More on Markov Properties

Lecture 2
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

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 $X \perp\!\!\!\perp (Y, Z)$.

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 $A \perp_\sigma B \,|\, C$ then $B \perp_\sigma A \,|\, C$;

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

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

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

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

Semigraphoid if only (S1)–(S4) holds.

Semigraphoid examples

- *Graph separation* $\perp_{\mathcal{G}}$ in undirected graph $\mathcal{G}$ forms a graphoid;
- *Variation independence* of projections for a subset U of a product space $\ddagger_{\mathcal{U}}$ forms a semigraphoid;
- *Uncorrelatedness* $\perp_2$ of residuals after linear regression (second order conditional independence) forms a semigraphoid;
- *Orthogonal meet* $\perp$ of closed subspaces of a Hilbert space yields a semigraphoid;
- *Probabilistic* conditional independence.

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.

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

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

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 $f(x) > 0$.

Factorisation and Markov properties

The distribution of X *factorizes w.r.t.* $\mathcal{G}$ or satisfies (F) if

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

$\mathcal{A}$ are *complete* subsets and $\mathcal{C}$ are the cliques of $\mathcal{G}$.

It then holds that

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

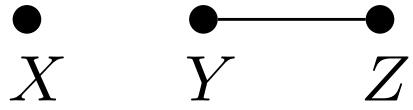
and further:

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:

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

Pairwise Markov but not local Markov



Let $X = Y = Z$ with $P\{X = 1\} = P\{X = 0\} = 1/2$.

This *satisfies (P) but not (L)*.

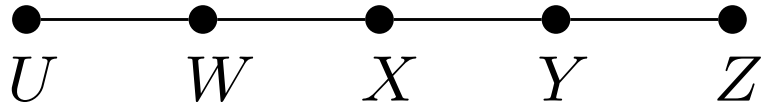
(P): $X \perp\!\!\!\perp Y \,|\, Z$ and $X \perp\!\!\!\perp Z \,|\, Y$.

(L): $\mathrm{bd}(X) = \emptyset$ so (L) would imply $X \perp\!\!\!\perp (Y, Z)$ which is false.

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

Dual graph: $\alpha \check{\sim} \beta$ if and only if $\alpha \not\sim \beta$

Local Markov but not global Markov



Let U and Z be independent with

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

$W = U$, $Y = Z$, and $X = WY$.

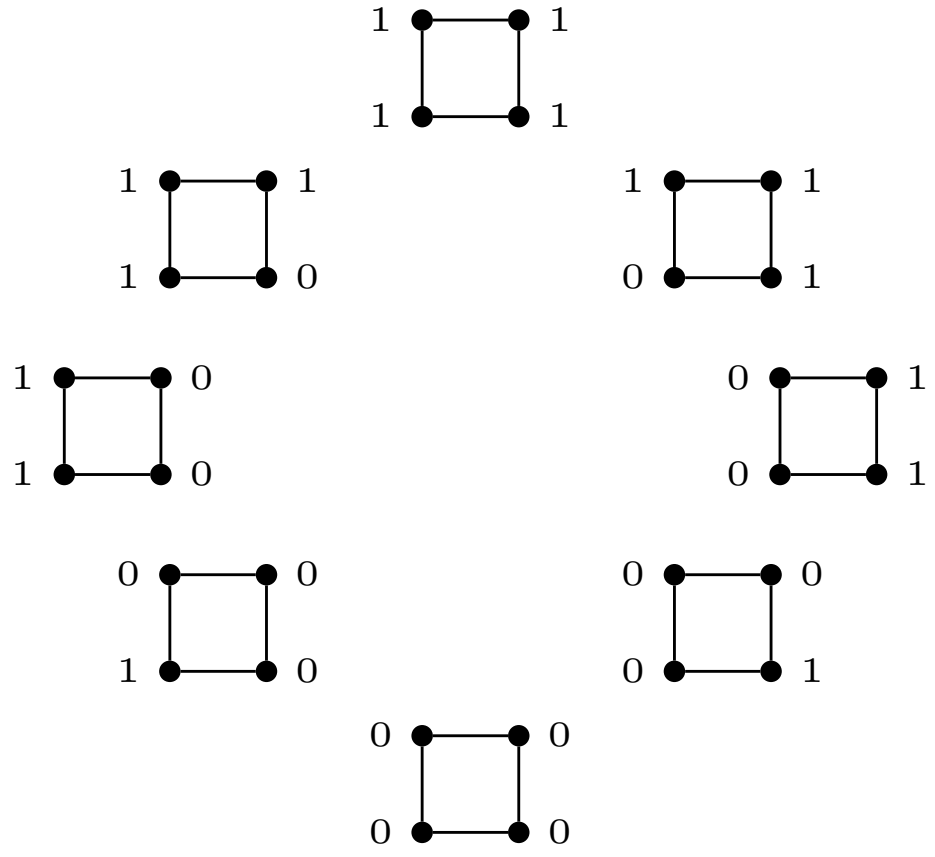
This satisfies (L) but not (G).

(L): Variables depend deterministically on their neighbours.

(G): False that $W \perp\!\!\!\perp Y \mid X$, for example when $X = 0$.

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

Global but not factorizing



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).

To see the latter, assume the density factorizes. Then e.g.

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

so these factors are all positive.

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

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

In fact, (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.

To be shown later.

Instability under limits

Consider a sequence $P_n, n = 1, 2, \ldots$ of probability measures on $\mathcal{X}$ and assume that $A \perp\!\!\!\perp_{P_n} B \,|\, C$.

If $P_n \to P$ (weakly, say) it does *not* hold in general that $A \perp\!\!\!\perp_P B \,|\, C$.

A simple counterexample is as follows: Consider $X = (X_1, X_2, X_3) \sim \mathcal{N}_3(0, \Sigma_n)$ with

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

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

concentration matrix K_n is

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

so for all n it holds that $1 \perp\!\!\!\perp_{P_n} 3 \,|\, 2$.

The critical feature seems to be that K_n does not converge, hence the densities do not converge.

What is a reasonable general additional condition for ensuring closure under limits?

The answer seems to be *convergence in total variation* (A. Klenke, St Flour 2006).

Stability under limits

If $\mathcal{X}$ is discrete and finite and $P_n \to P$ pointwise, *conditional independence is preserved:*

This follows from the fact that

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

and this relation is clearly stable under pointwise limits.

Hence *(G), (L) and (P) are closed under pointwise limits in the discrete case*.

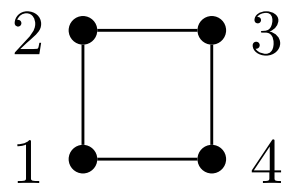
Instability under limits

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

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

This factorizes w.r.t. the graph



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

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

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

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

Markov faithfulness

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

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

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 $\perp\!\!\!\perp_P$ are isomorphic.

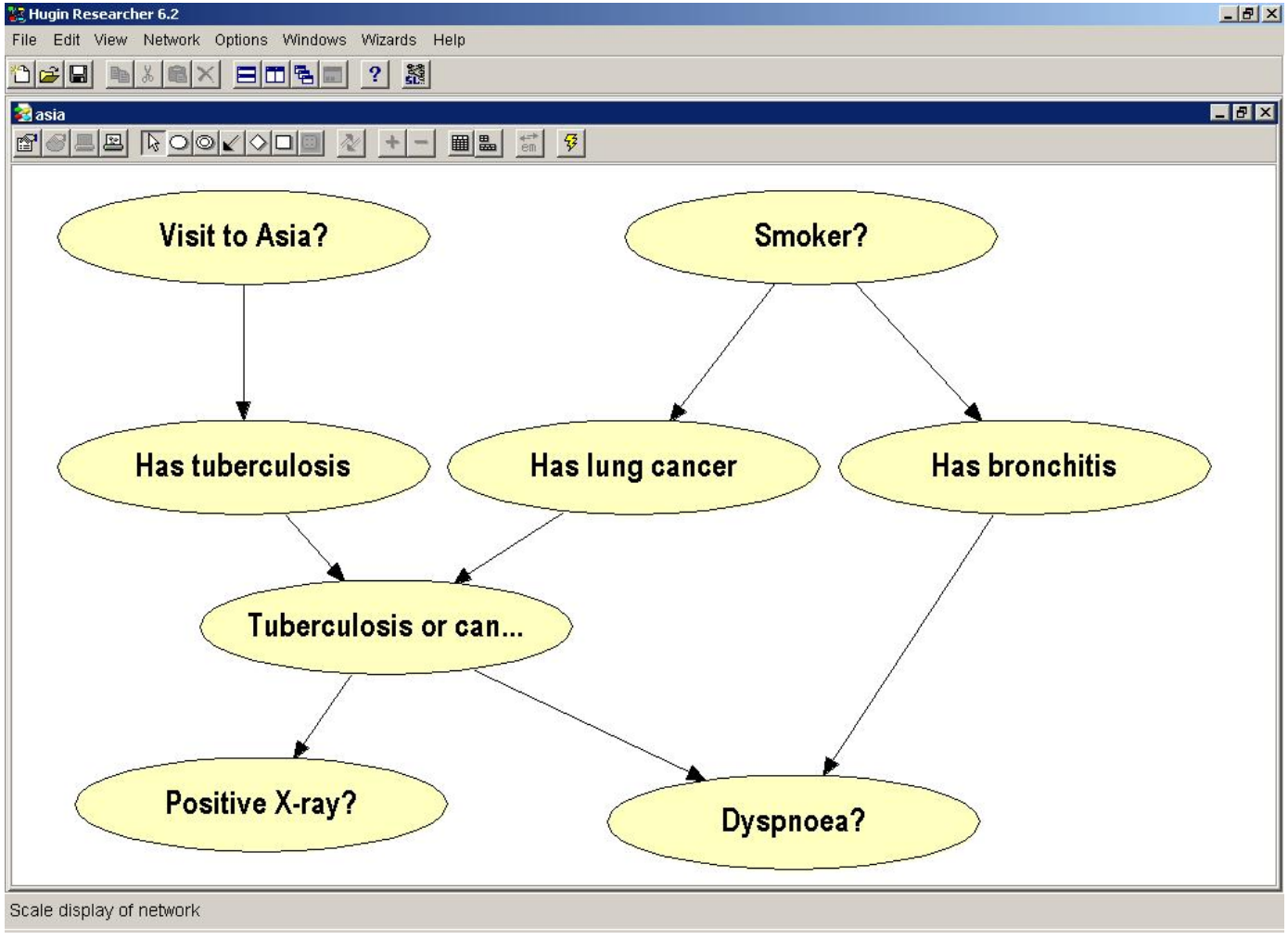
Directed acyclic graphs

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.

Example of a directed graphical model



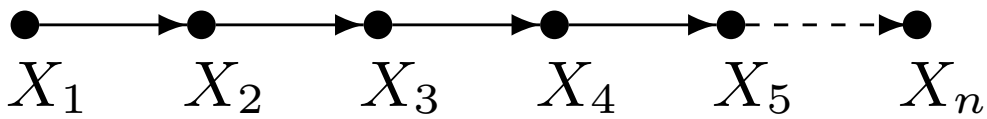
Local directed Markov property

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

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

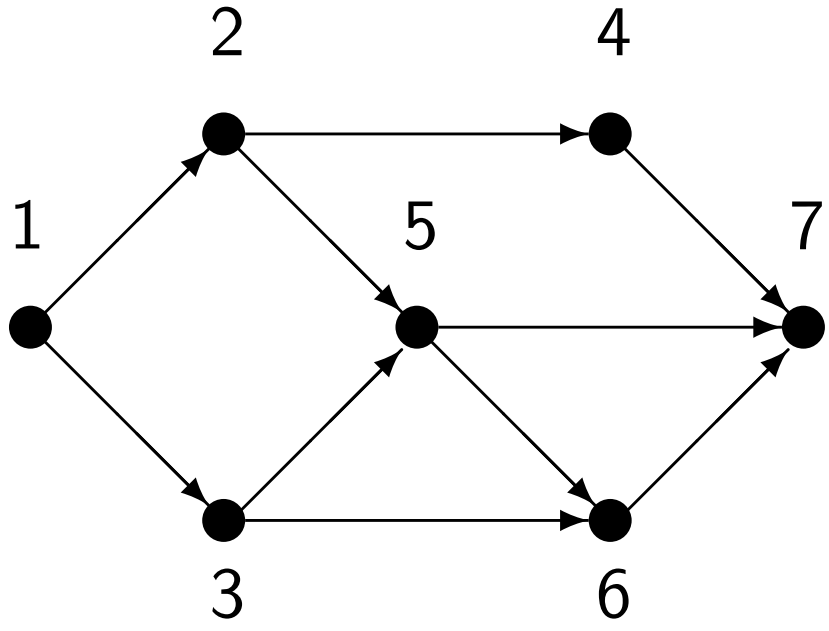
Here $\mathrm{nd}(\alpha)$ are the *non-descendants* of α.

A well-known example is a Markov chain:



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

Local directed Markov property



For example, the local Markov property says

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

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

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

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

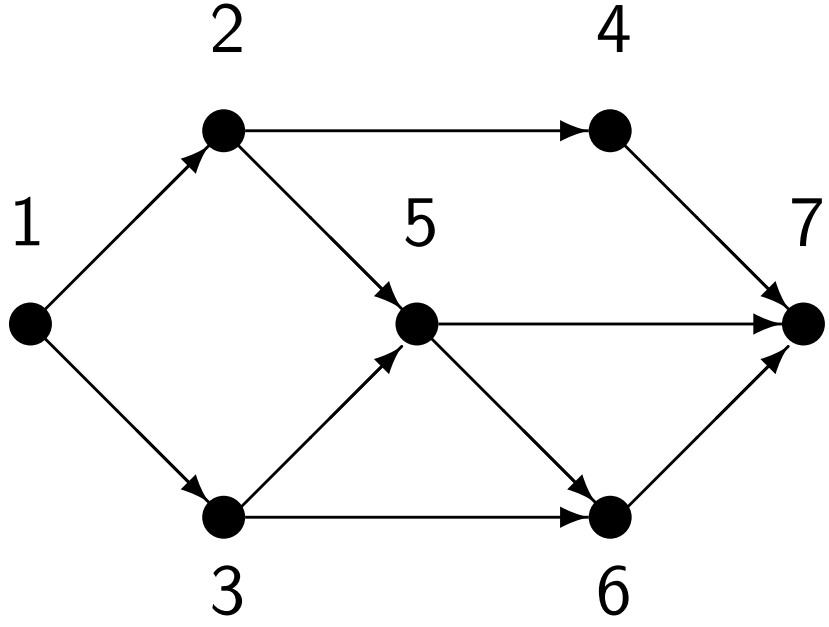
$$\alpha \in \mathrm{pa}(\beta) \implies \alpha < \beta.$$

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

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

Here $\mathrm{pr}(\alpha)$ 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

$4 \perp_\sigma \{1,3\} \,|\, 2,$

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

$3 \perp_\sigma \{2\} \,|\, 1.$

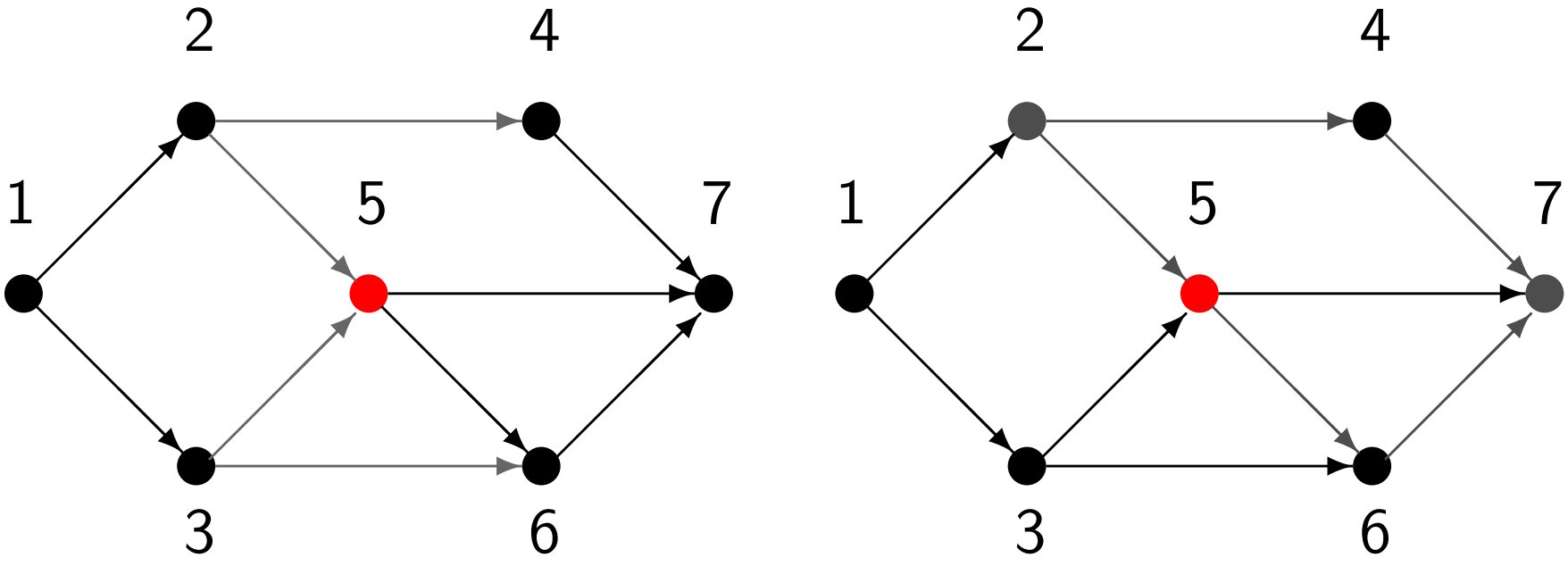
Separation in DAGs

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 $A \perp_{\mathcal{D}} B \,|\, S$.

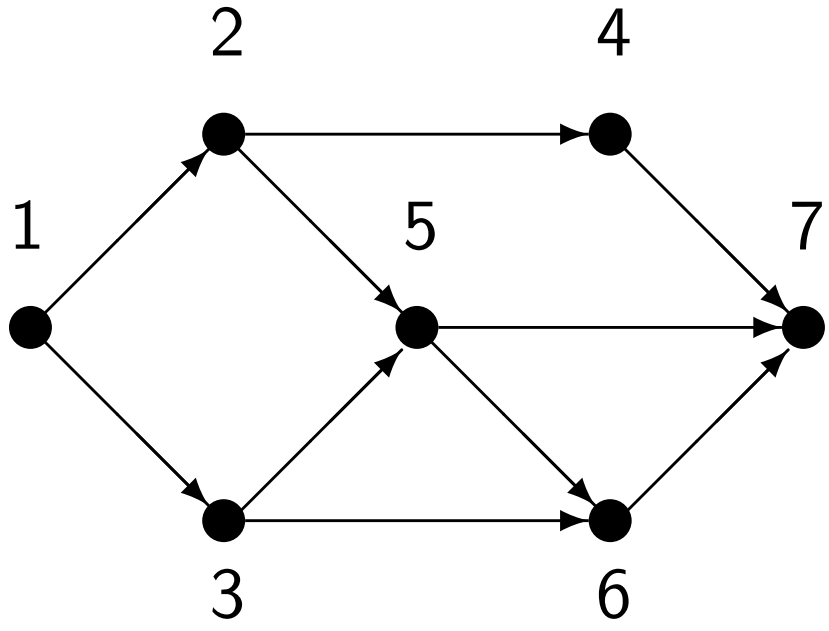
Separation by example



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.

Returning to example



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

Equivalence of Markov properties

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

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

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

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

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

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

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

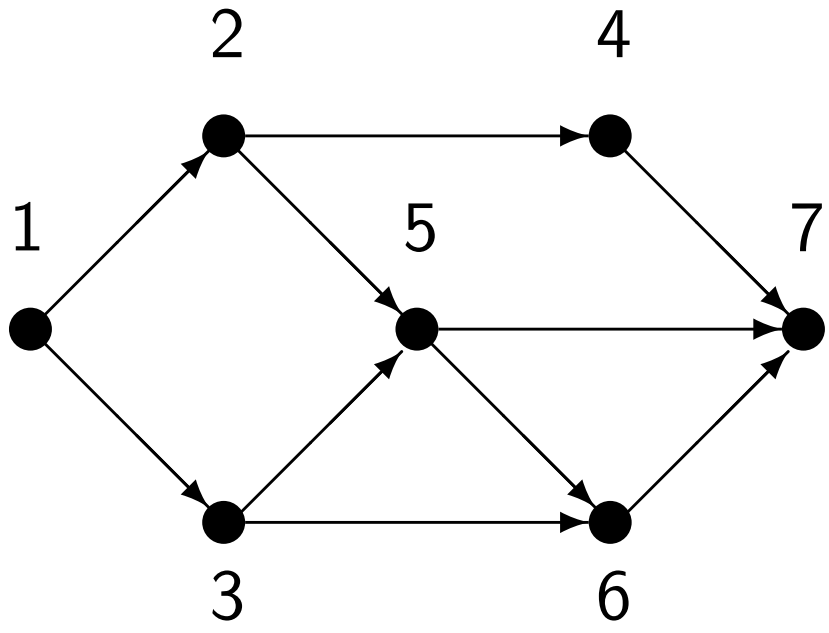
where $k_v \geq 0$ and $\int_{\mathcal{X}_v} k_v(x_v \,|\, x_{\mathrm{pa}(v)})\, \mu_v(dx_v) = 1$.

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

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

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

Example of DAG factorization



The above graph corresponds to the factorization

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

Markov properties and factorization

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

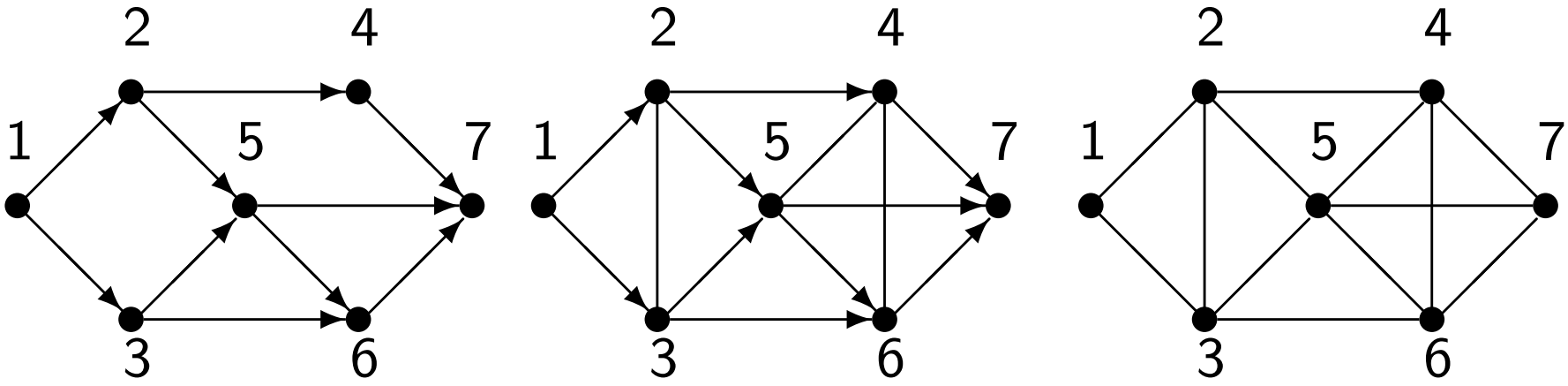
It is then always true that (F) holds if and only if $\perp\!\!\!\perp_P$ satisfies (G),

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

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

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:



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:

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

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

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

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 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.*

Alternative equivalent separation

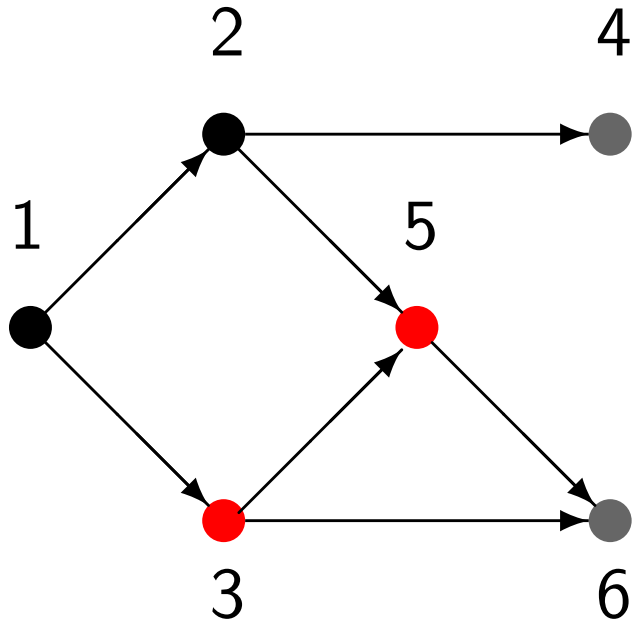
To resolve query involving three sets A, B, S:

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

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

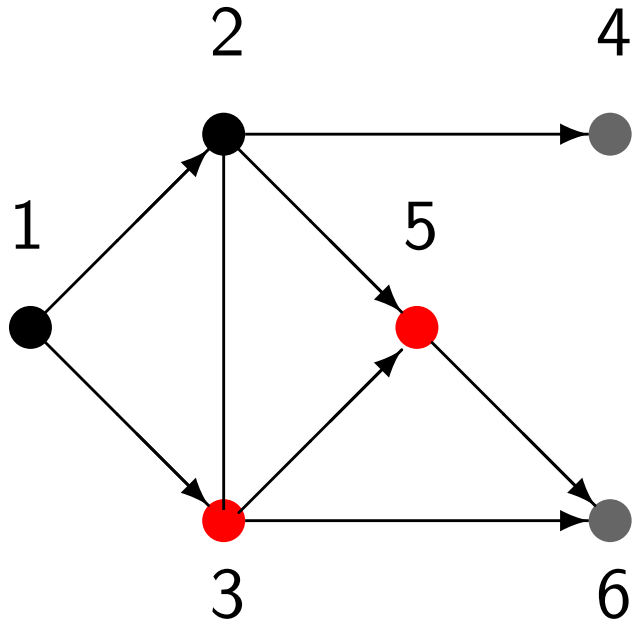
Proof in Lauritzen (1996) needs to allow self-intersecting paths to be correct.

Forming ancestral set



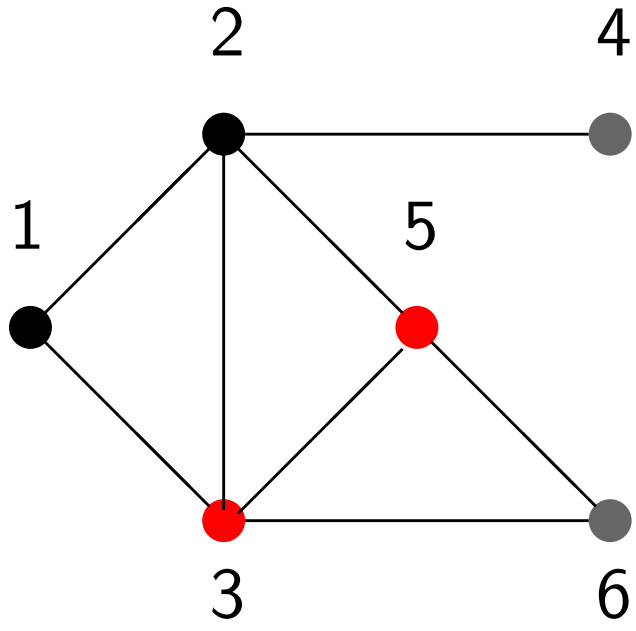
The subgraph induced by all ancestors of nodes involved in the query $4 \perp_m 6 \,|\, 3, 5$?

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 $4 \perp_m 6 \,|\, 3, 5$?

Dropping directions



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

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

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

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. $\mathcal{D}^m_{\mathrm{An}(A \cup B \cup S)}$ and hence satisfies (G) for this graph.

Faithfulness

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

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

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 $\perp\!\!\!\perp_P$ are isomorphic.

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

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

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

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.