of sets $\mathscr{A}$ is an undirected tree with $\mathscr{A}$ as a vertex set, satisfying the *junction tree property:*

If $A,B \in \mathscr{A}$ and C is on the unique path in $\mathscr{T}$ between A and B it holds that $A \cap B \subset C$.

If the sets in $\mathscr{A}$ are pairwise incomparable, *they can be arranged in a junction tree if and only if $\mathscr{A} = \mathscr{C}$ where $\mathscr{C}$ are the cliques of a chordal graph.*

The junction tree can be *constructed directly from the MCS ordering $C_\lambda, \lambda \in \Lambda$.*

Message passing

Initially the junction tree has potentials $\phi_C, c \in \mathscr{C} \cup \mathscr{S}$ so that the joint distribution of interest satisfies

$$p(x|x_E^*) \propto \frac{\prod_{C \in \mathscr{C}} \phi_C(x_C)}{\prod_{S \in \mathscr{S}} \phi_S(x_S)}.$$

The expression on the right-hand side is invariant under message passing. A message sent from a clique which has already received message from all other of its neighbours is *live. When exactly two live messages have been sent along every branch of the junction tree it holds that*

$$\phi_B(x_B) = p(x_B, x_E^*) \text{ for all } B \in \mathscr{C} \cup \mathscr{S},$$

from which most quantities of interest can be directly calculated.

Chapter 4
Specific Graphical Models

4.1 Log-linear Models

4.1.1 Interactions and factorization

Let $\mathscr{A}$ be a set of subsets of V. A density f or function is said to *factorize* w.r.t. $\mathscr{A}$ if there exist functions $\psi_a(x)$ which depend on x through x_a only and

$$f(x) = \prod_{a \in \mathscr{A}} \psi_a(x).$$

The set of distributions $\mathscr{P}_{\mathscr{A}}$ which factorize w.r.t. $\mathscr{A}$ is the *hierarchical log–linear model* generated by $\mathscr{A}$. The set $\mathscr{A}$ is the *generating class* of the log-linear model.

Typically the sets in $\mathscr{A}$ are taken to be pairwise incomparable under inclusion, so that no set in $\mathscr{A}$ is a subset of another set in $\mathscr{A}$. This need not necessarily be so but avoids redundancy in the representation.

The traditional notation used for contingency tables lets m_{ijk} denote the mean of the counts N_{ijk} in the cell (i, j, k) which is then expanded as e.g.

$$\log m_{ijk} = \alpha_i + \beta_j + \gamma_k \tag{4.1}$$

or

$$\log m_{ijk} = \alpha_{ij} + \beta_{jk} \tag{4.2}$$

or

$$\log m_{ijk} = \alpha_{ij} + \beta_{jk} + \gamma_{ik}, \tag{4.3}$$

or (with redundancy)

$$\log m_{ijk} = \gamma + \delta_i + \phi_j + \eta_k + \alpha_{ij} + \beta_{jk} + \gamma_{ik}, \tag{4.4}$$

To make the connection between this notation and the one used here, we assume that we have observations $X^1 = x^1, \ldots, X^n = x^n$ and $V = \{I, J, K\}$. We then write

$i = 1, \ldots, |I|$ for the possible values of X_I etc. and

$$N_{ijk} = |\{\nu : x^\nu = (i,j,k)\}|.$$

Then we have $m_{ijk} = nf(x)$ and if f is strictly positive and factorizes w.r.t. $\mathscr{A} = \{\{I,J\},\{J,K\}\}$, it holds that

$$\log f(x) = \log \psi_{IJ}(x_I, x_J) + \log \psi_{JK}(x_J, x_K).$$

Thus if we let

$$\alpha_{ij} = \log n + \log \psi_{IJ}(x_I, x_J), \quad \beta_{jk} = \log \psi_{JK}(x_J, x_K)$$

we have

$$\log m_{ijk} = \alpha_{ij} + \beta_{jk}.$$

The main difference is the assumption of positivity needed for the logarithm to be well defined. This is not necessary when using the multiplicative definition above. It is typically an advantage to relax the restriction of positivity although it also creates technical difficulties.

The logarithm of the factors $\phi_a = \log \psi_a$ are known as *interaction terms of order* $|a| - 1$ or $|a|$*-factor interactions*. Interaction terms of 0th order are called *main effects*. In the following we also refer to the factors themselves as interactions and main effects, rather than their logarithms.

4.1.2 Dependence graphs and factor graphs

Any joint probability distribution P of $X = (X_v, v \in V)$ has a *dependence graph* $G = G(P) = (V, E_P)$. This is defined by letting $\alpha \not\sim \beta$ in $G(P)$ exactly when

$$\alpha \perp\!\!\!\perp_P \beta \,|\, V \setminus \{\alpha, \beta\}.$$

X will then satisfy the pairwise Markov w.r.t. $G(P)$ and $G(P)$ is the smallest graph with this property, i.e. *P is pairwise Markov w.r.t. $\mathscr{G}$ iff*

$$G(P) \subseteq \mathscr{G}.$$

The *dependence graph $G(\mathscr{P})$ for a family $\mathscr{P}$* of probability measures is the smallest graph $\mathscr{G}$ so that all $P \in \mathscr{P}$ are pairwise Markov w.r.t. $\mathscr{G}$:

$$\alpha \perp\!\!\!\perp_P \beta \,|\, V \setminus \{\alpha, \beta\} \text{ for all } P \in \mathscr{P}.$$

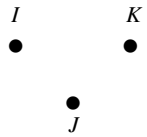
For any generating class $\mathscr{A}$ we construct the dependence graph $G(\mathscr{A}) = G(\mathscr{P}_\mathscr{A})$ of the log–linear model $\mathscr{P}_\mathscr{A}$. This is determined by the relation

$$\alpha \sim \beta \iff \exists a \in \mathscr{A} : \alpha, \beta \in a.$$

Sets in $\mathscr{A}$ are clearly complete in $G(\mathscr{A})$ and therefore *distributions in $\mathscr{P}_{\mathscr{A}}$ factorize according to $G(\mathscr{A})$*. They are thus also global, local, and pairwise Markov w.r.t. $G(\mathscr{A})$.

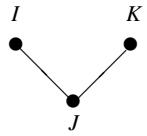
Some simple examples

Example 4.1 (Independence). The log–linear model specified by (4.1) is known as the *main effects model.* It has generating class $\mathscr{A} = \{\{I\},\{J\},\{K\}\}$ consisting of singletons only and dependence graph



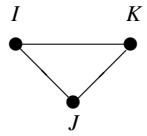
Thus it corresponds to *complete independence*.

Example 4.2 (Conditional independence). The log–linear model specified by (4.2) has no interaction between I and K. It has generating class $\mathscr{A} = \{\{I,J\},\{J,K\}\}$ and dependence graph



Thus it corresponds to the *conditional independence* $I \perp\!\!\!\perp K \,|\, J$.

Example 4.3 (No second-order interaction). The log–linear model specified by (4.3) has no second-order interaction. It has generating class $\mathscr{A} = \{\{I,J\},\{J,K\},\{I,K\}\}$ and its dependence graph



is the complete graph. Thus it has no conditional independence interpretation.

Conformal log-linear models

As a generating class defines a dependence graph $G(\mathscr{A})$, the reverse is also true. The set $\mathscr{C}(\mathscr{G})$ of cliques of $\mathscr{G}$ is a generating class for the log–linear model of distributions which factorize w.r.t. $\mathscr{G}$.

If the dependence graph completely summarizes the restrictions imposed by $\mathscr{A}$, i.e. if $\mathscr{A} = \mathscr{C}(G(\mathscr{A}))$, we say that $\mathscr{A}$ is *conformal*. The generating classes for the models given by (4.1) and (4.2) are conformal, whereas this is not the case for (4.3).

Factor graphs

The *factor graph* of $\mathscr{A}$ is the bipartite graph with vertices $V \cup \mathscr{A}$ and edges define by

$$\alpha \sim a \iff \alpha \in a.$$

Using this graph even non-conformal log–linear models admit a simple visual representation, as illustrated in Figure 4.1. which displays the factor graph of the non-conformal model in Example 4.3 with no second-order interaction.

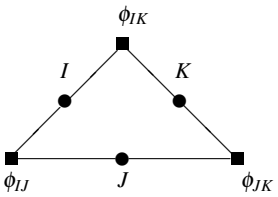


Fig. 4.1 The factor graph of the model in Example 4.3 with no second-order interaction.

If $\mathscr{F} = F(\mathscr{A})$ is the factor graph for $\mathscr{A}$ and $\mathscr{G} = G(\mathscr{A})$ the corresponding dependence graph, it is not difficult to see that for A, B, S being subsets of V

$$A \perp_{\mathscr{G}} B \,|\, S \iff A \perp_{\mathscr{F}} B \,|\, S$$

and hence *conditional independence properties can be read directly off the factor graph* also. In that sense, the factor graph is more informative than the dependence graph.

4.1.3 Data and likelihood function

Data in list form

Consider a sample $X^1 = x^1, \ldots, X^n = x^n$ from a distribution with probability mass function p. We refer to such data as being in *list form,* e.g. as

Case	Admitted?	Sex
1	Yes	Male
2	Yes	Female
3	No	Male
4	Yes	Male
⋮	⋮	⋮

Contingency Table

Data often presented in the form of a *contingency table* or *cross-classification*, obtained from the list by sorting according to category:

	Sex	
Admitted?	Male	Female
Yes	1198	557
No	1493	1278

This is a *two-way table* (or two-way classification) with categorical variables A: Admitted? and S: Sex. In this case it is a 2×2-*table*. The numerical entries are *cell counts*

$$n(x) = |\{\nu : x^\nu = x\}|$$

and the total number of observations is $n = \sum_{x \in \mathscr{X}} n_(x)$.

Likelihood function

Assume now $p \in \mathscr{P}_\mathscr{A}$ but otherwise unknown. The likelihood function can be expressed as

$$L(p) = \prod_{\nu=1}^{n} p(x^\nu) = \prod_{x \in \mathscr{X}} p(x)^{n(x)}.$$

In contingency table form the data follow a multinomial distribution

$$P\{N(x) = n(x), x \in \mathscr{X}\} = \frac{n!}{\prod_{x \in X} n(x)!} \prod_{x \in \mathscr{X}} p(x)^{n(x)}$$

but this only affects the likelihood function by a constant factor. The likelihood function is clearly continuous as a function of the ($|\mathscr{X}|$-dimensional vector) unknown probability distribution p. Since the *closure* $\overline{\mathscr{P}_\mathscr{A}}$ is compact (bounded and closed), *L attains its maximum on* $\overline{\mathscr{P}_\mathscr{A}}$ (not necessarily on $\mathscr{P}_\mathscr{A}$ itself).

Uniqueness of the MLE

Indeed, it is also true that *L has a unique maximum over* $\overline{\mathscr{P}_\mathscr{A}}$, essentially because the likelihood function is log-concave.The proof is indirect: Assume $p_1, p_2 \in \overline{\mathscr{P}_\mathscr{A}}$ with $p_1 \neq p_2$ and

$$L(p_1) = L(p_2) = \sup_{p \in \overline{\mathscr{P}_\mathscr{A}}} L(p). \tag{4.5}$$

Define

$$p_{12}(x) = c\sqrt{p_1(x)p_2(x)},$$

where $c^{-1} = \{\sum_x \sqrt{p_1(x)p_2(x)}\}$ is a normalizing constant. Then $p_{12} \in \overline{\mathscr{P}_\mathscr{A}}$ because

$$\begin{aligned} p_{12}(x) &= c\sqrt{p_1(x)p_2(x)} \\ &= \lim_{n\to\infty} c \prod_{a\in\mathscr{A}} \sqrt{\psi^1_{an}(x)\psi^2_{an}(x)} = \lim_{n\to\infty} \prod_{a\in\mathscr{A}} \psi^{12}_{an}(x), \end{aligned}$$

where e.g. $\psi^{12}_{an} = c^{1/|\mathscr{A}|}\sqrt{\psi^1_{an}(x)\psi^2_{an}(x)}$. The Cauchy–Schwarz inequality yields

$$c^{-1} = \sum_x \sqrt{p_1(x)p_2(x)} < \sqrt{\sum_x p_1(x)}\sqrt{\sum_x p_2(x)} = 1.$$

Hence

$$\begin{aligned} L(p_{12}) &= \prod_x p_{12}(x)^{n(x)} = \prod_x \Big\{ c\{\sqrt{p_1(x)p_2(x)}\Big\}^{n(x)} \\ &= c^n \prod_x \sqrt{p_1(x)}^{n(x)} \prod_x \sqrt{p_2(x)}^{n(x)} \\ &= c^n \sqrt{L(p_1)L(p_2)} > \sqrt{L(p_1)L(p_2)} = L(p_1) = L(p_2), \end{aligned}$$

which contradicts (4.5). Hence we conclude $p_1 = p_2$.

Likelihood equations

A simple application of the information inequality yields:

Proposition 4.1. *The maximum likelihood estimate $\hat{p}$ of p is the unique element of $\overline{\mathscr{P}_{\mathscr{A}}}$ which satisfies the system of equations*

$$n\hat{p}(x_a) = n(x_a), \forall a \in \mathscr{A}, x_a \in \mathscr{X}_a. \tag{4.6}$$

*Here $g(x_a) = \sum_{y:y_a=x_a} g(y)$ is the a-*marginal *of the function g.*

Proof. See Lauritzen (1996, Thm. 4.8) for a formal proof of this fact.

The system of equations (4.6) expresses the *fitting of the marginals* in $\mathscr{A}$. This is also an instance of the familiar result that in an exponential family (log-linear $\sim$ exponential), the MLE is found by equating the sufficient statistics (marginal counts) to their expectation.

Iterative proportional scaling

To show that the equations (4.6) indeed have a solution, we simply describe a convergent algorithm which solves it. This cycles (repeatedly) through all the a-marginals in $\mathscr{A}$ and fit them one by one. For $a \in \mathscr{A}$ define the following *scaling* operation on p:

$$(T_a p)(x) \leftarrow p(x)\frac{n(x_a)}{np(x_a)}, \quad x \in \mathscr{X}$$

where $0/0 = 0$ and $b/0$ is undefined if $b \neq 0$. Fitting the marginals The operation T_a *fits the a-marginal* if $p(x_a) > 0$ when $n(x_a) > 0$:

$$\begin{aligned} n(T_a p)(x_a) &= n \sum_{y:y_a=x_a} p(y) \frac{n(y_a)}{np(y_a)} \\ &= n \frac{n(x_a)}{np(x_a)} \sum_{y:y_a=x_a} p(y) \\ &= n \frac{n(x_a)}{np(x_a)} p(x_a) = n(x_a). \end{aligned}$$

Next, we make an ordering of the generators $\mathscr{A} = \{a_1, \ldots, a_k\}$. We define S by a full cycle of scalings

$$Sp = T_{a_k} \cdots T_{a_2} T_{a_1}$$

and consider the iteration

$$p_0(x) \leftarrow 1/|\mathscr{X}|, \quad p_n = Sp_{n-1}, n = 1, \ldots.$$

Proposition 4.2. *The iteration specified is convergent:*

$$\lim_{n\to\infty} p_n = \hat{p},$$

where $\hat{p}$ is the unique maximum likelihood estimate of $p \in \overline{\mathscr{P}_{\mathscr{A}}}$.

In other words, the limit $\hat{p}$ is the unique solution of the equation system (4.6).

Proof. The key elements in the proof of this result are:

1. If $p \in \overline{\mathscr{P}_{\mathscr{A}}}$, so is $T_a p$;
2. T_a is continuous at any point p of $\overline{\mathscr{P}_{\mathscr{A}}}$ with $p(x_a) \neq 0$ whenever $n(x_a) = 0$;
3. $L(T_a p) \geq L(p)$ with equality if and only if (4.6) is satisfied, so lthe ikelihood always increases at very step;
4. $\hat{p}$ is the unique fixpoint for T (and S);
5. $\overline{\mathscr{P}_{\mathscr{A}}}$ is compact.

We abstain from giving further details here.

The algorithm is known as *Iterative Proportional Scaling*, the *IPS*-algorithm, *Iterative Proportional Fitting* or the *IPF*-algorithm. It has numerous implementations, for example in *R* (inefficiently) in `loglin` with front end `loglm` in MASS (Venables and Ripley 2002).

Example 4.4. We illustrate the steps of the algorithm by a simple example:

	Admitted?		
Sex	Yes	No	*S*-marginal
Male	1198	1493	2691
Female	557	1278	1835
A-marginal	1755	2771	4526

These data are concerned with student admissions from Berkeley (Bickel et al 1973) and adapted by Edwards (2000). We consider the model with $A \perp\!\!\!\perp S$, corresponding to $\mathscr{A} = \{\{A\},\{S\}\}$. We should then fit the A-marginal and the S-marginal. For illustration we shall do so iteratively. The initial values are uniform:

	Admitted?		
Sex	Yes	No	S-marginal
Male	1131.5	1131.5	2691
Female	1131.5	1131.5	1835
A-marginal	1755	2771	4526

Initially all entries are equal to $4526/4$. Gives initial values of np_0. Next, we fit the S-marginal:

	Admitted?		
Sex	Yes	No	S-marginal
Male	1345.5	1345.5	2691 .
Female	917.5	917.5	1835
A-marginal	1755	2771	4526

We have calculated the entries as

$$1345.5 = 1131.5\frac{2691}{1131.5+1131.5}$$

and so on. Subsequently we fit the A-marginal:

	Admitted		
Sex	Yes	No	S-marginal
Male	1043.46	1647.54	2691 .
Female	711.54	1123.46	1835
A-marginal	1755	2771	4526

For example

$$711.54 = 917.5\frac{1755}{917.5+1345.5}$$

and so on. *The algorithm has now converged,* so there is no need to use more steps. If we wish, we can normalize to obtain probabilities. Dividing everything by 4526 yields $\hat{p}$.

	Admitted?		
Sex	Yes	No	S-marginal
Male	0.231	0.364	0.595
Female	0.157	0.248	0.405
A-marginal	0.388	0.612	1

In this example it is unnecessary to use the IPS algorithm as there is an explicit formula. We shall later elaborate on that issue.

IPS by probability propagation

The IPS-algorithm performs the scaling operations T_a:

$$p(x) \leftarrow p(x)\frac{n(x_a)}{np(x_a)}, \quad x \in \mathscr{X}. \tag{4.7}$$

This moves through all possible values of $x \in \mathscr{X}$, which in general can be *huge*, hence impossible.

Jiroušek and Přeučil (1995) realized that the algorithm could be implemented using probability propagation as follows: A chordal graph $\mathscr{G}$ with cliques $\mathscr{C}$ so that for all $a \in \mathscr{A}$, a are complete subsets of $\mathscr{G}$ is a *chordal cover* of $\mathscr{A}$. The steps of the efficient implementation are now:

1. Find chordal cover $\mathscr{G}$ of $\mathscr{A}$;
2. Arrange cliques $\mathscr{C}$ of $\mathscr{G}$ in a junction tree;
3. Represent p *implicitly* as

$$p(x) = \frac{\prod_{C\in\mathscr{C}} \psi_C(x)}{\prod_{S\in\mathscr{S}} \psi_S(x)};$$

4. Replace the step (4.7) with

$$\psi_C(x_C) \leftarrow \psi_C(x_C)\frac{n(x_a)}{np(x_a)}, \quad x_C \in \mathscr{X}_C,$$

where $a \subseteq C$ and $p(x_a)$ is calculated by *probability propagation*.

Since the scaling only involves $\mathscr{X}_C$, this is feasible if $\max_{C\in\mathscr{C}} |\mathscr{X}_C|$ is of a reasonable size.

Closed form maximum likelihood

In some cases the IPS algorithm converges after a finite number of cycles. An explicit formula is then available for the MLE of $p \in \mathscr{P}_{\mathscr{A}}$.

A generating class $\mathscr{A}$ is called *decomposable* if $\mathscr{A} = \mathscr{C}$ (i.e. $\mathscr{A}$ is conformal) and $\mathscr{C}$ are the cliques of a chordal graph $\mathscr{G}$. It can be shown that *the IPS-algorithm converges after a finite number of cycles (at most two) if and only if $\mathscr{A}$ is decomposable.*

Thus $\mathscr{A} = \{\{1,2\},\{2,3\},\{1,3\}\}$ is the smallest non-conformal generating class, demanding proper iteration. Since the IPS-algorithm converges in a finite number of steps, there must be an explicit expression for calculating the MLE in this case, to be given below.

Let $\mathscr{S}$ be the set of *minimal separators* of the chordal graph $\mathscr{G}$. The MLE for p under the log-linear model with generating class $\mathscr{A} = \mathscr{C}(\mathscr{G})$ is

$$\hat{p}(x) = \frac{\prod_{C\in\mathscr{C}} n(x_C)}{n\prod_{S\in\mathscr{S}} n(x_S)^{\nu(S)}} \tag{4.8}$$

where $\nu(S)$ is the number of times S appears as an intersection $a \cap b$ of neighbours in a junction tree $\mathscr{T}$ with $\mathscr{A}$ as vertex set.

A simple inductive argument shows that $\hat{p}$ given above indeed satisfies the likelihood equation (4.6) and hence this must be the MLE. Contrast this result with the factorization of the probability function itself:

$$p(x) = \frac{\prod_{C\in\mathscr{C}} p(x_C)}{\prod_{S\in\mathscr{S}} p(x_S)^{\nu(S)}}.$$

For the specific case where $\mathscr{G}$ is a tree, (4.8) reduces to

$$\hat{p}(x) = \frac{\prod_{e\in E} n(x_e)}{n\prod_{v\in V} n(x_v)^{\deg(v)-1}} = \frac{1}{n}\prod_{uv\in E}\frac{n(x_{uv})}{n(x_u)n(x_v)}\prod_{v\in V} n(x_v), \tag{4.9}$$

where we have used that the degree of a vertex exactly is equal to the number of times this vertex occurs as an endpoint of an edge.

4.2 Gaussian Graphical Models

4.2.1 The multivariate Gaussian distribution

Definition and density

A d-dimensional random vector $X = (X_1,\ldots,X_d)$ has a *multivariate Gaussian distribution* or *normal* distribution on $\mathscr{R}^d$ if there is a vector $\xi \in \mathscr{R}^d$ and a $d\times d$ matrix Σ such that

$$\lambda^\top X \sim \mathscr{N}(\lambda^\top\xi, \lambda^\top\Sigma\lambda) \quad \text{for all } \lambda \in R^d. \tag{4.10}$$

We then write $X \sim \mathscr{N}_d(\xi,\Sigma)$. Taking $\lambda = e_i$ or $\lambda = e_i + e_j$ where e_i is the unit vector with i-th coordinate 1 and the remaining equal to zero yields:

$$X_i \sim \mathcal{N}(\xi_i, \sigma_{ii}), \quad \mathrm{Cov}(X_i, X_j) = \sigma_{ij}.$$

The definition (4.10) makes sense if and only if $\lambda^\top \Sigma \lambda \geq 0$, i.e. if Σ is *positive semidefinite*.

If Σ is *positive definite*, i.e. if $\lambda^\top \Sigma \lambda > 0$ for $\lambda \neq 0$, the multivariate distribution has density w.r.t. Lebesgue measure on $\mathscr{R}^d$

$$f(x \,|\, \xi, \Sigma) = (2\pi)^{-d/2} (\det K)^{1/2} e^{-(x-\xi)^\top K (x-\xi)/2}, \tag{4.11}$$

where $K = \Sigma^{-1}$ is the *concentration matrix* of the distribution. We then also say that Σ is *regular*.

Marginal and conditional distributions

Partition X into X_1 and X_2, where $X_1 \in \mathscr{R}^r$ and $X_2 \in \mathscr{R}^s$ with $r+s=d$ and partition mean vector, concentration and covariance matrix accordingly as

$$\xi = \begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix}, \quad K = \begin{pmatrix} K_{11} & K_{12} \\ K_{21} & K_{22} \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$$

so that Σ_{11} is $r \times r$ and so on. *Then, if* $X \sim \mathcal{N}_d(\xi, \Sigma)$ it holds that

$$X_2 \sim \mathcal{N}_s(\xi_2, \Sigma_{22})$$

and

$$X_1 \,|\, X_2 = x_2 \sim \mathcal{N}_r(\xi_{1|2}, \Sigma_{1|2}),$$

where

$$\xi_{1|2} = \xi_1 + \Sigma_{12}\Sigma_{22}^{-}(x_2 - \xi_2) \quad \text{and} \quad \Sigma_{1|2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-}\Sigma_{21}.$$

Here Σ_{22}^{-} is an arbitrary *generalized inverse* to Σ_{22}, i.e. any symmetric matrix which satisfies

$$\Sigma_{22}^{-}\Sigma_{22} = \Sigma_{22}\Sigma_{22}^{-} = I.$$

In the regular case it also holds that

$$K_{11}^{-1} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} \tag{4.12}$$

and

$$K_{11}^{-1}K_{12} = -\Sigma_{12}\Sigma_{22}^{-1}, \tag{4.13}$$

so then,

$$\xi_{1|2} = \xi_1 - K_{11}^{-1}K_{12}(x_2 - \xi_2) \quad \text{and} \quad \Sigma_{1|2} = K_{11}^{-1}.$$

In particular, *if* $\Sigma_{12} = 0$, X_1 *and* X_2 *are independent.*

Factorization of the multivariate Gaussian

Consider a multivariate Gaussian random vector $X = \mathcal{N}_V(\xi, \Sigma)$ with Σ regular so it has density

$$f(x \mid \xi, \Sigma) = (2\pi)^{-|V|/2} (\det K)^{1/2} e^{-(x-\xi)^\top K (x-\xi)/2},$$

where $K = \Sigma^{-1}$ is the *concentration matrix* of the distribution. Thus *the Gaussian density factorizes w.r.t. $\mathcal{G}$ if and only if*

$$\alpha \not\sim \beta \Rightarrow k_{\alpha\beta} = 0$$

i.e. if the concentration matrix has zero entries for non-adjacent vertices.

Gaussian likelihood functions

Consider $\xi = 0$ and a sample $X^1 = x^1, \ldots, X^n = x^n$ $\mathcal{N}_d(0, \Sigma)$ with Σ regular. Using (4.11), we get the likelihood function

$$\begin{aligned} L(K) &= (2\pi)^{-nd/2} (\det K)^{n/2} e^{-\sum_{v=1}^n (x^v)^\top K x^v/2} \\ &\propto (\det K)^{n/2} e^{-\sum_{v=1}^n \operatorname{tr}\{K x^v (x^v)^\top\}/2} \\ &= (\det K)^{n/2} e^{-\operatorname{tr}\{K \sum_{v=1}^n x^v (x^v)^\top\}/2} \\ &= (\det K)^{n/2} e^{-\operatorname{tr}(KW)/2}. \end{aligned} \tag{4.14}$$

where

$$W = \sum_{v=1}^n x^v (x^v)^\top$$

is the matrix of *sums of squares and products*.

4.2.2 The Wishart distribution

The Wishart distribution is the sampling distribution of the matrix of sums of squares and products. More precisely, a random $d \times d$ matrix S has a *d-dimensional Wishart distribution* with parameter Σ and *n degrees of freedom* if

$$W \stackrel{\mathcal{D}}{=} \sum_{i=1}^n X^v (X^v)^\top$$

where $X^v \sim \mathcal{N}_d(0, \Sigma)$. We then write

$$W \sim \mathcal{W}_d(n, \Sigma).$$

The Wishart distribution is the multivariate analogue of the χ^2:

$$\mathscr{W}_1(n,\sigma^2) = \sigma^2\chi^2(n).$$

Basic properties of the Wishart distribution

If $W \sim \mathscr{W}_d(n,\Sigma)$ its mean is $\mathbb{E}(W) = n\Sigma$. If W_1 and W_2 are independent with $W_i \sim \mathscr{W}_d(n_i,\Sigma)$, then

$$W_1 + W_2 \sim \mathscr{W}_d(n_1+n_2,\Sigma).$$

If A is an $r \times d$ matrix and $W \sim \mathscr{W}_d(n,\Sigma)$, then

$$AWA^\top \sim \mathscr{W}_r(n, A\Sigma A^\top).$$

For $r = 1$ we get that when $W \sim \mathscr{W}_d(n,\Sigma)$ and $\lambda \in R^d$,

$$\lambda^\top W \lambda \sim \sigma_\lambda^2 \chi^2(n),$$

where $\sigma_\lambda^2 = \lambda^\top \Sigma \lambda$.

Wishart density

If $W \sim \mathscr{W}_d(n,\Sigma)$, where Σ is regular, then W *is regular with probability one if and only if* $n \geq d$. When $n \geq d$ the Wishart distribution has density

$$\begin{aligned} &f_d(w\,|\,n,\Sigma) \\ &= c(d,n)^{-1}(\det\Sigma)^{-n/2}(\det w)^{(n-d-1)/2}e^{-\operatorname{tr}(\Sigma^{-1}w)/2} \end{aligned}$$

w.r.t. Lebesgue measure on the set of positive definite matrices. The *Wishart constant* $c(d,n)$ is

$$c(d,n) = 2^{nd/2}(2\pi)^{d(d-1)/4}\prod_{i=1}^{d}\Gamma\{(n+1-i)/2\}.$$

4.2.3 Gaussian graphical models

Conditional independence in the multivariate Gaussian distribution

Consider $X = (X_1,\ldots,X_V) \sim \mathscr{N}_{|V|}(0,\Sigma)$ with Σ regular and $K = \Sigma^{-1}$. The concentration matrix of the conditional distribution of (X_α, X_β) given $X_{V\setminus\{\alpha,\beta\}}$ is

$$K_{\{\alpha,\beta\}} = \begin{pmatrix} k_{\alpha\alpha} & k_{\alpha\beta} \\ k_{\beta\alpha} & k_{\beta\beta} \end{pmatrix}.$$

Hence

$$\alpha \perp\!\!\!\perp \beta \,|\, V \setminus \{\alpha,\beta\} \iff k_{\alpha\beta} = 0.$$

Thus *the dependence graph $\mathscr{G}(K)$ of a regular Gaussian distribution is given by*

$$\alpha \not\sim \beta \iff k_{\alpha\beta} = 0.$$

Graphical models

$\mathscr{S}(\mathscr{G})$ denotes the symmetric matrices A with $a_{\alpha\beta} = 0$ unless $\alpha \sim \beta$ and $\mathscr{S}^+(\mathscr{G})$ their positive definite elements.

A *Gaussian graphical model* for X specifies X as multivariate normal with $K \in \mathscr{S}^+(\mathscr{G})$ and otherwise unknown. Note that the density then factorizes as

$$\log f(x) = \text{constant} - \frac{1}{2}\sum_{\alpha\in V} k_{\alpha\alpha}x_\alpha^2 - \sum_{\{\alpha,\beta\}\in E} k_{\alpha\beta}x_\alpha x_\beta,$$

hence *no interaction terms involve more than pairs.* This is different from the discrete case and generally makes things easier.

Likelihood function

The likelihood function based on a sample of size n is

$$L(K) \propto (\det K)^{n/2} e^{-\operatorname{tr}(KW)/2},$$

where W is the Wishart matrix of sums of squares and products, $W \sim \mathscr{W}_{|V|}(n,\Sigma)$ with $\Sigma^{-1} = K \in \mathscr{S}^+(\mathscr{G})$. For any matrix A we let $A(\mathscr{G}) = \{a(\mathscr{G})_{\alpha\beta}\}$ where

$$a(\mathscr{G})_{\alpha\beta} = \begin{cases} a_{\alpha\beta} \text{ if } \alpha = \beta \text{ or } \alpha \sim \beta \\ 0 \qquad \text{otherwise.} \end{cases}$$

Then, as $K \in \mathscr{S}(\mathscr{G})$ it holds for any A that

$$\operatorname{tr}(KA) = \operatorname{tr}\{KA(\mathscr{G})\}. \tag{4.15}$$

Using this fact for $A = W$ we can identify the family as a (regular and canonical) exponential family with elements of $W(\mathscr{G})$ as canonical sufficient statistics and the maximum likelihood estimate is therefore given as the unique solution to the system of likelihood equations

$$\mathbb{E}\{W(\mathscr{G})\} = n\Sigma(\mathscr{G}) = w(\mathscr{G})_{\text{obs}}.$$

Alternatively we can write the equations as

$$n\hat{\sigma}_{vv} = w_{vv}, \quad n\hat{\sigma}_{\alpha\beta} = w_{\alpha\beta}, \quad v \in V, \{\alpha,\beta\} \in E,$$

with the model restriction $\Sigma^{-1} \in \mathscr{S}^+(\mathscr{G})$. This 'fits variances and covariances along nodes and edges in $\mathscr{G}$' so we can write the equations as

$$n\hat{\Sigma}_{cc} = w_{cc} \text{ for all cliques } c \in \mathscr{C}(\mathscr{G}),$$

hence making the equations analogous to the discrete case. From (4.15) it follows that we for $\hat{K}$ have

$$\text{tr}\{\hat{K}W\} = \text{tr}\{\hat{K}W(\mathscr{G})\} = \text{tr}\{\hat{K}n\hat{\Sigma}(\mathscr{G})\} = n\,\text{tr}\{\hat{K}\hat{\Sigma}\} = nd$$

so that the maximized likelihood function becomes

$$L(\hat{K}) = (2\pi)^{-nd/2}(\det\hat{K})^{n/2}e^{-n/2} \propto (\det\hat{K})^{n/2}. \tag{4.16}$$

Iterative Proportional Scaling

For $K \in \mathscr{S}^+(\mathscr{G})$ and $c \in \mathscr{C}$, define the operation of 'adjusting the c-marginal' as follows. Let $a = V \setminus c$ and

$$T_cK = \begin{pmatrix} n(w_{cc})^{-1} + K_{ca}(K_{aa})^{-1}K_{ac} & K_{ca} \\ K_{ac} & K_{aa} \end{pmatrix}. \tag{4.17}$$

This operation is clearly well defined if w_{cc} is positive definite. Exploiting that it holds in general that

$$(K^{-1})_{cc} = \Sigma_{cc} = \left\{K_{cc} - K_{ca}(K_{aa})^{-1}K_{ac}\right\}^{-1},$$

we find the covariance $\tilde{\Sigma}_{cc}$ corresponding to the adjusted concentration matrix as

$$\begin{aligned}\tilde{\Sigma}_{cc} &= \{(T_cK)^{-1}\}_{cc} \\ &= \left\{n(w_{cc})^{-1} + K_{ca}(K_{aa})^{-1}K_{ac} - K_{ca}(K_{aa})^{-1}K_{ac}\right\}^{-1} \\ &= w_{cc}/n,\end{aligned}$$

hence T_cK *does indeed fit the marginals*.

From (4.17) it is seen that the pattern of zeros in K is preserved under the operation T_c, and it can also be seen to stay positive definite. In fact, T_c *scales proportionally* in the sense that

$$f\{x\,|\,(T_cK)^{-1}\} = f(x\,|\,K^{-1})\frac{f(x_c\,|\,w_{cc}/n)}{f(x_c\,|\,\Sigma_{cc})}.$$

This clearly demonstrates the analogy to the discrete case.

Next we choose any ordering $(c_1,\ldots,c_k)$ of the cliques in $\mathscr{G}$. Choose further $K_0 = I$ and define for $r = 0,1,\ldots$

$$K_{r+1} = (T_{c_1}\cdots T_{c_k})K_r.$$

The following now holds

Proposition 4.3. *Consider a sample from a covariance selection model with graph $\mathscr{G}$. Then*

$$\hat{K} = \lim_{r\to\infty} K_r,$$

provided the maximum likelihood estimate $\hat{K}$ of K exists.

Proof. This is Theorem 5.4 of Lauritzen (1996).

The problem of existence of the MLE is non-trivial:

(i) *If $n < \sup_{a\in\mathscr{A}} |a|$ the MLE does not exist.*
(ii) *if $n > \sup_{C\in\mathscr{C}} |C| - 1$, where $\mathscr{C}$ are the cliques of a chordal cover of $\mathscr{A}$ the MLE exists with probability one.*

The quantity $\tau(\mathscr{G})$ being the smallest possible value of the right hand-side of (ii) $\sup_{C\in\mathscr{C}(\mathscr{G}^*)} |C| - 1$, is known as the *tree-width* of the graph $\mathscr{G}$. Calculation of the tree-width is NP-complete, but for any fixed k it can be decided in linear time whether $\tau \le k$.

For n between these values the general situation is unclear. For the k-cycle it holds Buhl (1993) that for $n = 2$,

$$P\{\text{MLE exists}\,|\,\Sigma = I\} = 1 - \frac{2}{k-1!},$$

whereas for $n = 1$ the MLE does not exist and for $n \ge 3$ the MLE exists with probability one, as a k-cycle has tree-width 2.

Chordal graphs

If the graph $\mathscr{G}$ is chordal, we say that the graphical model is *decomposable*. We then have the familiar *factorization of densities*

$$f(x\,|\,\Sigma) = \frac{\prod_{C\in\mathscr{C}} f(x_C\,|\,\Sigma_C)}{\prod_{S\in\mathscr{S}} f(x_S\,|\,\Sigma_S)^{\nu(S)}} \tag{4.18}$$

where $\nu(S)$ is the number of times S appears as an intersection between neighbouring cliques of a junction tree for $\mathscr{C}$.

Relations for trace and determinant

Using the factorization (4.18) we can match the expressions for the trace and determinant to obtain that for a chordal graph $\mathscr{G}$ it holds that

$$\operatorname{tr}(KW) = \sum_{C\in\mathscr{C}} \operatorname{tr}(K_C W_C) - \sum_{S\in\mathscr{S}} \nu(S)\operatorname{tr}(K_S W_S)$$

and further

$$\begin{aligned}\det\Sigma = \{\det(K)\}^{-1} &= \frac{\prod_{C\in\mathscr{C}}\det\{(K^{-1})_C\}}{\prod_{S\in\mathscr{S}}[\det\{(K^{-1})_S\}]^{\nu(S)}} \\ &= \frac{\prod_{C\in\mathscr{C}}\det\{\Sigma_C\}}{\prod_{S\in\mathscr{S}}\{\det(\Sigma_S)\}^{\nu(S)}}.\end{aligned}$$

If we let $K = W = I$ in the first of these equations we obtain the identity

$$|V| = \sum_{C\in\mathscr{C}} |C| - \sum_{S\in\mathscr{S}} \nu(S)|S|,$$

which is also a special case of (3.2).

Maximum likelihood estimates

For a $|d|\times|e|$ matrix $A = \{a_{\gamma\mu}\}_{\gamma\in d,\mu\in e}$ we let $[A]^V$ denote the matrix obtained from A by filling up with zero entries to obtain full dimension $|V|\times|V|$, i.e.

$$\left([A]^V\right)_{\gamma\mu} = \begin{cases} a_{\gamma\mu} & \text{if } \gamma\in d, \mu\in e \\ 0 & \text{otherwise.}\end{cases}$$

For a chordal graph it holds that *the maximum likelihood estimates exists if and only if $n \geq C$ for all $C\in\mathscr{C}$*. As in the discrete case, then *the IPS-algorithm converges in a finite number of steps*.

The following simple formula then holds for the maximum likelihood estimate of K:

$$\hat{K} = n\left\{\sum_{C\in\mathscr{C}}\left[(w_C)^{-1}\right]^V - \sum_{S\in\mathscr{S}}\nu(S)\left[(w_S)^{-1}\right]^V\right\} \tag{4.19}$$

and the determinant of the MLE is

$$\det(\hat{K}) = \frac{\prod_{S\in\mathscr{S}}\{\det(w_S)\}^{\nu(S)}}{\prod_{C\in\mathscr{C}}\det(w_C)} n^d. \tag{4.20}$$

Note that setting $W = I$ in the first identity yields another variant of (3.2) to

$$1 = \sum_{C\in\mathscr{C}} \chi_C - \sum_{S\in\mathscr{S}} \nu(S)\chi_S, \tag{4.21}$$

where χ_A is the indicator function for the set A.

4.3 Summary

A brief summary of the contents of this chapter is given below.

Log–linear models

A density *f factorizes* w.r.t. a set $\mathscr{A}$ of subsets of V if

$$f(x) = \prod_{a \in \mathscr{A}} \psi_a(x).$$

The set of distributions $\mathscr{P}_{\mathscr{A}}$ which factorize w.r.t. a set of $\mathscr{A}$ is the *hierarchical log–linear model* with *generating class* $\mathscr{A}$.

Dependence graph

The *dependence graph* $\mathscr{G}(\mathscr{P})$ for a family of distributions $\mathscr{P}$ is the smallest graph $\mathscr{G}$ so that

$$\alpha \perp\!\!\!\perp_P \beta \,|\, V \setminus \{\alpha, \beta\} \text{ for all } P \in \mathscr{P}.$$

The *dependence graph of a log-linear model* $\mathscr{P}_{\mathscr{A}}$ is determined by

$$\alpha \sim \beta \iff \exists a \in \mathscr{A} : \alpha, \beta \in a.$$

Distributions in $\mathscr{P}_{\mathscr{A}}$ *factorize* according to $\mathscr{G}(\mathscr{A})$ and are all global, local, and pairwise Markov w.r.t. $\mathscr{G}(\mathscr{A})$.

Conformal log-linear model

The set $\mathscr{C}(\mathscr{G})$ of cliques of $\mathscr{G}$ is a generating class for the log–linear model of distributions which factorize w.r.t. $\mathscr{G}$. If the dependence graph completely summarizes the restrictions imposed by $\mathscr{A}$, i.e. if $\mathscr{A} = \mathscr{C}(\mathscr{G}(\mathscr{A}))$, $\mathscr{A}$ is *conformal*.

Likelihood equations

For any generating class $\mathscr{A}$ it holds that *the maximum likelihood estimate* $\hat{p}$ *of* p *is the unique element of* $\overline{\mathscr{P}_{\mathscr{A}}}$ *which satisfies the system of equations*

$$n\hat{p}(x_a) = n(x_a), \forall a \in \mathscr{A}, x_a \in \mathscr{X}_a.$$

The equations are solved by *Iterative Proportional Scaling*. For $a \in \mathscr{A}$ we let

$$(T_a p)(x) \leftarrow p(x)\frac{n(x_a)}{np(x_a)}, \quad x \in \mathscr{X}.$$

and define S by

$$Sp = T_{a_k} \cdots T_{a_2} T_{a_1} p.$$

Let $p_0(x) \leftarrow 1/|\mathscr{X}|, \quad p_n = Sp_{n-1}, n = 1, \ldots$. *It then holds that* $\lim_{n\to\infty} p_n = \hat{p}$ *where* $\hat{p}$ *is the unique maximum likelihood estimate of* $p \in \overline{\mathscr{P}_{\mathscr{A}}}$.

Closed form maximum likelihood

The generating class $\mathscr{A}$ is *decomposable* if $\mathscr{A} = \mathscr{C}$ where $\mathscr{C}$ are the cliques of a chordal graph.

The IPS-algorithm converges after at a finite number of cycles (at most two) if and only if $\mathscr{A}$ *is decomposable.*

The MLE for p under the decomposable log-linear model $\mathscr{A} = \mathscr{C}(\mathscr{G})$ *is*

$$\hat{p}(x) = \frac{\prod_{C\in\mathscr{C}} n(x_C)}{n\prod_{S\in\mathscr{S}} n(x_S)^{\nu(S)}},$$

where $\nu(S)$ is the usual multiplicity of a separator.

Gaussian graphical models

The likelihood function based on a sample of size n is

$$L(K) \propto (\det K)^{n/2} e^{-\operatorname{tr}(KW)/2},$$

where W is the Wishart matrix of sums of squares and products, $W \sim \mathscr{W}_{|V|}(n, \Sigma)$ with $\Sigma^{-1} = K \in \mathscr{S}^+(\mathscr{G})$, where $\mathscr{S}^+(\mathscr{G})$ are the positive definite matrices with $\alpha \not\sim \beta \Rightarrow k_{\alpha\beta} = 0$.

The MLE of $\hat{K}$ *is the unique element of* $\mathscr{S}^+(\mathscr{G})$ *satisfying*

$$n\hat{\Sigma}_{cc} = w_{cc} \text{ for all cliques } c \in \mathscr{C}(\mathscr{G}).$$

These equations are also solved by Iterative Proportional Scaling: For $K \in \mathscr{S}^+(\mathscr{G})$ and $c \in \mathscr{C}$, let

$$T_c K = \begin{pmatrix} n(w_{cc})^{-1} + K_{ca}(K_{aa})^{-1}K_{ac} & K_{ca} \\ K_{ac} & K_{aa} \end{pmatrix}.$$

Next choose an ordering $(c_1, \ldots, c_k)$ of the cliques in $\mathscr{G}$, let $K_0 = I$ and define for $r = 0, 1, \ldots$

$$K_{r+1} = (T_{c_1} \cdots T_{c_k}) K_r.$$

It then holds that $\hat{K} = \lim_{r\to\infty} K_r$, *provided the maximum likelihood estimate* $\hat{K}$ *of* K *exists.*

If the graph $\mathscr{G}$ is chordal, we say that the graphical model is *decomposable*. In this case, *the IPS-algorithm converges in at most two cycles*, as in the discrete case. The MLE for decomposable models is given as

$$\hat{K} = n\left\{\sum_{C\in\mathscr{C}}\left[(w_C)^{-1}\right]^V - \sum_{S\in\mathscr{S}} \nu(S)\left[(w_S)^{-1}\right]^V\right\}.$$

Chapter 5
Further Statistical Theory

5.1 Hyper Markov Laws

Special Wishart distributions

The formula for the maximum likelihood estimate (4.19) derived in the previous chapter specifies $\hat{\Sigma}$ as a random matrix. As we shall see, the sampling distribution of this random Wishart-type matrix is partly reflecting Markov properties of the graph $\mathscr{G}$. Before we delve further into this, we shall need some more terminology.

Laws and distributions

Families of distributions may not always be simply parameterized, or we may want to describe the families without specific reference to a parametrization. Generally we think of a family of the form

$$\mathscr{P} = \{P_\theta, \theta \in \Theta\}$$

and sometimes identify $\mathscr{P}$ with Θ which is justified when the parametrization

$$\theta \to P_\theta$$

is one-to-one and onto. For example, in a Gaussian graphical model $\theta = K \in \mathscr{S}^+(\mathscr{G})$ is uniquely identifying any regular Gaussian distribution satisfying the Markov properties w.r.t. $\mathscr{G}$.

Parametrization of a hierarchical log-linear model when $\mathscr{P} = \mathscr{P}_\mathscr{A}$ is more subtle, and specific choices must be made to ensure a one-to-one correspondence between the parameters; we omit the details here.

In any case, any probability measure on $\mathscr{P}$ (or on Θ) represents a random element of $\mathscr{P}$, i.e. a random distribution. The sampling distribution of a maximum

likelihood estimate such as $\hat{p}$ is an example of such a measure, as are Bayesian prior distributions on Θ (or $\mathscr{P}$).

In the following we generally refer to a probability measure on $\mathscr{P}$ as a *law*, whereas a *distribution* is used to signify a probability measure on $\mathscr{X}$. Thus we shall e.g. speak of the *Wishart law* as we emphasize that it specifies a distribution of $f(\cdot\,|\,\Sigma)$, by considering Σ to be random.

Hyper Markov laws

We identify $\theta \in \Theta$ with $P_\theta \in \mathscr{P}$, so e.g. θ_A for $A \subseteq V$ denotes the distribution of X_A under P_θ and $\theta_{A|B}$ the family of conditional distributions of X_A given X_B, etc. For a law $\mathscr{L}$ on Θ we write

$$A \perp\!\!\!\perp_{\mathscr{L}} B\,|\,S \iff \theta_{A\cup S} \perp\!\!\!\perp_{\mathscr{L}} \theta_{B\cup S}\,|\,\theta_S.$$

A law $\mathscr{L}$ on Θ is said to be *hyper Markov* w.r.t. $\mathscr{G}$ if

(i) All $\theta \in \Theta$ are globally Markov w.r.t. $\mathscr{G}$;
(ii) $A \perp\!\!\!\perp_{\mathscr{L}} B\,|\,S$ whenever S is complete and $A \perp_{\mathscr{G}} B\,|\,S$.

Note that the conditional independence in (ii) is *only required to hold for graph decompositions* as S is assumed to be complete. This implies in particular that a law can be hyper Markov on $\mathscr{G}$ without being hyper Markov on $\mathscr{G}^* = (V, E^*)$ even if this has more edges than $\mathscr{G}$, i.e. $E^* \supseteq E$. This is because $\mathscr{G}^*$ may have more complete subsets than $\mathscr{G}$ and hence more independence statements are required to be true. The complication is a consequence of the fact that we have deviated from Dawid and Lauritzen (1993) and defined hyper Markov laws for graphs that are not necessarily chordal; see Corollary 5.2 below.

If θ follows a hyper Markov law for the graph in Fig. 3.3 it holds for example that

$$\theta_{1235} \perp\!\!\!\perp \theta_{24567}\,|\,\theta_{25}.$$

We shall later show that *this is true for $\hat{\theta} = \hat{p}$ and also for $\hat{\Sigma}$ in the graphical model with this graph,* i.e. if $W \sim \mathscr{W}_7(f, \Sigma)$ with $\Sigma^{-1} = K \in \mathscr{S}^+(\mathscr{G})$ then it holds, for example, for the maximum likelihood estimate that

$$\hat{\Sigma} = \frac{1}{n}\left\{\sum_{C\in\mathscr{C}}\left[(w_C)^{-1}\right]^V - \sum_{S\in\mathscr{S}} \nu(S)\left[(w_S)^{-1}\right]^V\right\}^{-1}$$

that

$$\hat{\Sigma}_{1235} \perp\!\!\!\perp \hat{\Sigma}_{24567}\,|\,\hat{\Sigma}_{25}.$$

Consequences of the hyper Markov property

If $A \perp\!\!\!\perp_{\mathscr{L}} B \,|\, S$ we may further deduce that $\theta_A \perp\!\!\!\perp_{\mathscr{L}} \theta_B \,|\, \theta_S$, since θ_A and θ_B are functions of $\theta_{A\cup S}$ and $\theta_{B\cup S}$ respectively. But *the converse is false.* The relation $\theta_A \perp\!\!\!\perp_{\mathscr{L}} \theta_B \,|\, \theta_S$ does *not* imply $\theta_{A\cup S} \perp\!\!\!\perp_{\mathscr{L}} \theta_{B\cup S} \,|\, \theta_S$, since $\theta_{A\cup S}$ is *not* in general a function of (θ_A, θ_S). In contrast, $X_{A\cup B}$ is a (one-to-one) function of (X_A, X_B). However since $\theta_{A\,|\,S}$ and $\theta_{B\,|\,S}$ are functions of $(\theta_{A\cup S}, \theta_{B\cup S})$, *it generally holds that*

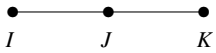
$$A \perp\!\!\!\perp_{\mathscr{L}} B \,|\, S \iff \theta_{A\,|\,S} \perp\!\!\!\perp_{\mathscr{L}} \theta_{B\,|\,S} \,|\, \theta_S. \tag{5.1}$$

Under some circumstances it is of interest to consider the notion of a *strong hyper Markov law*, demanding complete mutual independence of conditional and marginal distributions:

$$\theta_{A\,|\,S} \perp\!\!\!\perp_{\mathscr{L}} \theta_{B\,|\,S} \perp\!\!\!\perp_{\mathscr{L}} \theta_S$$

whenever S is complete and separates A from B. This is clearly stronger than (5.1). The notion is of particular importance for Bayesian analysis of graphical models with chordal graphs.

Example 5.1. This little example is a special case where we can directly demonstrate the hyper Markov property of the law of the maximum likelihood estimate. Consider the conditional independence model with graph



Here the MLE based on data $X^{(n)} = (X^1, \ldots, X^n)$ is

$$\hat{p}_{ijk} = \frac{N_{ij+}N_{+jk}}{nN_{+j+}}$$

and

$$\hat{p}_{ij+} = \frac{N_{ij+}}{n}, \quad \hat{p}_{+jk} = \frac{N_{+jk}}{n}, \quad \hat{p}_{+j+} = \frac{N_{+j+}}{n}.$$

Clearly, it holds that $\hat{p}$ is Markov on $\mathscr{G}$ and

$$\{N_{ij+}\} \perp\!\!\!\perp \{N_{+jk}\} \,|\, \{X_j^{(n)}\}.$$

But since e.g.

$$P(\{N_{ij+} = n_{ij}\} \,|\, \{X_j^{(n)}\}) = \prod_j \left(\frac{n_{+j+}!}{\prod_i n_{ij+}!} \prod_i p_{ij+}^{n_{ij+}} \right),$$

we have

$$\{N_{ij+}\} \perp\!\!\!\perp \{X_j^{(n)}\} \,|\, \{N_{+j+}\}$$

and hence

$$\{N_{ij+}\} \perp\!\!\!\perp \{N_{+jk}\} \,|\, \{N_{+j+}\},$$

which yields the hyper Markov property of $\hat{p}$. The law does not satisfy the strong hyper Markov property as the range of, say, $\{N_{ij+}\}$ is constrained by the value of $\{N_{+j+}\}$.

Chordal graphs

For chordal graphs the hyper Markov and ordinary Markov property are less different. For example, it is true for chordal graphs that the Markov property is preserved when (chordal) supergraphs are formed.

Proposition 5.1. *If $\mathscr{G} = (V,E)$ and $\mathscr{G}^* = (V,E^*)$ are both chordal graphs and $E \subseteq E^*$, then any hyper Markov law $\mathscr{L}$ over $\mathscr{G}$ is hyper Markov over $\mathscr{G}^*$.*

Proof. This result is Theorem 3.10 of Dawid and Lauritzen (1993) but we shall give a direct argument here. Firstly, as any Markov distribution over $\mathscr{G}$ is Markov over the supergraph $\mathscr{G}^*$, we only have to show the second condition for the law to be hyper Markov.

Lemma 3.2 implies that it is sufficient to consider the case where E and E^* differ by a single edge with endpoints $\{\alpha,\beta\}$ then contained in a single clique C^* of $\mathscr{G}^*$ according to Lemma 3.1. The clique C^* is the only complete separator in $\mathscr{G}^*$ which is not a complete separator in $\mathscr{G}$. So we have to show that for any hyper Markov law $\mathscr{L}$ on $\mathscr{G}$ it holds that

$$A \perp_{\mathscr{G}^*} B \,|\, C^* \Rightarrow \theta_{A|C^*} \perp\!\!\!\perp \theta_{B|C^*} \,|\, \theta_{C^*}. \tag{5.2}$$

We let $C = C^* \setminus \{\alpha,\beta\}$ and realize that we must have $\alpha \perp_{\mathscr{G}} \beta \,|\, C$ since $\mathscr{G}$ and $\mathscr{G}^*$ is chordal and any path in from α to β circumventing C would create a cycle in $\mathscr{G}$ or in $\mathscr{G}^*$. Let A_α be the vertices in A which are not separated from α by $\alpha \cup C$, $A_{\bar{\alpha}} = A \setminus A_\alpha$, and similarly with $B_\alpha, B_{\bar{\alpha}}$. The same argument implies the separations

$$\begin{aligned}
A_\alpha &\perp_{\mathscr{G}} (A_{\bar{\alpha}} \cup B \cup \beta) \,|\, \alpha \cup C,\\
A_{\bar{\alpha}} &\perp_{\mathscr{G}} (A_\alpha \cup B \cup \alpha) \,|\, \beta \cup C\\
B_\alpha &\perp_{\mathscr{G}} (B_{\bar{\alpha}} \cup A \cup \beta) \,|\, \alpha \cup C\\
B_{\bar{\alpha}} &\perp_{\mathscr{G}} (B_\alpha \cup A \cup \alpha) \,|\, \beta \cup C
\end{aligned}$$

In summary this means that the entire joint distribution θ can be represented as

$$\theta = \theta_C \, \theta_{\alpha|C} \, \theta_{\beta|C} \, \theta_{A_\alpha|\alpha\cup C} \, \theta_{B_\alpha|\alpha\cup C} \, \theta_{A_{\bar{\alpha}}|\beta\cup C} \, \theta_{B_{\bar{\alpha}}|\beta\cup C}$$

and also that its constituents satisfy the Markov property w.r.t. the graph in Fig. 5.1. Using this Markov property in combination with the fact that

$$\theta_{A|C^*} = \theta_{A_\alpha|\alpha\cup C} \, \theta_{A_{\bar{\alpha}}|\beta\cup C}, \; \theta_{B|C^*} = \theta_{B_\alpha|\alpha\cup C} \, \theta_{B_{\bar{\alpha}}|\beta\cup C}, \; \theta_{C^*} = \theta_{\alpha|C} \, \theta_{\beta|C} \, \theta_C,$$

yields (5.2) and the proof is complete.

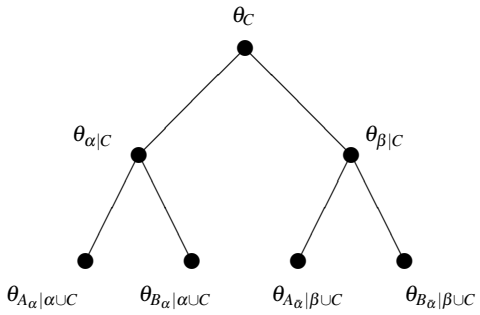


Fig. 5.1 The Markov structure of the joint law of the constituents of θ.

A consequence of this result is the following corollary, stating that for chordal graphs it is not necessary to demand that S is a complete separator to obtain the relevant conditional independence.

Proposition 5.2. *If $\mathcal{G}$ is chordal and θ is hyper Markov on $\mathcal{G}$, it holds that*

$$A \perp_{\mathcal{G}} B \,|\, S \Rightarrow A \perp\!\!\!\perp_{\mathcal{L}} B \,|\, S.$$

Proof. Again, this is Theorem 2.8 of Dawid and Lauritzen (1993). It follows by forming the graph $\mathcal{G}[S]$ connecting all pairs of vertices in S and connecting any other pair α, β if and only if $\neg(\alpha \perp_{\mathcal{G}} \beta \,|\, S)$. Then $\mathcal{G}[S]$ is a chordal graph with $\mathcal{G}[S] \geq \mathcal{G}$ so that $A \perp_{\mathcal{G}[S]} B \,|\, S$, and Proposition 5.1 applies.

If $\mathcal{G}$ is not chordal, we can form a chordal cover $\mathcal{G}^*$ by completing all prime components of $\mathcal{G}$. Then if θ is hyper Markov on $\mathcal{G}$, it is also hyper Markov on $\mathcal{G}^*$ and thus

$$A \perp_{\mathcal{G}^*} B \,|\, S \Rightarrow A \perp\!\!\!\perp_{\mathcal{L}} B \,|\, S.$$

But the similar result would be *false* for an arbitrary chordal cover of $\mathcal{G}$. The hyper Markov property thus has a simple formulation in terms of junction trees: Arrange the prime components $\mathcal{Q}$ of $\mathcal{G}$ in a junction tree $\mathcal{T}$ with complete separators $\mathcal{S}$ and consider the *extended junction tree* $\mathcal{T}^*$ which is the (bipartite) tree with $\mathcal{Q} \cup \mathcal{S}$ as vertices and edges from separators to prime components so that $C \sim S \sim D$ in $\mathcal{T}^*$ if and only if $C \sim D$ in $\mathcal{T}$. Next, associate θ_A to A for each $A \in \mathcal{Q} \cup \mathcal{S}$. It now holds that

$$A \perp_{\mathcal{T}^*} B \,|\, S \iff A \perp_{\mathcal{G}^*} B \,|\, S \iff \exists S^* \subseteq S : A \perp_{\mathcal{G}} B \,|\, S^* \text{ with } S^* \text{ complete},$$

implying that *$\mathcal{L}$ is hyper Markov on $\mathcal{G}$ if and only if $\{\theta_A, A \in \mathcal{Q} \cup \mathcal{S}\}$ is globally Markov w.r.t. the extended junction tree $\mathcal{T}^*$.*

Directed hyper Markov property

We have similar notions and results in the directed case. Say that $\mathscr{L} = \mathscr{L}(\theta)$ is *directed hyper Markov* w.r.t. a DAG $\mathscr{D}$ if θ is directed Markov on $\mathscr{D}$ for all $\theta \in \Theta$ and

$$\theta_{v \cup \mathrm{pa}(v)} \perp\!\!\!\perp_{\mathscr{L}} \theta_{\mathrm{nd}(v)} \,|\, \theta_{\mathrm{pa}(v)},$$

or equivalently $\theta_{v|\,\mathrm{pa}(v)} \perp\!\!\!\perp_{\mathscr{L}} \theta_{\mathrm{nd}(v)} \,|\, \theta_{\mathrm{pa}(v)}$, or equivalently for a well-ordering

$$\theta_{v \cup \mathrm{pa}(v)} \perp\!\!\!\perp_{\mathscr{L}} \theta_{\mathrm{pr}(v)} \,|\, \theta_{\mathrm{pa}(v)}.$$

It clearly holds that if v^* is a terminal vertex in V and $\mathscr{L}$ is directed hyper Markov over $\mathscr{D}$, then $\mathscr{L}_{V \setminus \{v^*\}}$ is directed hyper Markov over $\mathscr{D}_{V \setminus \{v^*\}}$. Repeated use of this fact yields that *if $\mathscr{L}$ is directed hyper Markov over $\mathscr{D}$ and A is an ancestral set, then $\mathscr{L}_A$ is directed hyper Markov over $\mathscr{D}_A$.*

Indeed, *if $\mathscr{D}$ is perfect, $\mathscr{L}$ is directed hyper Markov w.r.t. $\mathscr{D}$ if and only if $\mathscr{L}$ is hyper Markov w.r.t. $\mathscr{G} = \sigma(\mathscr{D}) = \mathscr{D}^m$.*

5.2 Meta Markov Models

Meta independence

Stochastic independence and conditional independence of parameters of marginal and conditional distributions can only occur when the associated parameters are variation independent. In the following we formalize such relationships among parameters of graphical models. We shall for $A, B \subseteq V$ identify

$$\theta_{A \cup B} = (\theta_{B|A}, \theta_A) = (\theta_{A|B}, \theta_B),$$

i.e. any joint distribution of $X_{A \cup B}$ is identified with a pair of further marginal and conditional distributions. Define for $S \subseteq V$ the S-section $\Theta^{\theta_S^*}$ of Θ as

$$\Theta^{\theta_S^*} = \{\theta \in \Theta : \theta_S = \theta_S^*, \theta \in \Theta\}.$$

The *meta independence relation* $\ddagger_{\mathscr{P}}$ is defined as

$$A \ddagger_{\mathscr{P}} B \,|\, S \iff \forall \theta_S^* \in \Theta_S : \Theta^{\theta_S^*} = \Theta_{A|S}^{\theta_S^*} \times \Theta_{B|S}^{\theta_S^*},$$

In words, A and B are *meta independent* w.r.t. $\mathscr{P}$ given S, if the pair of conditional distributions $(\theta_{A|S}, \theta_{B|S})$ vary in a product space when θ_S is fixed. Equivalently, fixing the values of $\theta_{B|S}$ and θ_S places the same restriction on $\theta_{A|S}$ as just fixing θ_S.

The relation $\ddagger_{\mathscr{P}}$ satisfies the semigraphoid axioms as it is a special instance of variation independence.

Meta Markov models

We say that a model determined by a family of distributions $\mathscr{P}$, or its parametrization Θ, is *meta Markov* w.r.t. $\mathscr{G}$ if

(i) All $\theta \in \Theta$ are globally Markov w.r.t. $\mathscr{G}$;
(ii) $A \perp_{\mathscr{G}} B \,|\, S \Rightarrow A \ddagger_{\mathscr{P}} B \,|\, S$ whenever S is complete.

Thus, *a Markov model is meta Markov if and only if*

$$A \perp_{\mathscr{G}^*} B \,|\, S \Rightarrow A \ddagger_{\mathscr{P}} B \,|\, S,$$

where $\mathscr{G}^*$ is obtained from $\mathscr{G}$ by completing all prime components. Note that if $\mathscr{G}$ is chordal, we have $\mathscr{G}^* = \mathscr{G}$ and hence *it holds for any meta Markov model $\mathscr{P}$ that*

$$A \perp_{\mathscr{G}} B \,|\, S \Rightarrow A \ddagger_{\mathscr{P}} B \,|\, S.$$

Hyper Markov laws and meta Markov models

Note that *for any triple (A,B,S) and any law $\mathscr{L}$ on Θ it holds that*

$$A \perp\!\!\!\perp_{\mathscr{L}} B \,|\, S \Rightarrow A \ddagger_{\mathscr{P}} B \,|\, S$$

for if $\theta_{A|S} \perp\!\!\!\perp_{\mathscr{L}} \theta_{B|S} \,|\, \theta_S$ it must in particular be true that $(\theta_{A|S}, \theta_{B|S})$ vary in a product space for every fixed value of θ_S. Thus hyper Markov laws live on meta Markov models: *If a law $\mathscr{L}$ on Θ is hyper Markov w.r.t. $\mathscr{G}$, Θ is meta Markov w.r.t. $\mathscr{G}$.*

In particular, *if a Markov model is not meta Markov, it cannot carry a hyper Markov law without further restricting to $\Theta_0 \subset \Theta$.*

A Gaussian graphical model with graph $\mathscr{G}$ is meta Markov on $\mathscr{G}$. This follows for example from results of collapsibility of Gaussian graphical models in Frydenberg (1990b), who show that in such a model, the conditional distribution $\theta_{V\setminus C|C}$ is variation independent of the marginal distribution θ_C if and only if the boundary of every connected component of $V \setminus C$ is complete, which trivially holds when C itself is complete. **FiXme Fatal: Hertil er jeg kommet**

Log-linear meta Markov models

Using results on collapsibility of log-linear models (Asmussen and Edwards 1983), it follows that *a log-linear model $\mathscr{P}_{\mathscr{A}}$ is meta Markov on its dependence graph $\mathscr{G}(\mathscr{A})$ if and only if for any minimal complete separator S of $\mathscr{G}(\mathscr{A})$ there there is an $a \in \mathscr{A}$ with $S \subseteq a$.* **FiXme Note: check this result, please** In particular, *if $\mathscr{A}$ is conformal, $\mathscr{P}_{\mathscr{A}}$ is meta Markov.* **FiXme Note: give argument**

Example 5.2. The log-linear model with generating class

$$\mathscr{A} = \{ab, ac, ad, bc, bd, be, cd, ce, de\}$$

has dependence graph with cliques $\mathscr{C} = \{abcd, bcde\}$, displayed in Fig. 5.2. Since the complete separator *bcd* is not in $\mathscr{A}$, this model is *not* meta Markov.

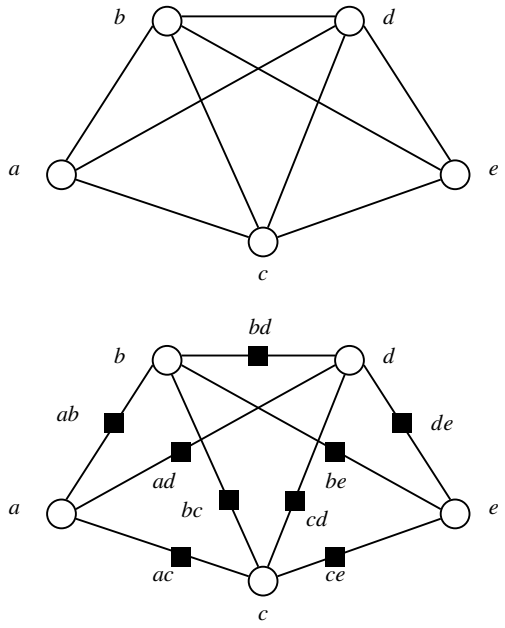


Fig. 5.2 Dependence and factor graph of the generating class $\mathscr{A}$ in Example 5.2.

Example 5.3. The model with generating class

$$\mathscr{A}' = \{ab, ac, ad, bcd, be, ce, de\}$$

has the same dependence graph $\mathscr{G}(\mathscr{A}') = \mathscr{G}(\mathscr{A})$ but even though $\mathscr{A}'$ is not conformal, $\mathscr{P}_{\mathscr{A}'}$ is meta Markov on $\mathscr{G}(\mathscr{A}')$.

Example 5.4. The model with generating class

$$\mathscr{A}'' = \{ab, ac, bc, bd, cd, ce, de\}$$

has a different dependence graph $\mathscr{G}(\mathscr{A}'')$, see Fig. 5.4. The separator *bcd* is not in $\mathscr{A}''$, but $\mathscr{P}_{\mathscr{A}''}$ is meta Markov on $\mathscr{G}(\mathscr{A}'')$, as both *minimal* separators *bc* and *cd* are in $\mathscr{A}''$.

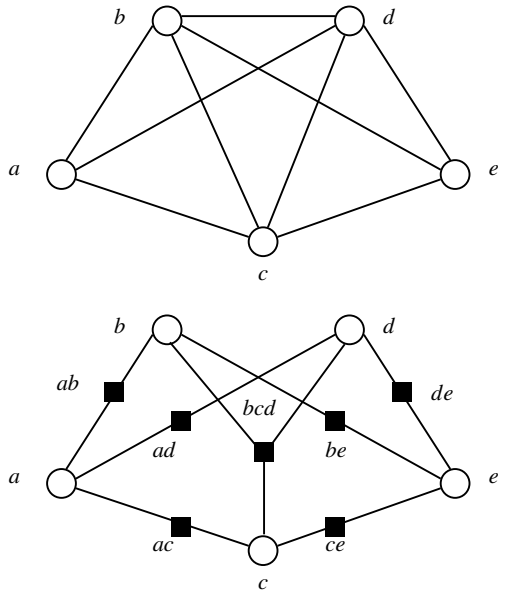


Fig. 5.3 Dependence and factor graph of the generating class $\mathscr{A}'$ in Example 5.3.

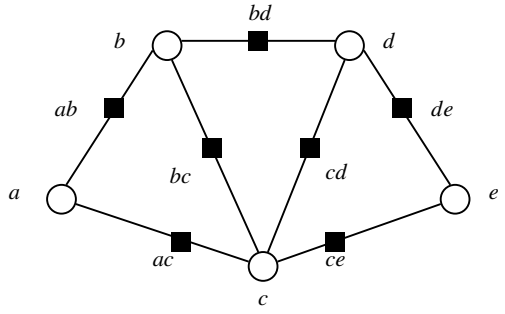


Fig. 5.4 Factor graph of the generating class $\mathscr{A}''$ in Example 5.4. The dependence graph looks identical to the factor graph when edge labels are removed.

Meta Markov properties on supergraphs

If θ is globally Markov w.r.t. the graph $\mathscr{G}$, it is also Markov w.r.t. any super graph $\mathscr{G}' = (V, E')$ with $E \subseteq E'$.

The similar fact is *not* true for meta Markov models. For example, the Gaussian graphical model for the 4-cycle $\mathscr{G}$ with adjacencies $1 \sim 2 \sim 3 \sim 4 \sim 1$, is meta Markov on $\mathscr{G}$, because it has no complete separators.

But the same model is *not* meta Markov w.r.t. the larger graph $\mathscr{G}'$ with cliques $\{124, 234\}$, since for any $K \in \mathscr{S}^+(\mathscr{G})$,

$$\sigma_{24} = \frac{\sigma_{12}\sigma_{14}}{\sigma_{11}} + \frac{\sigma_{13}\sigma_{34}}{\sigma_{33}}.$$

So fixing the value of σ_{24} restricts the remaining parameters in a complex way.

Maximum likelihood in meta Markov models

Under certain conditions, *the MLE $\hat{\theta}$ of the unknown distribution θ will follow a hyper Markov law over Θ under P_θ*. These are

(i) Θ is meta Markov w.r.t. $\mathscr{G}$;
(ii) For any prime component Q of $\mathscr{G}$, the MLE $\hat{\theta}_Q$ for θ_Q based on $X_Q^{(n)}$ is *sufficient* for Θ_Q and *boundedly complete*.

A sufficient condition for (ii) is that *Θ_Q is a full and regular exponential family in the sense of Barndorff-Nielsen (1978).* In particular, *these conditions are satisfied for any Gaussian graphical model and any meta Markov log-linear model.*

Canonical construction of hyper Markov laws

The distributions of maximum likelihood estimators are important examples of hyper Markov laws. But for *chordal graphs* there is a canonical construction of such laws.

Let $\mathscr{C}$ be the cliques of a chordal graph $\mathscr{G}$ and let $\mathscr{L}_C, C \in \mathscr{C}$ be a family of laws over $\Theta_C \subseteq \mathbb{P}(\mathscr{X}_C)$. The family of laws are *hyperconsistent* if for any C and D with $C \cap D = S \neq \emptyset$, $\mathscr{L}_C$ and $\mathscr{L}_D$ induce the same law for θ_S.

If $\mathscr{L}_C, C \in \mathscr{C}$ are hyperconsistent, there is a unique hyper Markov law $\mathscr{L}$ over $\mathscr{G}$ with $\mathscr{L}(\theta_C) = \mathscr{L}_C, C \in \mathscr{C}$.

Strong hyper and meta Markov properties

In some cases it is of interest to consider a stronger version of the hyper and meta Markov properties.

A meta Markov model is *strongly meta Markov* if $\theta_{A|S} \ddagger_{\mathscr{P}} \theta_S$ for all complete separators S.

Similarly, a hyper Markov model is *strongly hyper Markov* if $\theta_{A|S} \perp\!\!\!\perp_{\mathscr{L}} \theta_S$ for all complete separators S.

A directed hyper Markov model is *strongly directed hyper Markov* if $\theta_{v|\,\mathrm{pa}(v)} \perp\!\!\!\perp_{\mathscr{L}} \theta_{\mathrm{pa}(v)}$ for all $v \in V$.

Gaussian graphical models and log-linear meta Markov models are strong meta Markov models.

5.2.1 Bayesian inference

Parameter $\theta \in \Theta$, data $X = x$, likelihood

$$L(\theta \,|\, x) \propto p(x \,|\, \theta) = \frac{dP_\theta(x)}{d\mu(x)}.$$

Express knowledge about θ through a *prior* π on θ. Use also π to denote density of prior w.r.t. some measure ν on Θ.

Inference about θ from x is then represented through *posterior distribution* $\pi^*(\theta) = p(\theta \,|\, x)$. Then, from Bayes' formula

$$\pi^*(\theta) = p(x \,|\, \theta)\pi(\theta)/p(x) \propto L(\theta \,|\, x)\pi(\theta)$$

so the *likelihood function is equal to the density of the posterior w.r.t. the prior* modulo a constant.

Example 5.5 (Bernoulli experiments). Data $X_1 = x_1, \ldots, X_n = x_n$ independent and Bernoulli distributed with parameter θ, i.e.

$$P(X_i = 1 \,|\, \theta) = 1 - P(X_i = 0) = \theta.$$

Represent as a directed acyclic graph with θ as only parent to all nodes $x_i, i = 1, \ldots, n$. Use a beta prior:

$$\pi(\theta \,|\, a, b) \propto \theta^{a-1}(1-\theta)^{b-1}.$$

If we let $x = \sum x_i$, we get the posterior:

$$\begin{aligned} \pi^*(\theta) &\propto \theta^x(1-\theta)^{n-x}\theta^{a-1}(1-\theta)^{b-1} \\ &= \theta^{x+a-1}(1-\theta)^{n-x+b-1}. \end{aligned}$$

So the posterior is also beta with parameters $(a+x, b+n-x)$.

Closure under sampling

A family $\mathscr{P}$ of laws on Θ is said to be *closed under sampling* from x if

$$\pi \in \mathscr{P} \Rightarrow \pi^* \in \mathscr{P}.$$

The family of beta laws is closed under Bernoulli sampling. If the family of priors is parametrised:

$$\mathscr{P} = \{P_\alpha, \alpha \in \mathscr{A}\}$$

we sometimes say that α is a *hyperparameter.* Then, Bayesian inference can be made by just updating hyperparameters. The terminology of hyperparameter breaks down in more complex models, corresponding to large directed graphs, where all

parent variables can be seen as 'parameters' for their children. Thus the division into three levels, with data, parameters, and hyperparameters is not helpful.

For a k-dimensional exponential family

$$p(x \mid \theta) = b(x)e^{\theta^\top t(x) - \psi(\theta)}$$

the *standard conjugate family* (Diaconis and Ylvisaker 1979) is

$$\pi(\theta \mid a, \kappa) \propto e^{\theta^\top a - \kappa\psi(\theta)}$$

for $(a, \kappa) \in \mathscr{A} \subseteq \mathscr{R}^k \times \mathscr{R}_+$, where $\mathscr{A}$ is determined so that the normalisation constant is finite. Posterior updating from $(x_1, \ldots, x_n)$ with $t = \sum_i t(x_i)$ is then made as $(a^*, \kappa^*) = (a + t, \kappa + n)$.

Closure under sampling of hyper Markov properties

The hyper Markov property is in wide generality closed under sampling. For $\mathscr{L}$ being a prior law over Θ and $X = x$ is an observation from θ, let $\mathscr{L}^* = \mathscr{L}(\theta \mid X = x)$ denote the *posterior law* over Θ. It then holds that *If $\mathscr{L}$ is hyper Markov w.r.t. $\mathscr{G}$ so is $\mathscr{L}^*$.*

And further, *if $\mathscr{L}$ is strongly hyper Markov w.r.t. $\mathscr{G}$ so is $\mathscr{L}^*$,* so also the strong hyper Markov property is preserved.

In the latter case, *the update of $\mathscr{L}$ is even local to prime components*, i.e.

$$\mathscr{L}^*(\theta_Q) = \mathscr{L}_Q^*(\theta_Q) = \mathscr{L}_Q(\theta_Q \mid X_Q = x_Q)$$

and *the marginal distribution p of X is globally Markov w.r.t. $\mathscr{G}^*$,* where

$$p(x) = \int_\Theta P(X = x \mid \theta)\mathscr{L}(d\theta).$$

FiXme Fatal: write more about the strong hyper Markov property, either here or earlier

5.2.2 Hyper inverse Wishart and Dirichlet laws

Gaussian graphical models are canonical exponential families. The standard family of conjugate priors have densities

$$\pi(K \mid \Phi, \delta) \propto (\det K)^{\delta/2} e^{-\operatorname{tr}(K\Phi)}, K \in \mathscr{S}^+(\mathscr{G}).$$

These laws are termed *hyper inverse Wishart laws* as Σ follows an inverse Wishart law for complete graphs.

For chordal graphs, each marginal law $\mathscr{L}_C$ of Σ_C is inverse Wishart.

For any meta Markov model where Θ and Θ_Q are full and regular exponential families for all prime components Q, it follows directly from Barndorff-Nielsen (1978), page 149, that *the standard conjugate prior law is strongly hyper Markov w.r.t.* $\mathscr{G}$.

This is in particular true for the hyper inverse Wishart laws.

The analogous prior distribution for log-linear meta Markov models are likewise termed *hyper Dirichlet laws*.

They are also strongly hyper Markov and if $\mathscr{G}$ is *chordal, each induced marginal law* $\mathscr{L}_C$ *is a standard Dirichlet law*.

Conjugate prior laws are strong hyper Markov

If Θ *is meta Markov and* Θ_Q *are full and regular exponential families for all prime components* Q*, the standard conjugate prior law is strongly hyper Markov w.r.t.* $\mathscr{G}$.

This is in particular true for the hyper inverse Wishart laws and the hyper Dirichlet laws.

Thus, for the hyper inverse and hyper Dirichlet laws we have simple *local updating* based on *conjugate priors* for Bayesian inference.

5.3 Summary

Laws and distributions

A statistical model involves a family $\mathscr{P}$ of distributions, often parametrized as

$$\mathscr{P} = \{P_\theta, \theta \in \Theta\}.$$

We refer to a probability measure on $\mathscr{P}$ or Θ as a *law*, whereas a *distribution* is a probability measure on $\mathscr{X}$.

Hyper Markov Laws

A law $\mathscr{L}$ on Θ is *hyper Markov* w.r.t. $\mathscr{G}$ if

(i) All $\theta \in \Theta$ are globally Markov w.r.t. $\mathscr{G}$;
(ii) $A \perp\!\!\!\perp_{\mathscr{L}} B \,|\, S$ whenever S is *complete* and $A \perp_{\mathscr{G}} B \,|\, S$.

Directed hyper Markov property

$\mathscr{L} = \mathscr{L}(\theta)$ is *directed hyper Markov* w.r.t. a DAG $\mathscr{D}$ if θ is directed Markov on $\mathscr{D}$ for all $\theta \in \Theta$ and

$$\theta_{v\cup \mathrm{pa}(v)} \perp\!\!\!\perp_{\mathscr{L}} \theta_{\mathrm{nd}(v)} \,|\, \theta_{\mathrm{pa}(v)},$$

or equivalently

$$\theta_{v|\,\mathrm{pa}(v)} \perp\!\!\!\perp_{\mathscr{L}} \theta_{\mathrm{nd}(v)} \,|\, \theta_{\mathrm{pa}(v)}.$$

If $\mathscr{D}$ is perfect, $\mathscr{L}$ is directed hyper Markov w.r.t. $\mathscr{D}$ if and only if $\mathscr{L}$ is hyper Markov w.r.t. $\mathscr{G} = \sigma(\mathscr{D}) = \mathscr{D}^m$.

Meta Markov models

For $A, B \subseteq V$ identify

$$\theta_{A\cup B} = (\theta_{B|A}, \theta_A) = (\theta_{A|B}, \theta_B).$$

A and B are *meta independent* w.r.t. $\mathscr{P}$ given S, denoted $A \ddagger_{\mathscr{P}} B \,|\, S$, if the pair of conditional distributions $(\theta_{A|S}, \theta_{B|S})$ vary in a product space when θ_S is fixed.

The family $\mathscr{P}$, or Θ, is *meta Markov* w.r.t. $\mathscr{G}$ if

(i) All $\theta \in \Theta$ are globally Markov w.r.t. $\mathscr{G}$;
(ii) $A \perp_{\mathscr{G}} B \,|\, S \Rightarrow A \ddagger_{\mathscr{P}} B \,|\, S$ whenever S is complete.

Hyper Markov laws and meta Markov models

Hyper Markov laws live on meta Markov models.

A Gaussian graphical model with graph $\mathscr{G}$ is meta Markov on $\mathscr{G}$.

A log-linear model $\mathscr{P}_{\mathscr{A}}$ is meta Markov on its dependence graph $\mathscr{G}(\mathscr{A})$ if and only if $S \in \mathscr{A}$ for any minimal complete separator S of $\mathscr{G}(\mathscr{A})$.

In particular, *if $\mathscr{A}$ is conformal, $\mathscr{P}_{\mathscr{A}}$ is meta Markov.*

Maximum likelihood in meta Markov models

If the following conditions are satisfied:

(i) Θ is meta Markov w.r.t. $\mathscr{G}$;
(ii) For any prime component Q of $\mathscr{G}$, Θ_Q is a full and regular exponential family,

the MLE $\hat{\theta}$ of the unknown distribution θ will follow a hyper Markov law over Θ under P_θ.

In particular, *this holds for any Gaussian graphical model and any meta Markov log-linear model.*

Strong hyper and meta Markov properties

Similarly, a hyper Markov law is *strongly hyper Markov* if $\theta_{A|S} \perp\!\!\!\perp_{\mathscr{L}} \theta_S$ for all complete separators S.

A directed hyper Markov lawis *strongly directed hyper Markov* if $\theta_{v|\,\mathrm{pa}(v)} \perp\!\!\!\perp_{\mathscr{L}} \theta_{\mathrm{pa}(v)}$ for all $v \in V$.

A meta Markov model is *strongly meta Markov* if $\theta_{A|S} \ddagger_{\mathscr{P}} \theta_S$ for all complete separators S.

Gaussian graphical models and log-linear meta Markov models are strong meta Markov models.

Closure under sampling of hyper Markov properties

If $\mathscr{L}$ is a prior law over Θ and $X = x$ is an observation from θ, $\mathscr{L}^* = \mathscr{L}(\theta \,|\, X = x)$ denotes the *posterior law* over Θ.

If $\mathscr{L}$ is hyper Markov w.r.t. $\mathscr{G}$ so is $\mathscr{L}^*$.

If $\mathscr{L}$ is strongly hyper Markov w.r.t. $\mathscr{G}$ so is $\mathscr{L}^*$.

In the latter case, *the update of $\mathscr{L}$ is local to prime components*, i.e.

$$\mathscr{L}^*(\theta_Q) = \mathscr{L}_Q^*(\theta_Q) = \mathscr{L}_Q(\theta_Q \,|\, X_Q = x_Q)$$

and *the marginal distribution p of X is globally Markov w.r.t. $\mathscr{G}^*$,* where

$$p(x) = \int_\Theta P(X = x \,|\, \theta) \mathscr{L}(d\theta).$$

Hyper inverse Wishart and Dirichlet laws

Gaussian graphical models are canonical exponential families. The standard family of conjugate priors have densities

$$\pi(K \,|\, \Phi, \delta) \propto (\det K)^{\delta/2} e^{-\mathrm{tr}(K\Phi)}, K \in \mathscr{S}^+(\mathscr{G}).$$

These laws are termed *hyper inverse Wishart laws* as Σ follows an inverse Wishart law for complete graphs. *For chordal graphs, each marginal law $\mathscr{L}_C, \mathscr{C}$ of Σ_C is inverse Wishart.*

The standard conjugate prior law for log-linear meta Markov models are termed *hyper Dirichlet laws*. If $\mathscr{G}$ is *chordal, each induced marginal law $\mathscr{L}_C, C \in \mathscr{C}$ is a standard Dirichlet law*.

Chapter 6
Estimation of Structure

6.1 Estimation of Structure and Bayes Factors

Previous chapters have considered the situation where the graph $\mathscr{G}$ defining the model has been known and the inference problems were concerned with an unknown P_θ with $\theta \in \Theta$. This chapter discusses inference concerning the graph $\mathscr{G}$, specifying only a family Γ of possible graphs.

It is important to ensure that any methods used must scale well with data size we typically need to consider many structures and also huge collections of high-dimensional data.

What we here choose to term *structure estimation* is also known under other names as *model selection* (mainstream statistics), *system identification* (engineering), or *structural learning* (AI or machine learning.) Different situations occur depending on the type of assumptions concerning Γ Common assumptions include that Γ is the set of *undirected graphs* over V; the set of *chordal graphs* over V; the set of *forests* over V; the set of *trees* over V; the set of *directed acyclic graphs* over V; or potentially other types of conditional independence structure.

Why estimation of structure?

It may be worthwhile to dwell somewhat on the rationale behind structure estimation. We think of it as a method to get a quick overview of relations between a huge set of variables in a complex stochastic system and see it in many ways as a parallel to e.g. histograms or density estimation which gives a rough overview of the features of univariate data. It will typically be used in areas such as, for example, general data mining, identification of gene regulatory networks, or for reconstructing family trees from DNA information. Established methods exist and are in daily routine use, but there is a clear need for better understanding of their statistical properties.

We begin by showing a few simple examples of structure estimation to motivate that the issue is not a priori impossible.

Example 6.1 (Markov mesh model). Figure 6.1 shows the graph of a so-called Markov mesh model with 36 variables. All variables are binary and the only variable

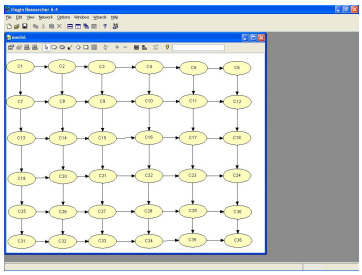


Fig. 6.1 Graph of a Markov mesh model with 36 binary variables.

without parents, in the upper left-hand corner is uniformly distributed. The remaining variables on the upper and left sides of the 6×6 square have a single parent and the conditional probability that it is in a given state is $3/4$ if the state is the same as its parent. The remaining nodes have two parents and if these are identical, the child with have that state with probability $3/4$ whereas it will otherwise follow the upper parent with probability 2/3.

Figure 6.2 shows two different attempts of estimating the structure based on the same 10,000 simulated cases. The two methods are to be described in more detail later, but it is apparent that the estimated structure in both cases have a strong similarity to the true one. In fact, one of the methods reconstructs the Markov mesh model perfectly. Both methods used search for a DAG structure which is compatible with the data.

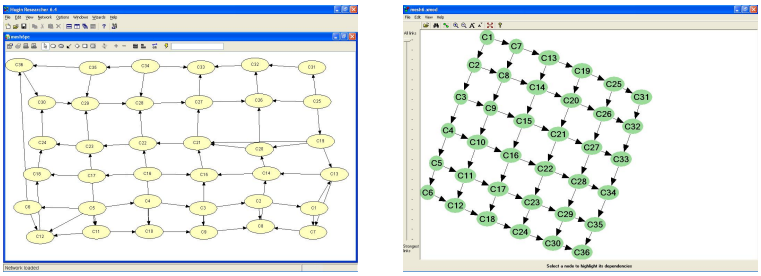


Fig. 6.2 Structure estimate of Markov mesh model from 10000 simulated cases. The left-hand side shows the estimate using the crudest algorithm (PC) implemented in HUGIN. The right-hand side the Bayesian estimate using greedy equivalence search (GES) as implemented in WINMINE.

Example 6.2 (Tree model). The graph of this example has a particular simple structure which is that of a rooted tree. Since a rooted tree with arrows pointing away

from a root is a perfect DAG, the associated structure is equivalent to the corresponding undirected tree. The state at the root is uniformly distributed and any other node reproduces the state of the parent node with probability $3/4$.

Figure 6.3 shows the structure estimate of the tree based on 10,000 simulated cases and using the same methods as for the Markov mesh model. In both cases, the method has attempted to estimate the structure based on the assumption that the structure was a DAG. Note that in this case it is the first method which reconstruncts correctly whereas there are too many links in the second case.

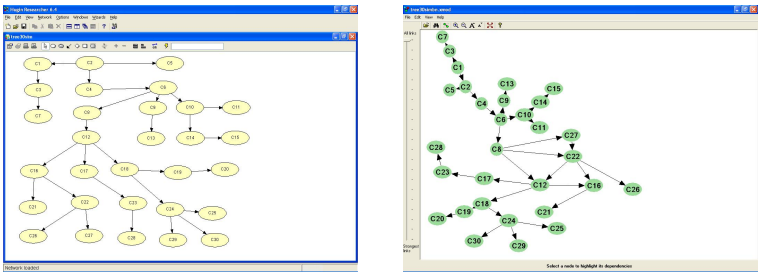


Fig. 6.3 Estimates of a tree model with 30 variables based on 10000 observations. The graph to the left represents the estimate using the PC algorithm and yields a 100% correct reconstruction. The graph to the right represents the Bayesian estimate using GES.

Example 6.3 (Chest clinic). The next example is taken from Lauritzen and Spiegelhalter (1988) and reflects the structure involving risk factors and symptoms for lung-disease. The (fictitious) description given by the authors of the associated medical knowledge is as follows

> "Shortness–of–breath (dyspnoea) may be due to tuberculosis, lung cancer or bronchitis, or none of them, or more than one of them. A recent visit to Asia increases the chances of tuberculosis, while smoking is known to be a risk factor for both lung cancer and bronchitis. The results of a single chest X–ray do not discriminate between lung cancer and tuberculosis, as neither does the presence or absence of dyspnoea."

The actual probabilities involved in this example are given in the original reference and we abstain from repeating them here.

Figure 6.4 displays the network structure reflecting the knowledge as given above and three different structure estimates. Note that this problem is obviously more difficult than the previous examples, in particular because some of the diseases are rare and larger data sets as well as more refined structure estimators are needed to even get close to the original structure.

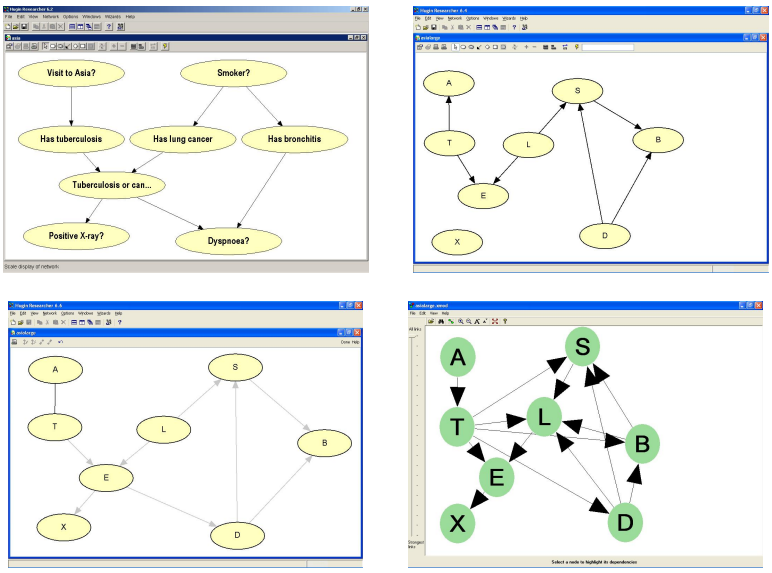


Fig. 6.4 A Bayesian network model for lung disease and estimates of the model based on simulated cases. The structure generating the data is in the upper left corner. Then, clockwise, estimates using the same data but different estimation algorithms: the PC algorithm, Bayesian GES, the NPC algorithm. In the latter case 100,000 cases were used.

Types of approach

Essentially all structure estimation methods combine a specification of potentially interesting structures with a way of judging the *adequacy of structure* and a *search strategy,* which evaluates a large number of space of possible structures.

As detailed further in the following sections, methods of judging adequacy include using

- tests of significance;
- penalised likelihood scores;

$$I_\kappa(\mathscr{G}) = \log \hat{L} - \kappa \dim(\mathscr{G})$$

  with $\kappa = 1$ for AIC Akaike (1974), or $\kappa = \frac{1}{2}\log N$ for BIC Schwarz (1978);
- Bayesian posterior probabilities.

The search strategies are more or less based on heuristics, which all attempt to overcome the fundamental problem that a crude global search among all potential structures is not feasible as the number of structures is astronomical.

FiXme Fatal: elaborate on each of these or rearrange

Bayes factors

For $\mathscr{G} \in \Gamma$, $\Theta_{\mathscr{G}}$ is associated parameter space so that P factorizes w.r.t. $\mathscr{G}$ if and only if $P = P_\theta$ for some $\theta \in \Theta_{\mathscr{G}}$. $\mathscr{L}_{\mathscr{G}}$ is prior law on $\Theta_{\mathscr{G}}$.

The *Bayes factor* (likelihood ratio) for discriminating between $\mathscr{G}_1$ and $\mathscr{G}_2$ based on observations $X^{(n)} = x^{(n)}$ is

$$\mathrm{BF}(\mathscr{G}_1 : \mathscr{G}_2) = \frac{f(x^{(n)} \,|\, \mathscr{G}_1)}{f(x^{(n)} \,|\, \mathscr{G}_2)},$$

where

$$f(x^{(n)} \,|\, \mathscr{G}) = \int_{\Theta_{\mathscr{G}}} f(x^{(n)} \,|\, \mathscr{G}, \theta)\, \mathscr{L}_{\mathscr{G}}(d\theta)$$

is known as the *marginal likelihood* of $\mathscr{G}$.

Posterior distribution over graphs

If $\pi(\mathscr{G})$ is a prior probability distribution over a given set of graphs Γ, the posterior distribution is determined as

$$\pi^*(\mathscr{G}) = \pi(\mathscr{G} \,|\, x^{(n)}) \propto f(x^{(n)} \,|\, \mathscr{G})\pi(\mathscr{G})$$

or equivalently

$$\frac{\pi^*(\mathscr{G}_1)}{\pi^*(\mathscr{G}_2)} = \mathrm{BF}(\mathscr{G}_1 : \mathscr{G}_2)\frac{\pi(\mathscr{G}_1)}{\pi(\mathscr{G}_2)}.$$

Bayesian analysis looks for the *MAP estimate* $\mathscr{G}^*$ maximizing $\pi^*(\mathscr{G})$ over Γ, or attempts to *sample from the posterior* using e.g. Monte-Carlo methods.

6.2 Estimating Trees and Forests

Estimating trees

Let us assume that the distribution P of $X = X_v, v \in V$ over a discrete state space $\mathscr{X}$ factorizes w.r.t. an unknown *tree* τ and that we have observations $X^1 = x_1, \ldots, X^n = x^n$ as independent and identically distributed according to P.

Chow and Liu (1968) showed that *the maximum likelihood estimate* $\hat{\tau}$ *of* τ *is a maximal weight spanning tree* (MWST), where the *weight* of a tree τ is

$$\lambda(\tau) = \sum_{e \in E(\tau)} \lambda_n(e) = \sum_{e \in E(\tau)} H_n(e)$$

and $H_n(e)$ is the empirical *cross-entropy* or *mutual information* between endpoint variables of the edge $e = \{u, v\}$:

$$H_n(e) = \sum_{x_u x_v} \frac{n(x_u,x_v)}{n} \log \frac{n(x_u,x_v)/n}{n(x_u)n(x_v)/n^2} = \sum_{x_u,x_v} n(x_u,x_v) \log \frac{n(x_u,x_v)}{n(x_u)n(x_v)}.$$

This result is easily *extended to Gaussian graphical models,* just with the weight $\lambda_n(e)$ of an edge in a tree determined as any strictly increasing function of the empirical cross-entropy along the edge

$$H_n(e) = -\frac{1}{2}\log(1-r_e^2),$$

where r_e^2 is *empirical correlation coeffient* along edge $e = \{u,v\}$

$$r_e^2 = \frac{(\sum_{i=1}^n x_u^i x_v^i)^2}{(\sum_{i=1}^n (x_u^i)^2)(\sum_{i=1}^n (x_v^i)^2)} = \frac{w_{uv}^2}{w_{uu}w_{vv}}.$$

To see this, use the expression (4.20) for the determinant of the MLE which in the case of a tree reduces to

$$\begin{aligned}\det(\hat{K}) &= \frac{\prod_{v\in V}(w_{vv})^{\deg(v)-1}}{\prod_{e\in E}\det(w_e)} n^d \\ &\propto \prod_{v\in V}(w_{vv})^{-1} \prod_{\{u,v\}\in E} \frac{w_{uu}w_{vv}}{w_{uu}w_{vv}-w_{uv}^2} \propto (1-r_e^2)^{-1}.\end{aligned}$$

From (4.16) we know that the maximized likelihood function for a fixed tree is proportional to a power of this determinant and hence is maximized when the logarithm of the determinant is maximized. But since we then have

$$\log\det\hat{K}(\tau) = 2\sum_{e\in E(\tau)} H_n(e) = 2\lambda(\tau),$$

maximizing $\hat{L}(\tau)$ over all possible trees is equivalent to maximizing $\lambda(\tau)$.

Highest AIC or BIC scoring forest *also available as MWSF,* with modified weights

$$w_n^{\text{pen}}(e) = nw_n(e) - \kappa_n \text{df}_e,$$

with $\kappa_n = 2$ for AIC, $\kappa_n = \log n$ for BIC and df_e the *degrees of freedom for independence* along e.

Fast algorithms Kruskal Jr. (1956) compute maximal weight spanning tree (or forest) from weights $W = (w_{uv}, u,v \in V)$.

Chow and Wagner (1978) show *a.s. consistency in total variation of $\hat{P}$:* If P factorises w.r.t. τ, then

$$\sup_x |p(x) - \hat{p}(x)| \to 0 \text{ for } n \to \infty,$$

so *if τ is unique for P, $\hat{\tau} = \tau$ for all $n > N$ for some N.*

If P does not factorize w.r.t. a tree, *$\hat{P}$ converges to closest tree-approximation $\tilde{P}$ to P* (Kullback-Leibler distance).

Strong hyper Markov prior laws

For strong hyper Markov prior laws, $X^{(n)}$ is itself marginally Markov so

$$f(x^{(n)} \,|\, \mathscr{G}) = \frac{\prod_{Q \in \mathscr{Q}} f(x_Q^{(n)} \,|\, \mathscr{G})}{\prod_{S \in \mathscr{S}} f(x_S^{(n)} \,|\, \mathscr{G})^{\nu_{\mathscr{G}}(S)}}, \tag{6.1}$$

where $\mathscr{Q}$ are the prime components and $\mathscr{S}$ the minimal complete separators of $\mathscr{G}$.

Hyper inverse Wishart laws

Denote the normalisation constant of the hyper inverse Wishart density as

$$h(\delta, \Phi; \mathscr{G}) = \int_{\mathscr{S}^+(\mathscr{G})} (\det K)^{\delta/2} e^{-\operatorname{tr}(K\Phi)} \, dK,$$

i.e. the usual Wishart constant if $Q = C$ is a clique.

Combining with the Gaussian likelihood, it is easily seen that for Gaussian graphical models we have

$$f(x^{(n)} \,|\, \mathscr{G}) = \frac{h(\delta + n, \Phi + W^n; \mathscr{G})}{h(\delta, \Phi; \mathscr{G})}.$$

Comparing with (6.1) leads to a similar factorization of the normalising constant

$$h(\delta, \Phi; \mathscr{G}) = \frac{\prod_{Q \in \mathscr{Q}} h(\delta, \Phi_Q; \mathscr{G}_Q)}{\prod_{S \in \mathscr{S}} h(\delta, \Phi_S; S)^{\nu_{\mathscr{G}}(S)}}.$$

For *chordal graphs* all terms in this expression reduce to known Wishart constants, and we can thus calculate the normalization constant explicitly.

In general, Monte-Carlo simulation or similar methods must be used Atay-Kayis and Massam (2005).

The marginal distribution of $W^{(n)}$ is (weak) *hyper Markov* w.r.t. $\mathscr{G}$. It was termed the *hyper matrix F law* by Dawid and Lauritzen (1993).

Bayes factors for forests

Trees and forests are decomposable graphs, so for a forest ϕ we get

$$f(\phi \,|\, x^{(n)}) \propto \frac{\prod_{e \in E(\phi)} f(x_e^{(n)})}{\prod_{v \in V} f(x_v^{(n)})^{d_\phi(v)-1}},$$

since all minimal complete separators are singletons and $\nu_\phi(\{v\}) = d_\phi(v) - 1$.

Multiplying the right-hand side with $\prod_{v \in V} f(x_v^{(n)})$ yields

$$\frac{\prod_{e\in E(\phi)} f(x_e^{(n)})}{\prod_{v\in V} f(x_v^{(n)})^{d_\phi(v)-1}} = \prod_{v\in V} f(x_v^{(n)}) \prod_{e\in\phi} \mathrm{BF}(e),$$

where $\mathrm{BF}(e)$ is the *Bayes factor* for independence along the edge e:

$$\mathrm{BF}(e) = \frac{f(x_u^{(n)}, x_v^{(n)})}{f(x_u^{(n)}) f(x_v^{(n)})}.$$

Thus the *posterior distribution* of ϕ is

$$\pi^*(\phi) \propto \prod_{e\in E(\phi)} \mathrm{BF}(e).$$

In the case where ϕ is restricted to contain a *single tree*, the normalization constant for this distribution can be explicitly obtained via the *Matrix Tree Theorem,* see e.g. Bollobás (1998).

Bayesian analysis

MAP estimates of forests can thus be computed using an MWSF algorithm, using $w(e) = \log BF(e)$ as weights.

Algorithms exist for generating random spanning trees Aldous (1990), so *full posterior analysis is in principle possible for trees*.

These work less well for weights occurring with typical Bayes factors, as most of these are essentially zero, so methods based on the *Matrix Tree Theorem* seem currently more useful.

Only heuristics available for MAP estimators or maximizing penalized likelihoods such as AIC or BIC, for other than trees.

Some challenges for undirected graphs

- Find *feasible algorithm for (perfect) simulation* from a distribution over chordal graphs as
$$p(\mathscr{G}) \propto \frac{\prod_{C\in\mathscr{C}} w(C)}{\prod_{S\in\mathscr{S}} w(S)^{\nu_{\mathscr{G}}(S)}},$$
where $w(A), A \subseteq V$ are a prescribed set of positive weights.
- Find *feasible algorithm for obtaining MAP* in decomposable case. This may not be universally possible as problem most likely is NP-complete.

6.3 Learning Bayesian networks

6.3.1 Model search methods

Directed hyper Markov property

$\mathscr{L} = \mathscr{L}(\theta)$ is *directed hyper Markov* w.r.t. a DAG $\mathscr{D}$ if θ is directed Markov on $\mathscr{D}$ for all $\theta \in \Theta$ and

$$\theta_{v|\,\mathrm{pa}(v)} \perp\!\!\!\perp_{\mathscr{L}} \theta_{\mathrm{nd}(v)} \,|\, \theta_{\mathrm{pa}(v)}.$$

A law $\mathscr{L}$ is directed hyper Markov on $\mathscr{D}$ if and only if $\mathscr{L}_A$ is hyper Markov on $(\mathscr{D}_A)^m$ for any ancestral set $A \subseteq V$.

$\mathscr{L}$ is *strongly directed hyper Markov* if in addition $\theta_{v|\,\mathrm{pa}(v)} \perp\!\!\!\perp_{\mathscr{L}} \theta_{\mathrm{pa}(v)}$ for all v or, equivalently *if the conditional distributions $\theta_{v|\,\mathrm{pa}(v)}, v \in V$ are mutually independent.*

Graphically, this is most easily displayed by introducing one *additional parent* $\theta_{v|\,\mathrm{pa}(v)}$ for every vertex V in $\mathscr{D}$, so then

$$f(x\,|\,\theta) = \prod_{v \in V} f(x_v \,|\, x_{\mathrm{pa}(v)}, \theta_{v|\,\mathrm{pa}(v)}).$$

Exploiting independence and taking expectations over θ yields that *also marginally*,

$$f(x\,|\,\mathscr{D}) = \int_{\Theta_{\mathscr{D}}} f(x\,|\,\theta)\, \mathscr{L}_{\mathscr{D}}(\theta) = \prod_{v \in V} f(x_v \,|\, x_{\mathrm{pa}(v)}).$$

If $\mathscr{L}$ is strongly directed hyper Markov and $\mathscr{L}^*$ it holds that *also the posterior law $\mathscr{L}^*$ is is strongly directed hyper Markov* and

$$\mathscr{L}^*(\theta_{v|\,\mathrm{pa}(v)}) \propto f(x_v \,|\, x_{\mathrm{pa}(v)}, \theta_{v|\,\mathrm{pa}(v)}) \mathscr{L}(\theta_{v|\,\mathrm{pa}(v)})$$

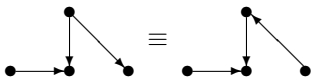
Spiegelhalter and Lauritzen (1990).

Markov equivalence

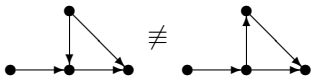
$\mathscr{D}$ and $\mathscr{D}'$ are equivalent if and only if:

1. $\mathscr{D}$ and $\mathscr{D}'$ have same *skeleton* (ignoring directions)
2. $\mathscr{D}$ and $\mathscr{D}'$ have same unmarried parents

so



but



Searching equivalence classes

In general, there is no hope of distinguishing Markov equivalent DAGs, so $\mathscr{D}$ *can at best be identified up to Markov equivalence.*

The number D_n of unlabelled DAGs with n vertices is given by the recursion Robinson (1977)

$$D_n = \sum_{i=1}^{n} (-1)^{i+1} \binom{n}{i} 2^{i(n-i)} D_{n-i}$$

which grows superexponentially. For $n = 10$, $D_n \approx 4.2 \times 10^{18}$. The number of equivalence classes is smaller, but is conjectured still to grow superexponentially.

Conjugate priors for DAGs

In the discrete case, the obvious conjugate prior is for fixed v to let

$$\{\theta_{v|\,\mathrm{pa}_{\mathscr{D}}(v)}(x_v \,|\, x^*_{\mathrm{pa}_{\mathscr{D}}(v)}), x_v \in \mathscr{X}_v\}$$

be *Dirichlet distributed* and independent for $v \in V$ and $x^*_{\mathrm{pa}_{\mathscr{D}}(v)} \in \mathscr{X}_{\mathrm{pa}_{\mathscr{D}}(v)}$ Spiegelhalter and Lauritzen (1990).

We can derive these Dirichlet distributions from a fixed *master Dirichlet* distribution $\mathscr{D}(\alpha)$, where $\alpha = \alpha(x), x \in \mathscr{X}$, by letting

$$\{\theta_{v|\,\mathrm{pa}(v)}(x_v \,|\, x^*_{\mathrm{pa}_{\mathscr{D}}(v)})\} \sim \mathscr{D}(\alpha(x_v, x^*_{\mathrm{pa}_{\mathscr{D}}(v)}),$$

where as usual $\alpha(x_a) = \sum_{y:y_a=x_a} \alpha(y)$.

Typically, α is specified by letting $\alpha = \lambda p_0(x)$ where p_0 is an initial guess on the joint distribution, for example specified through a DAG $\mathscr{D}_0$, and λ is the *equivalent sample size* for the prior information.

The values $\alpha(x_v, x^*_{\mathrm{pa}_{\mathscr{D}}(v)}) = \lambda p_0(x_v, x^*_{\mathrm{pa}_{\mathscr{D}}(v)})$ can then be calculated by *probability propagation*.

Common default values is $\lambda = 1$ and $\alpha(x) = |\mathscr{X}|^{-1}$.

A similar construction is possible in the Gaussian case using the Wishart distribution Geiger and Heckerman (1994) and for mixed discrete Gaussian networks Bøttcher (2001), the latter implemented in the R-package DEAL Bøttcher and Dethlefsen (2003).

In all cases, it was shown Geiger and Heckerman (1997, 2002) that *prior distributions constructed in this way are the* only *distributions which are*

1. *modular:*
$$\mathrm{pa}_{\mathscr{D}}(v) = \mathrm{pa}_{\mathscr{D}'}(v) \Rightarrow \theta_{v|\,\mathrm{pa}_{\mathscr{D}}(v)} \sim \theta_{v|\,\mathrm{pa}_{\mathscr{D}'}(v)};$$
2. *score equivalent:*
$$\mathscr{D} \equiv \mathscr{D}' \Rightarrow f(x^{(n)} \,|\, \mathscr{D}) = f(x^{(n)} \,|\, \mathscr{D}').$$

Marginal likelihood Bayes factors derived from these *strongly directed hyper Dirichlet priors* have a simple form

$$f(x^{(n)} \mid \mathscr{D}) = \prod_v \prod_{x_{\mathrm{pa}(v)}} \frac{\Gamma(\alpha(x_{\mathrm{pa}_{\mathscr{D}}(v)}))}{\Gamma(\alpha(x_{\mathrm{pa}_{\mathscr{D}}(v)}) + n(x_{\mathrm{pa}_{\mathscr{D}}(v)}))} \times \prod_{x_v} \frac{\Gamma(\alpha(x_{v \cup \mathrm{pa}_{\mathscr{D}}(v)}) + n(x_{v \cup \mathrm{pa}_{\mathscr{D}}(v)}))}{\Gamma(\alpha(x_{v \cup \mathrm{pa}_{\mathscr{D}}(v)}))}.$$

Cooper and Herskovits (1992); Heckerman et al (1995)

Challenge: Find *good algorithm for sampling* from the full posterior over DAGs or equivalence classes of DAGs. *Issue:* prior uniform over equivalence classes or over DAGs?

Greedy equivalence class search

1. Initialize with empty DAG
2. Repeatedly search among equivalence classes with a *single additional edge* and go to class with highest score - until no improvement.
3. Repeatedly search among equivalence classes with a *single edge less* and move to one with highest score - until no improvement.

For BIC or Bayesian posterior score with directed hyper Dirichlet priors, this algorithm yields consistent estimate of equivalence class for P. Chickering (2002)

6.3.2 Constraint-based search

Another alternative search algorithm is known as *constraint based search*.

Essentially, the search methods generate queries of the type "$A \perp\!\!\!\perp B \mid S$?", and the answer to such a query divides Γ into those graphs conforming with the query and those that do not.

These type of methods were originally designed by computer scientists in the context where P was fully available, so queries could be answered without error.

The advantage of this type of method is that relatively few queries are needed to identify a DAG $\mathscr{D}$ (or rather its equivalence class).

The disadvantage is that there seems to be no coherent and principled method to answer the query in the presence of statistical uncertainty, which is computable.

SGS and PC algorithms

SGS-algorithm Spirtes et al (1993):

Step 1: Identify *skeleton* using that, for P faithful,

$$u \not\sim v \iff \exists S \subseteq V \setminus \{u,v\} : X_u \perp\!\!\!\perp X_v \mid X_S.$$

Begin with complete graph, check for $S = \emptyset$ and remove edges when independence holds. Then continue for increasing $|S|$.
PC-algorithm (same reference) exploits that only S with $S \subseteq \mathrm{bd}(u) \setminus v$ or $S \subseteq \mathrm{bd}(v) \setminus u$ needs checking where bd refers to current skeleton.

Step 2: Identify directions to be consistent with independence relations found in Step 1.

Exact properties of PC-algorithm

If P is faithful to DAG $\mathscr{D}$, PC-algorithm finds $\mathscr{D}'$ equivalent to $\mathscr{D}$.
It uses N independence checks where N is at most

$$N \leq 2\binom{|V|}{2}\sum_{i=0}^{d}\binom{|V|-1}{i} \leq \frac{|V|^{d+1}}{(d-1)!},$$

where d is the maximal degree of any vertex in $\mathscr{D}$.
So worst case complexity is exponential, but *algorithm fast for sparse graphs.*
Sampling properties are less well understood although consistency results exist.
The general idea has these elements:

1. When a query is decided negatively, $\neg(A \perp\!\!\!\perp B \,|\, S)$, it is *taken at face value*; When a query is decided positively, $A \perp\!\!\!\perp B \,|\, S$, it is *recorded with care*;
2. If at some later stage, the PC algorithm would remove an edge so that a negative query $\neg(A \perp\!\!\!\perp B \,|\, S)$ would conflict with $A \perp_{\mathscr{D}} B \,|\, S$, the removal of this edge is suppressed.
   This leads to *unresolved queries* which are then passed to the user.

6.4 Summary

Types of approach

- Methods for *judging adequacy of structure* such as
  - Tests of significance
  - Penalised likelihood scores

    $$I_\kappa(\mathscr{G}) = \log \hat{L} - \kappa \dim(\mathscr{G})$$

    with $\kappa = 1$ for AIC Akaike (1974), or $\kappa = \frac{1}{2}\log n$ for BIC Schwarz (1978).
  - *Bayesian posterior probabilities.*

- *Search strategies* through space of possible structures, more or less based on *heuristics.*

Bayes factors For $\mathscr{G} \in \Gamma$, $\Theta_{\mathscr{G}}$ is associated parameter space so that P factorizes w.r.t. $\mathscr{G}$ if $P = P_\theta$ for some $\theta \in \Theta_{\mathscr{G}}$. $\mathscr{L}_{\mathscr{G}}$ is prior law on $\Theta_{\mathscr{G}}$.

The *Bayes factor* for discriminating between $\mathscr{G}_1$ and $\mathscr{G}_2$ based on $X^{(n)} = x^{(n)}$ is

$$\mathrm{BF}(\mathscr{G}_1 : \mathscr{G}_2) = \frac{f(x^{(n)} \,|\, \mathscr{G}_1)}{f(x^{(n)} \,|\, \mathscr{G}_2)},$$

where

$$f(x^{(n)} \,|\, \mathscr{G}) = \int_{\Theta_{\mathscr{G}}} f(x^{(n)} \,|\, \mathscr{G}, \theta)\, \mathscr{L}_{\mathscr{G}}(d\theta)$$

is known as the *marginal likelihood* of $\mathscr{G}$. Posterior distribution over graphs If $\pi(\mathscr{G})$ is a prior probability distribution over a given set of graphs Γ, the posterior distribution is determined as

$$\pi^*(\mathscr{G}) = \pi(\mathscr{G} \,|\, x^{(n)}) \propto f(x^{(n)} \,|\, \mathscr{G}) \pi(\mathscr{G})$$

or equivalently

$$\frac{\pi^*(\mathscr{G}_1)}{\pi^*(\mathscr{G}_2)} = \mathrm{BF}(\mathscr{G}_1 : \mathscr{G}_2) \frac{\pi(\mathscr{G}_1)}{\pi(\mathscr{G}_2)}.$$

The *BIC is an $O(1)$-approximation to* $\log \mathrm{BF}$ using Laplace's method of integrals on the marginal likelihood.

Bayesian analysis looks for the *MAP estimate* $\mathscr{G}^*$ maximizing $\pi^*(\mathscr{G})$ over Γ, or attempts to *sample from the posterior* using e.g. Monte-Carlo methods. Estimating trees Assume P factorizes w.r.t. an unknown *tree* $\mathscr{T}$. *MLE* $\hat{\tau}$ *of* $\mathscr{T}$ *has maximal weight*, where the *weight* of τ is

$$w(\tau) = \sum_{e \in E(\tau)} w_n(e) = \sum_{e \in E(\tau)} H_n(e)$$

and $H_n(e)$ is the empirical *cross-entropy* or *mutual information* between endpoint variables of the edge $e = \{u, v\}$. For *Gaussian trees* this becomes

$$w_n(e) = -\frac{1}{2} \log(1 - r_e^2),$$

where r_e^2 is *correlation coeffient* along edge $e = \{u, v\}$.

Highest AIC or BIC scoring forest *also available as MWSF,* with modified weights

$$w_n^{\mathrm{pen}}(e) = n w_n(e) - \kappa_n \mathrm{df}_e,$$

with $\kappa_n = 1$ for AIC, $\kappa_n = \frac{1}{2} \log n$ for BIC and df_e the *degrees of freedom for independence* along e.

Use *maximal weight spanning tree* (or forest) algorithm from weights $W = (w_{uv}, u, v \in V)$.

Hyper inverse Wishart laws Denote the normalisation constant of the hyper inverse Wishart density as

$$h(\delta, \Phi; \mathscr{G}) = \int_{\mathscr{S}^+(\mathscr{G})} (\det K)^{\delta/2} e^{-\operatorname{tr}(K\Phi)}\, dK,$$

The marginal likelihood is then

$$f(x^{(n)} \,|\, \mathscr{G}) = \frac{h(\delta + n, \Phi + W^n; \mathscr{G})}{h(\delta, \Phi; \mathscr{G})}.$$

where

$$h(\delta, \Phi; \mathscr{G}) = \frac{\prod_{Q \in \mathscr{Q}} h(\delta, \Phi_Q; \mathscr{G}_Q)}{\prod_{S \in \mathscr{S}} h(\delta, \Phi_S; S)^{\nu_{\mathscr{G}}(S)}}.$$

For *chordal graphs* all terms reduce to known Wishart constants.

In general, Monte-Carlo simulation or similar methods must be used Atay-Kayis and Massam (2005).

Bayes factors for forests Trees and forests are decomposable graphs, so for a forest ϕ we get

$$\begin{aligned} \pi^*(\phi) &\propto \frac{\prod_{e \in E(\phi)} f(x_e^{(n)})}{\prod_{v \in V} f(x_v^{(n)})^{d_\phi(v)-1}} \\ &\propto \prod_{e \in E(\phi)} \mathrm{BF}(e), \end{aligned}$$

where $\mathrm{BF}(e)$ is the *Bayes factor* for independence along the edge e:

$$\mathrm{BF}(e) = \frac{f(x_u^{(n)}, x_v^{(n)})}{f(x_u^{(n)}) f(x_v^{(n)})}.$$

MAP estimates of forests can thus be computed using an MWSF algorithm, using $w(e) = \log BF(e)$ as weights.

When ϕ is restricted to contain a *single tree*, the normalization constant can be explicitly obtained via the *Matrix Tree Theorem,* see e.g. Bollobás (1998).

Algorithms exist for generating random spanning trees Aldous (1990), so *full posterior analysis is in principle possible for trees*.

Only heuristics available for MAP estimators or maximizing penalized likelihoods such as AIC or BIC, for other than trees.

List of Corrections

References

Akaike H (1974) A new look at the statistical model identification. IEEE Transactions on Automatic Control 19:716–723

Aldous D (1990) A random walk construction of uniform spanning trees and uniform labelled trees. SIAM Journal on Discrete Mathematics 3(4):450–465

Andersen SK, Olesen KG, Jensen FV, Jensen F (1989) Hugin - a shell for building Bayesian belief universes for expert systems. In: Sridharan NS (ed) Proceedings of the 11th International Joint Conference on Artificial Intelligence, Morgan Kaufmann Publishers, San Mateo, CA, pp 1080–1085

Asmussen S, Edwards D (1983) Collapsibility and response variables in contingency tables. Biometrika 70:567–578

Atay-Kayis A, Massam H (2005) A Monte Carlo method for computing the marginal likelihood in non-decomposable graphical Gaussian models. Biometrika 92:317–335

Bahl L, Cocke J, Jelinek F, Raviv J (1974) Optimal decoding of linear codes for minimizing symbol error rate. IEEE Transactions on Information Theory 20:284–287

Barndorff-Nielsen OE (1978) Information and Exponential Families in Statistical Theory. John Wiley and Sons, New York

Baum LE (1972) An equality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes. Inequalities 3:1–8

Berge C (1973) Graphs and Hypergraphs. North-Holland, Amsterdam, The Netherlands, translated from French by E. Minieka

Berry A, Bordat JP, Cogis O (2000) Generating all the minimal separators of a graph. International Journal of Foundations of Computer Science 11:397–403

Bickel PJ, Hammel EA, O'Connell JW (1973) Sex bias in graduate admissions: Data from Berkeley. Science 187(4175):398–404

Bollobás B (1998) Modern Graph Theory. Springer-Verlag, New York

Bøttcher SG (2001) Learning Bayesian networks with mixed variables. In: Proceedings of the Eighth International Workshop in Artificial Intelligence and Statistics, pp 149–156

Bøttcher SG, Dethlefsen C (2003) `deal`: A package for learning Bayesian networks. Journal of Statistical Software 8:1–40

Bouchitté V, Todinca I (2001) Treewidth and minimum fill-in: Grouping the minimal separators. SIAM Journal on Computing 31:212–232

Buhl SL (1993) On the existence of maximum likelihood estimators for graphical Gaussian models. Scandinavian Journal of Statistics 20:263–270

Cannings C, Thompson EA, Skolnick MH (1976) Recursive derivation of likelihoods on pedigrees of arbitrary complexity. Advances in Applied Probability 8:622–625

Chickering DM (2002) Optimal structure identification with greedy search. Journal of Machine Learning Research 3:507–554

Chow CK, Liu CN (1968) Approximating discrete probability distributions with dependence trees. IEEE Transactions on Information Theory 14:462–467

Chow CK, Wagner TJ (1978) Consistency of an estimate of tree-dependent probability distributions. IEEE Transactions on Information Theory 19:369–371

Cooper GF, Herskovits E (1992) A Bayesian method for the induction of probabilistic networks from data. Machine Learning 9:309–347

Cowell RG, Dawid AP, Lauritzen SL, Spiegelhalter DJ (1999) Probabilistic Networks and Expert Systems. Springer-Verlag, New York

Dawid AP (1979) Conditional independence in statistical theory (with discussion). Journal of the Royal Statistical Society, Series B 41:1–31

Dawid AP (1980) Conditional independence for statistical operations. The Annals of Statistics 8:598–617

Dawid AP, Lauritzen SL (1993) Hyper Markov laws in the statistical analysis of decomposable graphical models. The Annals of Statistics 21:1272–1317

Diaconis P, Ylvisaker D (1979) Conjugate priors for exponential families. The Annals of Statistics 7:269–281

Diestel R (1987) Simplicial decompositions of graphs – some uniqueness results. Journal of Combinatorial Theory, Series B 42:133–145

Diestel R (1990) Graph Decompositions. Clarendon Press, Oxford, United Kingdom

Dirac GA (1961) On rigid circuit graphs. Abhandlungen Mathematisches Seminar Hamburg 25:71–76

Edwards D (2000) Introduction to Graphical Modelling, 2nd edn. Springer-Verlag, New York

Elston RC, Stewart J (1971) A general model for the genetic analysis of pedigree data. Human Heredity 21:523–542

Frydenberg M (1990a) The chain graph Markov property. Scandinavian Journal of Statistics 17:333–353

Frydenberg M (1990b) Marginalization and collapsibility in graphical interaction models. The Annals of Statistics 18:790–805

Geiger D, Heckerman D (1994) Learning Gaussian networks. In: de Mantaras RL, Poole D (eds) Proceedings of the 10th Conference on Uncertainty in Artificial Intelligence, Morgan Kaufmann Publishers, San Francisco, CA, pp 235–243

Geiger D, Heckerman D (1997) A characterization of the Dirichlet distribution through global and local independence. The Annals of Statistics 25:1344–1369

Geiger D, Heckerman D (2002) Parameter priors for directed acyclic graphical models and the characterization of several probability distributions. The Annals of Statistics 30:1412–1440

Geiger D, Verma TS, Pearl J (1990) Identifying independence in Bayesian networks. Networks 20:507–534

Heckerman D, Geiger D, Chickering DM (1995) Learning Bayesian networks: The combination of knowledge and statistical data. Machine Learning 20:197–243

Jensen F (2002) HUGIN API Reference Manual Version 5.4. HUGIN Expert Ltd., Aalborg, Denmark

Jensen F, Jensen FV, Dittmer SL (1994) From influence diagrams to junction trees. In: de Mantaras RL, Poole D (eds) Proceedings of the 10th Conference on Uncertainty in Artificial Intelligence, Morgan Kaufmann Publishers, San Francisco, CA, pp 367–373

Jensen FV, Jensen F (1994) Optimal junction trees. In: de Mantaras RL, Poole D (eds) Proceedings of the 10th Conference on Uncertainty in Artificial Intelligence, Morgan Kaufmann Publishers, San Francisco, CA, pp 360–366

Jensen FV, Lauritzen SL, Olesen KG (1990) Bayesian updating in causal probabilistic networks by local computation. Computational Statistics Quarterly 4:269–282

Jiroušek R, Přeučil R (1995) On the effective implementation of the iterative proportional fitting procedure. Computational Statistics and Data Analysis 19:177–189

Kalman RE, Bucy R (1961) New results in linear filtering and prediction. Journal of Basic Engineering 83 D:95–108

Kong A (1986) Multivariate belief functions and graphical models. Ph.D. Thesis, Department of Statistics, Harvard University, Massachusetts

Kruskal Jr JB (1956) On the shortest spanning subtree of a graph and the travelling salesman problem. Proceedings of the American Mathematical Society 7:48–50

Lauritzen SL (1996) Graphical Models. Clarendon Press, Oxford, United Kingdom

Lauritzen SL, Jensen FV (1997) Local computation with valuations from a commutative semigroup. Annals of Mathematics and Artificial Intelligence 21:51–69

Lauritzen SL, Nilsson D (2001) Representing and solving decision problems with limited information. Management Science 47:1238–1251

Lauritzen SL, Spiegelhalter DJ (1988) Local computations with probabilities on graphical structures and their application to expert systems (with discussion). Journal of the Royal Statistical Society, Series B 50:157–224

Lauritzen SL, Speed TP, Vijayan K (1984) Decomposable graphs and hypergraphs. Journal of the Australian Mathematical Society, Series A 36:12–29

Lauritzen SL, Dawid AP, Larsen BN, Leimer HG (1990) Independence properties of directed Markov fields. Networks 20:491–505

Leimer HG (1993) Optimal decomposition by clique separators. Discrete Mathematics 113:99–123

Matúš F (1992) On equivalence of Markov properties over undirected graphs. Journal of Applied Probability 29:745–749

Meek C (1995) Strong completeness and faithfulness in Bayesian networks. In: Besnard P, Hanks S (eds) Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence, Morgan Kaufmann Publishers, San Francisco, CA, pp 411–418

Moussouris J (1974) Gibbs and Markov random systems with constraints. Journal of Statistical Physics 10:11–33

Nilsson D (1998) An efficient algorithm for finding the M most probable configurations in a probabilistic expert system. Statistics and Computing 8:159–173

Parter S (1961) The use of linear graphs in Gauss elimination. SIAM Review 3:119–130

Pearl J (1986) Fusion, propagation and structuring in belief networks. Artificial Intelligence 29:241–288

Pearl J (1988) Probabilistic Inference in Intelligent Systems. Morgan Kaufmann Publishers, San Mateo, CA

Pearl J, Paz A (1987) Graphoids: A graph based logic for reasoning about relevancy relations. In: Boulay BD, Hogg D, Steel L (eds) Advances in Artificial Intelligence – II, North-Holland, Amsterdam, The Netherlands, pp 357–363

Richardson TS (2003) Markov properties for acyclic directed mixed graphs. Scandinavian Journal of Statistics 30:145–158

Robinson RW (1977) Counting unlabelled acyclic digraphs. In: Little CHC (ed) Lecture Notes in Mathematics: Combinatorial Mathematics V, vol 622, Springer-Verlag, New York

Rose DJ, Tarjan RE, Lueker GS (1976) Algorithmic aspects of vertex elimination on graphs. SIAM Journal on Computing 5:266–283

Schwarz G (1978) Estimating the dimension of a model. The Annals of Statistics 6:461–464

Shenoy PP, Shafer G (1986) Propagating belief functions using local propagation. IEEE Expert 1:43–52

Shenoy PP, Shafer G (1990) Axioms for probability and belief–function propagation. In: Shachter RD, Levitt TS, Kanal LN, Lemmer JF (eds) Uncertainty in Artificial Intelligence 4, North-Holland, Amsterdam, The Netherlands, pp 169–198

Shoiket K, Geiger D (1997) A practical algorithm for finding optimal triangulations. In: Proceedings of the Fourteenth National Conference on Artificial Intelligence, AAAI Press, Menlo Park, California, pp 185–190

Spiegelhalter DJ, Lauritzen SL (1990) Sequential updating of conditional probabilities on directed graphical structures. Networks 20:579–605

Spirtes P, Glymour C, Scheines R (1993) Causation, Prediction and Search. Springer-Verlag, New York, reprinted by MIT Press

Studený M (1992) Conditional independence relations have no finite complete characterization. In: Transactions of the 11th Prague Conference on Information Theory, Statistical Decision Functions and Random Processes, Academia, Prague, Czech Republic, pp 377–396

Studený M (1993) Structural semigraphoids. International Journal of General Systems 22:207–217

Tarjan RE (1985) Decomposition by clique separators. Discrete Mathematics 55:221–232

Tarjan RE, Yannakakis M (1984) Simple linear-time algorithms to test chordality of graphs, test acyclicity of hypergraphs, and selectively reduce acyclic hypergraphs. SIAM Journal on Computing 13:566–579

Thiele TN (1880) Om Anvendelse af mindste Kvadraters Methode i nogle Tilfælde, hvor en Komplikation af visse Slags uensartede tilfældige Fejlkilder giver Fejlene en 'systematisk' Karakter. Vidensk Selsk Skr 5 Rk, naturvid og mat Afd 12:381–408, french version: *Sur la Compensation de quelques Erreurs quasi-systématiques par la Méthode des moindres Carrés.* Reitzel, København, 1880.

Venables WN, Ripley BD (2002) Modern Applied Statistics with S, 4th edn. Springer-Verlag, New York

Verma T, Pearl J (1990) Equivalence and synthesis of causal models. In: Bonissone P, Henrion M, Kanal LN, Lemmer JF (eds) Proceedings of the 6th Conference on Uncertainty in Artificial Intelligence, North-Holland, Amsterdam, pp 255–270

Viterbi AJ (1967) Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. IEEE Transactions on Information Theory 13:260–269

Wagner K (1937) Über eine Eigenschaft der ebenen Komplexe. Mathematische Annalen 114:570–590

Yannakakis M (1981) Computing the minimum fill-in is NP-complete. SIAM Journal on Algebraic and Discrete Methods 2:77–79