Lecture 2:

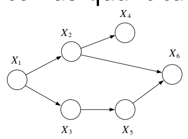
Directed Graphical Models

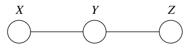
Sam Roweis

January 9, 2002

Joint Probabilities

- Out goal is to represent $\mathsf{P}(\mathbf{X}) = \mathsf{P}(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)$ efficiently when there are many variables.
- For n discrete variables of arity k, the naive (table) representation requires k^n entries.
- Graphical models represent joint distributions compactly using a set of "local" relationships.
- Each random variable in our probabilistic model corresponds to a node in the graph.
- There are directed/undirected edges between the nodes which tell us qualitatively about the factorization of the joint probability.
- There are functions stored at the nodes which tell us quantitatively the distribution





- Also known as Bayes(ian) (Belief) Net(work)s.

Outline

- Joint Probabilities & Graphical Models
- Directed Acyclic Graphical Models
- Missing Edges & Conditional Independence
- D-Separation
- Bayes-Ball Algorithm
- Hammersley-Clifford Theorem

Directed Graphical Models

- Consider *directed acyclic graphs* over n variables.
- Each node has (possibly empty) set of parents π_i.
- Each node maintains a function $f_i(\mathbf{x}_i; \mathbf{x}_{\pi_i})$ s.t. $f_i > 0$ and $\sum_{\mathbf{x}_i} f_i(\mathbf{x}_i; \mathbf{x}_{\pi_i}) = 1 \; \forall \pi_i$.
- Define the joint probability to be:

$$\mathsf{P}(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n) = \prod_i f_i(\mathbf{x}_i; \mathbf{x}_{\pi_i})$$

- Amazing fact: even with no restriction on the relationship between the functions f_i, it is always true that

$$f_i(\mathbf{x}_i; \mathbf{x}_{\pi_i}) = \mathsf{P}(\mathbf{x}_i | \mathbf{x}_{\pi_i})$$

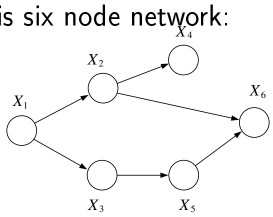
- Thus we will write

$$\mathsf{P}(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n) = \prod_i \mathsf{P}(\mathbf{x}_i | \mathbf{x}_{\pi_i})$$

- A factorization of the joint in terms of *local conditional probabilities*.
  Exponential in "fan-in" instead of n.

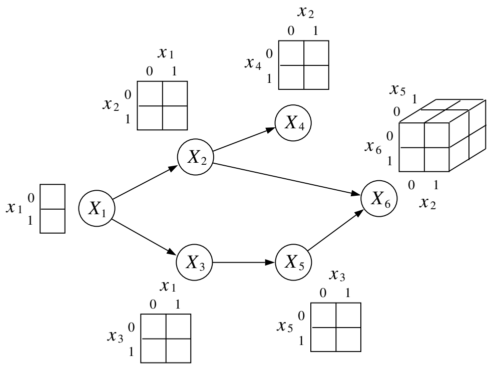
Example DAG

- Consider this six node network:



- The joint probability is now:

$$\mathsf{P}(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4, \mathbf{x}_5, \mathbf{x}_6) = \mathsf{P}(\mathbf{x}_1)\mathsf{P}(\mathbf{x}_2|\mathbf{x}_1)\mathsf{P}(\mathbf{x}_3|\mathbf{x}_1)\mathsf{P}(\mathbf{x}_4|\mathbf{x}_2)\mathsf{P}(\mathbf{x}_5|\mathbf{x}_3)\mathsf{P}(\mathbf{x}_6|\mathbf{x}_2, \mathbf{x}_5)$$



Missing Edges

- Key point about directed graphical models: *Missing edges imply conditional independence*
- Remember, that by the chain rule we can always write the full joint as a product of conditionals, given an ordering:

$$\mathsf{P}(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4, \ldots) = \mathsf{P}(\mathbf{x}_1)\mathsf{P}(\mathbf{x}_2|\mathbf{x}_1)\mathsf{P}(\mathbf{x}_3|\mathbf{x}_1, \mathbf{x}_2)\mathsf{P}(\mathbf{x}_4|\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3)\ldots$$

- If the joint is represented by a DAGM, then some of the conditioned variables on the right hand sides are missing. This is equivalent to enforcing conditional independence.
- Start with the "idiot's graph": each node has all previous nodes in the ordering as its parents.
- Now remove edges to get your DAG.
- Removing an edge into node i eliminiates a term from the conditional probability factor $p(\mathbf{x}_i|\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{i-1})$ in the chain rule.

Conditional Independence

- Recall: $\mathbf{x}_A \perp \mathbf{x}_B | \mathbf{x}_C$
  two (sets of) variables $\mathbf{x}_A$ and $\mathbf{x}_B$ are conditionally independent given a third $\mathbf{x}_C$ if:

$$\mathsf{P}(\mathbf{x}_A, \mathbf{x}_B|\mathbf{x}_C) = \mathsf{P}(\mathbf{x}_A|\mathbf{x}_C)\mathsf{P}(\mathbf{x}_B|\mathbf{x}_C) \qquad \forall \mathbf{x}_B$$

  which is equivalent to

$$\mathsf{P}(\mathbf{x}_A|\mathbf{x}_B, \mathbf{x}_C) = \mathsf{P}(\mathbf{x}_A|\mathbf{x}_C) \qquad \forall \mathbf{x}_B$$

- If we order the nodes so that your parents always come before you in the ordering ("topological") then a directed graphical model implies:

$$\{\mathbf{x}_i \perp \mathbf{x}_{\widetilde{\pi}_i} | \mathbf{x}_{\pi_i}\} \forall i$$

  where $\mathbf{x}_{\widetilde{\pi}_i}$ are the nodes that come before $\mathbf{x}_i$ that are not its parents.

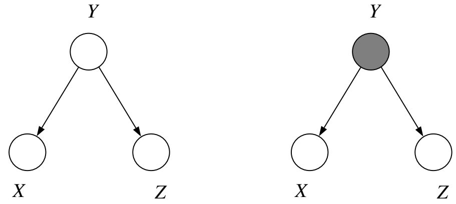
Even more structure

- Surprisingly, once you have specified the basic conditional independencies, there are other ones that follow.
- In general, it is a hard problem to say which extra CI statements follow from a basic set. However, in the case of DAGMs, we have an efficient way of generating *all* CI statements that *must be true* given the connectivity.
- This involves the idea of *d-separation* in a graph.

D-separation

- D-separation, or *directed-separation* is a notion of connectedness in DAGMs so that dependence implies connection and independence implies separation.
- In particular, we say that $\mathbf{x}_A \perp \mathbf{x}_B|\mathbf{x}_C$ if every variable in A is d-separated from every variable in B conditioned on all the variables in C.
- To check if an independence is true, we can cycle through each node in A, do a depth-first search to reach every node in B, and examine the path between them. If *all* of the paths are d-separated, then we can assert $\mathbf{x}_A \perp \mathbf{x}_B|\mathbf{x}_C$.
- Thus, it will be sufficient to consider triples of nodes. (Why?)

Common Cause



- Q: When we condition on $\mathbf{y}$, are $\mathbf{x}$ and $\mathbf{z}$ independent?

$$\mathsf{P}(\mathbf{x},\mathbf{y},\mathbf{z}) = \mathsf{P}(\mathbf{y})\mathsf{P}(\mathbf{x}|\mathbf{y})\mathsf{P}(\mathbf{z}|\mathbf{y})$$

which implies

$$\begin{aligned}\mathsf{P}(\mathbf{x},\mathbf{z}|\mathbf{y}) &= \frac{\mathsf{P}(\mathbf{x},\mathbf{y},\mathbf{z})}{\mathsf{P}(\mathbf{y})} \\ &= \frac{\mathsf{P}(\mathbf{y})\mathsf{P}(\mathbf{x}|\mathbf{y})\mathsf{P}(\mathbf{z}|\mathbf{y})}{\mathsf{P}(\mathbf{y})} \\ &= \mathsf{P}(\mathbf{x}|\mathbf{y})\mathsf{P}(\mathbf{z}|\mathbf{y})\end{aligned}$$

and therefore $\mathbf{x} \perp \mathbf{z}|\mathbf{y}$

- Think of $\mathbf{y}$ as the common cause of the two independent effects $\mathbf{x}$ and $\mathbf{z}$.

Chain

- Q: When we condition on $\mathbf{y}$, are $\mathbf{x}$ and $\mathbf{z}$ independent?

$$\mathsf{P}(\mathbf{x},\mathbf{y},\mathbf{z}) = \mathsf{P}(\mathbf{x})\mathsf{P}(\mathbf{y}|\mathbf{x})\mathsf{P}(\mathbf{z}|\mathbf{y})$$

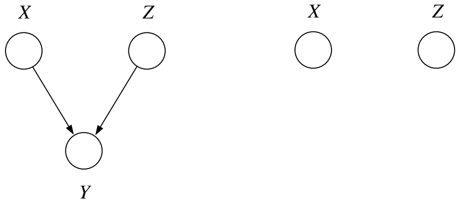
which implies

$$\begin{aligned}\mathsf{P}(\mathbf{z}|\mathbf{x},\mathbf{y}) &= \frac{\mathsf{P}(\mathbf{x},\mathbf{y},\mathbf{z})}{\mathsf{P}(\mathbf{x},\mathbf{y})} \\ &= \frac{\mathsf{P}(\mathbf{x})\mathsf{P}(\mathbf{y}|\mathbf{x})\mathsf{P}(\mathbf{z}|\mathbf{y})}{\mathsf{P}(\mathbf{x})\mathsf{P}(\mathbf{y}|\mathbf{x})} \\ &= \mathsf{P}(\mathbf{z}|\mathbf{y})\end{aligned}$$

and therefore $\mathbf{x} \perp \mathbf{z}|\mathbf{y}$

- Think of $\mathbf{x}$ as the past, $\mathbf{y}$ as the present and $\mathbf{z}$ as the future.

Explaining Away



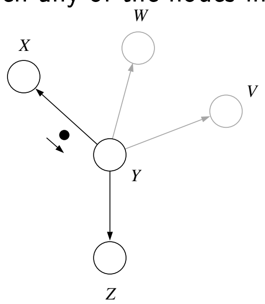
- Q: When we condition on $\mathbf{y}$, are $\mathbf{x}$ and $\mathbf{z}$ independent?

$$\mathsf{P}(\mathbf{x},\mathbf{y},\mathbf{z}) = \mathsf{P}(\mathbf{x})\mathsf{P}(\mathbf{z})\mathsf{P}(\mathbf{y}|\mathbf{x},\mathbf{z})$$

- $\mathbf{x}$ and $\mathbf{z}$ are *marginally independent*, but given $\mathbf{y}$ they are *conditionally dependent*.
- This important effect is called *explaining away*. (Also known as Berkson's paradox.)
- For example, flip two coins independently, and set $\mathbf{x}$=coin1,$\mathbf{z}$=coin2. Make $\mathbf{y}$=1 if the coins come up the same and $\mathbf{y}$=0 if they come up different.
- Now, if I tell you $\mathbf{y}$, $\mathbf{x}$ and $\mathbf{z}$ are coupled!

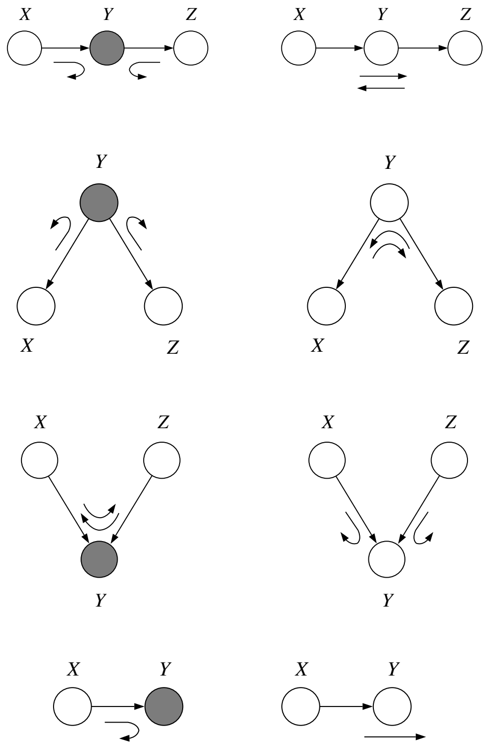
Bayes Ball Algorithm

- To check if $\mathbf{x}_A \perp \mathbf{x}_B | \mathbf{x}_C$ we need to check if every variable in A is d-separated from every variable in B conditioned on all variables in C.
- The *Bayes-Ball Algorithm* is a d-separation test. We shade all nodes $\mathbf{x}_C$, place balls at each node in $\mathbf{x}_A$ (or $\mathbf{x}_B$), let them bounce around according to some rules, and then ask if any of the balls reach any of the nodes in $\mathbf{x}_B$ (or $\mathbf{x}_A$).



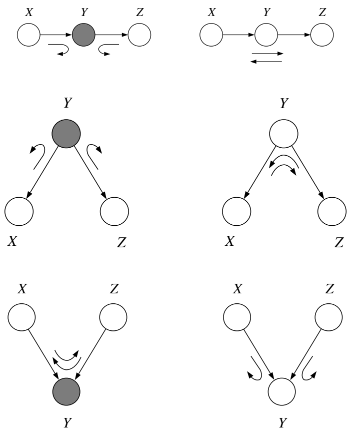
- So we need to know what happens when a ball arrives at a node $\mathbf{Y}$ on its way from $\mathbf{X}$ to $\mathbf{Z}$.
- Q: Given that $\mathbf{x}_C$ are clamped, when we wiggle nodes $\mathbf{x}_A$ can we change any of the node $\mathbf{x}_B$?

Canonical Micrographs

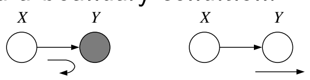


Bayes-Ball Rules

- The three cases we considered tell us rules:



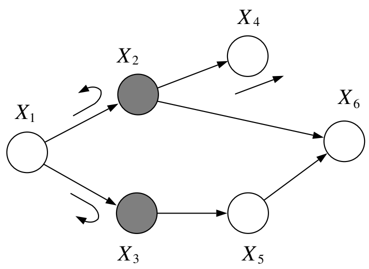
- We also need a boundary condition:



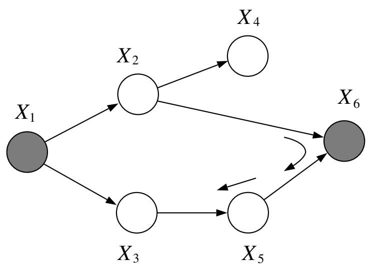
- Here's a trick for the explaining away case: If $\mathbf{y}$ *or any of its decendants* is shaded, the ball passes through.

Examples of Bayes-Ball

$$\mathbf{x}_1 \perp \mathbf{x}_6 | \{\mathbf{x}_2, \mathbf{x}_3\} \quad ?$$



$$\mathbf{x}_2 \perp \mathbf{x}_3 | \{\mathbf{x}_1, \mathbf{x}_6\} \quad ?$$



Notice: balls can travel opposite to edge directions.

Families of Distributions

- Consider two families of distributions.
- One is generated by all possible settings of the conditional probability tables in the DAGM form:

$$\mathsf{P}(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n) = \prod_i \mathsf{P}(\mathbf{x}_i | \mathbf{x}_{\pi_i})$$

- The other is generated by finding *all* the conditional independencies implied by a DAGM and eliminating any joint distributions which violate them.
- A version of the amazing *Hammersley-Clifford Theorem* (1971) states that these two families are exactly the same.