Lecture Outline

- Quick review of basics for conditional independence relations
- Semantics for *directed* and *undirected* graphical models
- *Hammersley-Clifford* Theorem (1971) for Markov Networks and undirected graphs
- *d-separation* for graphical models, and the related *Bayes Ball* and ***PC*** algorithms

Following A.P.Dawid's convenient notation:

Let $(\bullet \perp\!\!\!\perp \bullet)$ denote independence between (sets of) variables.

$(A \perp\!\!\!\perp B)$ means that $\boldsymbol{P}(A,B) = \boldsymbol{P}(A)\, P(B)$

or equivalently $\mathbf{P(A) = P(A \mid B)}$ **and** $\mathbf{P(B) = P(B \mid A)}$

And let $(\bullet \perp\!\!\!\perp \bullet \mid \bullet)$ denote conditional independence.

$(A \perp\!\!\!\perp B \mid C)$ means that $\boldsymbol{P}(\mathbf{A,B \mid C}) = \boldsymbol{P}(\mathbf{A \mid C})\, \boldsymbol{P}(\mathbf{B \mid C})$

or equivalently $\boldsymbol{P}(\mathbf{A \mid C}) = \boldsymbol{P}(\mathbf{A \mid B,C})$ **and** $\boldsymbol{P}(\mathbf{B \mid C}) = \boldsymbol{P}(\mathbf{B \mid A,C})$

We saw last lecture that you cannot conclude conditional independence from unconditional independence, or *vice versa*.

Recall *Simpson's Paradox* or that the *Linear Opinion Pool* for combining expert opinions does not support commuting "updating" with "pooling."

The elementary analysis behind these results is captured in the following inequality.

Let A, B be binary variables. Let $\boldsymbol{P}_1(\mathrm{A,B})$ and $\boldsymbol{P}_2(\mathrm{A,B})$ be two distributions, and let $\boldsymbol{P}_3(\mathrm{A,B})$ be a third distribution obtained by averaging the first two. That is,

fix $0 < \boldsymbol{x} < 1$, and define: $\boldsymbol{P}_3(\mathrm{A,B}) = \boldsymbol{x}\boldsymbol{P}_1(\mathrm{A,B}) + (1-\boldsymbol{x})\boldsymbol{P}_2(\mathrm{A,B})$.

Note that, unless either $\boldsymbol{P}_1(\mathrm{A}) = \boldsymbol{P}_2(\mathrm{A})$, or $\boldsymbol{P}_1(\mathrm{B}) = \boldsymbol{P}_2(\mathrm{B})$, we have this ***inequality***:

$$\boldsymbol{P}_3(\mathrm{A} \mid \mathrm{B}) = \frac{\boldsymbol{x}\boldsymbol{P}_1\boldsymbol{(A,B)} + (1\text{-}\boldsymbol{x})\boldsymbol{P}_2\boldsymbol{(A,B)}}{\boldsymbol{x}\boldsymbol{P}_1\boldsymbol{(B)} + (1\text{-}\boldsymbol{x})\boldsymbol{P}_2\boldsymbol{(B)}}$$

$$\neq \boldsymbol{x}\frac{\boldsymbol{P}_1\boldsymbol{(A,B)}}{\boldsymbol{P}_1\boldsymbol{(B)}} + (1\text{-}\boldsymbol{x})\frac{\boldsymbol{P}_2\boldsymbol{(A,B)}}{\boldsymbol{P}_2\boldsymbol{(B)}} = \boldsymbol{x}\boldsymbol{P}_1(\mathrm{A} \mid \mathrm{B}) + (1\text{-}\boldsymbol{x})\boldsymbol{P}_2(\mathrm{A} \mid \mathrm{B}).$$

This result may be visualized using Barycentric coordinates:

http://www.stat.cmu.edu/~eairoldi/classes/tetra.10702/tetraPlay.html

- **The average (marginalization) of two distributions on the surface of independence may be off the surface of independence.**
- **A distribution on the surface of independence may be the average of two distributions that are off the surface of independence.**

Now, let C be a third binary variable with P(A,B,C) a joint probability on {A,B,C}.

Define $\boldsymbol{P}_1$(A,B) = $\boldsymbol{P}$(A,B | C = 0), $\boldsymbol{P}_2$(A,B) = $\boldsymbol{P}$(A,B | C = 1), and let P(C = 0) = $\boldsymbol{x}$.

The result is that $\boldsymbol{P}$(A,B) = $\boldsymbol{P}_3$(A,B) and the following are seen to be invalid inferences:

1. $(\mathbf{A} \perp\!\!\!\perp \mathbf{B} \mid \mathbf{C}) \not\Rightarrow (\mathbf{A} \perp\!\!\!\perp \mathbf{B})$
2. $(\mathbf{A} \perp\!\!\!\perp \mathbf{B}) \not\Rightarrow (\mathbf{A} \perp\!\!\!\perp \mathbf{B} \mid \mathbf{C})$

Semantics for *DAGS* and *Undirected Graphs*

A *D*irected *A*cyclic *G*raph [*DAG*] is a set of *nodes* and *directed edges* between some of the nodes, where the relation of a directed path connecting nodes is a strict partial order. That is, there are no loops.

The nodes stand for (sets of) random variables, and the directed edges indicate how to factor the joint distribution over these variables, as follows:

Defn.: Let $Parents(X_i)$ be the set of immediate predecessors of node X_i.

$$P(X_1, \ldots, X_n) = \prod_i P(\, X_i \mid Parents(X_i)\,)$$

where $P(X \mid \emptyset) = P(X)$, by stipulation.

Consider this elementary DAG.

$$X \rightarrow Y \rightarrow Z$$

So, from this DAG, $P(X,Y,Z) = P(X)\, P(Y \mid X)\, P(Z \mid Y).$

Compare this to the generic factorization, which holds for all distributions,

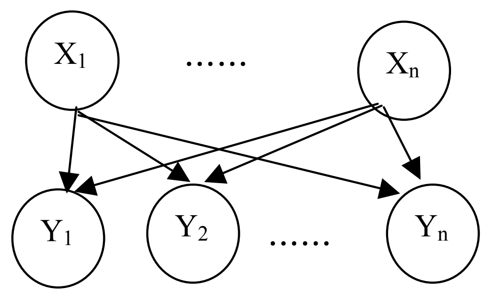
$$P(X,Y,Z) = P(X)\, P(Y \mid X)\, P(Z \mid X,Y).$$

Hence, this graph entails that $P(Z \mid Y) = P(Z \mid X,Y),$

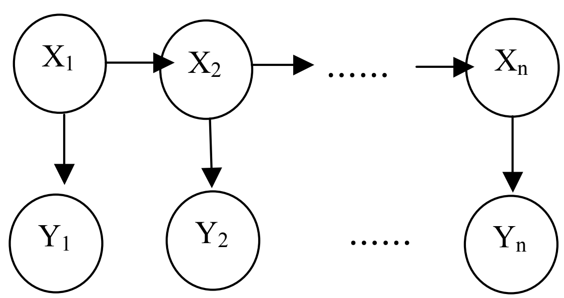
or $(Z \perp\!\!\!\perp X \mid Y).$

(We postpone characterizing the set of conditional independences in a DAG until after the next section of slides, dealing with undirected graphs.)

Two Examples



Factor Analysis



Hidden Markov Models [HMM's]

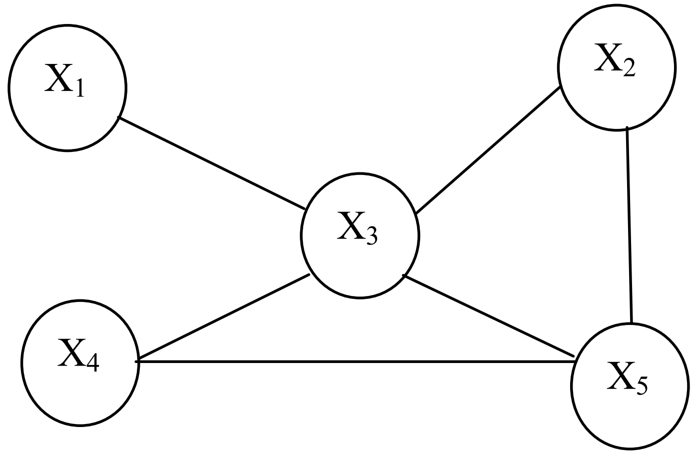
Markov Networks
Undirected Graphical Models

Next we consider simpler graphs in which edges are undirected.

Distinguish a node's immediate neighbors from other nodes.

The undirected graph provides information about conditional independence as follows:

> *Every node is conditionally independent of its non-neighbors, given its neighbors.*



$$(X_1 \perp\!\!\!\perp \{ X_2, X_4, X_5 \} \mid X_3).$$

Definitions:

- The (set) variable V is a *Markov Blanket* for (set) variable X if and only if

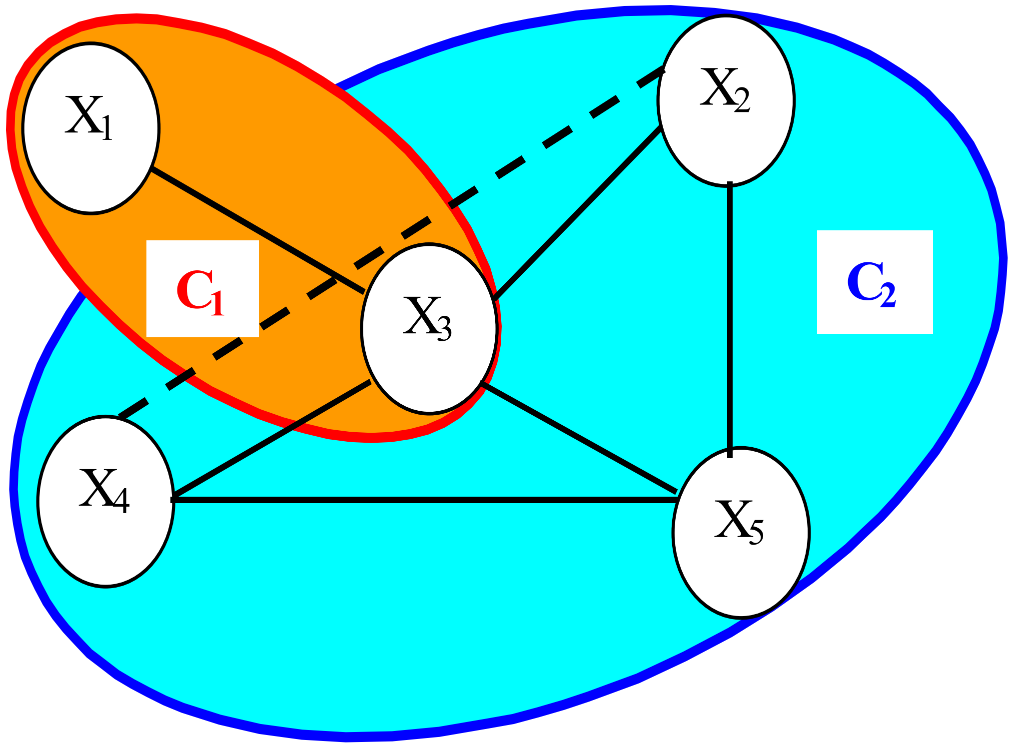
$$(X \perp\!\!\!\perp Y \mid V) \quad \text{for each } \; Y \notin V.$$

- and it is a *Markov Boundary* if it is a *minimal* Markov Blanket.

So X_3 is the Markov Boundary for X_1, and $\{X_2, X_4, X_5\}$ is the Markov Boundary for X_5.

An undirected graph provides a factorization of a joint probability as follows:

- A *clique* is a fully connected (maximal) subgraph, which we denote C_i.



For each clique, $\boldsymbol{C_i}$, define a non-negative function $\boldsymbol{g_i}$**:** $\boldsymbol{C_i} \rightarrow \Re^{\geq 0}$ called the *potential* – it might be used, for example, as an index of association of a node to that clique.

Let $\boldsymbol{Z}$ be the normalizing constant over all nodes, $\{X_1, \ldots., X_n\}$,

$$\boldsymbol{Z} = \sum_{\boldsymbol{j}=1}^{\boldsymbol{n}} \Pi_{\mathrm{i}}\, \mathrm{g_i}(X_{\mathrm{j}})$$

where we use only cases for which $\boldsymbol{g_i}$ is well defined.

Use the potential functions to define the factorization of a joint probability distn.

$$\boldsymbol{P(X_1, \ldots., X_n)} = \frac{1}{\boldsymbol{Z}} \boldsymbol{\Pi_{\mathrm{i}}\, \mathrm{g_i}(\mathrm{C_i})}.$$

Thus the factorization is by a product of functions defined over the cliques.

Definition:

Given a probability distribution ***P***, the undirected graph **G** ***represents the conditional independence relations in P*** (and then say that **G** is a ***Markov Field*** w.r.t. ***P***):

if graph-separation entails conditional independences for ***P***, i.e.,

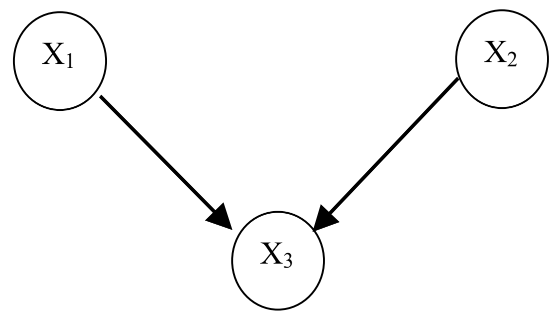
if the set of nodes ***Z*** lies in all paths in ***G*** connecting set ***X*** to set ***Y***, then

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

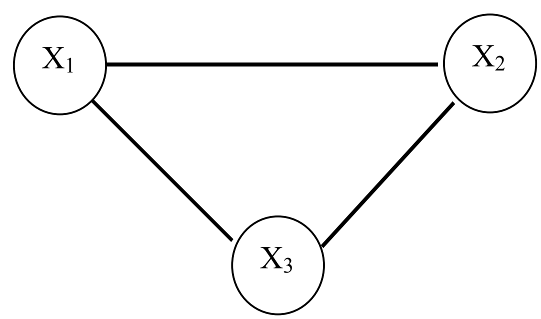
Hammersley-Clifford Theorem (1971):

- The graph ***G*** is a Markov Field w.r.t. ***P*** if it factors as the normalized product of non-negative functions of cliques.
- If the probability function is strictly positive, the converse holds as well for a strictly positive function of cliques.

However, not all probabilistic independence relations are captured by undirected graphs.

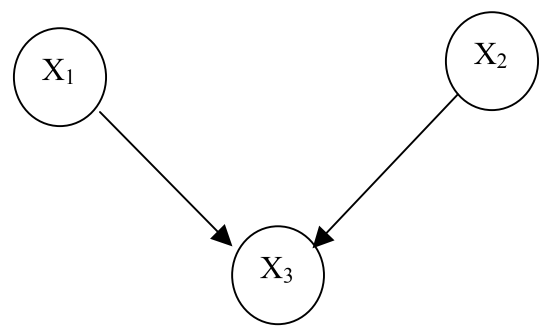


The best we can do with a directed graph is the degenerate “clique,” which entails no independences.



Now, we return to characterizing independence relations in DAGs. Consider a path through the graph, ignoring the direction of arrows and attending solely to connected nodes.

Defn. A node in DAG is a ***collider*** on a path if there are incoming directed arrows.



X_3 is a *collider* on the path connecting X_1 and X_2.

In this DAG, though $(X_1 \perp\!\!\!\perp X_2)$, it is ***not*** the case that $(X_1 \perp\!\!\!\perp X_2 \mid X_3)$.

So, *colliders* do not entail conditional independence, given the *collider*.

Definition (***d-separation***):

The set of variables ***Z*** ***d-separates*** the set ***X*** from the set ***Y*** if and only if for each undirected path between ***X*** and ***Y***, there is a node ***W*** such that

either ***W*** is a collider on this path and neither ***W*** nor any of its descendents belongs to ***Z***,

or ***W*** is not a collider on this path and ***W*** belongs to ***Z***.

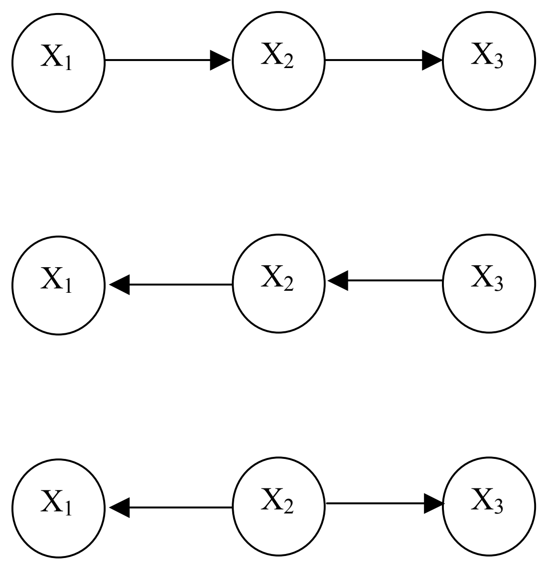
Theorem (Pearl – Spirtes – Glymour – Scheines ?) If a joint distribution ***P*** factors according to a DAG, and if ***X***, ***Y***, and ***Z*** are disjoint subsets of nodes such that ***Z*** ***d-separates*** ***X*** from ***Y*** in the graph, then ***P*** satisfies

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

The "Bayes-Ball" algorithm: implement ***d-separation.*** (See p.17, chpt. 2 Jordan's book)

Equivalent DAGs and the PC algorithm

These 3 DAGs all have the same indpendence relation **$(X_1 \perp\!\!\!\perp X_3 \mid X_2)$.**



We can determine the equivalence class for a given DAG with the PC algorithm.

http://www.phil.cmu.edu/projects/tetrad_download/

Example of an undirected graph whose conditional independences cannot be obtained with a directed graph