Markov properties for undirected graphs

Steffen Lauritzen, University of Oxford

Graphical Models, Lecture 2, Michaelmas Term 2010

October 14, 2010

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:

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

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

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

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

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

(C4) If $X \perp\!\!\!\perp Y \mid Z$ and $X \perp\!\!\!\perp W \mid (Y, Z)$, then $X \perp\!\!\!\perp (Y, W) \mid Z$;

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

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

Conditional independence can be seen as encoding abstract irrelevance. With the interpretation: *Knowing C, A is irrelevant for learning B,* (C1)–(C4) translate into:

- (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 analogous to (C5) 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.

Graphoid axioms

Ternary relation $\perp_\sigma$ 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.

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.

Systems of random variables

For a system V of *labeled random variables* $X_v, v \in V$, we use the shorthand

$$A \perp\!\!\!\perp B \mid C \iff X_A \perp\!\!\!\perp X_B \mid X_C,$$

where $X_A = (X_v, v \in A)$ denotes the variables with labels in A.

The properties (C1)–(C4) imply that $\perp\!\!\!\perp$ *satisfies the semi-graphoid axioms* for such a system, and the graphoid axioms if the joint density of the variables is strictly positive.

Geometric orthogonality

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

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

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

The analogue of (C5) does not hold in general.

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

(P) *the pairwise Markov property* if

$$\alpha \not\sim \beta \Rightarrow \alpha \perp_\sigma \beta \mid V \setminus \{\alpha, \beta\};$$

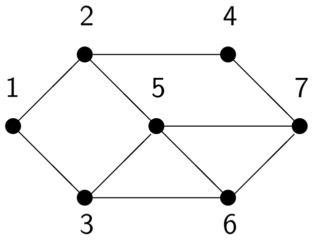
(L) *the local Markov property* if

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

(G) *the global Markov property* if

$$A \perp_\mathcal{G} B \mid S \Rightarrow A \perp_\sigma B \mid S.$$

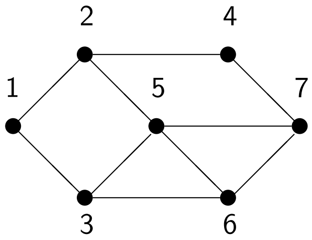
Pairwise Markov property



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

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

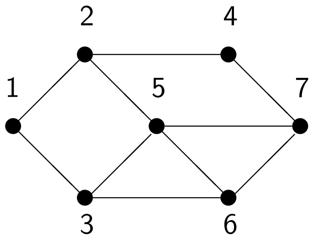
Local Markov property



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

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

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

For any semigraphoid it holds that

$$(\mathrm{G}) \Rightarrow (\mathrm{L}) \Rightarrow (\mathrm{P})$$

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

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

so that *in the graphoid case*

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

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

(G) ⇒ (L) ⇒ (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 $\alpha \not\sim \beta$. Thus

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

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

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

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

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

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). Thus by the induction hypothesis

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

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

(P) $\Rightarrow$ (G) for graphoids, continued

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.

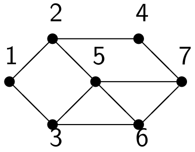
Assume density f w.r.t. product measure on $\mathcal{X}$.
For $a \subseteq V$, $\psi_a(x)$ denotes a function which depends on x_a only, i.e.

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

We can then write $\psi_a(x) = \psi_a(x_a)$ without ambiguity.
The distribution of X *factorizes w.r.t. $\mathcal{G}$* or satisfies (F) if

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

where $\mathcal{A}$ are *complete* subsets of $\mathcal{G}$.
Complete subsets of a graph are sets with all elements pairwise neighbours.



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

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

Let (F) denote the property that f factorizes w.r.t. $\mathcal{G}$ and let (G), (L) and (P) denote Markov properties w.r.t. $\perp\!\!\!\perp$. *It then holds that*

$$(F) \Rightarrow (G)$$

and further: *If $f(x) > 0$ for all x*, $(P) \Rightarrow (F)$.

The former of these is a simple direct consequence of the factorization whereas the second implication is more subtle and known as the *Hammersley–Clifford Theorem*.

Thus in the case of positive density (but typically only then), *all the properties coincide:*

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