Conditional Independence and Markov Properties

Lecture 1
Saint Flour Summerschool, July 5, 2006

Steffen L. Lauritzen, University of Oxford

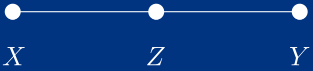
Overview of lectures

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.

Conditional independence

The notion of conditional independence is fundamental for graphical models.

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

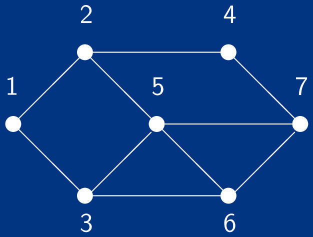


If the random variables have density w.r.t. a product measure μ, the conditional independence is reflected in the relation

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

where f is a generic symbol for the densities involved.

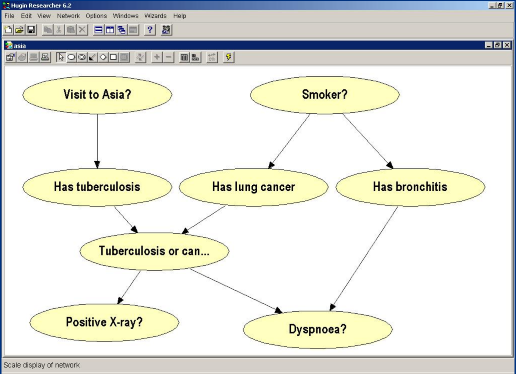
Graphical models



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

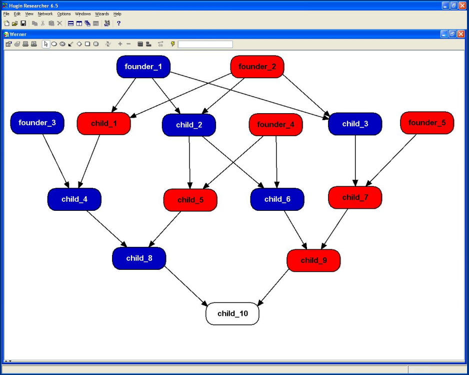
Then a set of variables A is conditionally independent of set B, given the values of a set of variables C if C *separates A from B*.

A directed graphical model



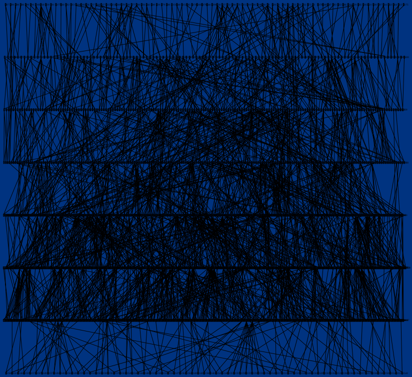
Directed model showing relations between risk factors, diseases, and symptoms.

A pedigree



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

A highly complex pedigree



Family relationship of 1641 members of Greenland Eskimo population.

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:

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

Fundamental properties

For random variables X, Y, Z, and W it holds

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

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

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

(C4) if $X \perp\!\!\!\perp Y \,|\, Z$ and $X \perp\!\!\!\perp W \,|\, (Y, Z)$, then $X \perp\!\!\!\perp (Y, W) \,|\, Z$;

If density w.r.t. product measure $f(x, y, z) > 0$ also

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

Additional note on (C5)

$f(x,y,z) > 0$ is *not necessary* for (C5). Enough e.g. that $f(y,z) > 0$ for all (y,z) or $f(x,z) > 0$ for all .

In discrete and finite case it is even enough that the bipartite graphs $\mathcal{G}_+ = (\mathcal{Y} \cup \mathcal{Z}, E_+)$ defined by

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

are all connected.

Alternatively it is sufficient if the same condition is satisfied with X replacing Y.

Is there a simple necessary and sufficient condition?

Graphoid axioms

Ternary relation $\perp_\sigma$ among subsets of a finite set V is *graphoid* if for all disjoint subsets A, B, C, and D of V:

(S1) if $A \perp_\sigma B \mid C$ then $B \perp_\sigma A \mid C$;

(S2) if $A \perp_\sigma B \mid C$ and $D \subseteq B$, then $A \perp_\sigma D \mid C$;

(S3) if $A \perp_\sigma B \mid C$ and $D \subseteq B$, then $A \perp_\sigma B \mid (C \cup D)$;

(S4) if $A \perp_\sigma B \mid C$ and $A \perp_\sigma D \mid (B \cup C)$, then $A \perp_\sigma (B \cup D) \mid C$;

(S5) if $A \perp_\sigma B \mid (C \cup D)$ and $A \perp_\sigma C \mid (B \cup D)$ then $A \perp_\sigma (B \cup C) \mid D$.

Semigraphoid if only (S1)–(S4) holds.

Irrelevance

Conditional independence can be seen as encoding irrelevance in a fundamental way. With the interpretation: *Knowing C, A is irrelevant for learning B,* (S1)–(S4) translate to:

(I1) If, knowing C, learning A is irrelevant for learning B, then B is irrelevant for learning A;

(I2) If, knowing C, learning A is irrelevant for learning B, then A is irrelevant for learning any part D of B;

(I3) If, knowing C, learning A is irrelevant for learning B, it remains irrelevant having learnt any part D of B;

(I4) If, knowing C, learning A is irrelevant for learning B and, having also learnt A, D remains irrelevant for learning B, then both of A and D are irrelevant for learning B.

The property (S5) is slightly more subtle and not generally obvious.

Also the symmetry (C1) is a special property of probabilistic conditional independence, rather than of general irrelevance, where (I1) could appear dubious.

Probabilistic semigraphoids

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

Then basic properties of conditional independence imply:

The relation $\perp\!\!\!\perp$ on subsets of V is a semigraphoid.

If $f(x) > 0$ for all x, $\perp\!\!\!\perp$ is also a graphoid.

Not all (semi)graphoids are probabilistically representable.

Second order conditional independence

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

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

in other words, if the partial correlations are zero

$$\rho_{AB\cdot C} = 0.$$

We then write $A \perp_2 B \,|\, C$.

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.

Separation in undirected graphs

Let $\mathcal{G} = (V, E)$ be finite and simple undirected graph (no self-loops, no multiple edges).

For subsets A, B, S of V, let $A \perp_{\mathcal{G}} B \mid S$ denote that S *separates A from B in $\mathcal{G}$*, i.e. that all paths from A to B intersect S.

Fact: *The relation $\perp_{\mathcal{G}}$ on subsets of V is a graphoid.*

This fact is the reason for choosing the name 'graphoid' for such separation relations.

Geometric Orthogonality

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

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

where $L \ominus N = L \cap N^{\perp}$. Then L and M are said to *meet orthogonally in N.* This has properties

(O1) If $L \perp M \mid N$ then $M \perp L \mid N$;

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

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

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

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

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

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

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 $\ddagger_{\mathcal{U}}$ as

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

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

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

Markov properties for semigraphoids

$\mathcal{G} = (V, E)$ simple undirected graph; $\perp_\sigma$ (semi)graphoid relation. Say $\perp_\sigma$ satisfies

(P) *the pairwise Markov property* if

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

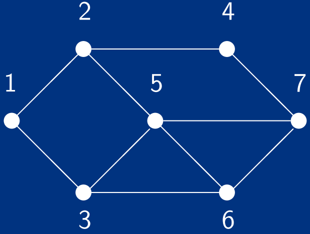
(L) *the local Markov property* if

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

(G) *the global Markov property* if

$$A \perp_\mathcal{G} B \,|\, S \implies A \perp_\sigma B \,|\, S.$$

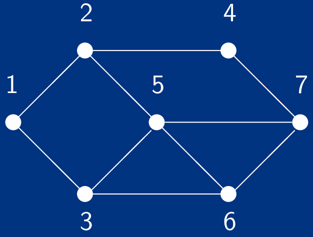
Pairwise Markov property



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

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

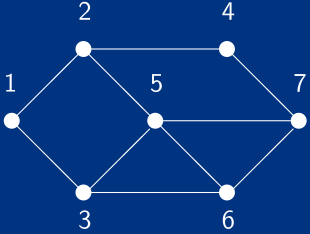
Local Markov property



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

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

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

Structural relations among Markov properties

For any semigraphoid it holds that

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

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

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

so that *in the graphoid case*

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

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

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

(G) implies (L) because $\mathrm{bd}(\alpha)$ separates α from $V \setminus \mathrm{cl}(\alpha)$.

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

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

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

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

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

(P) $\implies$ (G) for graphoids

Asuume (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).

Thus by the induction hypothesis

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

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

For $A \cup B \cup S \subset V$ we choose $\alpha \in V \setminus (A \cup B \cup S)$. Then $A \perp_\mathcal{G} B \,|\, (S \cup \{\alpha\})$ and hence the induction hypothesis yields $A \perp_\sigma B \,|\, (S \cup \{\alpha\})$.

Further, either $A \cup S$ separates B from $\{\alpha\}$ or $B \cup S$ separates A from $\{\alpha\}$. Assuming the former gives $\alpha \perp_\sigma B \,|\, A \cup S$.

Using (S5) we get $(A \cup \{\alpha\}) \perp_\sigma B \,|\, S$ and from (S2) we derive that $A \perp_\sigma B \mid S$.

The latter case is similar.

Factorisation and Markov properties

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

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

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

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

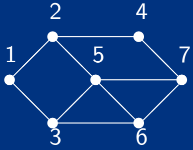
$$f(x) = \prod_{a \in \mathcal{A}} \psi_a(x),$$

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

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

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

Factorization example



The *cliques* 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.

The graph above corresponds to a factorization as

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

Factorisation of the multivariate Gaussian

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

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

where $K = \Sigma^{-1}$ is the *concentration matrix* of the distribution.

Thus *the Gaussian density factorizes w.r.t. $\mathcal{G}$ if and only if*

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

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

Factorization theorem

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

$$(\mathsf{F}) \implies (\mathsf{G})$$

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:

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