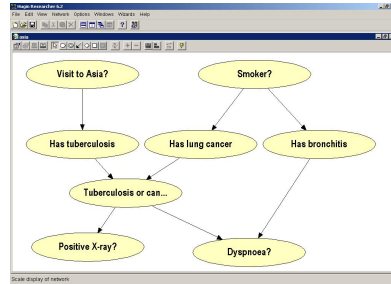


Conditional Independence and Markov Properties

Lecture 1 Saint Flour Summerschool, July 5, 2006

Steffen L. Lauritzen, University of Oxford

A directed graphical model

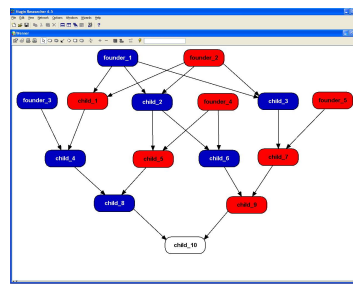


Directed model showing relations between risk factors, diseases, and symptoms.

Overview of lectures

1. Conditional independence and Markov properties
2. More on Markov properties
3. Graph decompositions and junction trees
4. Probability propagation and similar algorithms
5. Log-linear and Gaussian graphical models
6. Conjugate prior families for graphical models
7. Hyper Markov laws
8. Structure learning and Bayes factors
9. More on structure learning.

A pedigree

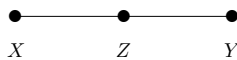


Graphical model for a pedigree from study of Werner's syndrome. Each node is itself a graphical model.

Conditional independence

The notion of conditional independence is fundamental for graphical models.

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

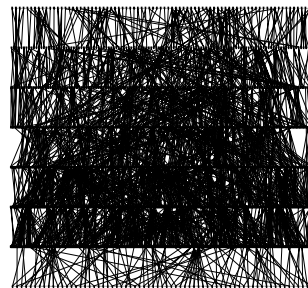


If the random variables have density w.r.t. a product measure μ , the conditional independence is reflected in the relation

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

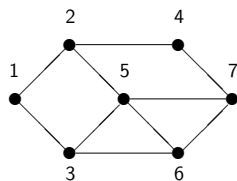
where f is a generic symbol for the densities involved.

A highly complex pedigree



Family relationship of 1641 members of Greenland Eskimo population.

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 .

Conditional independence

Random variables X and Y are *conditionally independent* given the random variable Z if

$$\mathcal{L}(X \mid Y, Z) = \mathcal{L}(X \mid Z).$$

We then write $X \perp\!\!\!\perp Y \mid Z$ (or $X \perp\!\!\!\perp_P Y \mid Z$)

Intuitively:

Knowing Z renders Y *irrelevant* for predicting X .

Factorisation of densities w.r.t. product measure:

$$\begin{aligned} X \perp\!\!\!\perp Y \mid Z &\iff f(x, y, z)f(z) = f(x, z)f(y, z) \\ &\iff \exists a, b : f(x, y, z) = a(x, z)b(y, z). \end{aligned}$$

<p style="text-align: center;">Fundamental properties</p> <p>For random variables $X, Y, Z,$ and W it holds</p> <p>(C1) if $X \perp\!\!\!\perp Y Z$ then $Y \perp\!\!\!\perp X Z;$</p> <p>(C2) if $X \perp\!\!\!\perp Y Z$ and $U = g(Y),$ then $X \perp\!\!\!\perp U Z;$</p> <p>(C3) if $X \perp\!\!\!\perp Y Z$ and $U = g(Y),$ then $X \perp\!\!\!\perp Y (Z, U);$</p> <p>(C4) if $X \perp\!\!\!\perp Y Z$ and $X \perp\!\!\!\perp W (Y, Z),$ then $X \perp\!\!\!\perp (Y, W) Z;$</p> <p>If density w.r.t. product measure $f(x, y, z) > 0$ also</p> <p>(C5) if $X \perp\!\!\!\perp Y Z$ and $X \perp\!\!\!\perp Z Y$ then $X \perp\!\!\!\perp (Y, Z).$</p>	<p>(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 $B.$</p> <p>The property (S5) is slightly more subtle and not generally obvious.</p> <p>Also the symmetry (C1) is a special property of probabilistic conditional independence, rather than of general irrelevance, where (I1) could appear dubious.</p>
<p style="text-align: center;">Additional note on (C5)</p> <p>$f(x, y, z) > 0$ is <i>not necessary</i> for (C5). Enough e.g. that $f(y, z) > 0$ for all (y, z) or $f(x, z) > 0$ for all .</p> <p>In discrete and finite case it is even enough that the bipartite graphs $\mathcal{G}_+ = (\mathcal{Y} \cup \mathcal{Z}, E_+)$ defined by</p> $y \sim_+ z \iff f(y, z) > 0,$ <p>are all connected.</p> <p>Alternatively it is sufficient if the same condition is satisfied with X replacing $Y.$</p> <p>Is there a simple necessary and sufficient condition?</p>	<p style="text-align: center;">Probabilistic semigraphoids</p> <p>V finite set, $X = (X_v, v \in V)$ random variables.</p> <p>For $A \subseteq V,$ let $X_A = (X_v, v \in A).$</p> <p>Let \mathcal{X}_v denote state space of $X_v.$</p> <p>Similarly $x_A = (x_v, v \in A) \in \mathcal{X}_A = \times_{v \in A} \mathcal{X}_v.$</p> <p>Abbreviate: $A \perp\!\!\!\perp B S \iff X_A \perp\!\!\!\perp X_B X_S.$</p> <p>Then basic properties of conditional independence imply:</p> <p><i>The relation $\perp\!\!\!\perp$ on subsets of V is a semigraphoid.</i></p> <p><i>If $f(x) > 0$ for all $x,$ $\perp\!\!\!\perp$ is also a graphoid.</i></p> <p><i>Not all (semi)graphoids are probabilistically representable.</i></p>
<p style="text-align: center;">Graphoid axioms</p> <p>Ternary relation \perp_σ among subsets of a finite set V is <i>graphoid</i> if for all disjoint subsets $A, B, C,$ and D of $V:$</p> <p>(S1) if $A \perp_\sigma B C$ then $B \perp_\sigma A C;$</p> <p>(S2) if $A \perp_\sigma B C$ and $D \subseteq B,$ then $A \perp_\sigma D C;$</p> <p>(S3) if $A \perp_\sigma B C$ and $D \subseteq B,$ then $A \perp_\sigma B (C \cup D);$</p> <p>(S4) if $A \perp_\sigma B C$ and $A \perp_\sigma D (B \cup C),$ then $A \perp_\sigma (B \cup D) C;$</p> <p>(S5) if $A \perp_\sigma B (C \cup D)$ and $A \perp_\sigma C (B \cup D)$ then $A \perp_\sigma (B \cup C) D.$</p> <p><i>Semigraphoid</i> if only (S1)–(S4) holds.</p>	<p style="text-align: center;">Second order conditional independence</p> <p>Sets of random variables A and B are <i>partially uncorrelated</i> for fixed C if their residuals after <i>linear regression</i> on X_C are uncorrelated:</p> $\text{Cov}\{X_A - \mathbf{E}^*(X_A X_C), X_B - \mathbf{E}^*(X_B X_C)\} = 0,$ <p>in other words, if the partial correlations are zero</p> $\rho_{AB \cdot C} = 0.$ <p>We then write $A \perp_2 B C.$</p> <p>Also \perp_2 satisfies the semigraphoid axioms (S1)–(S4) and the graphoid axioms if there is no non-trivial linear relation between the variables in $V.$</p>
<p style="text-align: center;">Irrelevance</p> <p>Conditional independence can be seen as encoding irrelevance in a fundamental way. With the interpretation: <i>Knowing $C,$ A is irrelevant for learning $B,$ (S1)–(S4) translate to:</i></p> <p>(I1) If, knowing $C,$ learning A is irrelevant for learning $B,$ then B is irrelevant for learning $A;$</p> <p>(I2) If, knowing $C,$ learning A is irrelevant for learning $B,$ then A is irrelevant for learning any part D of $B;$</p> <p>(I3) If, knowing $C,$ learning A is irrelevant for learning $B,$ it remains irrelevant having learnt any part D of $B;$</p>	<p style="text-align: center;">Separation in undirected graphs</p> <p>Let $\mathcal{G} = (V, E)$ be finite and simple undirected graph (no self-loops, no multiple edges).</p> <p>For subsets A, B, S of $V,$ let $A \perp_{\mathcal{G}} B S$ denote that S separates A from B in $\mathcal{G},$ i.e. that all paths from A to B intersect $S.$</p> <p>Fact: <i>The relation $\perp_{\mathcal{G}}$ on subsets of V is a graphoid.</i></p> <p>This fact is the reason for choosing the name ‘graphoid’ for such separation relations.</p>

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 $L \ominus N = L \cap N^\perp$. Then L and M are said to *meet orthogonally in N* . This 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 $U \perp M | N$;

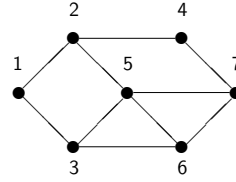
- (O3) If $L \perp M | N$ and U is a linear subspace of M , then $L \perp U | (N + U)$;
- (O4) If $L \perp M | N$ and $L \perp R | (M + N)$, then $L \perp (M + R) | N$.

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

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

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

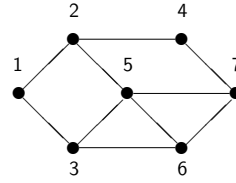
Pairwise Markov property



Any non-adjacent pair of random variables are conditionally independent given the remaining.

For example, $1 \perp\!\!\!\perp 5 | \{2, 3, 4, 6, 7\}$ and $4 \perp\!\!\!\perp 6 | \{1, 2, 3, 5, 7\}$.

Local Markov property



Every variable is conditionally independent of the remaining, given its neighbours.

For example, $5 \perp\!\!\!\perp \{1, 4\} | \{2, 3, 6, 7\}$ and $7 \perp\!\!\!\perp \{1, 2, 3\} | \{4, 5, 6\}$.

Variation independence

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

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

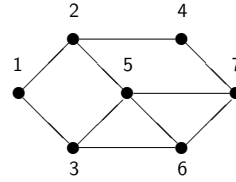
Define further the conditional independence relation $\perp_{\mathcal{U}}$ as

$$A \perp_{\mathcal{U}} B | C \iff \forall u_C^* : \mathcal{U}^{u_C^*} = \{\mathcal{U}^{u_C^*}\}_A \times \{\mathcal{U}^{u_C^*}\}_B$$

i.e. if and only if the C -sections all have the form of a product space.

The relation $\perp_{\mathcal{U}}$ satisfies the *semigraphoid axioms*. In particular $\perp_{\mathcal{U}}$ holds if \mathcal{U} is the support of a probability measure satisfying the similar conditional independence restriction.

Global Markov property



To find conditional independence relations, one should look for separating sets, such as $\{2, 3\}$, $\{4, 5, 6\}$, or $\{2, 5, 6\}$

For example, it follows that $1 \perp\!\!\!\perp 7 | \{2, 5, 6\}$ and $2 \perp\!\!\!\perp 6 | \{3, 4, 5\}$.

Markov properties for semigraphoids

$G = (V, E)$ simple undirected graph; \perp_σ (semi)graphoid relation. Say \perp_σ satisfies

(P) the pairwise Markov property if

$$\alpha \not\sim \beta \implies \alpha \perp_\sigma \beta | V \setminus \{\alpha, \beta\};$$

(L) the local Markov property if

$$\forall \alpha \in V : \alpha \perp_\sigma V \setminus \text{cl}(\alpha) | \text{bd}(\alpha);$$

(G) the global Markov property if

$$A \perp_G B | S \implies A \perp_\sigma B | S.$$

Structural relations among Markov properties

For any semigraphoid it holds that

$$(G) \implies (L) \implies (P)$$

If \perp_σ satisfies graphoid axioms it further holds that

$$(P) \implies (G)$$

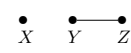
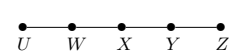
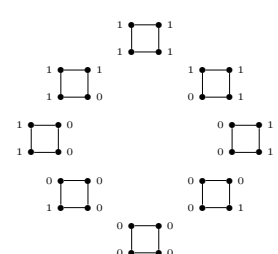
so that in the graphoid case

$$(G) \iff (L) \iff (P).$$

The latter holds in particular for \perp_f , when $f(x) > 0$.

<p style="text-align: center;">(G) \implies (L) \implies (P)</p> <p>(G) implies (L) because $\text{bd}(\alpha)$ separates α from $V \setminus \text{cl}(\alpha)$. Assume (L). Then $\beta \in V \setminus \text{cl}(\alpha)$ because $\alpha \not\sim \beta$. Thus</p> $\text{bd}(\alpha) \cup ((V \setminus \text{cl}(\alpha)) \setminus \{\beta\}) = V \setminus \{\alpha, \beta\},$ <p>Hence by (L) and (S3) we get that</p> $\alpha \perp_{\sigma} (V \setminus \text{cl}(\alpha)) \mid V \setminus \{\alpha, \beta\}.$ <p>(S2) then gives $\alpha \perp_{\sigma} \beta \mid V \setminus \{\alpha, \beta\}$ which is (P).</p>	<p style="text-align: center;">Factorization example</p> <p>The <i>cliques</i> of this graph are the maximal complete subsets $\{1, 2\}$, $\{1, 3\}$, $\{2, 4\}$, $\{2, 5\}$, $\{3, 5, 6\}$, $\{4, 7\}$, and $\{5, 6, 7\}$. A complete set is any subset of these sets.</p> <p>The graph above corresponds to a factorization as</p> $f(x) = \psi_{12}(x_1, x_2)\psi_{13}(x_1, x_3)\psi_{24}(x_2, x_4)\psi_{25}(x_2, x_5) \\ \times \psi_{356}(x_3, x_5, x_6)\psi_{47}(x_4, x_7)\psi_{567}(x_5, x_6, x_7).$
<p style="text-align: center;">(P) \implies (G) for graphoids</p> <p>Assume (P) and $A \perp_{\mathcal{G}} B \mid S$. We must show $A \perp_{\sigma} B \mid S$. Wlog assume A and B non-empty. Proof is reverse induction on $n = S$. If $n = V - 2$ then A and B are singletons and (P) yields $A \perp_{\sigma} B \mid S$ directly. Assume $S = n < V - 2$ and conclusion established for $S > n$. First assume $V = A \cup B \cup S$. Then either A or B has at least two elements, say A. If $\alpha \in A$ then $B \perp_{\mathcal{G}} (A \setminus \{\alpha\}) \mid (S \cup \{\alpha\})$ and also $\alpha \perp_{\mathcal{G}} B \mid (S \cup A \setminus \{\alpha\})$ (as $\perp_{\mathcal{G}}$ is a semi-graphoid).</p>	<p style="text-align: center;">Factorisation of the multivariate Gaussian</p> <p>Consider a multivariate Gaussian random vector $X = \mathcal{N}_V(\xi, \Sigma)$ with Σ regular so it has density</p> $f(x \mid \xi, \Sigma) = (2\pi)^{- V /2} (\det K)^{1/2} e^{-(x-\xi)^{\top} K (x-\xi)/2},$ <p>where $K = \Sigma^{-1}$ is the <i>concentration matrix</i> of the distribution. Thus the Gaussian density factorizes w.r.t. \mathcal{G} if and only if</p> $\alpha \not\sim \beta \implies k_{\alpha\beta} = 0$ <p>i.e. if the concentration matrix has zero entries for non-adjacent vertices.</p>
<p>Thus by the induction hypothesis</p> $(A \setminus \{\alpha\}) \perp_{\sigma} B \mid (S \cup \{\alpha\}) \text{ and } \{\alpha\} \perp_{\sigma} B \mid (S \cup A \setminus \{\alpha\}).$ <p>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 $A \perp_{\mathcal{G}} B \mid (S \cup \{\alpha\})$ and hence the induction hypothesis yields $A \perp_{\sigma} B \mid (S \cup \{\alpha\})$. Further, either $A \cup S$ separates B from $\{\alpha\}$ or $B \cup S$ separates A from $\{\alpha\}$. Assuming the former gives $\alpha \perp_{\sigma} B \mid A \cup S$. Using (S5) we get $(A \cup \{\alpha\}) \perp_{\sigma} B \mid S$ and from (S2) we derive that $A \perp_{\sigma} B \mid S$. The latter case is similar.</p>	<p style="text-align: center;">Factorization theorem</p> <p>Consider a distribution with density f w.r.t. a product measure and let (G), (L) and (P) denote Markov properties w.r.t. the semigraphoid relation \perp_{\perp}. It then holds that</p> $(F) \implies (G)$ <p>and further: If $f(x) > 0$ for all x: (P) \implies (F). Thus in the case of positive density (but typically only then), all the properties coincide:</p> $(F) \iff (G) \iff (L) \iff (P).$
<p style="text-align: center;">Factorisation and Markov properties</p> <p>For $a \subseteq V$, $\psi_a(x)$ is a function depending on x_a only, i.e.</p> $x_a = y_a \implies \psi_a(x) = \psi_a(y).$ <p>We can then write $\psi_a(x) = \psi_a(x_a)$ without ambiguity. The distribution of X factorizes w.r.t. \mathcal{G} or satisfies (F) if its density f w.r.t. product measure on \mathcal{X} has the form</p> $f(x) = \prod_{a \in \mathcal{A}} \psi_a(x),$ <p>where \mathcal{A} are complete subsets of \mathcal{G} or, equivalently, if</p> $f(x) = \prod_{c \in \mathcal{C}} \tilde{\psi}_c(x),$ <p>where \mathcal{C} are the cliques of \mathcal{G}.</p>	

<p style="text-align: center;">More on Markov Properties</p> <p style="text-align: center;">Lecture 2 Saint Flour Summerschool, July 5, 2006</p> <p style="text-align: center;">Steffen L. Lauritzen, University of Oxford</p>	<p style="text-align: center;">Semigraphoid examples</p> <ul style="list-style-type: none"> • <i>Graph separation</i> \perp_G in undirected graph G forms a graphoid; • <i>Variation independence</i> of projections for a subset U of a product space $\ddagger_{\mathcal{U}}$ forms a semigraphoid; • <i>Uncorrelatedness</i> \perp_2 of residuals after linear regression (second order conditional independence) forms a semigraphoid; • <i>Orthogonal meet</i> \perp of closed subspaces of a Hilbert space yields a semigraphoid; • <i>Probabilistic conditional independence</i>.
<p style="text-align: center;">Overview of lectures</p> <ol style="list-style-type: none"> 1. Conditional independence and Markov properties 2. More on Markov properties 3. Graph decompositions and junction trees 4. Probability propagation and similar algorithms 5. Log-linear and Gaussian graphical models 6. Conjugate prior families for graphical models 7. Hyper Markov laws 8. Structure learning and Bayes factors 9. More on structure learning. 	<p style="text-align: center;">Probabilistic semigraphoids</p> <p>V finite set, $X = (X_v, v \in V)$ random variables. For $A \subseteq V$, let $X_A = (X_v, v \in A)$. Let \mathcal{X}_v denote state space of X_v. Similarly $x_A = (x_v, v \in A) \in \mathcal{X}_A = \times_{v \in A} \mathcal{X}_v$. Abbreviate: $A \perp\!\!\!\perp B \mid S \iff X_A \perp\!\!\!\perp X_B \mid X_S$. Then basic properties of conditional independence imply: <i>The relation $\perp\!\!\!\perp$ on subsets of V is a semigraphoid.</i> <i>If $f(x) > 0$ for all x, $\perp\!\!\!\perp$ is also a graphoid.</i> <i>Not all (semi)graphoids are probabilistically representable.</i></p>
<p style="text-align: center;">Conditional Independence</p> <p>For random variables X, Y, Z, and W it holds</p> <p>(C1) if $X \perp\!\!\!\perp Y \mid Z$ then $Y \perp\!\!\!\perp X \mid Z$;</p> <p>(C2) if $X \perp\!\!\!\perp Y \mid Z$ and $U = g(Y)$, then $X \perp\!\!\!\perp U \mid Z$;</p> <p>(C3) if $X \perp\!\!\!\perp Y \mid Z$ and $U = g(Y)$, then $X \perp\!\!\!\perp Y \mid (Z, U)$;</p> <p>(C4) if $X \perp\!\!\!\perp Y \mid Z$ and $X \perp\!\!\!\perp W \mid (Y, Z)$, then $X \perp\!\!\!\perp (Y, W) \mid Z$;</p> <p>If density w.r.t. product measure $f(x, y, z) > 0$ also</p> <p>(C5) if $X \perp\!\!\!\perp Y \mid Z$ and $X \perp\!\!\!\perp Z \mid Y$ then $X \perp\!\!\!\perp (Y, Z)$.</p>	<p style="text-align: center;">Markov properties for semigraphoids</p> <p>$\mathcal{G} = (V, E)$ simple undirected graph; \perp_σ (semi)graphoid relation. Say \perp_σ satisfies</p> <p>(P) <i>the pairwise Markov property</i> if $\alpha \not\sim \beta \implies \alpha \perp_\sigma \beta \mid V \setminus \{\alpha, \beta\}$;</p> <p>(L) <i>the local Markov property</i> if $\forall \alpha \in V : \alpha \perp_\sigma V \setminus \text{cl}(\alpha) \mid \text{bd}(\alpha)$;</p> <p>(G) <i>the global Markov property</i> if $A \perp_G B \mid S \implies A \perp_\sigma B \mid S$.</p>
<p style="text-align: center;">Graphoid axioms</p> <p>Ternary relation \perp_σ among subsets of a finite set V is graphoid if for all disjoint subsets A, B, C, and D of V:</p> <p>(S1) if $A \perp_\sigma B \mid C$ then $B \perp_\sigma A \mid C$;</p> <p>(S2) if $A \perp_\sigma B \mid C$ and $D \subseteq B$, then $A \perp_\sigma D \mid C$;</p> <p>(S3) if $A \perp_\sigma B \mid C$ and $D \subseteq B$, then $A \perp_\sigma B \mid (C \cup D)$;</p> <p>(S4) if $A \perp_\sigma B \mid C$ and $A \perp_\sigma D \mid (B \cup C)$, then $A \perp_\sigma (B \cup D) \mid C$;</p> <p>(S5) if $A \perp_\sigma B \mid (C \cup D)$ and $A \perp_\sigma C \mid (B \cup D)$ then $A \perp_\sigma (B \cup C) \mid D$.</p> <p><i>Semigraphoid</i> if only (S1)–(S4) holds.</p>	<p style="text-align: center;">Structural relations among Markov properties</p> <p><i>For any semigraphoid it holds that</i></p> <p style="text-align: center;">(G) \implies (L) \implies (P)</p> <p><i>If \perp_σ satisfies graphoid axioms it further holds that</i></p> <p style="text-align: center;">(P) \implies (G)</p> <p><i>so that in the graphoid case</i></p> <p style="text-align: center;">(G) \iff (L) \iff (P).</p> <p><i>The latter holds in particular for $\perp\!\!\!\perp$, when $f(x) > 0$.</i></p>

<p style="text-align: center;">Factorisation and Markov properties</p> <p>The distribution of X factorizes w.r.t. \mathcal{G} or satisfies (F) if</p> $f(x) = \prod_{a \in \mathcal{A}} \psi_a(x) = \prod_{c \in \mathcal{C}} \tilde{\psi}_c(x)$ <p>\mathcal{A} are complete subsets and \mathcal{C} are the cliques of \mathcal{G}.</p> <p>It then holds that</p> $(F) \implies (G)$ <p>and further:</p> <p>If $f(x) > 0$ for all x: (P) \implies (F).</p> <p>Thus in the case of positive density (but typically only then), all the properties coincide:</p> $(F) \iff (G) \iff (L) \iff (P).$	<p>To see the latter, assume the density factorizes. Then e.g.</p> $0 \neq 1/8 = f(0,0,0,0) = \psi_{12}(0,0)\psi_{23}(0,0)\psi_{34}(0,0)\psi_{41}(0,0)$ <p>so these factors are all positive.</p> <p>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.</p> <p>But this contradicts the fact that only 8 out of 16 possible configurations have positive probability.</p> <p>In fact, (F) \iff (G) if and only if \mathcal{G} is chordal, i.e. does not have an n-cycle with $n \geq 4$ as an induced subgraph.</p> <p>To be shown later.</p>
<p style="text-align: center;">Pairwise Markov but not local Markov</p>  <p>Let $X = Y = Z$ with $P\{X = 1\} = P\{X = 0\} = 1/2$.</p> <p>This satisfies (P) but not (L).</p> <p>(P): $X \perp\!\!\!\perp Y \mid Z$ and $X \perp\!\!\!\perp Z \mid Y$.</p> <p>(L): $\text{bd}(X) = \emptyset$ so (L) would imply $X \perp\!\!\!\perp (Y, Z)$ which is false.</p> <p>(L) \iff (P) if and only if $\tilde{\mathcal{G}}$ has no induced subgraph $\tilde{\mathcal{G}}_A = (A, \tilde{E}_A)$ with $A = 3$ and $\tilde{E}_A \in \{2, 3\}$ (Matúš 1992).</p> <p>Dual graph: $\alpha \sim \beta$ if and only if $\alpha \not\sim \beta$</p>	<p style="text-align: center;">Instability under limits</p> <p>Consider a sequence $P_n, n = 1, 2, \dots$ of probability measures on \mathcal{X} and assume that $A \perp\!\!\!\perp_{P_n} B \mid C$.</p> <p>If $P_n \rightarrow P$ (weakly, say) it does <i>not</i> hold in general that $A \perp\!\!\!\perp_P B \mid C$.</p> <p>A simple counterexample is as follows: Consider $X = (X_1, X_2, X_3) \sim \mathcal{N}_3(0, \Sigma_n)$ with</p> $\Sigma_n = \begin{pmatrix} 1 & \frac{1}{\sqrt{n}} & \frac{1}{2} \\ \frac{1}{\sqrt{n}} & \frac{2}{n} & \frac{1}{\sqrt{n}} \\ \frac{1}{2} & \frac{1}{\sqrt{n}} & 1 \end{pmatrix} \rightarrow \begin{pmatrix} 1 & 0 & \frac{1}{2} \\ 0 & 0 & 0 \\ \frac{1}{2} & 0 & 1 \end{pmatrix}$ <p>so in the limit it is not true that $1 \perp\!\!\!\perp_P 3 \mid 2$. The</p>
<p style="text-align: center;">Local Markov but not global Markov</p>  <p>Let U and Z be independent with</p> $P(U = 1) = P(Z = 1) = P(U = 0) = P(Z = 0) = 1/2,$ <p>$W = U, Y = Z$, and $X = WY$.</p> <p>This satisfies (L) but not (G).</p> <p>(L): Variables depend deterministically on their neighbours.</p> <p>(G): False that $W \perp\!\!\!\perp Y \mid X$, for example when $X = 0$.</p> <p>(G) \iff (L) if and only if the dual graph $\tilde{\mathcal{G}}$ does not have the 4-cycle as an induced subgraph (Matúš 1992).</p>	<p>concentration matrix K_n is</p> $K_n = \Sigma_n^{-1} = \begin{pmatrix} 2 & -\sqrt{n} & 0 \\ -\sqrt{n} & \frac{3n}{2} & -\sqrt{n} \\ 0 & -\sqrt{n} & 2 \end{pmatrix}$ <p>so for all n it holds that $1 \perp\!\!\!\perp_{P_n} 3 \mid 2$.</p> <p>The critical feature seems to be that K_n does not converge, hence the densities do not converge.</p> <p>What is a reasonable general additional condition for ensuring closure under limits?</p> <p>The answer seems to be convergence in total variation (A. Klenke, St Flour 2006).</p>
<p style="text-align: center;">Global but not factorizing</p>  <p>Uniform on these 8 configurations is (G) w.r.t. the 4-cycle. Conditioning on opposite corners renders one corner deterministic. Yet, (F) is not satisfied (Moussouris 1974).</p>	<p style="text-align: center;">Stability under limits</p> <p>If \mathcal{X} is discrete and finite and $P_n \rightarrow P$ pointwise, conditional independence is preserved:</p> <p>This follows from the fact that</p> $X \perp\!\!\!\perp_{P_n} Y \mid Z \iff f_n(x, y, z) f_n(z) = f_n(x, z) f_n(y, z)$ <p>and this relation is clearly stable under pointwise limits.</p> <p>Hence (G), (L) and (P) are closed under pointwise limits in the discrete case.</p>

Instability under limits

Even in the discrete case, (F) is not in general closed under pointwise limits.

Consider four binary variables X_1, X_2, X_3, X_4 with joint distribution

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

This factorizes w.r.t. the graph



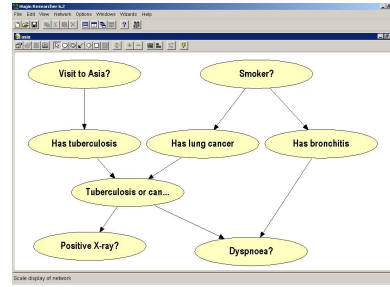
and $f_n(x) = n/(8 + 8n)$ for each of the configurations below

$$\begin{matrix} (0, 0, 0, 0) & (1, 0, 0, 0) & (1, 1, 0, 0) & (1, 1, 1, 0) \\ (0, 0, 0, 1) & (0, 0, 1, 1) & (0, 1, 1, 1) & (1, 1, 1, 1), \end{matrix}$$

whereas $f_n(x) = 1/(8 + 8n)$ for the remaining 8 configurations.

When $n \rightarrow \infty$, the density converges to $f(x) = 1/8$ for each of the configurations above and $f(x) = 0$ otherwise, i.e. the Mousouris example, which is globally Markov but does not factorize.

Example of a directed graphical model



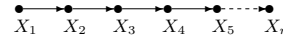
Local directed Markov property

A semigraphoid relation \perp_σ satisfies the local Markov property (L) w.r.t. a directed acyclic graph \mathcal{D} if

$$\forall \alpha \in V : \alpha \perp_\sigma \{ \text{nd}(\alpha) \setminus \text{pa}(\alpha) \} \mid \text{pa}(\alpha).$$

Here $\text{nd}(\alpha)$ are the non-descendants of α .

A well-known example is a Markov chain:



with $X_{i+1} \perp\!\!\!\perp (X_1, \dots, X_{i-1}) \mid X_i$ for $i = 3, \dots, n$.

Markov faithfulness

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

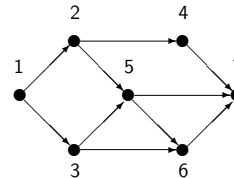
$$A \perp_{\mathcal{G}} B \mid S \iff A \perp\!\!\!\perp B \mid S.$$

It can be shown by a dimensional argument that if $|\mathcal{X}_v| \geq 2$ for all $v \in V$, then there is a distribution P which is Markov faithful to \mathcal{G} .

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.

For a Markov faithful P , the graphoids $\perp_{\mathcal{G}}$ and \perp_P are isomorphic.

Local directed Markov property



For example, the local Markov property says

$$4 \perp_{\sigma} \{1, 3, 5, 6\} \mid 2,$$

$$5 \perp_{\sigma} \{1, 4\} \mid \{2, 3\}$$

$$3 \perp_{\sigma} \{2, 4\} \mid 1.$$

Directed acyclic graphs

A directed acyclic graph \mathcal{D} over a finite set V is a simple graph with all edges directed and no directed cycles.

Absence of directed cycles means 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.

Ordered Markov property

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

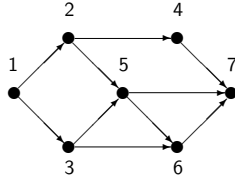
$$\alpha \in \text{pa}(\beta) \implies \alpha < \beta.$$

We then say semigraphoid relation \perp_σ satisfies the ordered Markov property (O) w.r.t. a well-ordered DAG \mathcal{D} if

$$\forall \alpha \in V : \alpha \perp_\sigma \{ \text{pr}(\alpha) \setminus \text{pa}(\alpha) \} \mid \text{pa}(\alpha).$$

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

Ordered Markov property



The numbering corresponds to a well-ordering. The ordered Markov property says for example

- $4 \perp_{\sigma} \{1, 3\} \mid 2,$
- $5 \perp_{\sigma} \{1, 4\} \mid \{2, 3\}$
- $3 \perp_{\sigma} \{2\} \mid 1.$

Equivalence of Markov properties

A semigraphoid relation \perp_{σ} satisfies the *global Markov property* (G) w.r.t. \mathcal{D} if

$$A \perp_{\mathcal{D}} B \mid S \implies A \perp_{\sigma} B \mid S.$$

It holds for any DAG \mathcal{D} and any semigraphoid relation \perp_{σ} that all directed Markov properties are equivalent:

$$(G) \iff (L) \iff (O).$$

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

Separation in DAGs

A trail τ from vertex α to vertex β in a DAG \mathcal{D} is *blocked* by 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 $A \perp_{\mathcal{D}} B \mid S$.

Factorisation with respect to a DAG

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

$$(F): f(x) = \prod_{v \in V} k_v(x_v \mid x_{\text{pa}(v)})$$

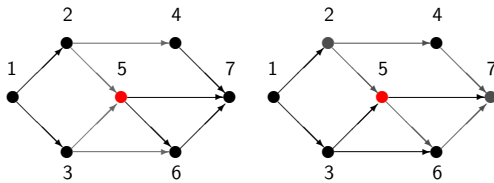
where $k_v \geq 0$ and $\int_{\mathcal{X}_v} k_v(x_v \mid x_{\text{pa}(v)}) \mu_v(dx_v) = 1$.

(F) is equivalent to (F*), where

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

i.e. it follows from (F) that k_v in fact are conditional densities. Proof by induction!

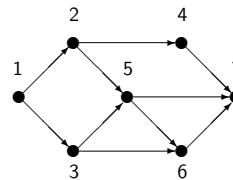
Separation by example



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 $S = \{3, 5\}$, they are all blocked.

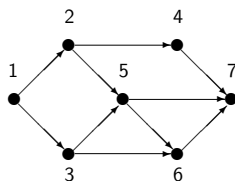
Example of DAG factorization



The above graph corresponds to the factorization

$$f(x) = f(x_1)f(x_2 \mid x_1)f(x_3 \mid x_1)f(x_4 \mid x_2) \\ \times f(x_5 \mid x_2, x_3)f(x_6 \mid x_3, x_5)f(x_7 \mid x_4, x_5, x_6).$$

Returning to example



Hence $4 \perp_{\mathcal{D}} 6 \mid 3, 5$, but it is *not* true that $4 \perp_{\mathcal{D}} 6 \mid 5$ nor that $4 \perp_{\mathcal{D}} 6$.

Markov properties and factorization

Assume that the probability distribution P has a density w.r.t. some product measure on \mathcal{X} .

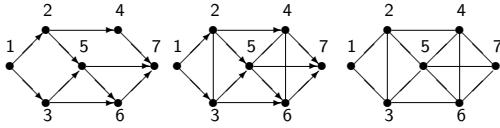
It is then always true that (F) holds if and only if $\perp_{\mathcal{D}}$ satisfies (G),

so all directed Markov properties are equivalent to the factorization property!

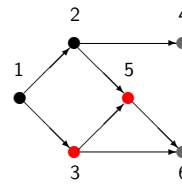
$$(F) \iff (G) \iff (L) \iff (O).$$

Moralization

The *moral graph* \mathcal{D}^m of a DAG \mathcal{D} is obtained by adding undirected edges between unmarried parents and subsequently dropping directions, as in the example below:



Forming ancestral set



The subgraph induced by all ancestors of nodes involved in the query $4 \perp_m 6 \mid 3, 5$?

Undirected factorizations

If P factorizes w.r.t. \mathcal{D} , it factorizes w.r.t. the moralised graph \mathcal{D}^m .

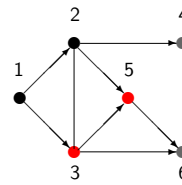
This is seen directly from the factorization:

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

since $\{v\} \cup \text{pa}(v)$ are all complete in \mathcal{D}^m .

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

Adding links between unmarried parents



Adding an undirected edge between 2 and 3 with common child 5 in the subgraph induced by all ancestors of nodes involved in the query $4 \perp_m 6 \mid 3, 5$?

Perfect DAGs

A DAG \mathcal{D} is *perfect* if all parents are married.

For a perfect DAG \mathcal{D} :

P satisfies (F) w.r.t. \mathcal{D} if and only if it satisfies (F) w.r.t. its skeleton $\sigma(\mathcal{D})$.

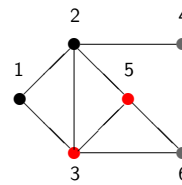
The *skeleton* is the undirected graph obtained from \mathcal{D} by ignoring directions.

For a perfect DAG \mathcal{D} we always have $\sigma(\mathcal{D}) = \mathcal{D}^m$.

A *rooted tree* with arrows pointing away from the root is a perfect DAG.

In particular, any Markov chain is also a Markov field.

Dropping directions



Since $\{3, 5\}$ separates 4 from 6 in this graph, we can conclude that $4 \perp_m 6 \mid 3, 5$

Alternative equivalent separation

To resolve query involving three sets A, B, S :

1. Reduce to subgraph induced by ancestral set $\mathcal{D}_{\text{An}(A \cup B \cup S)}$ of $A \cup B \cup S$;
2. Moralize to form $(\mathcal{D}_{\text{An}(A \cup B \cup S)})^m$;
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.

It then holds that $A \perp_m B \mid S$ if and only if $A \perp_{\mathcal{D}} B \mid S$.

Proof in Lauritzen (1996) needs to allow self-intersecting paths to be correct.

Properties of d -separation

It holds for any DAG \mathcal{D} that $\perp_{\mathcal{D}}$ satisfies graphoid axioms.

Clearly, this is then also true for \perp_m .

To show this is true, it is sometimes easy to use \perp_m , sometimes $\perp_{\mathcal{D}}$.

For example, (S2) is trivial for $\perp_{\mathcal{D}}$, whereas (S5) is trivial for \perp_m .

So, equivalence of $\perp_{\mathcal{D}}$ and \perp_m is useful.

Ancestral marginals

Consider a DAG \mathcal{D} and an *ancestral subset* $A \subseteq V$, i.e. one where

$$\alpha \in A \implies \text{pa}(\alpha) \in A.$$

If P factorizes w.r.t. \mathcal{D} , it factorizes w.r.t. \mathcal{D}_A .

Proof by induction, using that if A is ancestral and $A \neq V$, there is a terminal vertex α_0 with $\alpha_0 \notin A$.

It thus follows, that if P factorizes w.r.t. \mathcal{D} :

$$A \perp_m B | S \implies A \perp\!\!\!\perp B | S.$$

Because then P factorizes w.r.t. $\mathcal{D}_{\text{An}(A \cup B \cup S)}^m$ and hence satisfies (G) for this graph.

Markov equivalence of directed and undirected graphs

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

This happens if and only if \mathcal{D} is perfect and $\mathcal{G} = \sigma(\mathcal{D})$. So, these are all equivalent



but not equivalent to



Faithfulness

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

$$A \perp_{\mathcal{D}} B | S \iff A \perp\!\!\!\perp B | S.$$

It can be also here be shown that if $|\mathcal{X}_v| \geq 2$ for all $v \in V$, then there is a distribution P which is Markov faithful for \mathcal{D} , and the set of directed Markov distributions which are not faithful is a Lebesgue null-set in the set of directed Markov distributions.

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

Hence *d*-separation is indeed the strongest possible.

References

- Matúš, F. (1992). On equivalence of Markov properties over undirected graphs. *Journal of Applied Probability*, **29**, 745–9.
- Moussouris, J. (1974). Gibbs and Markov random systems with constraints. *Journal of Statistical Physics*, **10**, 11–33.

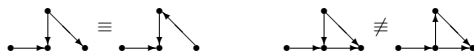
Markov equivalence

Two DAGS \mathcal{D} and \mathcal{D}' are *Markov equivalent* if the separation relations $\perp_{\mathcal{D}}$ and $\perp_{\mathcal{D}'}$ are identical.

\mathcal{D} and \mathcal{D}' are equivalent if and only if:

1. \mathcal{D} and \mathcal{D}' have same *skeleton* (ignoring directions)
2. \mathcal{D} and \mathcal{D}' have same unmarried parents

so

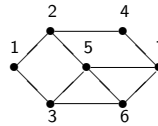


Graph Decompositions and Junction Trees

Lecture 3 Saint Flour Summerschool, July 6, 2006

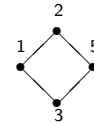
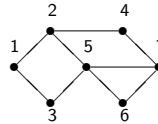
Steffen L. Lauritzen, University of Oxford

Examples



The graph to the left is prime

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



Overview of lectures

1. Conditional independence and Markov properties
2. More on Markov properties
3. Graph decompositions and junction trees
4. Probability propagation and similar algorithms
5. Log-linear and Gaussian graphical models
6. Conjugate prior families for graphical models
7. Hyper Markov laws
8. Structure learning and Bayes factors
9. More on structure learning.

Decomposition of Markov properties

Suppose P satisfies (F) w.r.t. \mathcal{G} and (A, B, S) is a decomposition. Then

(i) $P_{A \cup S}$ and $P_{B \cup S}$ satisfy (F) w.r.t. $\mathcal{G}_{A \cup S}$ and $\mathcal{G}_{B \cup S}$ respectively;

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

The first part of the statement is true when (F) is replaced by (G).

The second is also true for (G) if the relevant densities exist.

Some motivation

- Perfect DAGs are simple, because their directions can be ignored as they are Markov equivalent to their skeleton;
- Undirected graphs which can occur as skeletons of perfect DAGs are therefore particularly simple;
- An n -cycle with $n \geq 4$ cannot be oriented to form a perfect DAG:



- The important simplifying idea is that of graph decomposition and decomposability.

Markov combination

Let Q and R be distributions on $\mathcal{X}_{A \cup S}$ and $\mathcal{X}_{B \cup S}$ resp. and assume Q and R are consistent, i.e. $Q_S = R_S$.

Then there is a unique distribution $P = Q * R$ so that

(i) $P_{A \cup S} = Q$ and $P_{B \cup S} = R$;

(ii) $A \perp\!\!\!\perp B \mid S$.

$Q * R$ is the Markov combination of Q and R . If Q and R have densities q and r , so has P and

$$p(x)q_S(x_S) = p(x)r_S(x_S) = q(x_{A \cup S})r(x_{B \cup S}).$$

The Markov combination maximizes entropy among measures satisfying (i).

Graph decomposition

Consider an undirected graph $\mathcal{G} = (V, E)$. A partitioning of V into a triple (A, B, S) of subsets of V forms a decomposition of \mathcal{G} if

$$A \perp\!\!\!\perp B \mid S \text{ and } S \text{ is complete.}$$

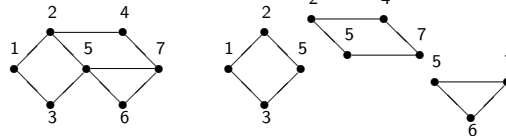
The decomposition is proper if $A \neq \emptyset$ and $B \neq \emptyset$.

The components of \mathcal{G} are the induced subgraphs $\mathcal{G}_{A \cup S}$ and $\mathcal{G}_{B \cup S}$.

A graph is prime if no proper decomposition exists.

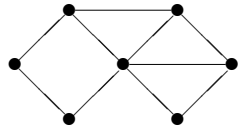
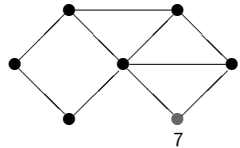
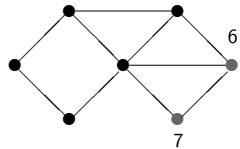
Decomposability

Any graph can be recursively decomposed into its maximal prime subgraphs:

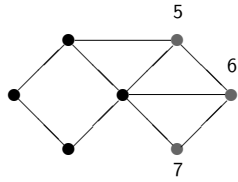


A graph is decomposable (or rather fully decomposable) if it is complete or admits a proper decomposition into decomposable subgraphs.

Definition is recursive. Alternatively this means that all maximal prime subgraphs are cliques.

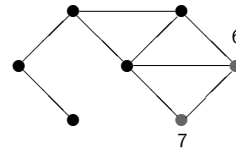
<p style="text-align: center;">Factorization of Markov distributions</p> <p>Recursive decomposition of a decomposable graph into cliques yields the formula:</p> $f(x) \prod_{S \in \mathcal{S}} f_S(x_S)^{\nu(S)} = \prod_{C \in \mathcal{C}} f_C(x_C).$ <p>Here \mathcal{S} is the set of <i>minimal complete separators</i> occurring in the decomposition process and $\nu(S)$ the number of times such a separator appears in this process.</p>	<p style="text-align: center;">Identifying chordal graphs</p> <p>Here is a (greedy) algorithm for checking chordality:</p> <ol style="list-style-type: none"> 1. Look for a vertex v^* with $\text{bd}(v^*)$ 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. <p>The complexity of this algorithm is $O(V ^2)$.</p>
<p style="text-align: center;">Combinatorial consequences</p> <p>Note that if we let $\mathcal{X}_v = \{0, 1\}$ and f be uniform, this yields</p> $2^{- V } \prod_{S \in \mathcal{S}} 2^{- S \nu(S)} = \prod_{C \in \mathcal{C}} 2^{- C }$ <p>and hence we must have</p> $\sum_{C \in \mathcal{C}} C - \sum_{S \in \mathcal{S}} S \nu(S) = V .$ <p>It also holds that</p> $\sum_{S \in \mathcal{S}} \nu(S) = V - 1.$	<p style="text-align: center;">Greedy algorithm</p>  <p>Is this graph chordal?</p>
<p style="text-align: center;">Properties associated with decomposability</p> <p>A numbering $V = \{1, \dots, V \}$ of the vertices of an undirected graph is <i>perfect</i> if the induced oriented graph is a perfect DAG or, equivalently, if</p> $\forall j = 2, \dots, V : \text{bd}(j) \cap \{1, \dots, j-1\} \text{ is complete in } \mathcal{G}.$ <p>An undirected graph \mathcal{G} is <i>chordal</i> if it has no chordless n-cycles with $n \geq 4$.</p> <p>These graphs are also known as <i>rigid circuit graphs</i> or <i>triangulated graphs</i>.</p> <p>A set S is an (α, β)-separator if $\alpha \perp_{\mathcal{G}} \beta \mid S$,</p>	<p style="text-align: center;">Greedy algorithm</p>  <p>Is this graph chordal?</p>
<p style="text-align: center;">Characterizing chordal graphs</p> <p>The following are equivalent for any undirected graph \mathcal{G}.</p> <ol style="list-style-type: none"> (i) \mathcal{G} is chordal; (ii) \mathcal{G} is decomposable; (iii) All maximal prime subgraphs of \mathcal{G} are cliques; (iv) \mathcal{G} admits a perfect numbering; (v) Every minimal (α, β)-separator are complete. <p>Trees are chordal graphs and thus decomposable.</p>	<p style="text-align: center;">Greedy algorithm</p>  <p>Is this graph chordal?</p>

Greedy algorithm



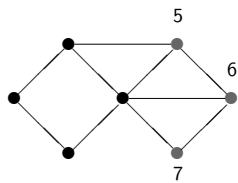
Is this graph chordal?

Greedy algorithm



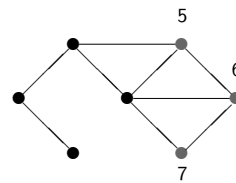
Is this graph chordal?

Greedy algorithm



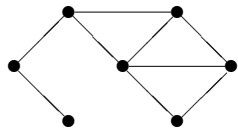
This graph is *not* chordal, as there is no candidate for number 4.

Greedy algorithm



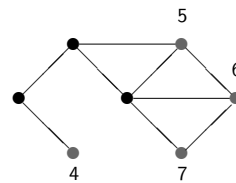
Is this graph chordal?

Greedy algorithm



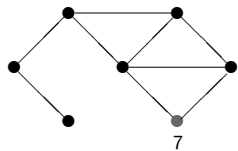
Is this graph chordal?

Greedy algorithm



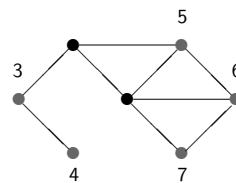
Is this graph chordal?

Greedy algorithm



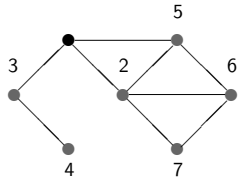
Is this graph chordal?

Greedy algorithm



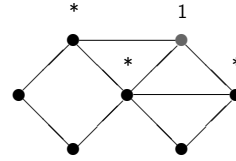
Is this graph chordal?

Greedy algorithm



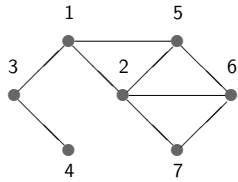
Is this graph chordal?

Maximum Cardinality Search



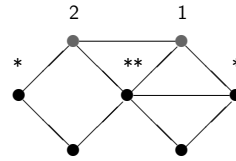
Is this graph chordal?

Greedy algorithm



This graph is chordal!

Maximum Cardinality Search



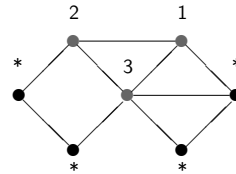
Is this graph chordal?

Maximum cardinality search

This simple algorithm has complexity $O(|V| + |E|)$:

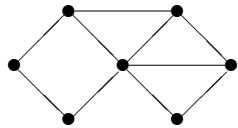
1. Choose $v_0 \in V$ arbitrary and let $v_0 = 1$;
2. When vertices $\{1, 2, \dots, j\}$ have been identified, choose $v = j + 1$ among $V \setminus \{1, 2, \dots, j\}$ with highest cardinality of its numbered neighbours;
3. If $\text{bd}(j + 1) \cap \{1, 2, \dots, j\}$ is not complete, \mathcal{G} is 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.

Maximum Cardinality Search



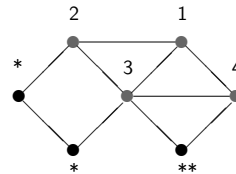
Is this graph chordal?

Maximum Cardinality Search



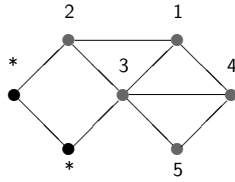
Is this graph chordal?

Maximum Cardinality Search



Is this graph chordal?

Maximum Cardinality Search



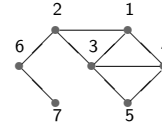
Is this graph chordal?

Finding the cliques of a chordal graph

From an MCS numbering $V = \{1, \dots, |V|\}$, let

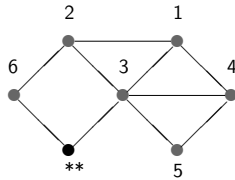
$$S_\lambda = \text{bd}(\lambda) \cap \{1, \dots, \lambda - 1\}$$

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



$\pi_\lambda: 0, 1, 2, 2, 2, 1, 1$. The cliques are $C_\lambda = \{\lambda\} \cup S_\lambda, \lambda \in \Lambda$.

Maximum Cardinality Search



Is this graph chordal?

Junction tree

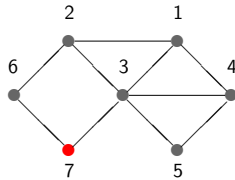
Let \mathcal{A} be a collection of finite subsets of a set V . A *junction tree* \mathcal{T} of sets in \mathcal{A} is an undirected tree with \mathcal{A} as a vertex set, satisfying the *junction tree property*:

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

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

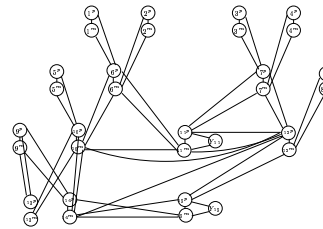
The junction tree can be constructed directly from the MCS ordering $C_\lambda, \lambda \in \Lambda$.

Maximum Cardinality Search



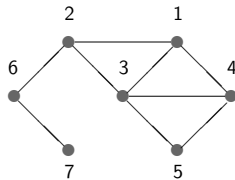
The graph is not chordal! because 7 does not have a complete boundary.

A chordal graph



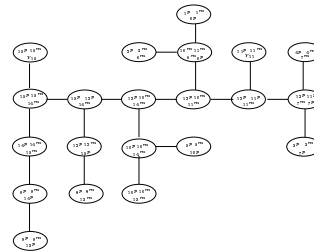
This graph is chordal, but it might not be that easy to see. ... Maximum Cardinality Search is handy!

Maximum Cardinality Search



MCS numbering for the chordal graph. Algorithm runs essentially as before.

Junction tree



Cliques of graph arranged into a tree with $C_1 \cap C_2 \subseteq D$ for all cliques D on path between C_1 and C_2 .

Probability Propagation and Related Algorithms

Lecture 4 Saint Flour Summerschool, July 8, 2006

Steffen L. Lauritzen, University of Oxford

Relation between different graphs

P directed Markov w.r.t. \mathcal{D} implies P factorizes w.r.t. \mathcal{D}^m .

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

$\mathcal{G} = \sigma(\mathcal{D})$ for some DAG \mathcal{D} if and only if \mathcal{G} is chordal.

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

Overview of lectures

1. Conditional independence and Markov properties
2. More on Markov properties
3. Graph decompositions and junction trees
4. Probability propagation and related algorithms
5. Log-linear and Gaussian graphical models
6. Conjugate prior families for graphical models
7. Hyper Markov laws
8. Structure learning and Bayes factors
9. More on structure learning.

Graph decomposition

Consider an undirected graph $\mathcal{G} = (V, E)$. A partitioning of V into a triple (A, B, S) of subsets of V forms a decomposition of \mathcal{G} if both of the following holds:

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

The decomposition is proper if $A \neq \emptyset$ and $B \neq \emptyset$.

The components of \mathcal{G} are the induced subgraphs \mathcal{G}_{AUS} and \mathcal{G}_{BUS} .

A graph is prime if no proper decomposition exists.

Markov properties for undirected graphs

- (P) pairwise Markov: $\alpha \not\sim \beta \implies \alpha \perp \beta | V \setminus \{\alpha, \beta\}$;
- (L) local Markov: $\alpha \perp V \setminus \text{cl}(\alpha) | \text{bd}(\alpha)$;
- (G) global Markov: $A \perp_{\mathcal{G}} B | S \implies A \perp B | S$;
- (F) Factorization: $f(x) = \prod_{a \in \mathcal{A}} \psi_a(x)$, \mathcal{A} being complete subsets of V .

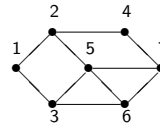
It then holds that

$$(F) \implies (G) \implies (L) \implies (P).$$

If $f(x) > 0$ even

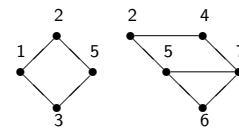
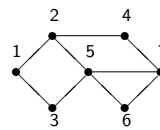
$$(F) \iff (G) \iff (L) \iff (P).$$

Examples



The graph to the left is prime

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



Markov properties for directed acyclic graphs

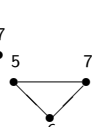
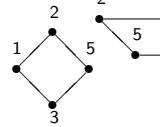
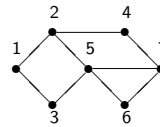
- (O) ordered Markov: $\alpha \perp \{\text{pr}(\alpha) \setminus \text{pa}(\alpha)\} | \text{pa}(\alpha)$;
- (L) local Markov: $\alpha \perp \{\text{nd}(\alpha) \setminus \text{pa}(\alpha)\} | \text{pa}(\alpha)$;
- (G) global Markov: $A \perp_{\mathcal{D}} B | S \implies A \perp B | S$.
- (F) Factorization: $f(x) = \prod_{v \in V} f(x_v | x_{\text{pa}(v)})$.

It then always holds that

$$(F) \iff (G) \iff (L) \iff (O).$$

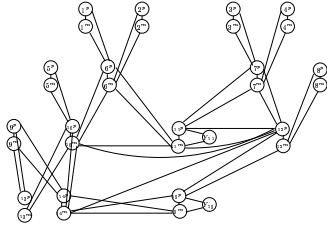
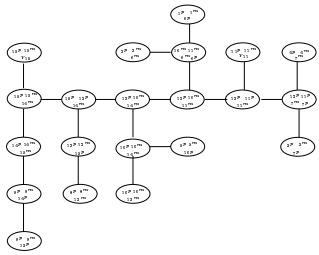
Decomposability

Any graph can be recursively decomposed into its uniquely defined prime components:

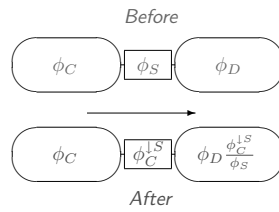
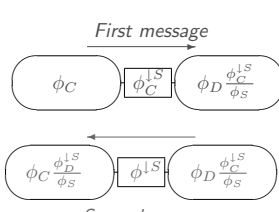


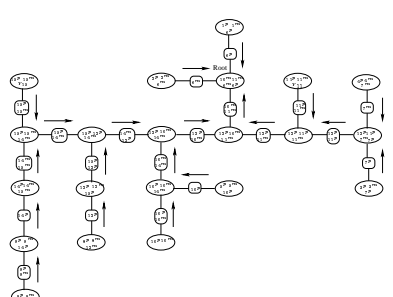
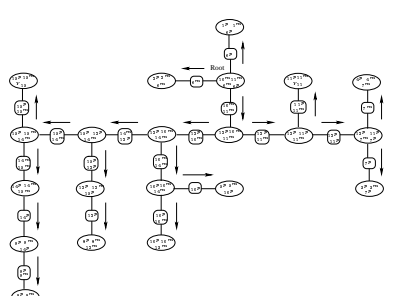
A graph is decomposable (or rather fully decomposable) if it is complete or admits a proper decomposition into decomposable subgraphs.

Definition is recursive. Alternatively this means that all prime components are cliques.

<p style="text-align: center;">Decomposition of Markov properties</p> <p>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:</p> <ul style="list-style-type: none"> (i) $P_{A US}$ and $P_{B US}$ factorize w.r.t. $\mathcal{G}_{A US}$ and $\mathcal{G}_{B US}$; (ii) $f(x)f_S(x_S) = f_{A US}(x_{A US})f_{B US}(x_{B US})$. <p>Recursive decomposition of a decomposable graph yields:</p> $f(x) \prod_{S \in \mathcal{S}} f_S(x_S)^{\nu(S)} = \prod_{C \in \mathcal{C}} f_C(x_C).$ <p>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.</p>	<p style="text-align: center;">Junction tree</p> <p>Let \mathcal{A} be a collection of finite subsets of a set V. A junction tree \mathcal{T} of sets in \mathcal{A} is an undirected tree with \mathcal{A} as a vertex set, satisfying the junction tree property:</p> <p>If $A, B \in \mathcal{A}$ and C is on the unique path in \mathcal{T} between A and B it holds that $A \cap B \subset C$.</p> <p>If the sets in \mathcal{A} are pairwise incomparable, they can be arranged in a junction tree if and only if $\mathcal{A} = \mathcal{C}$ where \mathcal{C} are the cliques of a chordal graph.</p> <p>The junction tree can be constructed directly from the MCS ordering $C_\lambda, \lambda \in \Lambda$.</p>
<p>More generally if \mathcal{Q} denotes the prime components of \mathcal{G}:</p> $f(x) \prod_{S \in \mathcal{S}} f_S(x_S)^{\nu(S)} = \prod_{Q \in \mathcal{Q}} f_Q(x_Q).$	<p style="text-align: center;">A chordal graph</p>  <p>This graph is chordal, but it might not be that easy to see... Maximum Cardinality Search is handy!</p>
<p style="text-align: center;">Characterizing chordal graphs</p> <p>The following are equivalent for any undirected graph \mathcal{G}.</p> <ul style="list-style-type: none"> (i) \mathcal{G} is chordal; (ii) \mathcal{G} is decomposable; (iii) All prime components of \mathcal{G} are cliques; (iv) \mathcal{G} admits a perfect numbering; (v) Every minimal (α, β)-separator are complete. <p>Trees are chordal graphs and thus decomposable.</p>	<p style="text-align: center;">Junction tree</p>  <p>Cliques of graph arranged into a tree with $C_1 \cap C_2 \subseteq D$ for all cliques D on path between C_1 and C_2.</p>
<p style="text-align: center;">Algorithms associated with chordality</p> <p>Maximum Cardinality Search (MCS) Tarjan and Yannakakis (1984) identifies whether a graph is chordal or not.</p> <p>If a graph \mathcal{G} is chordal, MCS yields a perfect numbering of the vertices. In addition it finds the cliques of \mathcal{G}:</p> <p>From an MCS numbering $V = \{1, \dots, V \}$, let</p> $S_\lambda = \text{bd}(\lambda) \cap \{1, \dots, \lambda - 1\}$ <p>and $\pi_\lambda = S_\lambda$. Call λ a ladder vertex if $\lambda = V$ or if $\pi_{\lambda+1} < \pi_\lambda + 1$ and let Λ be the set of ladder vertices.</p> <p>The cliques are $C_\lambda = \{\lambda\} \cup S_\lambda, \lambda \in \Lambda$.</p> <p>The numbers $\nu(S)$ in the decomposition formula are $\nu(S) = \{\lambda \in \Lambda : S_\lambda = S\}$.</p>	<p style="text-align: center;">Junction trees of prime components</p> <p>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).</p> <p>Then every pair of neighbours (C, D) in the junction tree represents a decomposition of \mathcal{G} into $\mathcal{G}_{\bar{C}}$ and $\mathcal{G}_{\bar{D}}$, where \bar{C} is the set of vertices in cliques connected to C but separated from D in the junction tree, and similarly with \bar{D}.</p> <p>Tarjan's algorithm is based on a slightly more sophisticated algorithm (Rose et al. 1976) known as Lexicographic Search (LEX) which runs in $O(V ^2)$ time.</p>

<p style="text-align: center;">Markov properties of junction tree</p> <p>Let $Q \in \mathcal{Q}$ be the prime components of a graph \mathcal{G}, arranged in a junction tree \mathcal{T}.</p> <p>Using that any graph decomposition also yields a decomposition of the Markov properties now gives that</p> <p><i>The distribution of $X = (X_v, v \in V)$ factorizes w.r.t. \mathcal{G} if and only if $X_Q, Q \in \mathcal{Q}$ factorizes w.r.t. \mathcal{T} and each of X_Q factorizes w.r.t. \mathcal{G}_Q.</i></p> <p>In particular, if \mathcal{G} is decomposable, $X = (X_v, v \in V)$ factorizes w.r.t. \mathcal{G} if and only if $X_C, C \in \mathcal{C}$ factorizes w.r.t. \mathcal{T}, i.e. the Markov property has essentially been transferred to that of a tree of cliques.</p>	<p style="text-align: center;">Computational challenge</p> <p>Calculate marginals $\psi_A = \phi^{\perp A}$ of joint valuation</p> $\phi = \otimes_{C \in \mathcal{C}} \phi_C$ <p>with domain $V = \cup_{C \in \mathcal{C}} C$.</p> <p><i>Direct computation of $\phi^{\perp A}$ is impossible if V is large.</i></p> <p><i>Challenge: calculate $\phi^{\perp A}$ using only local operations, i.e. operating on factors ψ_B with domain $B \subseteq C$ for some $C \in \mathcal{C}$.</i></p> <p>Typically also a <i>second purpose</i> of calculation.</p>
<p style="text-align: center;">Local computation</p> <p>Local computation algorithms similar to probability propagation have been developed independently in a number of areas with a variety of purposes. For example:</p> <ul style="list-style-type: none"> • Kalman filter and smoother (Thiele 1880; Kalman and Bucy 1961); • Solving sparse linear equations (Parter 1961); • Decoding digital signals (Viterbi 1967; Bahl <i>et al.</i> 1974); • Estimation in hidden Markov models (Baum 1972); • Peeling in pedigrees (Elston and Stewart 1971; Cannings <i>et al.</i> 1976); 	<p style="text-align: center;">A probability perspective</p> <p>Factorizing density on $\mathcal{X} = \times_{v \in V} \mathcal{X}_v$ with V and \mathcal{X}_v finite:</p> $p(x) = \prod_{C \in \mathcal{C}} \phi_C(x).$ <p>The <i>potentials</i> $\phi_C(x)$ depend on $x_C = (x_v, v \in C)$ only.</p> <p>Basic task to calculate <i>marginal</i> (likelihood)</p> $p^{\perp E}(x_E^*) = \sum_{y_{V \setminus E}} p(x_E^*, y_{V \setminus E})$ <p>for $E \subseteq V$ and fixed x_E^*, but sum has too many terms.</p> <p><i>A second purpose is to get the prediction</i> $p(x_v x_E^*) = p(x_v, x_E^*) / p(x_E^*)$ for $v \in V$.</p>
<ul style="list-style-type: none"> • Belief function evaluation (Kong 1986; Shenoy and Shafer 1986); • Probability propagation (Pearl 1986; Lauritzen and Spiegelhalter 1988; Jensen <i>et al.</i> 1990); • Abstract framework (Shenoy and Shafer 1990; Lauritzen and Jensen 1997). <p>Also dynamic programming, linear programming, optimizing decisions, calculating Nash equilibria in cooperative games, and many others. <i>List is far from exhaustive!</i></p> <p>All algorithms are using, explicitly or implicitly, a <i>graph decomposition</i> and a <i>junction tree</i> or similar to make the computations.</p>	<p style="text-align: center;">Sparse linear equations</p> <ul style="list-style-type: none"> • Valuations ϕ_C are <i>equation systems</i> involving variables with labels C; • $\phi_A \otimes \phi_B$ <i>concatenates</i> equation systems; • $\phi_B^{\perp A}$ <i>eliminates</i> variables in $B \setminus A$; • Marginal $\phi^{\perp A}$ of joint valuation <i>reduces</i> the system of equation to a smaller one; • Second computation finds a <i>solution</i> of the equation system by substitution.
<p style="text-align: center;">An abstract perspective</p> <p>V is large finite set and \mathcal{C} collection of small subsets of V.</p> <p>$\phi_C, C \in \mathcal{C}$ are <i>valuations</i> with domain C.</p> <p><i>Combination:</i> $\phi_A \otimes \phi_B$ has domain $A \cup B$.</p> <p>\otimes is assumed <i>commutative</i> and <i>associative</i>.</p> <p>For $A \subseteq V$ $\phi^{\perp A}$ denotes the <i>A-marginal</i> of ϕ. $\phi^{\perp A}$ has domain A.</p> <p>Assume <i>consonance</i>: $\phi^{\perp(A \cap B)} = (\phi^{\perp B})^{\perp A}$</p> <p>and <i>distributivity</i>: $(\phi \otimes \phi_C)^{\perp B} = (\phi^{\perp B}) \otimes \phi_C$, if $C \subseteq B$.</p>	<p style="text-align: center;">Constraint satisfaction</p> <ul style="list-style-type: none"> • ϕ_C represent <i>constraints</i> involving variables in C; • $\phi_A \otimes \phi_B$ represents <i>jointly feasible</i> configurations; • $\phi_B^{\perp A}$ finds <i>implied constraints</i>; • Marginal $\phi^{\perp A}$ finds <i>extendible</i> configurations; • Second computation <i>identifies</i> jointly feasible configurations. <p>If represented by indicator functions, \otimes is ordinary product and $\phi^{\perp E}(x_E^*) = \oplus_{y_{V \setminus E}} \phi(x_E^*, y_{V \setminus E})$, where $1 \oplus 1 = 1$ and $0 \oplus 1 = 1$ and $0 \oplus 0 = 0$.</p>

<p style="text-align: center;">Computational structure</p> <p>Algorithms all (implicitly or explicitly) arrange the collection of sets \mathcal{C} in a <i>junction tree</i> \mathcal{T}.</p> <p>Hence, this works <i>only</i> if \mathcal{C} are <i>cliques of chordal graph</i> \mathcal{G}.</p> <p>If this is not so from the outset, a <i>triangulation</i> is used to construct chordal graph \mathcal{G}' with $E \subseteq E'$.</p> <p>Clearly, in a probabilistic perspective, if P factorizes w.r.t. \mathcal{G} it factorizes w.r.t. \mathcal{G}'.</p> <p>Henceforth we assume this has been done and \mathcal{G} is chordal.</p> <p>Computations are executed by <i>message passing</i>.</p>	<p style="text-align: center;">Marginalization</p> <p>The A-<i>marginal</i> of a potential ϕ_B for $A \subseteq B$ is</p> $\phi_B^{\perp A}(x) = \sum_{y_B: y_A = x_A} \phi_B(y)$ <p>If ϕ_B depends on x through x_B only and $B \subseteq V$ is 'small', marginal can be computed easily.</p> <p>Marginalization satisfies</p> <p>Consonance For subsets A and B: $\phi^{\perp(A \cap B)} = (\phi^{\perp B})^{\perp A}$</p> <p>Distributivity If ϕ_C depends on x_C only and $C \subseteq B$: $(\phi_C \phi_B)^{\perp B} = (\phi^{\perp B}) \phi_C$.</p>
<p style="text-align: center;">Setting up the structure</p> <p>In many applications P is initially factorizing over a <i>directed acyclic graph</i> \mathcal{D}. The computational structure is then set up in several steps:</p> <ol style="list-style-type: none"> 1. <i>Moralisation</i>: Constructing \mathcal{D}^m, exploiting that if P factorizes on \mathcal{D}, it factorizes over \mathcal{D}^m. 2. <i>Triangulation</i>: Adding edges to find chordal graph \mathcal{G} with $\mathcal{D}^m \subseteq \mathcal{G}$. This step is non-trivial (NP-complete) to optimize; 3. <i>Constructing junction tree</i>: 4. <i>Initialization</i>: Assigning potential functions ϕ_C to cliques. 	<p style="text-align: center;">Messages</p> <p>When C sends message to D, the following happens:</p>  <p>Computation is <i>local</i>, involving only variables within cliques.</p>
<p style="text-align: center;">Basic computation</p> <p>This involves following steps</p> <ol style="list-style-type: none"> 1. <i>Incorporating observations</i>: 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: $p(x X_E = x_E^*) = \frac{\prod_{C \in \mathcal{C}} \phi_C(x_C)}{p(x_E^*)}.$ 2. Marginals $p(x_E^*)$ and $p(x_C x_E^*)$ are then calculated by a local <i>message passing</i> algorithm. 	<p>The expression</p> $\kappa(x) = \frac{\prod_{C \in \mathcal{C}} \phi_C(x_C)}{\prod_{S \in \mathcal{S}} \phi_S(x_S)}$ <p>is <i>invariant under the message passing</i> since $\phi_C \phi_D / \phi_S$ is:</p> $\frac{\phi_C \phi_D \frac{\phi_C^{\perp S}}{\phi_S}}{\phi_C^{\perp S}} = \frac{\phi_C \phi_D}{\phi_S}.$ <p>After the message has been sent, D contains the D-marginal of $\phi_C \phi_D / \phi_S$.</p> <p>To see this, calculate</p> $\left(\frac{\phi_C \phi_D}{\phi_S} \right)^{\perp D} = \frac{\phi_D}{\phi_S} \phi_C^{\perp D} = \frac{\phi_D}{\phi_S} \phi_C^{\perp S}.$
<p style="text-align: center;">Separators</p> <p>Between any two cliques C and D which are neighbours in the junction tree we introduce their intersection $S = C \cap D$. In fact, S are the <i>minimal separators</i> appearing in the decomposition sequence.</p> <p>We also assign potentials to separators, initially $\phi_S \equiv 1$ for all $S \in \mathcal{S}$, where \mathcal{S} is the set of separators.</p> <p>We also let</p> $\kappa(x) = \frac{\prod_{C \in \mathcal{C}} \phi_C(x_C)}{\prod_{S \in \mathcal{S}} \phi_S(x_S)}, \quad (1)$ <p>and now it holds that $p(x x_E^*) = \kappa(x) / p(x_E^*)$.</p> <p>The expression (1) will be <i>invariant</i> under the message passing.</p>	<p style="text-align: center;">Second message</p> <p>If D returns message to C, the following happens:</p>  <p>Now all sets contain the relevant marginal of $\phi = \phi_C \phi_D / \phi_S$:</p>

<p>The separator contains</p> $\phi^{\perp S} = \left(\frac{\phi_C \phi_D}{\phi_S} \right)^{\perp S} = (\phi^{\perp D})^{\perp S} = \left(\phi_D \frac{\phi_C^{\perp S}}{\phi_S} \right)^{\perp S} = \frac{\phi_C^{\perp S} \phi_D^{\perp S}}{\phi_S}$ <p>C contains</p> $\phi_C \frac{\phi^{\perp S}}{\phi_C^{\perp S}} = \frac{\phi_C}{\phi_S} \phi_D^{\perp S} = \phi^{\perp C}$ <p>since, as before</p> $\left(\frac{\phi_C \phi_D}{\phi_S} \right)^{\perp C} = \frac{\phi_D}{\phi_S} \phi_C^{\perp D} = \frac{\phi_C}{\phi_S} \phi_D^{\perp S}$ <p>Further messages between C and D are neutral! Nothing will change if a message is repeated.</p>	<h3>Alternative scheduling of messages</h3> <p><i>Local control:</i></p> <p>Allow clique to send message if and only if it has already received message from all other neighbours. Such messages are <i>live</i>.</p> <p>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.</p> <p>Additional messages never do any harm (ignoring efficiency issues) as κ is invariant under message passing.</p> <p><i>Exactly two live messages along every branch is needed.</i></p>
<h3>Message passing</h3> <p>Two phases:</p> <ul style="list-style-type: none"> • COLLINFO: messages are sent from leaves towards arbitrarily chosen root R. After COLLINFO, the root potential satisfies $\phi_R(x_R) = p(x_R, x_E^*)$. • DISTINFO: messages are sent from root R towards leaves. After COLLINFO and subsequent DISTINFO, it holds for all $B \in \mathcal{C} \cup \mathcal{S}$ that $\phi_B(x_B) = p(x_B, x_E^*)$. <p>Hence $p(x_E^*) = \sum_{x_S} \phi_S(x_S)$ for any $S \in \mathcal{S}$ and $p(x_v x_E^*)$ can readily be computed from any ϕ_S with $v \in S$.</p>	<h3>Maximization</h3> <p>Replace sum-marginal with A-maxmarginal:</p> $\phi_B^{\perp A}(x) = \max_{y_B: y_A = x_A} \phi_B(y)$ <p>Satisfies <i>consistency</i>: $\phi^{\perp(A \cap B)} = (\phi^{\perp B})^{\perp A}$ and <i>distributivity</i>: $(\phi \phi_C)^{\perp B} = (\phi^{\perp B}) \phi_C$, if ϕ_C depends on x_C only and $C \subseteq B$.</p> <p>COLLINFO yields maximal value of density f. DISTINFO yields configuration with maximum probability. Viterbi decoding for HMMs is special case.</p> <p>Since (1) remains invariant, one can switch freely between max- and sum-propagation.</p>
<h3>COLLINFO</h3>  <p>Messages are sent from leaves towards root.</p>	<h3>Random propagation</h3> <p>After COLLINFO, the root potential is $\phi_R(x) \propto p(x_R x_E)$</p> <p>Modify DISTINFO as follows:</p> <ol style="list-style-type: none"> 1. Pick random configuration \tilde{x}_R from ϕ_R. 2. Send message to neighbours C as $\tilde{x}_{R \cap C} = \tilde{x}_S$ where $S = C \cap R$ is the separator. 3. Continue by picking \tilde{x}_C according to $\phi_C(x_{C \setminus S}, \tilde{x}_S)$ and send message further away from root. <p>When the sampling stops at leaves of junction tree, a configuration \tilde{x} has been generated from $p(x x_E^*)$.</p>
<h3>DISTINFO</h3>  <p>After COLLINFO, messages are sent from root towards leaves.</p>	<h3>References</h3> <p>Bahl, L., Cocke, J., Jelinek, F., and Raviv, J. (1974). Optimal decoding of linear codes for minimizing symbol error rate. <i>IEEE Transactions on Information Theory</i>, 20, 284–7.</p> <p>Baum, L. E. (1972). An equality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes. <i>Inequalities</i>, 3, 1–8.</p> <p>Cannings, C., Thompson, E. A., and Skolnick, M. H. (1976). Recursive derivation of likelihoods on pedigrees of arbitrary complexity. <i>Advances in Applied Probability</i>, 8, 622–5.</p> <p>Elston, R. C. and Stewart, J. (1971). A general model for</p>

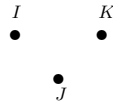
<p>the genetic analysis of pedigree data. <i>Human Heredity</i>, 21, 523–42.</p> <p>Jensen, F. V., Lauritzen, S. L., and Olesen, K. G. (1990). Bayesian updating in causal probabilistic networks by local computation. <i>Computational Statistics Quarterly</i>, 4, 269–82.</p> <p>Kalman, R. E. and Bucy, R. (1961). New results in linear filtering and prediction. <i>Journal of Basic Engineering</i>, 83 D, 95–108.</p> <p>Kong, A. (1986). <i>Multivariate belief functions and graphical models</i>. Ph.D. Thesis, Department of Statistics, Harvard University, Massachusetts.</p> <p>Lauritzen, S. L. and Jensen, F. V. (1997). Local computation with valuations from a commutative semigroup.</p>	<p>rithmic aspects of vertex elimination on graphs. <i>SIAM Journal on Computing</i>, 5, 266–83.</p> <p>Shenoy, P. P. and Shafer, G. (1986). Propagating belief functions using local propagation. <i>IEEE Expert</i>, 1, 43–52.</p> <p>Shenoy, P. P. and Shafer, G. (1990). Axioms for probability and belief-function propagation. In <i>Uncertainty in artificial intelligence 4</i>, (ed. R. D. Shachter, T. S. Levitt, L. N. Kanal, and J. F. Lemmer), pp. 169–98. North-Holland, Amsterdam, The Netherlands.</p> <p>Tarjan, R. E. (1985). Decomposition by clique separators. <i>Discrete Mathematics</i>, 55, 221–32.</p> <p>Tarjan, R. E. and Yannakakis, M. (1984). Simple linear-time algorithms to test chordality of graphs, test acyclicity of hypergraphs, and selectively reduce</p>
<p><i>Annals of Mathematics and Artificial Intelligence</i>, 21, 51–69.</p> <p>Lauritzen, S. L. and Spiegelhalter, D. J. (1988). Local computations with probabilities on graphical structures and their application to expert systems (with discussion). <i>Journal of the Royal Statistical Society, Series B</i>, 50, 157–224.</p> <p>Leimer, H.-G. (1993). Optimal decomposition by clique separators. <i>Discrete Mathematics</i>, 113, 99–123.</p> <p>Parter, S. (1961). The use of linear graphs in Gauss elimination. <i>SIAM Review</i>, 3, 119–30.</p> <p>Pearl, J. (1986). Fusion, propagation and structuring in belief networks. <i>Artificial Intelligence</i>, 29, 241–88.</p> <p>Rose, D. J., Tarjan, R. E., and Lueker, G. S. (1976). Algo-</p>	<p>acyclic hypergraphs. <i>SIAM Journal on Computing</i>, 13, 566–79.</p> <p>Thiele, T. N. (1880). Om Anvendelse af mindste Kvadraters Methode i nogle Tilfælde, hvor en Komplikation af visse Slags uensartede tilfældige Fejlkliller giver Fejlene en ‘systematisk’ Karakter. <i>Vidensk. Selsk. Skr. 5. Rk., naturvid. og mat. Afd.</i>, 12, 381–408. French version: <i>Sur la Compensation de quelques Erreurs quasi-systématiques par la Méthode des moindres Carrés</i>. Reitzel, København, 1880.</p> <p>Viterbi, A. J. (1967). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. <i>IEEE Transactions on Information Theory</i>, 13, 260–9.</p>

<p style="text-align: center;">Log-Linear and Gaussian Graphical Models</p> <p style="text-align: center;">Lecture 5 Saint Flour Summerschool, July 10, 2006</p> <p style="text-align: center;">Steffen L. Lauritzen, University of Oxford</p>	<p style="text-align: center;">Connecting to tradition</p> <p>This largely a matter of different notation.</p> <p>Assume data $X^1 = x^1, \dots, X^n = x^n$ and $V = \{I, J, K\}$.</p> <p>Write $i = 1, \dots, I$ for the possible values of X_I etc. and</p> $N_{ijk} = \{\nu : x^\nu = (i, j, k)\} ,$ <p>etc. Then $m_{ijk} = n f(x)$ and if $f(x) > 0$ and factorizes w.r.t. $\mathcal{A} = \{\{I, J\}, \{J, K\}\}$</p> $\log f(x) = \log \psi_{IJ}(x_I, x_J) + \log \psi_{JK}(x_J, x_K).$ <p>Thus if we let</p> $\alpha_{ij} = \log n + \log \psi_{IJ}(x_I, x_J), \quad \beta_{jk} = \log \psi_{JK}(x_J, x_K)$
<p style="text-align: center;">Overview of lectures</p> <ol style="list-style-type: none"> 1. Conditional independence and Markov properties 2. More on Markov properties 3. Graph decompositions and junction trees 4. Probability propagation and related algorithms 5. Log-linear and Gaussian graphical models 6. Conjugate prior families for graphical models 7. Hyper Markov laws 8. Structure learning and Bayes factors 9. More on structure learning. 	<p>we have</p> $\log m_{ijk} = \alpha_{ij} + \beta_{jk}.$ <p>The only difference is the assumption of positivity which is not necessary when using the multiplicative definition.</p> <p>It is typically an advantage to relax the restriction of positivity although it also creates technical difficulties.</p> <p>The logarithm of the factors $\phi_a = \log \psi_a$ are known as <i>interaction terms of order $a - 1$ or a-factor interactions</i>.</p> <p>Interaction terms of 0th order are called <i>main effects</i>.</p> <p>We also refer to the factors themselves using the same terms.</p>
<p style="text-align: center;">Log-linear models</p> <p>\mathcal{A} denotes a set of (pairwise incomparable) subsets of V.</p> <p>A density f (or function) <i>factorizes</i> w.r.t. \mathcal{A} if there exist functions $\psi_a(x)$ which depend on x_a only and</p> $f(x) = \prod_{a \in \mathcal{A}} \psi_a(x).$ <p>The set of distributions $\mathcal{P}_{\mathcal{A}}$ which factorize w.r.t. \mathcal{A} is the <i>hierarchical log-linear model</i> generated by \mathcal{A}.</p> <p>\mathcal{A} is the <i>generating class</i> of the log-linear model.</p> <p><i>No specific need to demand sets in \mathcal{A} to be incomparable. Only to avoid redundancy.</i></p>	<p style="text-align: center;">Dependence graph</p> <p>Any joint probability distribution P of $X = (X_v, v \in V)$ has a <i>dependence graph</i> $G = G(P) = (V, E(P))$.</p> <p>This is defined by letting $\alpha \not\perp \beta$ in $G(P)$ exactly when</p> $\alpha \perp\!\!\!\perp_P \beta \mid V \setminus \{\alpha, \beta\}.$ <p>X will then satisfy the pairwise Markov w.r.t. $G(P)$ and $G(P)$ is smallest with this property, i.e. P is <i>pairwise Markov</i> w.r.t. \mathcal{G} iff</p> $G(P) \subseteq \mathcal{G}.$ <p>The <i>dependence graph</i> $G(\mathcal{P})$ for a family \mathcal{P} is the smallest graph \mathcal{G} so that all $P \in \mathcal{P}$ are pairwise Markov w.r.t. \mathcal{G}:</p> $\alpha \perp\!\!\!\perp_P \beta \mid V \setminus \{\alpha, \beta\} \text{ for all } P \in \mathcal{P}.$
<p style="text-align: center;">Traditional notation</p> <p>Traditionally used for contingency tables, where e.g. m_{ijk} denotes the mean of the counts N_{ijk} in the cell (i, j, k) which has then been expanded as e.g.</p> $\log m_{ijk} = \alpha_i + \beta_j + \gamma_k \quad (1)$ <p>or</p> $\log m_{ijk} = \alpha_{ij} + \beta_{jk} \quad (2)$ <p>or</p> $\log m_{ijk} = \alpha_{ij} + \beta_{jk} + \gamma_{ik}, \quad (3)$ <p>or (with redundancy)</p> $\log m_{ijk} = \gamma + \delta_i + \phi_j + \eta_k + \alpha_{ij} + \beta_{jk} + \gamma_{ik}, \quad (4)$ <p>etc.</p>	<p style="text-align: center;">Dependence graph of log-linear model</p> <p>For any generating class \mathcal{A} we construct the dependence graph $G(\mathcal{A}) = G(\mathcal{P}_{\mathcal{A}})$ of the log-linear model $\mathcal{P}_{\mathcal{A}}$.</p> <p>This is determined by the relation</p> $\alpha \sim \beta \iff \exists a \in \mathcal{A} : \alpha, \beta \in a.$ <p>Sets in \mathcal{A} are clearly complete in $G(\mathcal{A})$ and therefore <i>distributions in $\mathcal{P}_{\mathcal{A}}$ factorize according to $G(\mathcal{A})$</i>.</p> <p>They are thus also global, local, and pairwise Markov w.r.t. $G(\mathcal{A})$.</p>

Independence

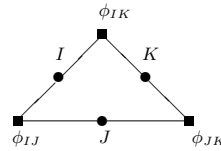
The log-linear model specified by (1) is known as the *main effects model*.

It has generating class consisting of singletons only $\mathcal{A} = \{\{I\}, \{J\}, \{K\}\}$. It has dependence graph



Thus it corresponds to *complete independence*.

Factor graphs



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

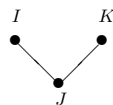
$$\alpha \sim a \iff \alpha \in a.$$

Using this graph even non-conformal log-linear models admit a simple visual representation.

Conditional independence

The log-linear model specified by (2) has no interaction between I and K .

It has generating class $\mathcal{A} = \{\{I, J\}, \{J, K\}\}$ and dependence graph



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

Separation in factor graphs

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

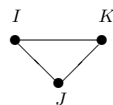
$$A \perp_{\mathcal{G}} B \mid S \iff A \perp_{\mathcal{F}} B \mid S$$

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.

No interaction of second order

The log-linear model specified by (3) has no second-order interaction. It has generating class $\mathcal{A} = \{\{I, J\}, \{J, K\}, \{I, K\}\}$ and its dependence graph



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

Data in list form

Consider a sample $X^1 = x^1, \dots, X^n = x^n$ 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
\vdots	\vdots	\vdots

Conformal log-linear models

As a generating class defines a dependence graph $G(\mathcal{A})$, the reverse is also true.

The set $\mathcal{C}(\mathcal{G})$ of cliques of \mathcal{G} is a generating class for the log-linear model of distributions which factorize w.r.t. \mathcal{G} .

If the dependence graph completely summarizes the restrictions imposed by \mathcal{A} , i.e. if $\mathcal{A} = \mathcal{C}(G(\mathcal{A}))$, \mathcal{A} is *conformal*.

The generating classes for the models given by (1) and (2) are conformal, whereas this is not the case for (3).

Contingency Table

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

Admitted	Sex	
	Male	Female
Yes	1198	557
No	1493	1278

The numerical entries are *cell counts*

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

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

<p style="text-align: center;">Likelihood function</p> <p>Assume now $p \in \mathcal{P}_{\mathcal{A}}$ but otherwise unknown. The likelihood function can be expressed as</p> $L(p) = \prod_{\nu=1}^n p(x^\nu) = \prod_{x \in \mathcal{X}} p(x)^{n(x)}.$ <p>In contingency table form the data follow a multinomial distribution</p> $P\{N(x) = n(x), x \in \mathcal{X}\} = \frac{n!}{\prod_{x \in \mathcal{X}} n(x)!} \prod_{x \in \mathcal{X}} p(x)^{n(x)}$ <p>but this only affects the likelihood function by a constant factor.</p>	<p>Hence</p> $\begin{aligned} L(p_{12}) &= \prod_x p_{12}(x)^{n(x)} \\ &= \prod_x \left\{ c \sqrt{p_1(x)p_2(x)} \right\}^{n(x)} \\ &= c^n \prod_x \sqrt{p_1(x)}^{n(x)} \prod_x \sqrt{p_2(x)}^{n(x)} \\ &= c^n \sqrt{L(p_1)L(p_2)} \\ &> \sqrt{L(p_1)L(p_2)} = L(p_1) = L(p_2), \end{aligned}$ <p>which contradicts (5). Hence we conclude $p_1 = p_2$.</p> <p>The extension to $\overline{\mathcal{P}_{\mathcal{A}}}$ is almost identical. It just needs a limit argument to establish $p_1, p_2 \in \mathcal{P}_{\mathcal{A}} \implies p_{12} \in \overline{\mathcal{P}_{\mathcal{A}}}$.</p>
<p style="text-align: center;">Properties of the likelihood function</p> <p>The likelihood function</p> $L(p) = \prod_{x \in \mathcal{X}} p(x)^{n(x)},$ <p>is continuous as a function of the $(\mathcal{X} -\text{dimensional vector})$ unknown probability distribution p.</p> <p>Since the closure $\overline{\mathcal{P}_{\mathcal{A}}}$ is compact (bounded and closed), L attains its maximum on $\overline{\mathcal{P}_{\mathcal{A}}}$ (not necessarily on $\mathcal{P}_{\mathcal{A}}$ itself).</p> <p>Indeed, it is also true that L has a unique maximum over $\overline{\mathcal{P}_{\mathcal{A}}}$, essentially because the likelihood function is log-concave.</p>	<p style="text-align: center;">Likelihood equations</p> <p>The maximum likelihood estimate \hat{p} of p is the unique element of $\overline{\mathcal{P}_{\mathcal{A}}}$ which satisfies the system of equations</p> $n\hat{p}(x_a) = n(x_a), \forall a \in \mathcal{A}, x_a \in \mathcal{X}_a. \quad (6)$ <p>Here $g(x_a) = \sum_{y: y_a = x_a} g(y)$ is the a-marginal of the function g.</p> <p>The system of equations (6) expresses the fitting of the marginals in \mathcal{A}.</p> <p>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.</p>
<p style="text-align: center;">Uniqueness of the MLE</p> <p>For simplicity, we only establish uniqueness within $\mathcal{P}_{\mathcal{A}}$. The proof is indirect.</p> <p>Assume $p_1, p_2 \in \mathcal{P}_{\mathcal{A}}$ with $p_1 \neq p_2$ and</p> $L(p_1) = L(p_2) = \sup_{p \in \mathcal{P}_{\mathcal{A}}} L(p). \quad (5)$ <p>Define</p> $p_{12}(x) = c \sqrt{p_1(x)p_2(x)},$ <p>where $c^{-1} = \{\sum_x \sqrt{p_1(x)p_2(x)}\}$ is a normalizing constant.</p>	<p style="text-align: center;">Proportional scaling</p> <p>To show that the equations (6) indeed have a solution, we simply describe a convergent algorithm which solves it. This cycles (repeatedly) through all the a-marginals in \mathcal{A} and fit them one by one.</p> <p>For $a \in \mathcal{A}$ define the following scaling operation on p:</p> $(T_a p)(x) \leftarrow p(x) \frac{n(x_a)}{np(x_a)}, \quad x \in \mathcal{X}$ <p>where $0/0 = 0$ and $b/0$ is undefined if $b \neq 0$.</p>
<p>Then $p_{12} \in \mathcal{P}_{\mathcal{A}}$ because</p> $\begin{aligned} p_{12}(x) &= c \sqrt{p_1(x)p_2(x)} \\ &= c \prod_{a \in \mathcal{A}} \sqrt{\psi_a^1(x)\psi_a^2(x)} = \prod_{a \in \mathcal{A}} \psi_a^{12}(x), \end{aligned}$ <p>where e.g. $\psi_a^{12} = c^{1/ \mathcal{A} } \sqrt{\psi_a^1(x)\psi_a^2(x)}$.</p> <p>The Cauchy-Schwarz inequality yields</p> $c^{-1} = \sum_x \sqrt{p_1(x)p_2(x)} < \sqrt{\sum_x p_1(x)} \sqrt{\sum_x p_2(x)} = 1.$	<p style="text-align: center;">Fitting the marginals</p> <p>The operation T_a fits the a-marginal if $p(x_a) > 0$ when $n(x_a) > 0$:</p> $\begin{aligned} n(T_a p)(x_a) &= n \sum_{y: y_a = x_a} p(y) \frac{n(y_a)}{np(y_a)} \\ &= n \frac{n(x_a)}{np(x_a)} \sum_{y: y_a = x_a} p(y) \\ &= n \frac{n(x_a)}{np(x_a)} p(x_a) = n(x_a). \end{aligned}$

Iterative Proportional Scaling

Make an ordering of the generators $\mathcal{A} = \{a_1, \dots, a_k\}$.
Define S by a full cycle of scalings

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

Define the iteration

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

It then holds that

$$\lim_{n \rightarrow \infty} p_n = \hat{p}$$

where \hat{p} is the unique maximum likelihood estimate of $p \in \overline{\mathcal{P}}_{\mathcal{A}}$, i.e. the solution of the equation system (6).

Fitting S -marginal

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

For example

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

and so on.

Iterative Proportional Fitting

Known as the *IPS*-algorithm or *IPF*-algorithm, or as a variety of other names. Implemented e.g. (inefficiently) in *R* in `loglin` with front end `loglm` in *MASS*.

Key elements in proof:

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

Fitting A -marginal

Sex	Admitted		S -marginal
	Yes	No	
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.

Algorithm has converged, as both marginals now fit!

A simple example

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

Admissions data from Berkeley. Consider $A \perp\!\!\!\perp S$, corresponding to $\mathcal{A} = \{\{A\}, \{S\}\}$.

We should fit A -marginal and S -marginal iteratively.

Normalised to probabilities

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

Dividing everything by 4526 yields \hat{p} .

It is overkill to use the IPS algorithm as there is an explicit formula in this case.

Initial values

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

Entries all equal to 4526/4. Gives initial values of np_0 .

IPS by probability propagation

The IPS-algorithm performs the scaling operations T_a :

$$p(x) \leftarrow p(x) \frac{n(x_a)}{np(x_a)}, \quad x \in \mathcal{X}. \quad (7)$$

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

Jiroušek and Přeučil (1995) realized that the algorithm could be implemented using probability propagation:

A chordal graph \mathcal{G} with cliques \mathcal{C} so that for all $a \in \mathcal{A}$, a are complete subsets of \mathcal{G} is a *chordal cover* of \mathcal{A} .

1. Find chordal cover \mathcal{G} of \mathcal{A} ;

<p>2. Arrange cliques \mathcal{C} of \mathcal{G} in a junction tree;</p> <p>3. Represent p implicitly as</p> $p(x) = \frac{\prod_{C \in \mathcal{C}} \psi_C(x)}{\prod_{S \in \mathcal{S}} \psi_S(x)};$ <p>4. Replace the step (7) with</p> $\psi_C(x_C) \leftarrow \psi_C(x_C) \frac{n(x_a)}{np(x_a)}, \quad x_C \in \mathcal{X}_C,$ <p>where $a \subseteq C$ and $p(x_a)$ is calculated by <i>probability propagation</i>.</p> <p>Since the scaling only involves \mathcal{X}_C, this is possible just if $\max_{C \in \mathcal{C}} \mathcal{X}_C$ is of a reasonable size.</p>	<p style="text-align: center;">Marginal and conditional distributions</p> <p>Partition X into X_1 and X_2, where $X_1 \in \mathcal{R}^r$ and $X_2 \in \mathcal{R}^s$ with $r + s = d$.</p> <p>Partition mean vector, concentration and covariance matrix accordingly as</p> $\xi = \begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix}, \quad K = \begin{pmatrix} K_{11} & K_{12} \\ K_{21} & K_{22} \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$ <p>so that Σ_{11} is $r \times r$ and so on. Then, if $X \sim \mathcal{N}_d(\xi, \Sigma)$</p> $X_2 \sim \mathcal{N}_s(\xi_2, \Sigma_{22})$ <p>and</p> $X_1 X_2 = x_2 \sim \mathcal{N}_r(\xi_{1 2}, \Sigma_{1 2}),$
<p style="text-align: center;">Closed form maximum likelihood</p> <p>In some cases the IPS algorithm converges after a finite number of cycles.</p> <p>An explicit formula is then available for the MLE of $p \in \mathcal{P}_A$.</p> <p>A generating class \mathcal{A} is called <i>decomposable</i> if $\mathcal{A} = \mathcal{C}$ (i.e. \mathcal{A} is conformal) and \mathcal{C} are the cliques of a chordal graph \mathcal{G}.</p> <p>The IPS-algorithm converges after a finite number of cycles (at most two) if and only if \mathcal{A} is decomposable.</p> <p>$\mathcal{A} = \{\{1, 2\}, \{2, 3\}, \{1, 3\}\}$ is the smallest non-conformal generating class, demanding proper iteration.</p>	<p>where</p> $\xi_{1 2} = \xi_1 + \Sigma_{12} \Sigma_{22}^{-1} (x_2 - \xi_2) \quad \text{and} \quad \Sigma_{1 2} = \Sigma_{11} - \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21}.$ <p>Σ_{22}^{-} is an arbitrary generalized inverse to Σ_{22}.</p> <p>In the regular case it also holds that</p> $K_{11}^{-1} = \Sigma_{11} - \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21} \quad (9)$ <p>and</p> $K_{11}^{-1} K_{12} = -\Sigma_{12} \Sigma_{22}^{-1}, \quad (10)$ <p>so then,</p> $\xi_{1 2} = \xi_1 - K_{11}^{-1} K_{12} (x_2 - \xi_2) \quad \text{and} \quad \Sigma_{1 2} = K_{11}^{-1}.$ <p>In particular, if $\Sigma_{12} = 0$, X_1 and X_2 are independent.</p>
<p style="text-align: center;">Explicit formula for MLE</p> <p>Let \mathcal{S} be the set of <i>minimal separators</i> of the chordal graph \mathcal{G}. The MLE for p under the log-linear model with generating class $\mathcal{A} = \mathcal{C}(\mathcal{G})$ is</p> $\hat{p}(x) = \frac{\prod_{C \in \mathcal{C}} n(x_C)}{n \prod_{S \in \mathcal{S}} n(x_S)^{\nu(S)}}$ <p>where $\nu(S)$ is the number of times S appears as an intersection $a \cap b$ of neighbours in a junction tree \mathcal{T} with \mathcal{A} as vertex set.</p> <p>Contrast this with the factorization of the probability function itself:</p> $p(x) = \frac{\prod_{C \in \mathcal{C}} p(x_C)}{\prod_{S \in \mathcal{S}} p(x_S)^{\nu(S)}}.$	<p style="text-align: center;">Gaussian likelihoods</p> <p>Consider $\xi = 0$ and a sample $X^1 = x^1, \dots, X^n = x^n \sim \mathcal{N}_d(0, \Sigma)$ with Σ regular.</p> <p>Using (8), we get the likelihood function</p> $\begin{aligned} L(K) &= (2\pi)^{-nd/2} (\det K)^{n/2} e^{-\sum_{\nu=1}^n (x^\nu)^\top K x^\nu / 2} \\ &\propto (\det K)^{n/2} e^{-\text{tr}\{K \sum_{\nu=1}^n x^\nu (x^\nu)^\top\} / 2} \\ &= (\det K)^{n/2} e^{-\text{tr}(KW) / 2}. \end{aligned} \quad (11)$ <p>where</p> $W = \sum_{\nu=1}^n x^\nu (x^\nu)^\top$ <p>is the matrix of <i>sums of squares and products</i>.</p>
<p style="text-align: center;">Density of multivariate Gaussian</p> <p>If Σ is <i>positive definite</i>, i.e. if $\lambda^\top \Sigma \lambda > 0$ for $\lambda \neq 0$, the distribution has density w.r.t. Lebesgue measure on \mathcal{R}^d</p> $f(x \xi, \Sigma) = (2\pi)^{-d/2} (\det K)^{1/2} e^{-(x-\xi)^\top K (x-\xi) / 2}, \quad (8)$ <p>where $K = \Sigma^{-1}$ is the <i>concentration matrix</i> of the distribution. We then also say that Σ is <i>regular</i>.</p>	<p style="text-align: center;">Wishart distribution</p> <p>The Wishart distribution is the sampling distribution of the matrix of sums of squares and products. More precisely:</p> <p>A random $d \times d$ matrix S has a <i>d-dimensional Wishart distribution</i> with parameter Σ and n degrees of freedom if</p> $W \stackrel{D}{=} \sum_{i=1}^n X^\nu (X^\nu)^\top$ <p>where $X^\nu \sim \mathcal{N}_d(0, \Sigma)$. We then write</p> $W \sim \mathcal{W}_d(n, \Sigma).$ <p>The Wishart is the multivariate analogue to the χ^2:</p> $\mathcal{W}_1(n, \sigma^2) = \sigma^2 \chi^2(n).$

<p>If $W \sim \mathcal{W}_d(n, \Sigma)$ its mean is $\mathbf{E}(W) = n\Sigma$.</p> <p>If W_1 and W_2 are independent with $W_i \sim \mathcal{W}_d(n_i, \Sigma)$, then</p> $W_1 + W_2 \sim \mathcal{W}_d(n_1 + n_2, \Sigma).$ <p>If A is an $r \times d$ matrix and $W \sim \mathcal{W}_d(n, \Sigma)$, then</p> $AWA^\top \sim \mathcal{W}_r(n, A\Sigma A^\top).$ <p>For $r = 1$ we get that when $W \sim \mathcal{W}_d(n, \Sigma)$ and $\lambda \in R^d$,</p> $\lambda^\top W \lambda \sim \sigma_\lambda^2 \chi^2(n),$ <p>where $\sigma_\lambda^2 = \lambda^\top \Sigma \lambda$.</p>	<p style="text-align: center;">Likelihood function</p> <p>The likelihood function based on a sample of size n is</p> $L(K) \propto (\det K)^{n/2} e^{-\text{tr}(KW)/2},$ <p>where W is the Wishart matrix of sums of squares and products, $W \sim \mathcal{W}_{ V }(n, \Sigma)$ with $\Sigma^{-1} = K \in \mathcal{S}^+(\mathcal{G})$.</p> <p>For any matrix A we let $A(\mathcal{G}) = \{a(\mathcal{G})_{\alpha\beta}\}$ where</p> $a(\mathcal{G})_{\alpha\beta} = \begin{cases} a_{\alpha\beta} & \text{if } \alpha = \beta \text{ or } \alpha \sim \beta \\ 0 & \text{otherwise.} \end{cases}$ <p>Then, as $K \in \mathcal{S}(\mathcal{G})$</p> $\text{tr}(KW) = \text{tr}\{KW(\mathcal{G})\}.$
<p style="text-align: center;">Wishart density</p> <p>If $W \sim \mathcal{W}_d(n, \Sigma)$, where Σ is regular, then</p> <p>W is regular with probability one if and only if $n \geq d$.</p> <p>When $n \geq d$ the Wishart distribution has density</p> $f_d(w n, \Sigma) = c(d, n)^{-1} (\det \Sigma)^{-n/2} (\det w)^{(n-d-1)/2} e^{-\text{tr}(\Sigma^{-1}w)/2}$ <p>w.r.t. Lebesgue measure on the set of positive definite matrices.</p> <p>The Wishart constant $c(d, n)$ is</p> $c(d, n) = 2^{nd/2} (2\pi)^{d(d-1)/4} \prod_{i=1}^d \Gamma\{(n+1-i)/2\}.$	<p>Hence we can identify the family as a (regular and canonical) exponential family with elements of $W(\mathcal{G})$ as canonical sufficient statistics and the likelihood equations</p> $\mathbf{E}\{W(\mathcal{G})\} = n\Sigma(\mathcal{G}) = w(\mathcal{G})_{\text{obs}}.$ <p>Alternatively we can write the equations as</p> $n\hat{\sigma}_{vv} = w_{vv}, \quad n\hat{\sigma}_{\alpha\beta} = w_{\alpha\beta}, \quad v \in V, \{\alpha, \beta\} \in E,$ <p>with the model restriction $\Sigma^{-1} \in \mathcal{S}^+(\mathcal{G})$.</p> <p>This 'fits variances and covariances along nodes and edges in \mathcal{G}' so we can write the equations as</p> $n\hat{\Sigma}_{cc} = w_{cc} \text{ for all cliques } c \in \mathcal{C}(\mathcal{G}),$ <p>hence making the equations analogous to the discrete case.</p>
<p style="text-align: center;">Conditional independence</p> <p>Consider $X = (X_1, \dots, X_V) \sim \mathcal{N}_{ V }(0, \Sigma)$ with Σ regular and $K = \Sigma^{-1}$.</p> <p>The concentration matrix of the conditional distribution of (X_α, X_β) given $X_{V \setminus \{\alpha, \beta\}}$ is</p> $\tilde{K}_{\{\alpha, \beta\}} = \begin{pmatrix} k_{\alpha\alpha} & k_{\alpha\beta} \\ k_{\beta\alpha} & k_{\beta\beta} \end{pmatrix}.$ <p>Hence</p> $\alpha \perp\!\!\!\perp \beta V \setminus \{\alpha, \beta\} \iff k_{\alpha\beta} = 0.$ <p>Thus the dependence graph $\mathcal{G}(K)$ of a regular Gaussian distribution is given by</p> $\alpha \not\sim \beta \iff k_{\alpha\beta} = 0.$	<p style="text-align: center;">Iterative Proportional Scaling</p> <p>For $K \in \mathcal{S}^+(\mathcal{G})$ and $c \in \mathcal{C}$, define the operation of 'adjusting the c-marginal' as follows. Let $a = V \setminus c$ and</p> $T_c K = \begin{pmatrix} n(w_{cc})^{-1} + K_{ca}(K_{aa})^{-1}K_{ac} & K_{ca} \\ K_{ac} & K_{aa} \end{pmatrix}. \quad (12)$ <p>This operation is clearly well defined if w_{cc} is positive definite.</p> <p>Exploiting that it holds in general that</p> $(K^{-1})_{cc} = \Sigma_{cc} = \{K_{cc} - K_{ca}(K_{aa})^{-1}K_{ac}\}^{-1},$ <p>we find the covariance $\tilde{\Sigma}_{cc}$ corresponding to the adjusted</p>
<p style="text-align: center;">Graphical models</p> <p>$\mathcal{S}(\mathcal{G})$ denotes the symmetric matrices A with $a_{\alpha\beta} = 0$ unless $\alpha \sim \beta$ and $\mathcal{S}^+(\mathcal{G})$ their positive definite elements.</p> <p>A Gaussian graphical model for X specifies X as multivariate normal with $K \in \mathcal{S}^+(\mathcal{G})$ and otherwise unknown.</p> <p>Note that the density then factorizes as</p> $\log f(x) = \text{constant} - \frac{1}{2} \sum_{\alpha \in V} k_{\alpha\alpha} x_\alpha^2 - \sum_{\{\alpha, \beta\} \in E} k_{\alpha\beta} x_\alpha x_\beta,$ <p>hence no interaction terms involve more than pairs..</p> <p>This is different from the discrete case and generally makes things easier.</p>	<p>concentration matrix becomes</p> $\begin{aligned} \tilde{\Sigma}_{cc} &= \{(T_c K)^{-1}\}_{cc} \\ &= \{n(w_{cc})^{-1} + K_{ca}(K_{aa})^{-1}K_{ac} - K_{ca}(K_{aa})^{-1}K_{ac}\}^{-1} \\ &= w_{cc}/n, \end{aligned}$ <p>hence $T_c K$ does indeed adjust the marginals.</p> <p>From (12) 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.</p> <p>In fact, T_c scales proportionally in the sense that</p> $f\{x (T_c K)^{-1}\} = f(x K^{-1}) \frac{f(x_c w_{cc}/n)}{f(x_c \Sigma_{cc})}.$ <p>This clearly demonstrates the analogy to the discrete case.</p>

<p>Next we choose any ordering (c_1, \dots, c_k) of the cliques in \mathcal{G}. Choose further $K_0 = I$ and define for $r = 0, 1, \dots$</p> $K_{r+1} = (T_{c_1} \cdots T_{c_k})K_r.$ <p>Then we have: Consider a sample from a covariance selection model with graph \mathcal{G}. Then</p> $\hat{K} = \lim_{r \rightarrow \infty} K_r,$ <p>provided the maximum likelihood estimate \hat{K} of K exists.</p> <p>The general problem of existence of the MLE is non-trivial: If $n < \sup_{a \in \mathcal{A}} a$ the MLE does not exist. If $n \geq \sup_{C \in \mathcal{C}} C$, where \mathcal{C} are the cliques of a chordal cover of \mathcal{A} the MLE exists with probability one.</p>	<p style="text-align: center;">Maximum likelihood estimates</p> <p>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.</p> $([A]^V)_{\gamma\mu} = \begin{cases} a_{\gamma\mu} & \text{if } \gamma \in d, \mu \in e \\ 0 & \text{otherwise.} \end{cases}$ <p>The maximum likelihood estimates exists if and only if $n \geq C$ for all $C \in \mathcal{C}$. Then the following simple formula holds for the maximum likelihood estimate of K:</p> $\hat{K} = n \left\{ \sum_{C \in \mathcal{C}} [(w_C)^{-1}]^V - \sum_{S \in \mathcal{S}} \nu(S) [(w_S)^{-1}]^V \right\}.$
<p>For n between these values the general situation is unclear. For the k-cycle it holds (Buhl 1993) that for $n = 2$,</p> $P\{\text{MLE exists} \mid \Sigma = I\} = 1 - \frac{2}{k-1!},$ <p>whereas for $n = 1$ the MLE does not exist and for $n \geq 3$ the MLE exists with probability one, as a k-cycle has a chordal cover with maximal clique size 3.</p>	<p>The determinant of the MLE is</p> $\det(\hat{K}) = \frac{\prod_{S \in \mathcal{S}} \{\det(w_S)\}^{\nu(S)}}{\prod_{C \in \mathcal{C}} \det(w_C)} n^{ V }.$
<p style="text-align: center;">Chordal graphs</p> <p>If the graph \mathcal{G} is chordal, we say that the graphical model is decomposable.</p> <p>In this case, the IPS-algorithm converges in a finite number of steps, as in the discrete case.</p> <p>We also have the familiar factorization of densities</p> $f(x \mid \Sigma) = \frac{\prod_{C \in \mathcal{C}} f(x_C \mid \Sigma_C)}{\prod_{S \in \mathcal{S}} f(x_S \mid \Sigma_S)^{\nu(S)}} \quad (13)$ <p>where $\nu(S)$ is the number of times S appear as intersection between neighbouring cliques of a junction tree for \mathcal{C}.</p>	<p style="text-align: center;">References</p> <p>Buhl, S. L. (1993). On the existence of maximum likelihood estimators for graphical Gaussian models. <i>Scandinavian Journal of Statistics</i>, 20, 263–70.</p> <p>Jiroušek, R. and Přeučil, R. (1995). On the effective implementation of the iterative proportional fitting procedure. <i>Computational Statistics and Data Analysis</i>, 19, 177–89.</p>
<p style="text-align: center;">Relations for trace and determinant</p> <p>Using the factorization (13) we can match the expressions for the trace and determinant to obtain</p> $\text{tr}(KW) = \sum_{C \in \mathcal{C}} \text{tr}(K_C W_C) - \sum_{S \in \mathcal{S}} \nu(S) \text{tr}(K_S W_S)$ <p>and further</p> $\begin{aligned} \det \Sigma &= \{\det(K)\}^{-1} = \frac{\prod_{C \in \mathcal{C}} \det\{(K^{-1})_C\}}{\prod_{S \in \mathcal{S}} [\det\{(K^{-1})_S\}]^{\nu(S)}} \\ &= \frac{\prod_{C \in \mathcal{C}} \det\{\Sigma_C\}}{\prod_{S \in \mathcal{S}} \{\det(\Sigma_S)\}^{\nu(S)}} \end{aligned}$	

<p style="text-align: center;">Hyper Markov Laws</p> <p style="text-align: center;">Lecture 6 Saint Flour Summerschool, July 13, 2006</p> <p style="text-align: center;">Steffen L. Lauritzen, University of Oxford</p>	<p style="text-align: center;">Conformal log-linear model</p> <p>The set $\mathcal{C}(\mathcal{G})$ of cliques of \mathcal{G} is a generating class for the log-linear model of distributions which factorize w.r.t. \mathcal{G}.</p> <p>If the dependence graph completely summarizes the restrictions imposed by \mathcal{A}, i.e. if $\mathcal{A} = \mathcal{C}(\mathcal{G}(\mathcal{A}))$, \mathcal{A} is <i>conformal</i>.</p> <p><i>Conformal log-linear models can be completely described in terms of conditional independence.</i></p> <p>For more general log-linear models <i>factor graphs</i> are needed to yield a faithful representation of the factorization. MIM (software by David Edwards www.hypergraph.dk), uses the term <i>interaction graph</i>.</p>
<p style="text-align: center;">Overview of lectures</p> <ol style="list-style-type: none"> 1. Conditional independence and Markov properties 2. More on Markov properties 3. Graph decompositions and junction trees 4. Probability propagation and related algorithms 5. Log-linear and Gaussian graphical models 6. <i>Hyper Markov laws</i> 7. More on hyper Markov laws 8. Structure estimation and Bayes factors 9. More on structure estimation. 	<p style="text-align: center;">Likelihood equations</p> <p>For any generating class \mathcal{A} it holds that <i>the maximum likelihood estimate \hat{p} of p is the unique element of $\mathcal{P}_{\mathcal{A}}$ which satisfies the system of equations</i></p> $n\hat{p}(x_a) = n(x_a), \forall a \in \mathcal{A}, x_a \in \mathcal{X}_a. \quad (1)$ <p>The system of equations (1) expresses the <i>fitting of the marginals</i> in \mathcal{A}.</p> <p>In general, the equations cannot be solved explicitly, but iterative methods are needed.</p>
<p style="text-align: center;">Log-linear models</p> <p>\mathcal{A} denotes a set of (pairwise incomparable) subsets of V. A density f factorizes w.r.t. \mathcal{A}</p> $f(x) = \prod_{a \in \mathcal{A}} \psi_a(x).$ <p>The set of distributions $\mathcal{P}_{\mathcal{A}}$ which factorize w.r.t. \mathcal{A} is the <i>hierarchical log-linear model</i> generated by the \mathcal{A}. \mathcal{A} is the <i>generating class</i> of the log-linear model.</p>	<p style="text-align: center;">Iterative Proportional Scaling</p> <p>For $a \in \mathcal{A}$ define the <i>scaling operation</i> on p:</p> $(T_a p)(x) \leftarrow p(x) \frac{n(x_a)}{np(x_a)}, \quad x \in \mathcal{X}. \quad (2)$ <p>The operation T_a fits the a-marginal. Now, make any ordering of the generators $\mathcal{A} = \{a_1, \dots, a_k\}$. Define S by</p> $Sp = T_{a_k} \cdots T_{a_2} T_{a_1} p.$ <p>Let $p_0(x) \leftarrow 1/ \mathcal{X}$, $p_n = Sp_{n-1}, n = 1, \dots$</p> <p><i>It then holds that $\lim_{n \rightarrow \infty} p_n = \hat{p}$ where \hat{p} is the unique maximum likelihood estimate of $p \in \mathcal{P}_{\mathcal{A}}$.</i></p> <p>It is easy to show that $\hat{p}(x) > 0$ for all $x \in \mathcal{X}$ if and only if $\hat{p} \in \mathcal{P}_{\mathcal{A}}$.</p>
<p style="text-align: center;">Dependence graph</p> <p>The <i>dependence graph</i> $\mathcal{G}(\mathcal{P})$ for a family of distributions \mathcal{P} is the smallest graph \mathcal{G} so that</p> $\alpha \perp\!\!\!\perp \beta \mid V \setminus \{\alpha, \beta\} \text{ for all } P \in \mathcal{P}.$ <p>The <i>dependence graph</i> of a log-linear model $\mathcal{P}_{\mathcal{A}}$ is then determined by</p> $\alpha \sim \beta \iff \exists a \in \mathcal{A} : \alpha, \beta \in a.$ <p>Sets in \mathcal{A} are complete in $\mathcal{G}(\mathcal{A})$ and therefore <i>distributions in $\mathcal{P}_{\mathcal{A}}$ factorize according to $\mathcal{G}(\mathcal{A})$.</i></p> <p>They are also global, local, and pairwise Markov w.r.t. $\mathcal{G}(\mathcal{A})$.</p>	<p style="text-align: center;">IPS by probability propagation</p> <p>A <i>chordal cover</i> of \mathcal{A} is a chordal graph \mathcal{G} so that for all $a \in \mathcal{A}$, a are complete subsets of \mathcal{G}.</p> <ol style="list-style-type: none"> 1. Find chordal cover \mathcal{G} of \mathcal{A} and arrange cliques C of \mathcal{G} in a junction tree; 2. Represent p implicitly as $p(x) = \frac{\prod_{C \in \mathcal{C}} \psi_C(x)}{\prod_{S \in \mathcal{S}} \psi_S(x)}$; 3. Replace (2) with $\psi_C(x_C) \leftarrow \psi_C(x_C) \frac{n(x_a)}{np(x_a)}, \quad x_C \in \mathcal{X}_C,$ <p>where $a \subseteq C$ and $p(x_a)$ is calculated by <i>probability propagation</i>.</p>

<p style="text-align: center;">Closed form maximum likelihood</p> <p>\mathcal{A} is decomposable if $\mathcal{A} = \mathcal{C}$ where \mathcal{C} are the cliques of a chordal graph.</p> <p>The IPS-algorithm converges after at a finite number of cycles (at most two) if and only if \mathcal{A} is decomposable.</p> <p>The MLE for p under the log-linear model $\mathcal{A} = \mathcal{C}(\mathcal{G})$ is</p> $\hat{p}(x) = \frac{\prod_{C \in \mathcal{C}} n(x_C)}{n \prod_{S \in \mathcal{S}} n(x_S)^{\nu(S)},}$ <p>where $\nu(S)$ is the usual multiplicity of a separator.</p> <p>In fact, with a suitably chosen ordering (e.g. MCS) of the cliques, the IPS-algorithm converges in a single cycle.</p>	<p style="text-align: center;">Existence of the MLE</p> <p>The general problem of existence of the MLE is non-trivial:</p> <p>If $n < \sup_{a \in \mathcal{A}} a$ the MLE does not exist.</p> <p>If $n \geq \sup_{C \in \mathcal{C}} C$, where \mathcal{C} are the cliques of a chordal cover of \mathcal{A} the MLE exists with probability one.</p> <p>For n between these values the general situation is unclear.</p> <p>For the k-cycle it holds (Buhl 1993) that for $n = 2$,</p> $P\{\text{MLE exists} \mid \Sigma = I\} = 1 - \frac{2}{(k-1)!},$ <p>whereas for $n = 1$ the MLE does not exist and for $n \geq 3$ the MLE exists with probability one, as a k-cycle has a chordal cover with maximal clique size 3.</p>
<p style="text-align: center;">Gaussian likelihood function</p> <p>The likelihood function based on a sample of size n is</p> $L(K) \propto (\det K)^{n/2} e^{-\text{tr}(KW)/2},$ <p>where W is the Wishart matrix of sums of squares and products, $W \sim \mathcal{W}_{ V }(n, \Sigma)$ with $\Sigma^{-1} = K \in \mathcal{S}^+(\mathcal{G})$, where $\mathcal{S}^+(\mathcal{G})$ are the positive definite matrices with $\alpha \not\sim \beta \implies k_{\alpha\beta} = 0$.</p> <p>The MLE of \hat{K} is the unique element of $\mathcal{S}^+(\mathcal{G})$ satisfying</p> $n \hat{\Sigma}_{cc} = w_{cc} \text{ for all cliques } c \in \mathcal{C}(\mathcal{G}).$	<p style="text-align: center;">Special Wishart distributions</p> <p>The formula</p> $\hat{\Sigma} = n \left\{ \sum_{C \in \mathcal{C}} [(W_C)^{-1}]^V - \sum_{S \in \mathcal{S}} \nu(S) [(W_S)^{-1}]^V \right\}^{-1}$ <p>specifies $\hat{\Sigma}$ as a random matrix.</p> <p>The distribution of this random Wishart-type matrix is partly reflecting Markov properties of the graph \mathcal{G}.</p> <p>This is also true for the distribution of $\hat{\Sigma}$ for a non-chordal graph \mathcal{G} but not to the same degree.</p> <p>Before we delve further into this, we shall need some more terminology.</p>
<p style="text-align: center;">Iterative Proportional Scaling</p> <p>For $K \in \mathcal{S}^+(\mathcal{G})$ and $c \in \mathcal{C}$, define the operation of 'adjusting the c-marginal' as follows. Let $a = V \setminus c$ and</p> $T_c K = \begin{pmatrix} n(w_{cc})^{-1} + K_{ca}(K_{aa})^{-1}K_{ac} & K_{ca} \\ K_{ac} & K_{aa} \end{pmatrix}. \quad (3)$ <p>Next we choose any ordering (c_1, \dots, c_k) of the cliques in \mathcal{G}. Choose further $K_0 = I$ and define for $r = 0, 1, \dots$</p> $K_{r+1} = (T_{c_1} \cdots T_{c_k}) K_r.$ <p>It then holds that $\hat{K} = \lim_{r \rightarrow \infty} K_r$, provided the maximum likelihood estimate \hat{K} of K exists.</p>	<p style="text-align: center;">Laws and distributions</p> <p>Families of distributions may not always be simply parameterized, or we may want to describe the families without specific reference to a parametrization.</p> <p>Generally we think of</p> $\mathcal{P} = \{P_\theta, \theta \in \Theta\}$ <p>and sometimes identify Θ with \mathcal{P} which is justified when the parametrization</p> $\theta \rightarrow P_\theta$ <p>is one-to-one and onto.</p> <p>In a Gaussian graphical model $\theta = K \in \mathcal{S}^+(\mathcal{G})$ is uniquely identifying any regular Gaussian distribution satisfying the Markov properties w.r.t. \mathcal{G}.</p>
<p style="text-align: center;">Chordal graphs</p> <p>If the graph \mathcal{G} is chordal, we say that the graphical model is decomposable.</p> <p>In this case, the IPS-algorithm converges in at most two cycles, as in the discrete case.</p> <p>The maximum likelihood estimates exists if and only if $n \geq \mathcal{C}$ for all $C \in \mathcal{C}$. Then</p> $\hat{K} = n \left\{ \sum_{C \in \mathcal{C}} [(w_C)^{-1}]^V - \sum_{S \in \mathcal{S}} \nu(S) [(w_S)^{-1}]^V \right\}.$ <p>the symbol $[A]^V$ denotes for $A = \{a_{\gamma\mu}\}_{\gamma \in d, \mu \in e}$ the matrix obtained from A by filling up with zero entries to obtain full dimension.</p>	<p>The case when $\mathcal{P} = \mathcal{P}_{\mathcal{A}}$ is more complex, and a specific parametrization needs to be chosen to make a simple and one-to-one correspondence.</p> <p>In any case, any probability measure on \mathcal{P} (or on Θ) represents a random element of \mathcal{P}, i.e. a random distribution. The sampling distribution of the MLE \hat{p} is an example of such a measure.</p> <p>To keep heads straight we refer to a probability measure on \mathcal{P} as a law, whereas a distribution is a probability measure on \mathcal{X}.</p> <p>Thus we shall e.g. speak of the Wishart law as we think of it specifying a distribution of $f(\cdot \mid \Sigma)$.</p>

Hyper Markov Laws

We identify $\theta \in \Theta$ and $P_\theta \in \mathcal{P}$, 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 \mathcal{L} on Θ we write

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

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

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

Note the conditional independence is only required to hold for *graph decompositions*.

Clearly, it holds that \hat{p} is Markov on \mathcal{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

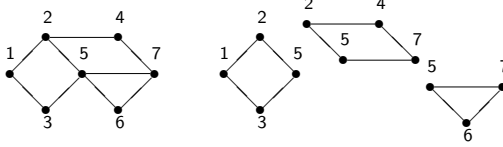
$$\{N_{ij+}\} \perp\!\!\!\perp \{X_j^{(n)}\} | \{N_{+j+}\}$$

and hence

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

which yields the hyper Markov property.

Hyper Markov property



If θ follows a hyper Markov law for this graph, it holds for example that

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

We shall later show that *this is indeed true for $\hat{\theta} = \hat{p}$ or $\hat{\Sigma}$ in the graphical model with this graph*, i.e.

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

Chordal graphs

If \mathcal{G} is chordal and θ is hyper Markov on \mathcal{G} , it holds that

$$A \perp\!\!\!\perp_{\mathcal{G}} B | S \implies A \perp\!\!\!\perp_{\mathcal{L}} B | S$$

i.e. it is not necessary to specify that S is a complete separator to obtain the relevant conditional independence.

This follows essentially because for a chordal graph it holds that

$$A \perp\!\!\!\perp_{\mathcal{G}} B | S \implies \exists S^* \subseteq S : A \perp\!\!\!\perp_{\mathcal{G}} B | S^* \text{ with } S^* \text{ complete.}$$

If \mathcal{G} is not chordal, we can form $\bar{\mathcal{G}}$ by completing all prime components of \mathcal{G} .

Consequences of the hyper Markov property

Clearly, if $A \perp\!\!\!\perp_{\mathcal{L}} B | S$, we have for example also (using property (C2) of conditional independence)

$$\theta_A \perp\!\!\!\perp_{\mathcal{L}} \theta_B | \theta_S$$

since θ_A and θ_B are functions of $\theta_{A \cup S}$ and $\theta_{B \cup S}$ respectively.

But *the converse is false!* $\theta_A \perp\!\!\!\perp_{\mathcal{L}} \theta_B | \theta_S$ does *not* imply $\theta_{A \cup S} \perp\!\!\!\perp_{\mathcal{L}} \theta_{B \cup S} | \theta_S$, since $\theta_{A \cup S}$ is *not* a function of (θ_A, θ_S) . In contrast, $X_{A \cup B}$ is indeed a (one-to-one) function of (X_A, X_B) .

However *it generally holds that*

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

Then if θ is hyper Markov on \mathcal{G} , it is also hyper Markov on $\bar{\mathcal{G}}$, and thus

$$A \perp\!\!\!\perp_{\bar{\mathcal{G}}} B | S \implies A \perp\!\!\!\perp_{\mathcal{L}} B | S.$$

But the similar result would be *false* for an arbitrary chordal cover of \mathcal{G} .

Simple example

Consider the conditional independence model with graph



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

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

and

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

Directed hyper Markov property

We have similar notions and results in the directed case.

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

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

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

$$\theta_{v \cup \text{pa}(v)} \perp\!\!\!\perp_{\mathcal{L}} \theta_{\text{pr}(v)} | \theta_{\text{pa}(v)}.$$

In general there is no similar statement corresponding to the global property and *d*-separation.

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

References

Buhl, S. L. (1993). On the existence of maximum likelihood estimators for graphical Gaussian models. *Scandinavian Journal of Statistics*, **20**, 263–70.

More on Hyper Markov Laws

Lecture 7 Saint Flour Summerschool, July 13, 2006

Steffen L. Lauritzen, University of Oxford

Hyper Markov Laws

We identify $\theta \in \Theta$ and $P_\theta \in \mathcal{P}$, so e.g. θ_A for $A \subseteq V$ denotes the marginal 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 \mathcal{L} on Θ we write

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

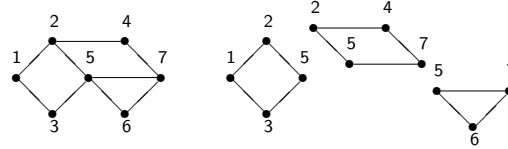
A law \mathcal{L} on Θ is *hyper Markov* w.r.t. \mathcal{G} if

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

Overview of lectures

1. Conditional independence and Markov properties
2. More on Markov properties
3. Graph decompositions and junction trees
4. Probability propagation and related algorithms
5. Log-linear and Gaussian graphical models
6. Hyper Markov laws
7. *More on Hyper Markov Laws*
8. Structure estimation and Bayes factors
9. More on structure estimation.

Hyper Markov property



If θ follows a hyper Markov law for this graph, it holds for example that

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

We shall later see that *this is indeed true for $\hat{\theta} = \hat{p}$ or $\hat{\Sigma}$ in the graphical model with this graph, i.e.*

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

Laws and distributions

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

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

We typically identify Θ with \mathcal{P} when the parametrization

$$\theta \rightarrow P_\theta$$

is one-to-one and onto.

In a Gaussian graphical model, $\theta = K \in S^+(\mathcal{G})$ is uniquely identifying any regular Gaussian distribution $\mathcal{N}_V(0, \Sigma)$, where $K = \Sigma^{-1}$, satisfying the Markov properties of \mathcal{G} .

Consequences of the hyper Markov property

We have

$$A \perp\!\!\!\perp_{\mathcal{L}} B | S \implies \theta_A \perp\!\!\!\perp_{\mathcal{L}} \theta_B | \theta_S,$$

but the converse is false!

It generally holds that

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

If \mathcal{G} is chordal and \mathcal{L} is hyper Markov on \mathcal{G} , it holds that

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

In general, if we form $\bar{\mathcal{G}}$ by completing all prime components of \mathcal{G} , then if \mathcal{L} is hyper Markov on $\bar{\mathcal{G}}$

$$A \perp_{\bar{\mathcal{G}}} B | S \implies A \perp\!\!\!\perp_{\mathcal{L}} B | S.$$

The case when $\mathcal{P} = \mathcal{P}_A$ is more complex, and a specific parametrization needs to be chosen to make a simple and one-to-one correspondence with a suitable parameter Θ .

A probability measure on \mathcal{P} (or on Θ) represents a random element of \mathcal{P} .

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

Thus we shall e.g. speak of the *Wishart law* as we think of W specifying a (random) distribution of X as $\mathcal{N}_V(0 | W)$.

Directed hyper Markov property

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

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

or equivalently

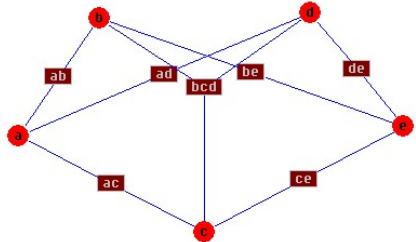
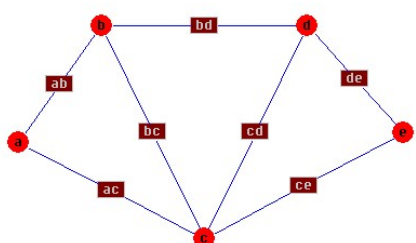
$$\theta_{v | \text{pa}(v)} \perp\!\!\!\perp_{\mathcal{L}} \theta_{\text{nd}(v)} | \theta_{\text{pa}(v)},$$

or equivalently for a well-ordering of \mathcal{D}

$$\theta_{v \cup \text{pa}(v)} \perp\!\!\!\perp_{\mathcal{L}} \theta_{\text{pr}(v)} | \theta_{\text{pa}(v)}.$$

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

<p style="text-align: center;">Meta independence</p> <p>In the following we shall for $A, B \subseteq V$ identify</p> $\theta_{A \cup B} = (\theta_{B A}, \theta_A) = (\theta_{A B}, \theta_B),$ <p>i.e. any joint distribution of $X_{A \cup B}$ is identified with a pair of further marginal and conditional distributions.</p> <p>Define for $S \subseteq V$ the S-section Θ^{θ_S} of Θ as</p> $\Theta^{\theta_S} = \{\theta \in \Theta : \theta_S = \theta_S^*, \theta \in \Theta\}.$ <p>The <i>meta independence relation</i> $\ddagger_{\mathcal{P}}$ is defined as</p> $A \ddagger_{\mathcal{P}} B S \iff \forall \theta_S^* \in \Theta_S : \Theta^{\theta_S^*} = \Theta_{A S}^{\theta_S^*} \times \Theta_{B S}^{\theta_S^*}.$	<p style="text-align: center;">Log-linear meta Markov models</p> <p>Using results on collapsibility of log-linear models (Asmussen and Edwards 1983) that</p> <p><i>A log-linear model $\mathcal{P}_{\mathcal{A}}$ is meta Markov on its dependence graph $\mathcal{G}(\mathcal{A})$ if and only if $S \in \mathcal{A}$ for any minimal complete separator S of $\mathcal{G}(\mathcal{A})$.</i></p> <p>In particular, if \mathcal{A} is conformal, $\mathcal{P}_{\mathcal{A}}$ is meta Markov.</p> <p>For example, the log-linear model with generating class</p> $\mathcal{A} = \{ab, ac, ad, bc, bd, be, cd, ce, de\}$ <p>has dependence graph with cliques $\mathcal{C} = \{abcd, bcde\}$. Since the complete separator bcd is not in \mathcal{A}, this model is <i>not</i> meta Markov.</p>
<p>In words, A and B are <i>meta independent</i> w.r.t. \mathcal{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.</p> <p>Equivalently, fixing the values of $\theta_{B S}$ and θ_S places the same restriction on $\theta_{A S}$ as just fixing θ_S.</p> <p><i>The relation $\ddagger_{\mathcal{P}}$ satisfies the semigraphoid axioms</i> as it is a special instance of variation independence.</p> <p>Note also that for any triple (A, B, S) and any law \mathcal{L} on Θ it holds that</p> $A \perp_{\mathcal{L}} B S \implies A \ddagger_{\mathcal{P}} B S$ <p>for if $\theta_{A S} \perp_{\mathcal{L}} \theta_{B S} \theta_S$ 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.</p>	<p>The model with generating class</p> $\mathcal{A}' = \{ab, ac, ad, bcd, be, ce, de\}$ <p>has the same dependence graph $\mathcal{G}(\mathcal{A}') = \mathcal{G}(\mathcal{A})$ but even though \mathcal{A}' is not conformal, $\mathcal{P}_{\mathcal{A}'}$ is meta Markov on $\mathcal{G}(\mathcal{A}')$.</p> <p>But also the model with generating class</p> $\mathcal{A}'' = \{ab, ac, bc, bd, cd, ce, de\}$ <p>has a different dependence graph $\mathcal{G}(\mathcal{A}'')$. The separator bcd is not in \mathcal{A}'', but $\mathcal{P}_{\mathcal{A}''}$ is meta Markov on $\mathcal{G}(\mathcal{A}'')$, as both <i>minimal separators</i> bc and cd are in \mathcal{A}''.</p>
<p style="text-align: center;">Meta Markov models</p> <p>The family \mathcal{P}, or Θ, is said to be <i>meta Markov</i> w.r.t. \mathcal{G} if</p> <ul style="list-style-type: none"> (i) All $\theta \in \Theta$ are globally Markov w.r.t. \mathcal{G}; (ii) $A \perp_{\mathcal{G}} B S \implies A \ddagger_{\mathcal{P}} B S$ whenever S is complete. <p><i>A Markov model is meta Markov if and only if</i></p> $A \perp_{\overline{\mathcal{G}}} B S \implies A \ddagger_{\mathcal{P}} B S,$ <p>where $\overline{\mathcal{G}}$ is obtained from \mathcal{G} by completing all prime components,</p> <p><i>If \mathcal{G} is chordal, $\overline{\mathcal{G}} = \mathcal{G}$ and hence for any meta Markov model \mathcal{P}</i></p> $A \perp_{\mathcal{G}} B S \implies A \ddagger_{\mathcal{P}} B S.$	<p style="text-align: center;">Dependence graph of \mathcal{A} and \mathcal{A}'</p>
<p style="text-align: center;">Hyper Markov laws and meta Markov models</p> <p>Since it for any law \mathcal{L} on Θ holds that</p> $A \perp_{\mathcal{L}} B S \implies A \ddagger_{\mathcal{P}} B S,$ <p>hyper Markov laws live on meta Markov models: <i>If a law \mathcal{L} on Θ is hyper Markov w.r.t. \mathcal{G}, Θ is meta Markov w.r.t. \mathcal{G}.</i></p> <p><i>In particular, if a Markov model is not meta Markov, it cannot carry a hyper Markov law without further restricting to $\Theta_0 \subset \Theta$.</i></p> <p><i>A Gaussian graphical model with graph \mathcal{G} is meta Markov on $\overline{\mathcal{G}}$.</i></p> <p>This follows for example from results of collapsibility of Gaussian graphical models (Frydenberg 1990).</p>	<p style="text-align: center;">Factor graph of \mathcal{A}</p>

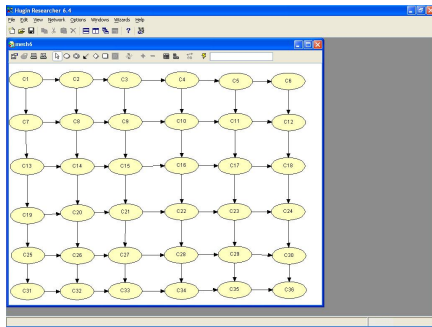
<p style="text-align: center;">Factor graph of \mathcal{A}'</p> 	<p style="text-align: center;">Canonical construction of hyper Markov laws</p> <p>The distributions of maximum likelihood estimators are important examples of hyper Markov laws. But for <i>chordal graphs</i> there is a canonical construction of such laws.</p> <p>Let \mathcal{C} be the cliques of a chordal graph \mathcal{G} and let $\mathcal{L}_{\mathcal{C}}, \mathcal{C} \in \mathcal{C}$ be a family of laws over $\Theta_{\mathcal{C}} \subseteq \mathbb{P}(\mathcal{X}_{\mathcal{C}})$.</p> <p>The family of laws are <i>hyperconsistent</i> if for any C and D with $C \cap D = S \neq \emptyset$, \mathcal{L}_C and \mathcal{L}_D induce the same law for θ_S.</p> <p>If $\mathcal{L}_{\mathcal{C}}, \mathcal{C} \in \mathcal{C}$ are hyperconsistent, there is a unique hyper Markov law \mathcal{L} over \mathcal{G} with $\mathcal{L}(\theta_{\mathcal{C}}) = \mathcal{L}_{\mathcal{C}}, \mathcal{C} \in \mathcal{C}$.</p>
<p style="text-align: center;">Factor graph of \mathcal{A}''</p> 	<p style="text-align: center;">Strong hyper and meta Markov properties</p> <p>In some cases it is of interest to consider a stronger version of the hyper and meta Markov properties.</p> <p>A meta Markov model is <i>strongly meta Markov</i> if $\theta_{A S} \perp\!\!\!\perp_P \theta_S$ for all complete separators S.</p> <p>Similarly, a hyper Markov model is <i>strongly hyper Markov</i> if $\theta_{A S} \perp\!\!\!\perp_{\mathcal{L}} \theta_S$ for all complete separators S.</p> <p>A directed hyper Markov model is <i>strongly directed hyper Markov</i> if $\theta_{v pa(v)} \perp\!\!\!\perp_{\mathcal{L}} \theta_{pa(v)}$ for all $v \in V$.</p> <p><i>Gaussian graphical models and log-linear meta Markov models are strong meta Markov models.</i></p>
<p style="text-align: center;">Meta Markov properties on supergraphs</p> <p>Clearly, if θ is globally Markov w.r.t. the graph \mathcal{G}, it is also Markov w.r.t. any super graph $\mathcal{G}' = (V, E')$ with $E \subseteq E'$.</p> <p>The similar fact is <i>not</i> true for meta Markov models. For example, the Gaussian graphical model for the 4-cycle \mathcal{G} with adjacencies $1 \sim 2 \sim 3 \sim 4 \sim 1$, is meta Markov on \mathcal{G}, because it has no complete separators.</p> <p>But the same model is <i>not</i> meta Markov w.r.t. the larger graph \mathcal{G}' with cliques $\{124, 234\}$, since for any $K \in \mathcal{S}^+(\mathcal{G})$,</p> $\sigma_{24} = \frac{\sigma_{12}\sigma_{14}}{\sigma_{11}} + \frac{\sigma_{13}\sigma_{34}}{\sigma_{33}}.$ <p>So fixing the value of σ_{24} restricts the remaining parameters in a complex way.</p>	<p style="text-align: center;">Bayesian inference</p> <p>Parameter $\theta \in \Theta$, data $X = x$, likelihood</p> $L(\theta x) \propto p(x \theta) = \frac{dP_{\theta}(x)}{d\mu(x)}.$ <p>Express knowledge about θ through a <i>prior law</i> π on θ. Use also π to denote density of the prior law w.r.t. some measure ν on Θ.</p> <p>Inference about θ from x is then represented through <i>posterior law</i> $\pi^*(\theta) = p(\theta x)$. Then, from Bayes' formula</p> $\pi^*(\theta) = p(x \theta)\pi(\theta)/p(x) \propto L(\theta x)\pi(\theta)$ <p>so the <i>likelihood function is equal to the density of the posterior w.r.t. the prior modulo a constant.</i></p>
<p style="text-align: center;">Maximum likelihood in meta Markov models</p> <p>Under certain conditions, the MLE $\hat{\theta}$ of the unknown distribution θ will follow a hyper Markov law over Θ under $P_{\hat{\theta}}$. These are</p> <ol style="list-style-type: none"> (i) Θ is meta Markov w.r.t. \mathcal{G}; (ii) For any prime component Q of \mathcal{G}, the MLE $\hat{\theta}_Q$ for θ_Q based on $X_Q^{(n)}$ is <i>sufficient</i> for Θ_Q and <i>boundedly complete</i>. <p>A sufficient condition for (ii) is that Θ_Q is a <i>full and regular exponential family</i> in the sense of Barndorff-Nielsen (1978).</p> <p>In particular, these conditions are satisfied for any <i>Gaussian graphical model</i> and any <i>meta Markov log-linear model</i>.</p>	<p style="text-align: center;">Bernoulli experiments</p> <p>Data $X_1 = x_1, \dots, X_n = x_n$ independent and Bernoulli distributed with parameter θ, i.e.</p> $P(X_i = 1 \theta) = 1 - P(X_i = 0) = \theta.$ <p>Use a beta prior:</p> $\pi(\theta a, b) \propto \theta^{a-1}(1-\theta)^{b-1}.$ <p>If we let $x = \sum x_i$, we get the posterior:</p> $\begin{aligned} \pi^*(\theta) &\propto \theta^x(1-\theta)^{n-x}\theta^{a-1}(1-\theta)^{b-1} \\ &= \theta^{x+a-1}(1-\theta)^{n-x+b-1} \end{aligned}$ <p>So the posterior is also beta with parameters $(a+x, b+n-x)$.</p>

<p style="text-align: center;">Conjugate families</p> <p>A family \mathcal{P} of laws on Θ is said to be <i>conjugate</i> under sampling from x if</p> $\pi \in \mathcal{P} \implies \pi^* \in \mathcal{P}.$ <p>The family of beta laws is conjugate under Bernoulli sampling.</p> <p>If the family of priors is parametrised:</p> $\mathcal{P} = \{P_\alpha, \alpha \in \mathcal{A}\}$ <p>we sometimes say that α is a <i>hyperparameter</i>. Then, Bayesian inference can be made by just updating hyperparameters. Terminology of hyperparameter breaks down in more complex models.</p>	<p style="text-align: center;">Hyper inverse Wishart and Dirichlet laws</p> <p>Gaussian graphical models are canonical exponential families. The standard family of conjugate priors have densities</p> $\pi(K \Phi, \delta) \propto (\det K)^{\delta/2} e^{-\text{tr}(K\Phi)}, K \in \mathcal{S}^+(\mathcal{G}).$ <p>These laws are termed <i>hyper inverse Wishart laws</i> as Σ follows an inverse Wishart law for complete graphs.</p> <p><i>For chordal graphs, each marginal law \mathcal{L}_C of Σ_C is inverse Wishart.</i></p> <p>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,</p>
<p style="text-align: center;">Conjugacy of hyper Markov properties</p> <p>If \mathcal{L} is a prior law over Θ and $X = x$ is an observation from θ, $\mathcal{L}^* = \mathcal{L}(\theta X = x)$ denotes the <i>posterior law</i> over Θ.</p> <p><i>If \mathcal{L} is hyper Markov w.r.t. \mathcal{G} so is \mathcal{L}^*.</i></p> <p><i>If \mathcal{L} is strongly hyper Markov w.r.t. \mathcal{G} so is \mathcal{L}^*.</i></p> <p>In the latter case, the update of \mathcal{L} is local to prime components, i.e.</p> $\mathcal{L}^*(\theta_Q) = \mathcal{L}_Q^*(\theta_Q) = \mathcal{L}_Q(\theta_Q X_Q = x_Q)$ <p>and the marginal distribution p of X is globally Markov w.r.t. $\overline{\mathcal{G}}$, where</p> $p(x) = \int_{\Theta} P(X = x \theta) \mathcal{L}(d\theta).$	<p>that <i>the standard conjugate prior law is strongly hyper Markov w.r.t. $\overline{\mathcal{G}}$.</i></p> <p>This is in particular true for the hyper inverse Wishart laws.</p> <p>The analogous prior distribution for log-linear meta Markov models are likewise termed <i>hyper Dirichlet laws</i>.</p> <p><i>They are also strongly hyper Markov and if \mathcal{G} is chordal, each induced marginal law \mathcal{L}_C is a standard Dirichlet law.</i></p>
<p style="text-align: center;">Conjugate exponential families</p> <p>For a k-dimensional exponential family</p> $p(x \theta) = b(x) e^{\theta^\top t(x) - \psi(\theta)}$ <p>the <i>standard conjugate family</i> is given as</p> $\pi(\theta a, \kappa) \propto e^{\theta^\top a - \kappa \psi(\theta)}$ <p>for $(a, \kappa) \in \mathcal{A} \subseteq \mathcal{R}^k \times \mathcal{R}_+$, where \mathcal{A} is determined so that the normalisation constant is finite.</p> <p>Posterior updating from (x_1, \dots, x_n) with $t = \sum_i t(x_i)$ is then made as $(a^*, \kappa^*) = (a + t, \kappa + n)$.</p> <p>The family of Beta laws is an example of a standard conjugate family.</p>	<p style="text-align: center;">References</p> <p>Asmussen, S. and Edwards, D. (1983). Collapsibility and response variables in contingency tables. <i>Biometrika</i>, 70, 567–78.</p> <p>Barndorff-Nielsen, O. E. (1978). <i>Information and exponential families in statistical theory</i>. John Wiley and Sons, New York.</p> <p>Frydenberg, M. (1990). Marginalization and collapsibility in graphical interaction models. <i>Annals of Statistics</i>, 18, 790–805.</p>

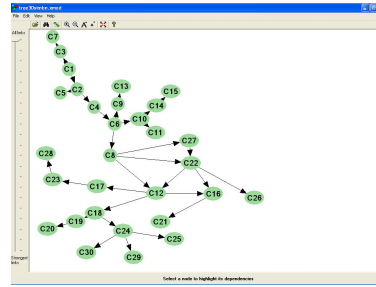
<p style="text-align: center;">Structure Estimation and Bayes Factors</p> <p style="text-align: center;">Lecture 8 Saint Flour Summerschool, July 14, 2006</p> <p style="text-align: center;">Steffen L. Lauritzen, University of Oxford</p>	<p style="text-align: center;">Directed hyper Markov property</p> <p>$\mathcal{L} = \mathcal{L}(\theta)$ is <i>directed hyper Markov</i> w.r.t. a DAG \mathcal{D} if θ is directed Markov on \mathcal{D} for all $\theta \in \Theta$ and</p> $\theta_v \perp_{\mathcal{L}} \theta_{\text{nd}(v)} \mid \theta_{\text{pa}(v)}.$ <p>If \mathcal{D} is perfect, \mathcal{L} is directed hyper Markov w.r.t. \mathcal{D} if and only if \mathcal{L} is hyper Markov w.r.t. $\mathcal{G} = \sigma(\mathcal{D}) = \mathcal{D}^m$.</p>
<p style="text-align: center;">Overview of lectures</p> <ol style="list-style-type: none"> 1. Conditional independence and Markov properties 2. More on Markov properties 3. Graph decompositions and junction trees 4. Probability propagation and related algorithms 5. Log-linear and Gaussian graphical models 6. Hyper Markov laws 7. More on Hyper Markov Laws 8. <i>Structure estimation and Bayes factors</i> 9. More on structure estimation. 	<p style="text-align: center;">Meta Markov models</p> <p>For $A, B \subseteq V$ identify</p> $\theta_{A \cup B} = (\theta_{B \mid A}, \theta_A) = (\theta_{A \mid B}, \theta_B).$ <p>A and B are <i>meta independent</i> w.r.t. \mathcal{P} given S, denoted $A \perp_{\mathcal{P}} B \mid S$, if the pair of conditional distributions $(\theta_{A \mid S}, \theta_{B \mid S})$ vary in a product space when θ_S is fixed.</p> <p>The family \mathcal{P}, or Θ, is <i>meta Markov</i> w.r.t. \mathcal{G} if</p> <ol style="list-style-type: none"> (i) All $\theta \in \Theta$ are globally Markov w.r.t. \mathcal{G}; (ii) $A \perp_{\mathcal{G}} B \mid S \implies A \perp_{\mathcal{P}} B \mid S$ whenever S is complete.
<p style="text-align: center;">Hyper Markov Laws</p> <p>Identify $\theta \in \Theta$ and $P_\theta \in \mathcal{P}$, so e.g. θ_A denotes the marginal distribution of X_A under P_θ and $\theta_{A \mid B}$ the family of conditional distributions of X_A given X_B, etc.</p> <p>For a law \mathcal{L} on Θ we write</p> $A \perp_{\mathcal{L}} B \mid S \iff \theta_{A \mid S} \perp_{\mathcal{L}} \theta_{B \mid S} \mid \theta_S.$ <p>A law \mathcal{L} on Θ is <i>hyper Markov</i> w.r.t. \mathcal{G} if</p> <ol style="list-style-type: none"> (i) All $\theta \in \Theta$ are globally Markov w.r.t. \mathcal{G}; (ii) $A \perp_{\mathcal{L}} B \mid S$ whenever S is complete and $A \perp_{\mathcal{G}} B \mid S$. 	<p style="text-align: center;">Hyper Markov laws and meta Markov models</p> <p><i>Hyper Markov laws live on meta Markov models.</i></p> <p><i>A Gaussian graphical model with graph \mathcal{G} is meta Markov on \mathcal{G}.</i></p> <p><i>A log-linear model \mathcal{P}_A is meta Markov on its dependence graph $\mathcal{G}(A)$ if and only if $S \in \mathcal{A}$ for any minimal complete separator S of $\mathcal{G}(A)$.</i></p> <p>In particular, if \mathcal{A} is conformal, \mathcal{P}_A is meta Markov.</p>
<p style="text-align: center;">Hyper Markov property</p> <p>The hyper Markov property has a simple formulation in terms of junction trees:</p> <p>Arrange the prime components Q of \mathcal{G} in a junction tree \mathcal{T} with complete separators S and consider the <i>extended junction tree</i> $\overline{\mathcal{T}}$ which is the (bipartite) tree with $Q \cup S$ as vertices and edges from separators to prime components so that $C \sim S \sim D$ in $\overline{\mathcal{T}}$ if and only if $C \sim D$ in \mathcal{T}.</p> <p>Next, associate θ_A to A for each $A \in Q \cup S$. It then holds that</p> <p>\mathcal{L} is hyper Markov on \mathcal{G} if and only if $\{\theta_A, A \in Q \cup S\}$ is globally Markov w.r.t. the extended junction tree $\overline{\mathcal{T}}$.</p>	<p style="text-align: center;">Maximum likelihood in meta Markov models</p> <p><i>If the following conditions are satisfied:</i></p> <ol style="list-style-type: none"> (i) Θ is meta Markov w.r.t. \mathcal{G}; (ii) For any prime component Q of \mathcal{G}, Θ_Q is a full and regular exponential family, <p><i>the MLE $\hat{\theta}$ of the unknown distribution θ will follow a hyper Markov law over Θ under P_θ.</i></p> <p>In particular, this holds for any Gaussian graphical model and any meta Markov log-linear model.</p>

<p>Strong hyper and meta Markov properties</p> <p>A meta Markov model is <i>strongly meta Markov</i> if $\theta_{A S} \perp\!\!\!\perp \theta_S$ for all complete separators S.</p> <p>Similarly, a hyper Markov model is <i>strongly hyper Markov</i> if $\theta_{A S} \perp\!\!\!\perp \theta_S$ for all complete separators S.</p> <p>A directed hyper Markov model is <i>strongly directed hyper Markov</i> if $\theta_v \perp\!\!\!\perp \theta_{pa(v)}$ for all $v \in V$.</p> <p>Gaussian graphical models and log-linear meta Markov models are strong meta Markov models.</p>	<p>Conjugate prior laws are strong hyper Markov</p> <p>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. \mathcal{G}.</p> <p>This is in particular true for the hyper inverse Wishart laws and the hyper Dirichlet laws.</p> <p>Thus, for the hyper inverse and hyper Dirichlet laws we have simple local updating based on conjugate priors for Bayesian inference.</p>
<p>Conjugacy of hyper Markov properties</p> <p>If \mathcal{L} is a prior law over Θ and $X = x$ is an observation from θ, $\mathcal{L}^* = \mathcal{L}(\theta X = x)$ denotes the posterior law over Θ.</p> <p>If \mathcal{L} is hyper Markov w.r.t. \mathcal{G} so is \mathcal{L}^*.</p> <p>If \mathcal{L} is strongly hyper Markov w.r.t. \mathcal{G} so is \mathcal{L}^*.</p> <p>In the latter case, the update of \mathcal{L} is local to prime components, i.e.</p> $\mathcal{L}^*(\theta_Q) = \mathcal{L}_Q^*(\theta_Q) = \mathcal{L}_Q(\theta_Q X_Q = x_Q)$ <p>and the marginal distribution p of X is globally Markov w.r.t. $\overline{\mathcal{G}}$, where</p> $p(x) = \int_{\Theta} P(X = x \theta) \mathcal{L}(d\theta).$	<p>Estimation of structure</p> <p>Previous lectures have considered the graph \mathcal{G} defining the model as known and inference was concerning an unknown P_{θ} with $\theta \in \Theta$.</p> <p>The last two lectures are concerned with inference concerning the graph \mathcal{G}, specifying only a family Γ of possible graphs.</p> <p>Methods must scale well with data size, as many structures and huge collections of data are to be considered.</p> <p>Structure estimation is also known as <i>model selection</i> (mainstream statistics) <i>system identification</i> (engineering), <i>structural learning</i> (AI or machine learning.)</p>
<p>Conjugate exponential families</p> <p>For a k-dimensional exponential family</p> $p(x \theta) = b(x) e^{\theta^T t(x) - \psi(\theta)}$ <p>the standard conjugate family is given as</p> $\pi(\theta a, \kappa) \propto e^{\theta^T a - \kappa \psi(\theta)}$ <p>for $(a, \kappa) \in \mathcal{A} \subseteq \mathcal{R}^k \times \mathcal{R}_+$, where \mathcal{A} is determined so that the normalisation constant is finite.</p> <p>Posterior updating from (x_1, \dots, x_n) with $t = \sum_i t(x_i)$ is then made as $(a^*, \kappa^*) = (a + t, \kappa + n)$.</p>	<p>Examples of structural assumptions</p> <p>Different situations occur depending on the type of assumptions concerning Γ.</p> <ol style="list-style-type: none"> 1. Γ is the set of <i>undirected graphs</i> over V; 2. Γ is the set of <i>chordal graphs</i> over V; 3. Γ is the set of <i>forests</i> over V; 4. Γ is the set of <i>trees</i> over V; 5. Γ is the set of <i>directed acyclic graphs</i> over V; 6. Other conditional independence structures
<p>Hyper inverse Wishart and Dirichlet laws</p> <p>Gaussian graphical models are canonical exponential families. The standard family of conjugate priors have densities</p> $\pi(K \Phi, \delta) \propto (\det K)^{\delta/2} e^{-\text{tr}(K\Phi)}, K \in S^+(\mathcal{G}).$ <p>These laws are termed <i>hyper inverse Wishart laws</i> as Σ follows an inverse Wishart law for complete graphs. For chordal graphs, each marginal law \mathcal{L}_C, C of Σ_C is inverse Wishart.</p> <p>The standard conjugate prior law for log-linear meta Markov models are termed <i>hyper Dirichlet laws</i>. If \mathcal{G} is chordal, each induced marginal law $\mathcal{L}_C, C \in \mathcal{C}$ is a standard Dirichlet law.</p>	<p>Why estimation of structure?</p> <ul style="list-style-type: none"> • Parallel to e.g. density estimation • Obtain quick overview of relations between variables in complex systems • Data mining • Gene regulatory networks • Reconstructing family trees from DNA information • Methods exist, but need better understanding of their statistical properties.

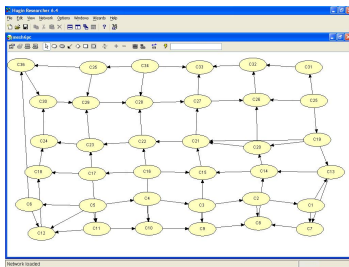
Markov mesh model



Bayesian GES on tree

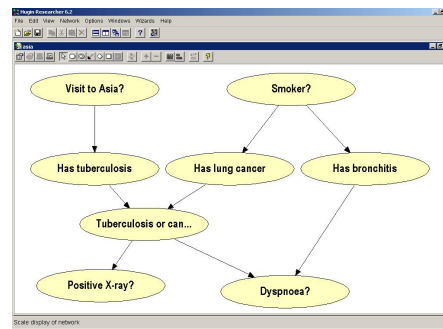


PC algorithm

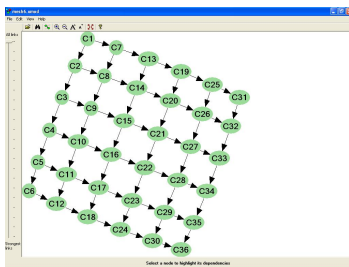


Crudest algorithm (HUGIN), 10000 simulated cases

Chest clinic

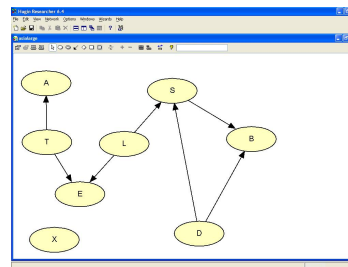


Bayesian GES



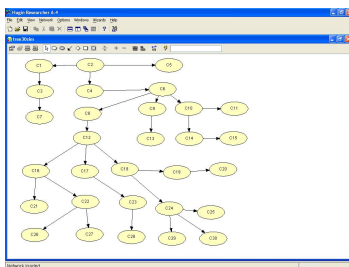
Crudest algorithm (WinMine), 10000 simulated cases

PC algorithm



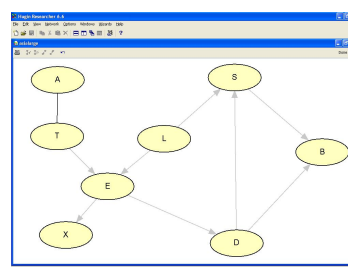
10000 simulated cases

Tree model



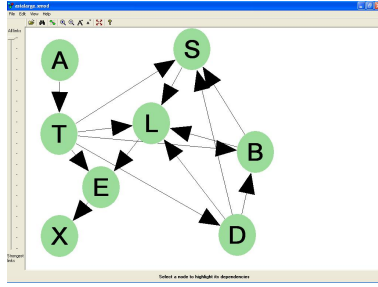
PC algorithm, 10000 cases, correct reconstruction

NPC algorithm



10000 simulated cases

Bayesian GES



More on trees

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

$$\sup_x |p(x) - \hat{p}(x)| \rightarrow 0 \text{ for } n \rightarrow \infty,$$

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).

Types of approach

- Methods for *judging adequacy of structure* such as
 - Tests of significance
 - Penalised likelihood scores

$$I_\kappa(\mathcal{G}) = \log \hat{L} - \kappa \dim(\mathcal{G})$$

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 $\mathcal{G} \in \Gamma$, $\Theta_{\mathcal{G}}$ is associated parameter space so that P factorizes w.r.t. \mathcal{G} if and only if $P = P_\theta$ for some $\theta \in \Theta_{\mathcal{G}}$. $\mathcal{L}_{\mathcal{G}}$ is prior law on $\Theta_{\mathcal{G}}$.

The *Bayes factor* (likelihood ratio) for discriminating between \mathcal{G}_1 and \mathcal{G}_2 based on observations $X^{(n)} = x^{(n)}$ is

$$\text{BF}(\mathcal{G}_1 : \mathcal{G}_2) = \frac{f(x^{(n)} | \mathcal{G}_1)}{f(x^{(n)} | \mathcal{G}_2)},$$

where

$$f(x^{(n)} | \mathcal{G}) = \int_{\Theta_{\mathcal{G}}} f(x^{(n)} | \mathcal{G}, \theta) \mathcal{L}_{\mathcal{G}}(d\theta)$$

is known as the *marginal likelihood* of \mathcal{G} .

Estimating trees

Assume P factorizes w.r.t. an unknown tree τ .

Chow and Liu (1968) showed MLE $\hat{\tau}$ of \mathcal{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 $e = \{u, v\}$:

$$H_n(e) = \sum_{x_u, x_v} \frac{n(x_u, x_v)}{n} \log \frac{n(x_u, x_v)/n}{n(x_u)n(x_v)/n^2}.$$

Posterior distribution over graphs

If $\pi(\mathcal{G})$ is a prior probability distribution over a given set of graphs Γ , the posterior distribution is determined as

$$\pi^*(\mathcal{G}) = \pi(\mathcal{G} | x^{(n)}) \propto f(x^{(n)} | \mathcal{G}) \pi(\mathcal{G})$$

or equivalently

$$\frac{\pi^*(\mathcal{G}_1)}{\pi^*(\mathcal{G}_2)} = \text{BF}(\mathcal{G}_1 : \mathcal{G}_2) \frac{\pi(\mathcal{G}_1)}{\pi(\mathcal{G}_2)}.$$

Bayesian analysis looks for the *MAP estimate* \mathcal{G}^* maximizing $\pi^*(\mathcal{G})$ over Γ , or attempts to *sample from the posterior* using e.g. Monte-Carlo methods.

Extensions

Results are easily *extended to Gaussian graphical models*, with the weight of a tree determined as

$$w_n(e) = -\frac{1}{2} \log(1 - r_e^2),$$

where r_e^2 is *correlation coefficient* along edge $e = \{u, v\}$.

Highest AIC or BIC scoring forest also available as *MWSF*, with modified weights

$$w_n^{\text{pen}}(e) = n w_n(e) - \kappa_n \text{df}_e,$$

with $\kappa_n = 2$ for AIC, $\kappa_n = \log n$ for BIC and df_e the *degrees of freedom for independence* along e .

Strong hyper Markov prior laws

For strong hyper Markov prior laws, $X^{(n)}$ is itself marginally Markov so



$$f(x^{(n)} | \mathcal{G}) = \frac{\prod_{Q \in \mathcal{Q}} f(x_Q^{(n)} | \mathcal{G})}{\prod_{S \in \mathcal{S}} f(x_S^{(n)} | \mathcal{G})^{\nu_{\mathcal{G}}(S)}}, \quad (1)$$

where \mathcal{Q} are the prime components and \mathcal{S} the minimal complete separators of \mathcal{G} .

<p style="text-align: center;">Hyper inverse Wishart laws</p> <p>Denote the normalisation constant of the hyper inverse Wishart density as</p> $h(\delta, \Phi; \mathcal{G}) = \int_{S^+(\mathcal{G})} (\det K)^{\delta/2} e^{-\text{tr}(K\Phi)} dK,$ <p>i.e. the usual Wishart constant if $Q = C$ is a clique.</p> <p>Combining with the Gaussian likelihood, it is easily seen that for Gaussian graphical models we have</p> $f(x^{(n)} \mathcal{G}) = \frac{h(\delta + n, \Phi + W^n; \mathcal{G})}{h(\delta, \Phi; \mathcal{G})}.$ <p>Comparing with (1) leads to a similar factorization of the</p>	<p style="text-align: center;">Bayesian analysis</p> <p><i>MAP estimates of forests can thus be computed using an MWSF algorithm, using $w(e) = \log BF(e)$ as weights.</i></p> <p>Algorithms exist for generating random spanning trees (Aldous 1990), so <i>full posterior analysis is in principle possible for trees.</i></p> <p>These work less well for weights occurring with typical Bayes factors, as most of these are essentially zero, so methods based on the <i>Matrix Tree Theorem</i> seem currently more useful.</p> <p><i>Only heuristics available for MAP estimators or maximizing penalized likelihoods such as AIC or BIC, for other than trees.</i></p>
<p>normalising constant</p> $h(\delta, \Phi; \mathcal{G}) = \frac{\prod_{Q \in \mathcal{Q}} h(\delta, \Phi_Q; \mathcal{G}_Q)}{\prod_{S \in \mathcal{S}} h(\delta, \Phi_S; S)^{\nu_{\mathcal{G}}(S)}}.$ <p>For <i>chordal graphs</i> all terms in this expression reduce to known Wishart constants, and we can thus calculate the normalization constant explicitly.</p> <p>In general, Monte-Carlo simulation or similar methods must be used (Atay-Kayis and Massam 2005).</p> <p>The marginal distribution of $W^{(n)}$ is (weak) <i>hyper Markov</i> w.r.t. \mathcal{G}. It was termed the <i>hyper matrix F law</i> by Dawid and Lauritzen (1993).</p>	<p style="text-align: center;">Some challenges for undirected graphs</p> <ul style="list-style-type: none"> Find <i>feasible algorithm for (perfect) simulation from a distribution over chordal graphs as</i> $p(\mathcal{G}) \propto \frac{\prod_{C \in \mathcal{C}} w(C)}{\prod_{S \in \mathcal{S}} w(S)^{\nu_{\mathcal{G}}(S)}},$ <p>where $w(A)$, $A \subseteq V$ are a prescribed set of positive weights.</p> <ul style="list-style-type: none"> Find <i>feasible algorithm for obtaining MAP in decomposable case. This may not be universally possible as problem most likely is NP-complete.</i>
<p style="text-align: center;">Bayes factors for forests</p> <p>Trees and forests are decomposable graphs, so for a forest ϕ we get</p> $f(\phi x^{(n)}) \propto \frac{\prod_{e \in E(\phi)} f(x_e^{(n)})}{\prod_{v \in V} f(x_v^{(n)})^{d_{\phi}(v)-1}},$ <p>since all minimal complete separators are singletons and $\nu_{\phi}(\{v\}) = d_{\phi}(v) - 1$.</p> <p>Multiplying the right-hand side with $\prod_{v \in V} f(x_v^{(n)})$ yields</p> $\frac{\prod_{e \in E(\phi)} f(x_e^{(n)})}{\prod_{v \in V} f(x_v^{(n)})^{d_{\phi}(v)-1}} = \prod_{v \in V} f(x_v^{(n)}) \prod_{e \in \phi} \text{BF}(e),$	<p style="text-align: center;">References</p> <p>Akaike, H. (1974). A new look at the statistical model identification. <i>IEEE Transactions on Automatic Control</i>, 19, 716–23.</p> <p>Aldous, D. (1990). A random walk construction of uniform spanning trees and uniform labelled trees. <i>SIAM Journal on Discrete Mathematics</i>, 3, (4), 450–65.</p> <p>Atay-Kayis, A. and Massam, H. (2005). A Monte Carlo method for computing the marginal likelihood in non-decomposable graphical Gaussian models. <i>Biometrika</i>, 92, 317–35.</p> <p>Bollobás, B. (1998). <i>Modern graph theory</i>. Springer-Verlag, New York.</p> <p>Chow, C. K. and Liu, C. N. (1968). Approximating dis-</p>
<p>where $\text{BF}(e)$ is the <i>Bayes factor</i> for independence along the edge e:</p> $\text{BF}(e) = \frac{f(x_u^{(n)}, x_v^{(n)})}{f(x_u^{(n)})f(x_v^{(n)})}.$ <p>Thus the <i>posterior distribution</i> of ϕ is</p> $\pi^*(\phi) \propto \prod_{e \in E(\phi)} \text{BF}(e).$ <p>In the case where ϕ is restricted to contain a <i>single tree</i>, the normalization constant for this distribution can be explicitly obtained via the <i>Matrix Tree Theorem</i>, see e.g. Bollobás (1998).</p>	<p>crete probability distributions with dependence trees. <i>IEEE Transactions on Information Theory</i>, 14, 462–7.</p> <p>Chow, C. K. and Wagner, T. J. (1978). Consistency of an estimate of tree-dependent probability distributions. <i>IEEE Transactions on Information Theory</i>, 19, 369–71.</p> <p>Dawid, A. P. and Lauritzen, S. L. (1993). Hyper Markov laws in the statistical analysis of decomposable graphical models. <i>Annals of Statistics</i>, 21, 1272–317.</p> <p>Kruskal Jr., J. B. (1956). On the shortest spanning subtree of a graph and the travelling salesman problem. <i>Proceedings of the American Mathematical Society</i>, 7, 48–50.</p> <p>Schwarz, G. (1978). Estimating the dimension of a model.</p>

Annals of Statistics, **6**, 461-4.

<p style="text-align: center;">More about Structure Estimation</p> <p style="text-align: center;">Lecture 9 Saint Flour Summerschool, July 17, 2006</p> <p style="text-align: center;">Steffen L. Lauritzen, University of Oxford</p>	<p>Highest AIC or BIC scoring forest also available as MWSF, with modified weights</p> $w_n^{\text{pen}}(e) = nw_n(e) - \kappa_n df_e,$ <p>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.</p> <p>Use maximal weight spanning tree (or forest) algorithm from weights $W = (w_{uv}, u, v \in V)$.</p>
<p style="text-align: center;">Overview of lectures</p> <ol style="list-style-type: none"> 1. Conditional independence and Markov properties 2. More on Markov properties 3. Graph decompositions and junction trees 4. Probability propagation and related algorithms 5. Log-linear and Gaussian graphical models 6. Hyper Markov laws 7. More on Hyper Markov Laws 8. Structure estimation and Bayes factors 9. <i>More on structure estimation.</i> 	<p style="text-align: center;">Bayes factors</p> <p>For $\mathcal{G} \in \Gamma$, $\Theta_{\mathcal{G}}$ is associated parameter space so that P factorizes w.r.t. \mathcal{G} if $P = P_{\theta}$ for some $\theta \in \Theta_{\mathcal{G}}$. $\mathcal{L}_{\mathcal{G}}$ is prior law on $\Theta_{\mathcal{G}}$.</p> <p>The Bayes factor for discriminating between \mathcal{G}_1 and \mathcal{G}_2 based on $X^{(n)} = x^{(n)}$ is</p> $\text{BF}(\mathcal{G}_1 : \mathcal{G}_2) = \frac{f(x^{(n)} \mathcal{G}_1)}{f(x^{(n)} \mathcal{G}_2)},$ <p>where</p> $f(x^{(n)} \mathcal{G}) = \int_{\Theta_{\mathcal{G}}} f(x^{(n)} \mathcal{G}, \theta) \mathcal{L}_{\mathcal{G}}(d\theta)$ <p>is known as the <i>marginal likelihood</i> of \mathcal{G}.</p>
<p style="text-align: center;">Types of approach</p> <ul style="list-style-type: none"> • Methods for <i>judging adequacy of structure</i> such as <ul style="list-style-type: none"> – Tests of significance – Penalised likelihood scores $I_{\kappa}(\mathcal{G}) = \log \hat{L} - \kappa \dim(\mathcal{G})$ <p>with $\kappa = 1$ for AIC Akaike (1974), or $\kappa = \frac{1}{2} \log n$ for BIC (Schwarz 1978).</p> <ul style="list-style-type: none"> – Bayesian posterior probabilities. • Search strategies through space of possible structures, more or less based on <i>heuristics</i>. 	<p style="text-align: center;">Posterior distribution over graphs</p> <p>If $\pi(\mathcal{G})$ is a prior probability distribution over a given set of graphs Γ, the posterior distribution is determined as</p> $\pi^*(\mathcal{G}) = \pi(\mathcal{G} x^{(n)}) \propto f(x^{(n)} \mathcal{G}) \pi(\mathcal{G})$ <p>or equivalently</p> $\frac{\pi^*(\mathcal{G}_1)}{\pi^*(\mathcal{G}_2)} = \text{BF}(\mathcal{G}_1 : \mathcal{G}_2) \frac{\pi(\mathcal{G}_1)}{\pi(\mathcal{G}_2)}.$ <p>The BIC is an $O(1)$-approximation to \log BF using Laplace's method of integrals on the marginal likelihood.</p> <p>Bayesian analysis looks for the MAP estimate \mathcal{G}^* maximizing $\pi^*(\mathcal{G})$ over Γ, or attempts to sample from the posterior using e.g. Monte-Carlo methods.</p>
<p style="text-align: center;">Estimating trees</p> <p>Assume P factorizes w.r.t. an unknown tree \mathcal{T}. MLE $\hat{\tau}$ of \mathcal{T} has maximal weight, where the weight of τ is</p> $w(\tau) = \sum_{e \in E(\tau)} w_n(e) = \sum_{e \in E(\tau)} H_n(e)$ <p>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</p> $w_n(e) = -\frac{1}{2} \log(1 - r_e^2),$ <p>where r_e^2 is correlation coefficient along edge $e = \{u, v\}$.</p>	<p style="text-align: center;">Hyper inverse Wishart laws</p> <p>Denote the normalisation constant of the hyper inverse Wishart density as</p> $h(\delta, \Phi; \mathcal{G}) = \int_{S^+(\mathcal{G})} (\det K)^{\delta/2} e^{-\text{tr}(K\Phi)} dK,$ <p>The marginal likelihood is then</p> $f(x^{(n)} \mathcal{G}) = \frac{h(\delta + n, \Phi + W^n; \mathcal{G})}{h(\delta, \Phi; \mathcal{G})}.$ <p>where</p> $h(\delta, \Phi; \mathcal{G}) = \frac{\prod_{Q \in \mathcal{Q}} h(\delta, \Phi_Q; \mathcal{G}_Q)}{\prod_{S \in \mathcal{S}} h(\delta, \Phi_S; S)^{\nu_{\mathcal{G}}(S)}}.$

<p>For <i>chordal graphs</i> all terms reduce to known Wishart constants.</p> <p>In general, Monte-Carlo simulation or similar methods must be used (Atay-Kayis and Massam 2005).</p>	<p>additional parent $\theta_v _{\text{pa}(v)}$ for every vertex V in \mathcal{D}, so then</p> $f(x \theta) = \prod_{v \in V} f(x_v x_{\text{pa}(v)}, \theta_v _{\text{pa}(v)}).$ <p>Exploiting independence and taking expectations over θ yields that <i>also marginally</i>,</p> $f(x \mathcal{D}) = \int_{\Theta_{\mathcal{D}}} f(x \theta) \mathcal{L}_{\mathcal{D}}(\theta) = \prod_{v \in V} f(x_v x_{\text{pa}(v)}).$ <p>If \mathcal{L} is strongly directed hyper Markov and \mathcal{L}^* it holds that <i>also the posterior law \mathcal{L}^* is strongly directed hyper Markov</i> and</p> $\mathcal{L}^*(\theta_v _{\text{pa}(v)}) \propto f(x_v x_{\text{pa}(v)}, \theta_v _{\text{pa}(v)}) \mathcal{L}(\theta_v _{\text{pa}(v)})$ <p>(Spiegelhalter and Lauritzen 1990).</p>
<p style="text-align: center;">Bayes factors for forests</p> <p>Trees and forests are decomposable graphs, so for a forest ϕ we get</p> $\begin{aligned} \pi^*(\phi) &\propto \frac{\prod_{e \in E(\phi)} f(x_e^{(n)})}{\prod_{v \in V} f(x_v^{(n)})^{d_{\phi}(v)-1}} \\ &\propto \prod_{e \in E(\phi)} \text{BF}(e), \end{aligned}$ <p>where $\text{BF}(e)$ is the <i>Bayes factor</i> for independence along the edge e:</p> $\text{BF}(e) = \frac{f(x_u^{(n)}, x_v^{(n)})}{f(x_u^{(n)})f(x_v^{(n)})}.$	<p style="text-align: center;">Markov equivalence</p> <p>\mathcal{D} and \mathcal{D}' are equivalent if and only if:</p> <ol style="list-style-type: none"> \mathcal{D} and \mathcal{D}' have same <i>skeleton</i> (ignoring directions) \mathcal{D} and \mathcal{D}' have same unmarried parents <p>so</p>  <p>but</p> 
<p><i>MAP estimates of forests can thus be computed using an MWSF algorithm, using $w(e) = \log \text{BF}(e)$ as weights.</i></p> <p>When ϕ is restricted to contain a <i>single tree</i>, the normalization constant can be explicitly obtained via the <i>Matrix Tree Theorem</i>, see e.g. Bollobás (1998).</p> <p>Algorithms exist for generating random spanning trees (Aldous 1990), so <i>full posterior analysis is in principle possible for trees</i>.</p> <p><i>Only heuristics available for MAP estimators or maximizing penalized likelihoods such as AIC or BIC, for other than trees.</i></p>	<p style="text-align: center;">Searching equivalence classes</p> <p>In general, there is no hope of distinguishing Markov equivalent DAGs, so \mathcal{D} can at best be identified up to <i>Markov equivalence</i>.</p> <p>The number D_n of unlabelled DAGs with n vertices is given by the recursion (Robinson 1977)</p> $D_n = \sum_{i=1}^n (-1)^{i+1} \binom{n}{i} 2^{i(n-i)} D_{n-i}$ <p>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.</p>
<p style="text-align: center;">Directed hyper Markov property</p> <p>$\mathcal{L} = \mathcal{L}(\theta)$ is <i>directed hyper Markov</i> w.r.t. a DAG \mathcal{D} if θ is directed Markov on \mathcal{D} for all $\theta \in \Theta$ and</p> $\theta_v _{\text{pa}(v)} \perp\!\!\!\perp_{\mathcal{L}} \theta_{\text{nd}(v)} \theta_{\text{pa}(v)}.$ <p>A law \mathcal{L} is <i>directed hyper Markov</i> on \mathcal{D} if and only if \mathcal{L}_A is hyper Markov on $(\mathcal{D}_A)^m$ for any ancestral set $A \subseteq V$.</p> <p>\mathcal{L} is <i>strongly directed hyper Markov</i> if in addition $\theta_v _{\text{pa}(v)} \perp\!\!\!\perp_{\mathcal{L}} \theta_{\text{pa}(v)}$ for all v or, equivalently if the conditional distributions $\theta_v _{\text{pa}(v)}, v \in V$ are mutually independent.</p> <p>Graphically, this is most easily displayed by introducing one</p>	<p style="text-align: center;">Conjugate priors for DAGs</p> <p>In the discrete case, the obvious conjugate prior is for fixed v to let</p> $\{\theta_v _{\text{pa}_{\mathcal{D}}(v)}(x_v x_{\text{pa}_{\mathcal{D}}(v)}^*), x_v \in \mathcal{X}_v\}$ <p>be <i>Dirichlet distributed</i> and independent for $v \in V$ and $x_{\text{pa}_{\mathcal{D}}(v)}^* \in \mathcal{X}_{\text{pa}_{\mathcal{D}}(v)}$ (Spiegelhalter and Lauritzen 1990).</p> <p>We can derive these Dirichlet distributions from a fixed <i>master Dirichlet</i> distribution $\mathcal{D}(\alpha)$, where $\alpha = \alpha(x), x \in \mathcal{X}$, by letting</p> $\{\theta_v _{\text{pa}(v)}(x_v x_{\text{pa}_{\mathcal{D}}(v)}^*)\} \sim \mathcal{D}(\alpha(x_v, x_{\text{pa}_{\mathcal{D}}(v)}^*)),$ <p>where as usual $\alpha(x_a) = \sum_{y: y_a = x_a} \alpha(y)$.</p>

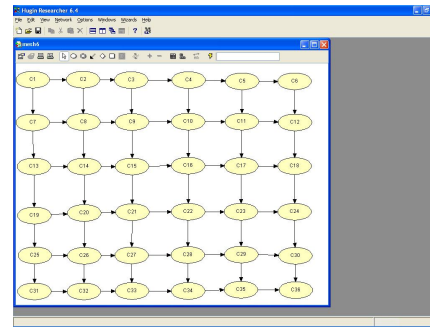
Typically, α is specified by letting $\alpha = \lambda p_0(x)$ where p_0 is an initial guess on the joint distribution, for example specified through a DAG \mathcal{D}_0 , and λ is the *equivalent sample size* for the prior information.

The values $\alpha(x_v, x_{\text{pa}_{\mathcal{D}}(v)}^*) = \lambda p_0(x_v, x_{\text{pa}_{\mathcal{D}}(v)}^*)$ can then be calculated by *probability propagation*.

Common default values is $\lambda = 1$ and $\alpha(x) = |\mathcal{X}|^{-1}$.

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).

Markov mesh model



Characterization of strong hyper priors

In all cases, it was shown (Geiger and Heckerman 1997, 2002) that *prior distributions constructed in this way are the only distributions which are*

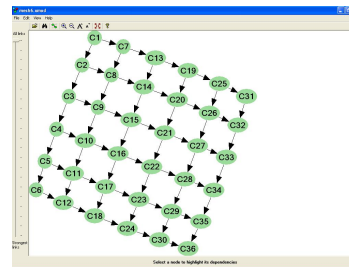
1. modular:

$$\text{pa}_{\mathcal{D}}(v) = \text{pa}_{\mathcal{D}'}(v) \implies \theta_v | \text{pa}_{\mathcal{D}}(v) \sim \theta_v | \text{pa}_{\mathcal{D}'}(v);$$

2. score equivalent:

$$\mathcal{D} \equiv \mathcal{D}' \implies f(x^{(n)} | \mathcal{D}) = f(x^{(n)} | \mathcal{D}').$$

Bayesian GES



Crudest algorithm (WinMine), 10000 simulated cases

Marginal likelihood

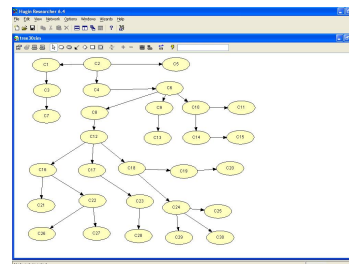
Bayes factors derived from these *strongly directed hyper Dirichlet priors* have a simple form

$$f(x^{(n)} | \mathcal{D}) = \prod_v \prod_{x_{\text{pa}_{\mathcal{D}}(v)}} \frac{\Gamma(\alpha(x_{\text{pa}_{\mathcal{D}}(v)}))}{\Gamma(\alpha(x_{\text{pa}_{\mathcal{D}}(v)}) + n(x_{\text{pa}_{\mathcal{D}}(v)}))} \times \prod_{x_v} \frac{\Gamma(\alpha(x_v \cup \text{pa}_{\mathcal{D}}(v)) + n(x_v \cup \text{pa}_{\mathcal{D}}(v)))}{\Gamma(\alpha(x_v \cup \text{pa}_{\mathcal{D}}(v)))}.$$

(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?

Tree model

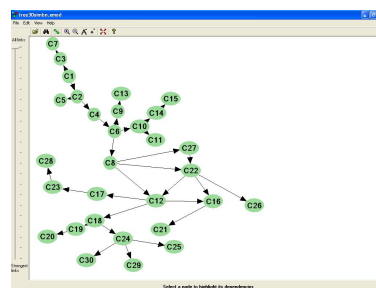


Greedy equivalence class search

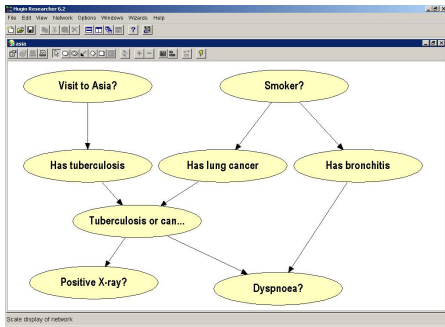
1. Initialize with empty DAG
2. Repeatedly search among equivalence classes with a *single additional edge* and go to class with highest score - until no improvement.
3. Repeatedly search among equivalence classes with a *single edge less* and move to one with highest score - until no improvement.

For *BIC* or *Bayesian posterior score* with *directed hyper Dirichlet priors*, this algorithm yields *consistent estimate of equivalence class for P*. (Chickering 2002)

Bayesian GES on tree



Chest clinic



SGS and PC algorithms

SGS-algorithm (Spirtes *et al.* 1993):

Step 1: Identify *skeleton* using that, for P faithful,

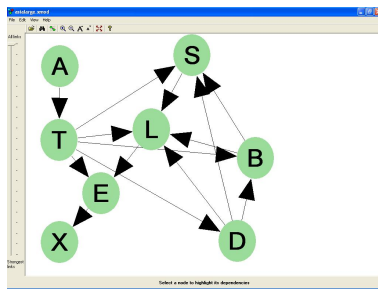
$$u \not\perp v \iff \exists S \subseteq V \setminus \{u, v\} : X_u \perp\!\!\!\perp X_v \mid X_S.$$

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 $S \subseteq \text{bd}(u) \setminus v$ or $S \subseteq \text{bd}(v) \setminus u$ needs checking where bd refers to current skeleton.

Step 2: Identify directions to be consistent with independence relations found in Step 1.

Bayesian GES



Exact properties of PC-algorithm

If P is faithful to DAG \mathcal{D} , PC-algorithm finds \mathcal{D}' equivalent to \mathcal{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 \mathcal{D} .

So worst case complexity is exponential, but *algorithm fast for sparse graphs*.

Sampling properties are less well understood although consistency results exist.

Constraint-based search

Another alternative search algorithm is known as *constraint based search*.

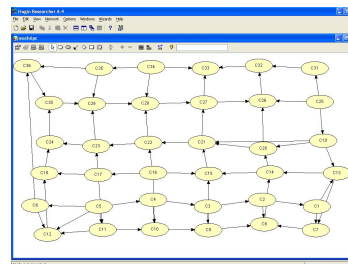
Essentially, the search methods generate queries of the type " $A \perp\!\!\!\perp B \mid S$ ", and the answer to such a query divides Γ into those graphs conforming with the query and those that do not.

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 \mathcal{D} (or rather its equivalence class).

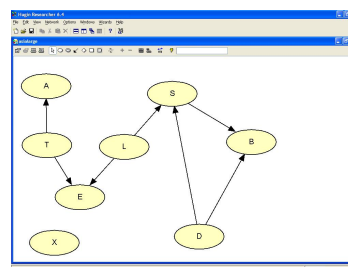
The disadvantage is that there seems to be no coherent and principled method to answer the query in the presence of statistical uncertainty, which is computable.

PC algorithm



Crudest algorithm (HUGIN), 10000 simulated cases

PC algorithm



10000 simulated cases

NPC algorithm

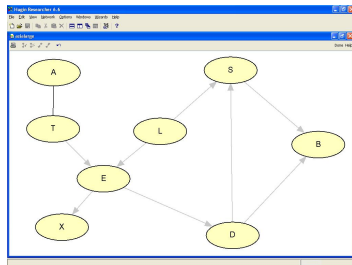
The NPC algorithm (Steck and Tresp 1996) stabilises the PC algorithm by adding a *necessary path condition*.

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 $A \perp_{\mathcal{D}} B | S$, the removal of this edge is suppressed.

This leads to *unresolved queries* which are then passed to the user.

NPC algorithm



10000 simulated cases

taras and D. Poole), pp. 235–43. Morgan Kaufmann Publishers, San Francisco, CA.

Geiger, D. and Heckerman, D. (1997). A characterization of the Dirichlet distribution through global and local independence. *Annals of Statistics*, **25**, 1344–69.

Geiger, D. and Heckerman, D. (2002). Parameter priors for directed acyclic graphical models and the characterization of several probability distributions. *Annals of Statistics*, **30**, 1412–40.

Heckerman, D., Geiger, D., and Chickering, D. M. (1995). Learning Bayesian networks: The combination of knowledge and statistical data. *Machine Learning*, **20**, 197–243.

Robinson, R. W. (1977). Counting unlabelled acyclic digraphs. In *Lecture notes in mathematics: Combinatorial mathematics V*, (ed. C. H. C. Little). Springer-Verlag, New York.

Blackburn, P. (1992). *Modal logic*. Wiley-Interscience, New York.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, **6**, 461–4.

Spiegelhalter, D. J. and Lauritzen, S. L. (1990). Sequential updating of conditional probabilities on directed graphical structures. *Networks*, **20**, 579–605.

Spirtes, P., Glymour, C., and Scheines, R. (1993). *Causation, prediction and search*. Springer-Verlag, New York. Reprinted by MIT Press.

Steck, H. and Tresp, V. (1996). Bayesian belief networks for data mining. In *Proceedings of 2nd workshop on Data Mining und Data Warehousing als Grundlage moderner entscheidungsunterstützender Systeme*,

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, **19**, 716–23.

Aldous, D. (1990). A random walk construction of uniform spanning trees and uniform labelled trees. *SIAM Journal on Discrete Mathematics*, **3**, (4), 450–65.

Atay-Kayis, A. and Massam, H. (2005). A Monte Carlo method for computing the marginal likelihood in non-decomposable graphical Gaussian models. *Biometrika*, **92**, 317–35.

Bollobás, B. (1998). *Modern graph theory*. Springer-Verlag, New York.

Böttcher, S. G. (2001). Learning Bayesian networks with

pp. 145–54. Magdeburg, Germany. University of Magdeburg.

mixed variables. In *Proceedings of the eighth international workshop in artificial intelligence and statistics*, pp. 149–56.

Böttcher, S. G. and Dethlefsen, C. (2003). *deal*: A package for learning Bayesian networks. *Journal of Statistical Software*, **8**, 1–40.

Chickering, D. M. (2002). Optimal structure identification with greedy search. *Journal of Machine Learning Research*, **3**, 507–54.

Cooper, G. F. and Herskovits, E. (1992). A Bayesian method for the induction of probabilistic networks from data. *Machine Learning*, **9**, 309–47.

Geiger, D. and Heckerman, D. (1994). Learning Gaussian networks. In *Proceedings of the 10th conference on uncertainty in artificial intelligence*, (ed. R. L. de Man-