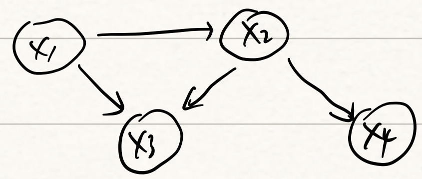
Lec 21 4270

Bayesian Network : $G = (V, E)$ — Directed acyclic Graph

$\begin{cases} V: \text{ set of node / vertices} \\ E: \text{ set of arcs (directed edges)} \end{cases}$

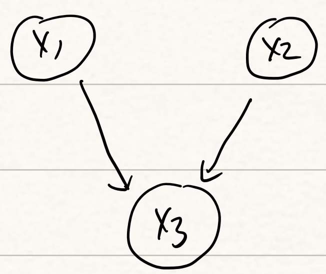
$P(\underline{x}) = \prod_{i \in V} P(x_i \mid x_{pa(i)})$ $\quad pa(i) \rightarrow$ direct parents of x_i

[E.g.]



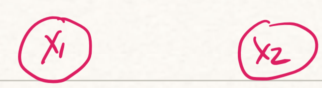
$P(\underline{X}) = P(X_1)\, P(X_2 \mid X_1)\, P(X_3 \mid X_1, X_2)\, P(X_4 \mid X_2)$

[E.g.]



Is $X_1 \perp\!\!\!\perp X_2$?

① Create ancestral graph of nodes of interest



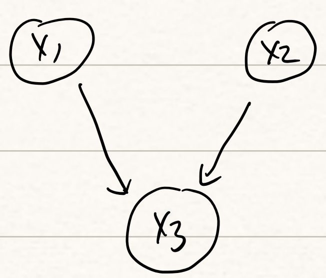
② Marry the parents

③ changes arcs to undirected edges

④ Examine separate sets

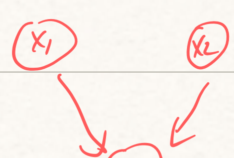
Ans: YES! $\Rightarrow$ $X_1 \perp\!\!\!\perp X_2$

[E.g.]

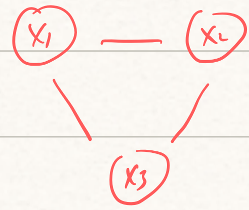


Is $X_1 \perp\!\!\!\perp X_2 \mid X_3$?

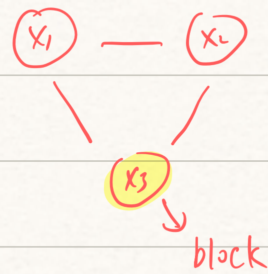
①



② + ③



④



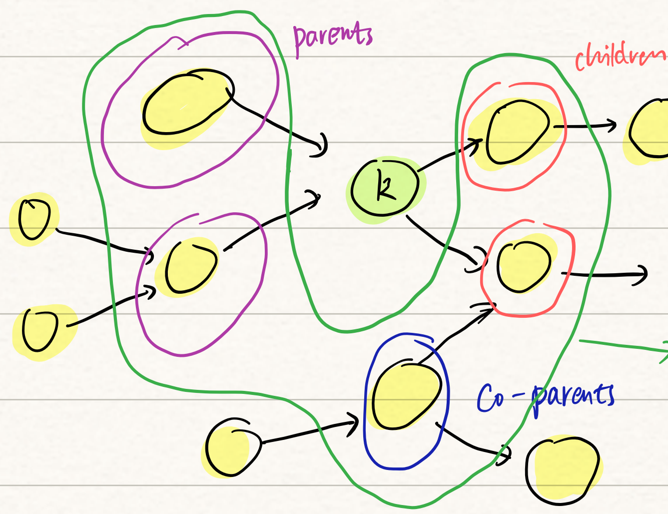
$\Rightarrow X_1 \not\perp X_2 \mid X_3$

Ans: NO!

Markov Blanket

Qn: Given a BN and any node $k \in V$, what is the smallest set $S \subset V \setminus \{k\}$ s.t.

$$(+) \quad P(X_k \mid X_S) = P(X_k \mid X_{V \setminus \{k\}}) \Rightarrow$$ find the smallest S to BLOCK k from outside



Markov Blanket

$$X_k \perp X_{V \setminus (\{k\} \cup MB_k)} \mid X_{MB_k}$$

MB_k: Markov Blanket

Defn: The smallest $S \subset V \setminus \{k\}$ s.t (+) holds is called the Markov Blanket of node k.

$\underline{k=1}$: RHS of (+)

大部分都约掉了！(不含 X_1)

$$P(X_1 \mid X_{V\setminus\{1\}}) = \frac{P(X)}{P(X_{V\setminus\{1\}})} = \boxed{\frac{\prod_{k=1}^{d} P(X_k \mid X_{pa(k)})}{\underbrace{\sum_{X_1} \prod_{k=1}^{d} P(X_k \mid X_{pa(k)})}_{\text{marginalization}}}}$$

Terms in the product in the denominator that do not depend on X_1 can be brought outside the sum and cancelled with the sum in the numerator.

Qn: which terms contain X_1 in $\prod_{k=1}^{n} P(X_k \mid X_{pa(k)})$

① $P(X_1 \mid X_{pa(1)})$ [involves the set of parents of X_1]

② If $X_{ch(1)}$ is a child of X_1, then there will be terms of the form:

$$P(X_{ch(1)} \mid X_1, \ldots)$$

[involves the children of X_1]

③ If X_1 & X_2 share a child, X_{ch}, then there will be terms of the form:

Moralization

$$P(X_{ch} \mid X_1, X_2, \ldots)$$

[involves co-parents of X_1]

Claim: The Markov Blanket of node 1 is the set

$$MB(1) = pa(1) \cup ch(1) \cup \underbrace{\text{co-parents}(1)}_{\text{come from marrying!}}$$

Learning Bayesian Networks

① Given a DAG $G=(V,E)$ $V=\{1,2,\dots,d\}$

tells us the structure of graph

② n complete data (observations) $X_t=(X_{t1},\dots,X_{td})^T$ $t=1,2,\dots,n$

$D=\{X_t\}_{t=1}^{n}$

Want to learn the conditional distribution $P(X_i \mid X_{pa(i)})=\theta_{x_i \mid x_{pa(i)}}$ $i=1,2,\dots,d$

(conditional probability tables)

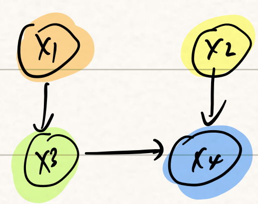
$$P(X)=\prod_{i=1}^{d}P(X_i \mid X_{pa(i)})=\prod_{i=1}^{d}\theta_{x_i \mid x_{pa(i)}}$$

$$\underline{\theta}=\{\{\theta_{x_i \mid x_{pa(i)}} : x_i\in[r_i],\ x_{pa(i)}\in[r_{pa(i)}]\}_{i=1}^{d}\}$$

① $\{1,\dots,r_i\}$ the set of value that X_i takes on

② $X_{pa(i)}=(X_j,X_k)$ $[r_{pa(i)}]=[r_j]\times[r_k]$

E.g.



$[r_i]=\{1,2\}$

1 $\theta_{x_1=1}$ $\theta_{x_1=2}$

1 $\theta_{x_2=1}$ $\theta_{x_2=2}$

2 $\theta_{x_3=1 \mid x_1=1}$ $\theta_{x_3=2 \mid x_1=1}$

$\theta_{x_3=1 \mid x_1=2}$ $\theta_{x_3=2 \mid x_1=2}$

4 $\theta_{x_4=1 \mid x_2=a,\,x_3=b}$ $\forall (a,b)\in[2]^2$

$\theta_{x_4=2 \mid x_2=a,\,x_3=b}$

4

8

Maximum Likelihood parameter estimation

Given $\mathcal{D} = \{\underline{x}_t\}_{t=1}^{n}$, want to learn $\underline{\theta}$.

$$\downarrow$$

$$\underline{x}_t \sim P$$

Log-likelihood:

$$\ell(\mathcal{D} : \underline{\theta}, G) = \log P(\mathcal{D} \mid \underline{\theta})$$

$$\overset{\text{i.i.d}}{=} \log \prod_{t=1}^{n} P(\underline{x}_t \mid \underline{\theta})$$

$$= \sum_{t=1}^{n} \log P(\underline{x}_t \mid \underline{\theta})$$

$$= \sum_{t=1}^{n} \log \prod_{i=1}^{d} P(x_{ti} \mid x_{t,pa(i)})$$

$$= \sum_{t=1}^{n} \sum_{i=1}^{d} \log \theta_{x_{ti} \mid x_{t,pa(i)}} \longleftrightarrow P(x_{ti} \mid x_{t,pa(i)})$$

个数 r_i 　　个数 $\underline{[r_j] \times [r_k] \times \cdots [r_{pa(i)}]}$.

$$= \sum_{i=1}^{d} \sum_{x_i, x_{pa(i)}} n(\underline{\underline{x_i}}, \underline{\underline{x_{pa(i)}}}) \log \theta_{x_i \mid x_{pa(i)}}$$

$$\Downarrow$$

\# of times $(x_i, x_{pa(i)})$ occurs in the dataset $\{(x_{ti}, x_{t,pa(i)}) : t = 1, 2, \ldots, n\}$

Now differentiate $\ell(\mathcal{D}; \underline{\theta}, G)$ w.r.t. $\underline{\theta}$.

(Notice that $\sum_{x_i} \theta_{x_i \mid x_{pa(i)}} = 1$, $\forall x_{pa(i)}$)

$$\Rightarrow \hat{\theta}_{x_i \mid x_{pa(i)}} = \frac{n(x_i, x_{pa(i)})}{\sum_{x_i'} n(x_i', x_{pa(i)})}$$

Empirical Frequency Counts

Difficult to learn in high degree! $\Rightarrow$ have many parents

☆ To provide one observation per configuration of parent variables require $\prod_{j \in pa(i)} r_j$ instances!

Model Selection

① Given $\hat{\theta}_{x_i | pa(i)}$ for all $(x_i, x_{pa(i)})$ and all $i \in [d]$,

we can compute

suppose we have good prior models

$$BIC(G) = \ell(\mathcal{D}; \hat{\theta}, G) - \frac{\dim(G)}{2} \log n$$

$\dim(G) =$ # of parameters used to describe the BN

that represents G

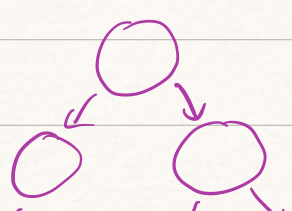
$$= \sum_{i=1}^{d} \underbrace{(r_i - 1) \prod_{j \in pa(i)} r_j}$$

$$\theta_{\underbrace{x_i}_{r_i - 1} | \underbrace{x_{pa(i)}}_{\prod r_j}}$$

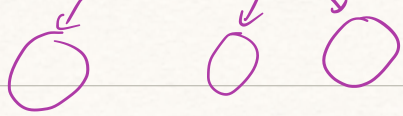
② Don't rely on outside prior information

If $G = (V, E)$ is a tree (connected) and every nodes have $\boxed{\text{in-degree} \le 1}$,

then there's a much better way to learn G! 入度



$\boxed{d \text{ nodes}}$ → Total # of such graphs



$\Downarrow$

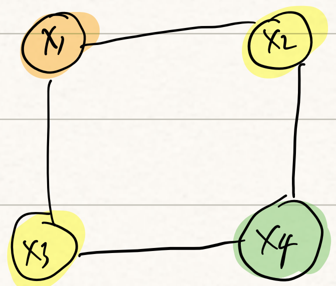
d^{d-2}

Markov Random Field (Undirected Graphic Model)

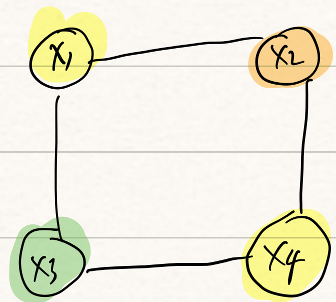
Undirected graph $G=(V,E)$ $\quad E \subset \binom{V}{2}$

$$= \{ \underbrace{\{i,j\}}_{\text{unordered}} : i,j \in V, i \neq j \}$$

Graph encodes conditional independence relations among x_i, $i=1,2,\dots,d$



$$\Longrightarrow X_1 \perp X_4 \mid \underbrace{(X_2, X_3)}_{\text{separator}} \quad \text{—}(*)$$

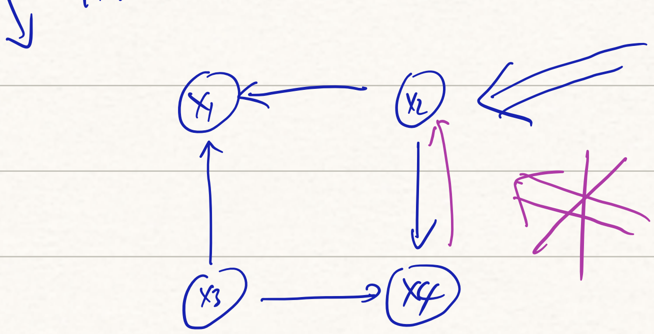


$$X_2 \perp X_3 \mid \underbrace{(X_1, X_4)}_{\text{separator}} \quad \text{—}(**)$$

Qn: Can you draw a Bayesian Network that encodes these two CI properties?

(* & **)

TRY



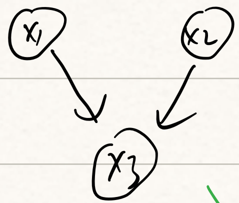
① $X_1 \perp X_4 \mid (X_2, X_3)$

② $X_2 \perp\!\!\!\perp X_3 \mid (X_1, X_4)$

Impossible to draw the BN that respects (*) & (**)

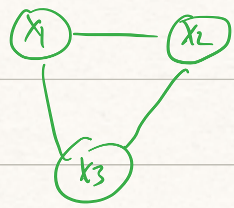
In terms of ability to explicate condition independence properties MRFs & BNs are not strict subsets to each other!

E.g.



$\Rightarrow X_1 \perp\!\!\!\perp X_2$

MRF



$X_1 \not\perp\!\!\!\perp X_2$

Hammersley - Clifford Thm

(Recall that for a BN, $p(\underline{x}) = \prod_{i=1}^{d} p(x_i \mid x_{pa(i)})$)

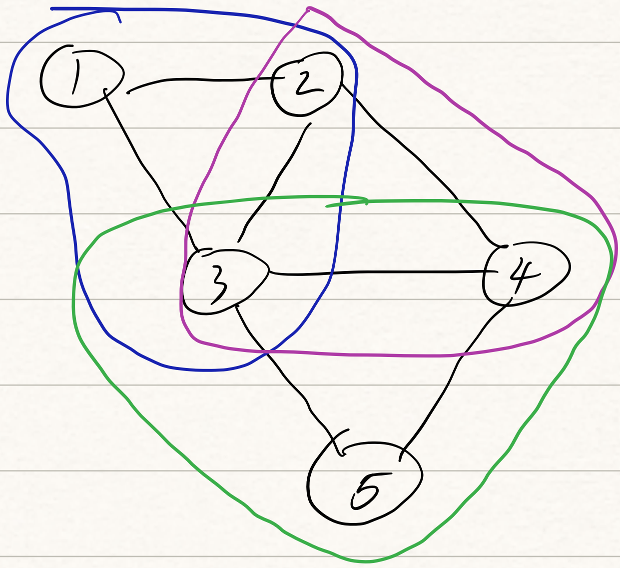
If $G=(V,E)$ is an undirected graph, can we write $p(\underline{x})$ in the form of products of terms defined on small subsets of nodes?

Defn: Given an undirected graph $G=(V,E)$, we say that

C is a clique if it is a fully connected set of nodes

完全图.

A maximal clique is one that cannot be extended by including one more node



Clique $C = \{1,2\}$ is not maximal!

Click $\tilde{C} = \{1,2,3\}$ is maximal

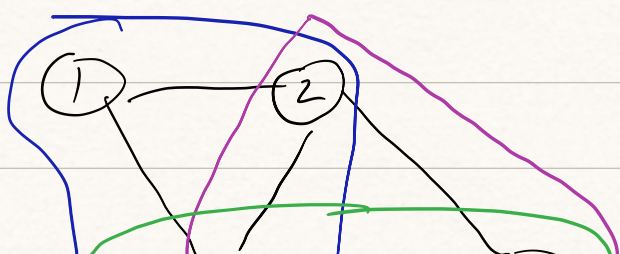
[Informal Version of the Hammersley – Clifford Thm]

For a MRF, we can write its joint distribution as follows

$$P(\underline{X}) = \frac{1}{Z} \prod_{c \in C} \phi_c(\underline{X}_c)$$

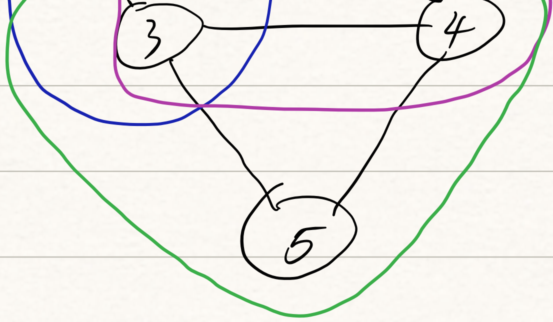
C: set of all maximal cliques of G.

ϕ_c: positive f^{n} defined on the domain of $\underline{X}_C = (X_i : i \in C)$



$$P(\underline{X}) = \frac{1}{Z} \phi_{1,2,3}(X_1, X_2, X_3)$$

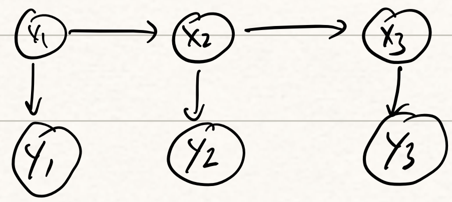
$$\phi_{2,3,4}(X_2, X_3, X_4)$$



$\phi_{3,4,5}(x_3, x_4, x_5)$

Gibbs distribution

E.g. HMM

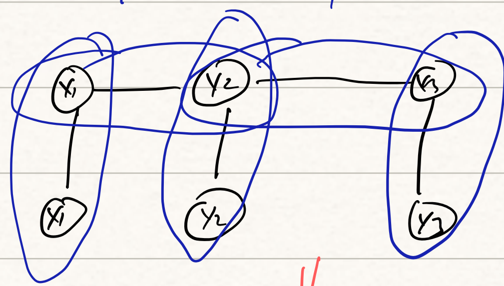


Bayesian Network

$\Downarrow$

$p(\underline{X}) = p(x_1)\, p(y_1 \mid x_1)\, p(x_2 \mid x_1)\, p(y_2 \mid x_2)\, p(x_3 \mid x_2)\, p(y_3 \mid x_3)$

Maximal Cliques



MRF

$\Downarrow$

No 2 parents!

$p(\underline{X}) = \frac{1}{Z}\, \phi_{x_1 y_1}(x_1, y_1)\, \phi_{x_1 x_2}(x_1, x_2)\, \phi_{x_2 y_2}(x_2, y_2)\, \phi_{x_2 x_3}(x_2, x_3)\, \phi_{x_3 y_3}(x_3, y_3)$