Bayesian Networks: Directed Markov Properties (Cont'd) and Markov Equivalent DAGs

Huizhen Yu

janey.yu@cs.helsinki.fi
Dept. Computer Science, Univ. of Helsinki

Probabilistic Models, Spring, 2010

Outline

Outline

Notation

For a DAG $G = (V, E)$,

- G^m: the moral graph.
- G_A: the subgraph induced by a subset A of vertices.
- $\text{pa}(\alpha)$: the set of parents of α.
- $\text{ch}(\alpha)$: the set of children of α.
- $\text{an}(\alpha)$: the set of ancestors of α.
- $\text{de}(\alpha)$: the set of descendants of α.
- An ancestral set: a set A such that $\text{pa}(v) \subseteq A, \forall v \in A$.
- $\text{An}(A)$: the minimal ancestral set containing A;
  $\text{An}(A) = A \cup \cup_{\alpha \in A} \text{an}(\alpha)$.
- A well-ordering: $(\alpha, \beta) \in E \ \Rightarrow \ \text{number}(\alpha) < \text{number}(\beta)$.
- $\text{pr}(\beta)$: the set of predecessors of β, i.e.,

$$\text{pr}(\beta) = \{\alpha \mid \text{number}(\alpha) < \text{number}(\beta)\}.$$

- $\text{bl}(v)$: the Markov blanket of v, equal to $\text{ne}(v)$, the set of neighbors of v in G^m; i.e.,

$$\text{bl}(v) = \text{ne}(v) = \text{pa}(v) \cup \text{ch}(v) \cup \{w \mid \text{ch}(w) \cap \text{ch}(v) \neq \emptyset\}.$$

Main Results from Last Lecture

For a distribution P that factorizes recursively according to a DAG G,

- P factorizes according to the moral graph G^m;
- the marginal distribution $P_A(X_A)$ factorizes recursively according to $G_{\mathrm{An}(A)}$, from which we obtain the global directed Markov property (DG): for any disjoint subsets A, B, S,

$$A \perp B \mid S \text{ in } \left(G_{\mathrm{An}(A \cup B \cup S)}\right)^m \quad \Rightarrow \quad X_A \perp X_B \mid X_S.$$

We have also shown

- The d-Separation concept;
- (DG) is equivalent to the d-separation based Markov property.

Local Directed Markov Property

Denote $\text{nd}(v)$ the set of *non-descendants* of v in G, i.e., $\text{nd}(v) = V \setminus (\text{de}(v) \cup \{v\})$.

Consider v, its parents $\text{pa}(v)$, and its non-descendants $\text{nd}(v)$.

- Because $\text{An}\big(\{v\} \cup \text{pa}(v) \cup \text{nd}(v)\big)$ is the minimal ancestral set containing these vertices,

$$\text{ch}(v) \cap \text{An}\big(\{v\} \cup \text{pa}(v) \cup \text{nd}(v)\big) = \emptyset.$$

- Edges in the moral graph $\big(G_{\text{An}(\{v\} \cup \text{pa}(v) \cup \text{nd}(v))}\big)^m$ are either from $G_{\text{An}(\{v\} \cup \text{pa}(v) \cup \text{nd}(v))}$ or from marriages.
- The two facts above imply that v is not connected directly to $\text{nd}(v) \setminus \text{pa}(v)$ in $\big(G_{\text{An}(\{v\} \cup \text{pa}(v) \cup \text{nd}(v))}\big)^m$.

So by (DG),

$$X_v \perp X_{\text{nd}(v)} \mid X_{\text{pa}(v)}. \tag{1}$$

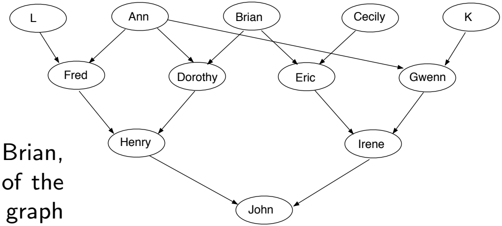
The property in (1) is called the *local directed Markov property* (DL). The discussion above shows also (DG) $\Rightarrow$ (DL).

Illustration of (DL)

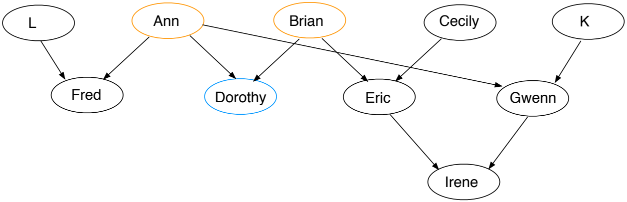
Example Stud-Farm:

$$\text{pa}(\text{Dorothy}) = \{\text{Ann, Brian}\}$$

$$\text{de}(\text{Dorothy}) = \{\text{Henry, John}\}$$



Given the genotypes of Ann and Brian, Dorothy's genotype is independent of the genotype of all the other nodes in the graph below.



Equivalence between Directed Markov Properties

Recall that for an undirected graph, (F) $\Rightarrow$ (G) $\Rightarrow$ (L) $\Rightarrow$ (P).

By contrast, for a DAG, the Markov properties are all equivalent:

$$\text{(DF)} \;\Leftrightarrow\; \text{(DG)} \;\Leftrightarrow\; \text{(DL)} \;\Leftrightarrow\; \text{(DO)},$$

where these properties are as defined earlier:

(DF) P factorizes recursively according to G, i.e.,

$$p(x) = \prod_{v \in V} p\big(x_v \,|\, x_{\mathrm{pa}(v)}\big).$$

(DG) P obeys the global directed Markov property with respect to G, i.e.,

$$A \perp B \,|\, S \quad \text{in} \quad \big(G_{\mathrm{An}(A \cup B \cup S)}\big)^m \quad \Rightarrow \quad X_A \perp X_B \,|\, X_S.$$

(DL) P obeys the local directed Markov property with respect to G, i.e.,

$$X_v \perp X_{\mathrm{nd}(v)} \mid X_{\mathrm{pa}(v)}.$$

(DO) P obeys the ordered directed Markov property with respect to G, i.e., with respect to any well-ordering,

$$X_v \perp X_{\mathrm{pr}(v)} \mid X_{\mathrm{pa}(v)}.$$

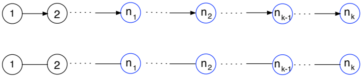
We verified "(DF) $\Rightarrow$ (DG) $\Rightarrow$ (DL)" and "(DO) $\Rightarrow$ (DF)." It is also straightforward to show "(DL) $\Rightarrow$ (DO)" (exercise).

Examples from Earlier Lectures Revisited

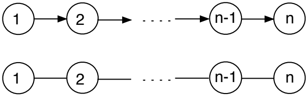
In Lec. 2, we verified the following statements using equations. We can now verify them using only graphs.

If $(X_1, \ldots, X_n)$ is a Markov chain, then

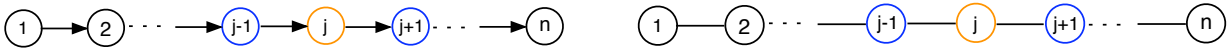
- $(X_{n_1}, X_{n_2}, \ldots, X_{n_k})$ is also a Markov chain, where $1 < n_1 < \ldots < n_k \le n$:



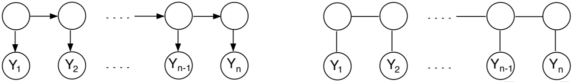
- $(X_n, \ldots, X_1)$ is also a Markov chain:



- $X_j \perp X_{-j} \,|\, (X_{j-1}, X_{j+1})$, where X_{-j} denotes the collection of all the variables but X_j:



In an HMM, the observation variables $Y_1, \ldots, Y_n$ are generally fully dependent:



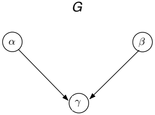
Outline

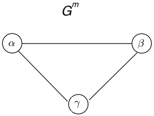
Some Observations

Recall that if P factorizes according to G^m, P does not necessarily factorize recursively according to G:

$$p(x) = \prod_{C \in \mathcal{C}(G^m)} \phi_C(x_C) \quad \not\Rightarrow \quad p(x) = \prod_{v \in V} p\big(x_v \,|\, x_{\mathrm{pa}(v)}\big).$$

Example:





$P(X_\alpha, X_\beta, X_\gamma)$ with

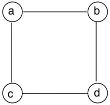
$$p(x) = c \exp\{x_\alpha x_\gamma + x_\beta x_\gamma\}$$

factorizes according to G^m, but

$$p(x) \neq p(x_\alpha)p(x_\beta)p(x_\gamma \,|\, x_\alpha, x_\beta)$$

because $X_\alpha \not\perp X_\beta$.

Conversely, undirected graphs can also represent conditional independence relations that cannot be represented by DAGs: e.g.,

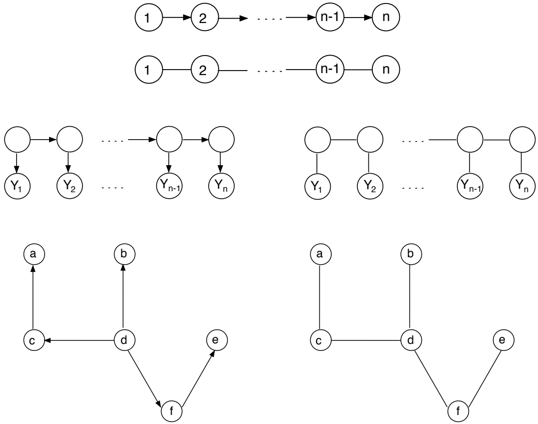


$$\{a\} \perp \{d\} \mid \{b, c\}$$

$$\{b\} \perp \{c\} \mid \{a, d\}$$

Some Observations

But for chains and more generally, rooted trees, the directed and undirected version seems to be equivalent: factorization w.r.t. one implies factorization w.r.t. the other.



Are there other types of graphs for which this is true?

Decomposable Graphs: Definition

A triple (A, B, C) of disjoint subsets of the vertex set V of an undirected graph G is said to *decompose* G, if $V = A \cup B \cup C$, and the following two conditions hold:

1. C separates A from B;
2. C is a complete subset of V.

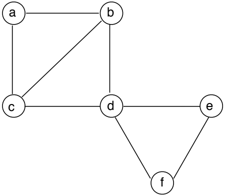
If both A and B are non-empty, we say that the decomposition is *proper*.

We say that an undirected graph G is *decomposable* if either:

(i) it is complete, or
(ii) it possesses a proper decomposition (A, B, C) such that both subgraphs $G_{A \cup C}$ and $G_{B \cup C}$ are decomposable.

Note that this definition is recursive.

Examples: chains, trees, and



Decomposable Graphs: Factorization of Distributions

Proposition 1: Assume that (A, B, S) decomposes an undirected graph $G = (V, E)$. Then a distribution P factorizes according to G if and only if both $P_{A \cup S}$ and $P_{B \cup S}$ factorize according to $G_{A \cup S}$ and $G_{B \cup S}$ respectively and P satisfies

$$p(x) = \frac{p_{A \cup S}(x_{A \cup S})\, p_{B \cup S}(x_{B \cup S})}{p_S(x_S)}.$$

Implication: If the graph is decomposable, then the marginal distributions $P_{A \cup S}$ and $P_{B \cup S}$ admit further factorization in the above form, and P factorizes recursively in terms of these marginals.

Example: For a Markov chain $X = (X_1, \ldots, X_n)$, $P(X)$ can be expressed as

$$p(x) = \frac{p(x_1, x_2) p(x_2, x_3) \cdots p(x_{n-1}, x_n)}{p(x_2) p(x_3) \cdots p(x_{n-1})}.$$

Proof of Proposition 1

Let $\mathcal{C}(G)$ denote the set of cliques of G. Suppose P factorizes according to G, i.e., ,

$$p(x) = \prod_{C \in \mathcal{C}(G)} \phi_C(x_C), \quad \text{for some functions } \phi_C.$$

Since (A, B, S) decomposes G, all the cliques that are not subsets of $A \cup S$ must be subsets of $B \cup S$, so that

$$p(x) = \prod_{C \in \mathcal{C}(G_{A\cup S})} \phi_C(x_C) \prod_{C \in \mathcal{C}(G_{B\cup S})} \phi_C(x_C) = h(x_{A\cup S})\, k(x_{B\cup S}),$$

for some functions h, k. It follows then

$$p(x_{A\cup S}) = \sum_{x_B} h(x_{A\cup S})\, k(x_{B\cup S}) = h(x_{A\cup S})\, \bar{k}(x_S), \quad \text{where } \bar{k}(x_S) = \sum_{x_B} k(x_{B\cup S});$$

$$p(x_{B\cup S}) = \sum_{x_A} h(x_{A\cup S})\, k(x_{B\cup S}) = \bar{h}(x_S)\, k(x_{B\cup S}), \quad \text{where } \bar{h}(x_S) = \sum_{x_A} h(x_{A\cup S});$$

$$p(x_S) = \sum_{x_A} \sum_{x_B} h(x_{A\cup S})\, k(x_{B\cup S}) = \bar{h}(x_S)\, \bar{k}(x_S).$$

So

$$\frac{p_{A\cup S}(x_{A\cup S})\, p_{B\cup S}(x_{B\cup S})}{p_S(x_S)} = \frac{h(x_{A\cup S})\, \bar{k}(x_S)\, \bar{h}(x_S)\, k(x_{B\cup S})}{\bar{h}(x_S)\, \bar{k}(x_S)} = h(x_{A\cup S})\, k(x_{B\cup S}) = p(x).$$

The converse is evident.

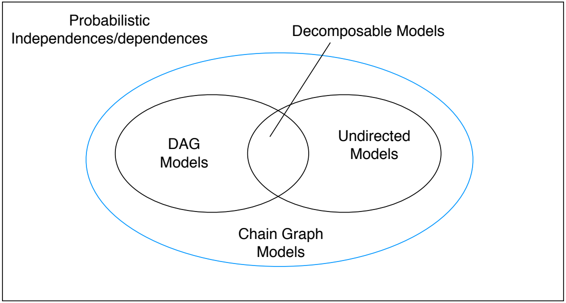
Decomposable Graphs: Directed or Undirected?

Definition: A DAG $G = (V, E)$ is said to be *perfect* if for all $v \in V$, $\text{pa}(v)$ is a complete subset of G, in other words, if its undirected version $G^{\sim} = G^{m}$.

Note: $G^{\sim}$ of a perfect DAG is decomposable. (How to see this?)

Proposition 2: Let G be a perfect DAG and $G^{\sim}$ its undirected version. Then a distribution P is directed Markov with respect to G if and only if it factorizes according to $G^{\sim}$.

Implication:



Proof of Proposition 2

Proof: Since $G^{\sim} = G^m$, if P factorizes recursively according to G, then it factorizes according to $G^{\sim}$.

Conversely, if P factorizes according to $G^{\sim}$, we show P recursively factorizes according to G by induction on the number of vertices $|V|$.

- For $|V| = 1$, the conclusion holds trivially.
- For $|V| > 1$, consider a terminal vertex α. Then,

  $$(V \setminus (\{\alpha\} \cup \mathrm{pa}(\alpha)), \{\alpha\}, \mathrm{pa}(\alpha)) \text{ decomposes } D^{\sim}.$$

  So by Proposition 1, P can be expressed as

  $$p(x) = \frac{p(x_{V\setminus\{\alpha\}})\, p(x_\alpha, x_{\mathrm{pa}(\alpha)})}{p(x_{\mathrm{pa}(\alpha)})} = p(x_{V\setminus\{\alpha\}})p(x_\alpha \,|\, x_{\mathrm{pa}(\alpha)}), \tag{2}$$

  and $P_{V\setminus\{\alpha\}}$ factorizes according to $G^{\sim}_{V\setminus\{\alpha\}}$. By induction, $P_{V\setminus\{\alpha\}}$ factorizes recursively according to $G_{V\setminus\{\alpha\}}$:

  $$p(x_{V\setminus\{\alpha\}}) = \prod_{v \in V\setminus\{\alpha\}} p(x_v \,|\, x_{\mathrm{pa}(v)}).$$

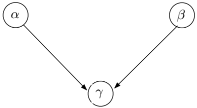
  Together with Eq. (2) this shows that P factorizes recursively according to G.

Outline

Markov Equivalence of DAGs

Terminologies for a DAG G:

- *Skeleton* of G: $G^{\sim}$, the undirected version of G.
- A *complex* (also called an *immorality*): a subgraph of G induced by three vertices (α, γ, β), of the form shown on the right. I.e., the two parents α, β are not joined in G.



Definition: We say that two DAG G and $\widetilde{G}$ are *Markov equivalent* if they represent the same conditional independence relations.

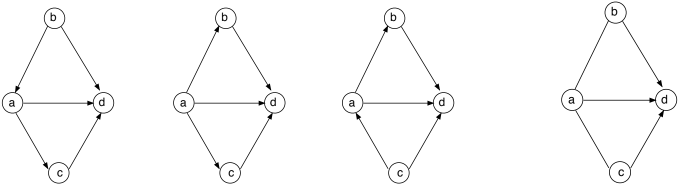
Theorem: *Two DAG G and $\widetilde{G}$ are Markov equivalent if and only if they have the same skeleton and the same complexes.*

- This theorem has important implication for automatic model selection:

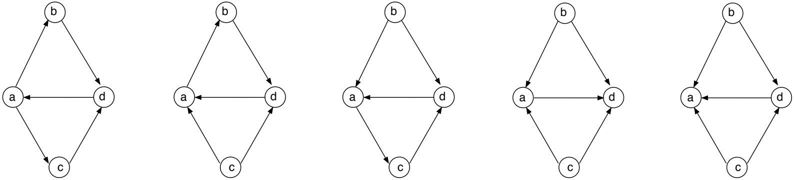
  Since (DG) $\Leftrightarrow$ (DF), the set of distributions associated with G is the same as that associated with any DAG Markov equivalent to G. When edges do not have causal interpretations, it would be in vain to try to distinguish between equivalent models, and we need to compare equivalent classes of DAGs instead. (The representative graph of an equivalent class is a chain graph, interestingly.)

Examples of Markov Equivalent DAGs

An equivalent class consisting of three DAGs and their "representative" (from Andersson et al. 1997):



What happened to the other 5 possibilities?



Proof of the Markov Equivalence Theorem for DAG

Proof of sufficiency: if G and $\widetilde{G}$ have the same skeleton and the same complexes, then they have the same moral graphs for all subgraphs induced by the same set of vertices. This means that for any disjoint subsets A, B, S of vertices,

$$A \perp B \mid S \text{ in } \left(G_{\mathrm{An}(A\cup B\cup S)}\right)^m \quad \Leftrightarrow \quad A \perp B \mid S \text{ in } \left(\widetilde{G}_{An(A\cup B\cup S)}\right)^m.$$

So G and $\widetilde{G}$ represent the same conditional independence relations.

Proof of necessity: we argue this by using counterexamples.

- Same skeleton: suppose there is an edge between α and β in $\widetilde{G}$ but not in G. Then α, β is non-adjacent in $\left(G_{\mathrm{An}(\{\alpha,\beta\})}\right)^m$, so $\alpha \perp \beta \mid \mathrm{An}(\{\alpha,\beta\}) \setminus \{\alpha,\beta\}$. Assume $x \in \{-1,1\}^{|V|}$. Let $p(x) \propto \exp\{x_\alpha x_\beta\}$. Then P factorizes according to $\widetilde{G}$, but X_α and X_β are not independent given $X_{\mathrm{An}(\{\alpha,\beta\})\setminus\{\alpha,\beta\}}$, a contradiction.
- Same complexes: suppose there is a complex induced by (α,γ,β) in G, but in $\widetilde{G}$, the connection between them is different. Again, α, β is non-adjacent in $\left(G_{\mathrm{An}(\{\alpha,\beta\})}\right)^m$, so $\alpha \perp \beta \mid \mathrm{An}(\{\alpha,\beta\}) \setminus \{\alpha,\beta\}$. Assume $x \in \{-1,1\}^{|V|}$. Let $p(x) \propto \exp\{x_\alpha x_\gamma + x_\beta x_\gamma\}$. Then P factorizes according to $\widetilde{G}$, but X_α and X_β are not independent without X_γ given, a contradiction.

Note: This theorem is shown by Verma and Pearl (1991) and independently by Frydenberg (1990) (for chain graphs). In the above we followed the latter proof, which uses (DG) instead of d-separation.

Further Readings

For directed Markov properties and decomposable graphs:

1. Robert G. Cowell et al. *Probabilistic Networks and Expert Systems*, Springer, 2007. Chap. 5.2, 5.3.