STA 414/2104:
Probabilistic Machine Learning

Week 2 : Graphical Models

Murat A. Erdogdu

University of Toronto

Overview

Summary of the content:

- Directed Graphical Models
- Markov Random Fields

Joint distributions

- The joint distribution of N random variables $(x_1, x_2, ..., x_N)$ is a very general way to encode knowledge about a system.
- Assume $x_i \in \{0, 1\}$ are binary, then it requires $2^N - 1$ parameters to specify the joint distribution

$$p(x_1, x_2, ..., x_N).$$

- This can be also written as

$$p(x_1, x_2, \ldots, x_N) = \prod_{j=1}^{N} p(x_j | x_1, x_2, \ldots, x_{j-1})$$

  for any ordering of the variables, where $p(x_1|x_0) = p(x_1)$.
- We can exploit dependencies among variables and reduce the number of parameters! (e.g. Naive Bayes)

Conditional Independence

- Assume there are N random variables $x_1, x_2, ..., x_N$.
- For set $A \subset \{1, 2, ..., N\}$, we denote by $x_A = \{x_i \ : \ i \in A\}$. Assume A, B, C are disjoint. In particular, we say that

$$x_A \perp x_B \mid x_C$$

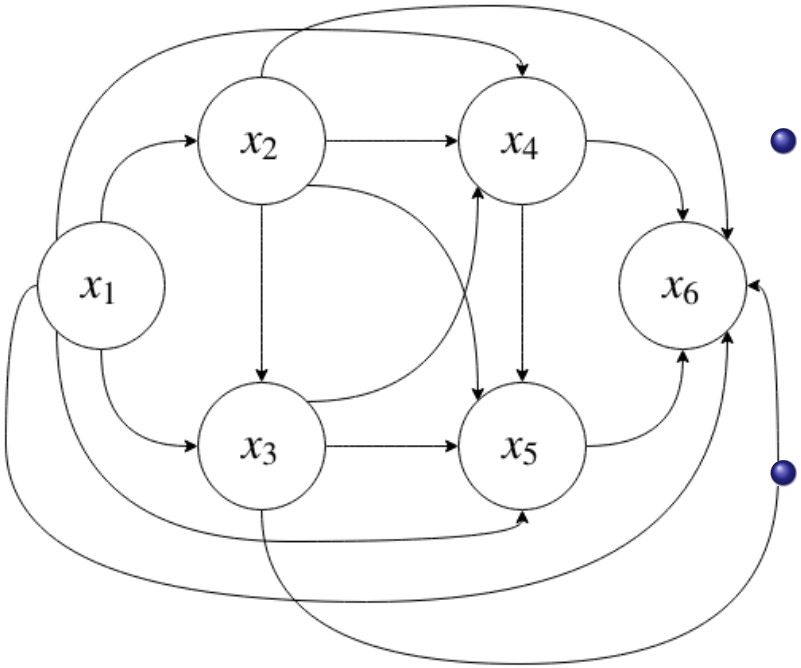
  if random variables x_A, x_B are conditionally independent given x_C.
- We have

$$x_A \perp x_B | x_C$$

  iff
  - $\Leftrightarrow p(x_A, x_B|x_C) = p(x_A|x_C)p(x_B|x_C)$
  - $\Leftrightarrow p(x_A|x_B, x_C) = p(x_A|x_C)$
  - $\Leftrightarrow p(x_B|x_A, x_C) = p(x_B|x_C)$

These are all equivalent!

Directed Acyclic Graphical Models (Bayes' Nets)



- A directed acyclic graphical model (DAG) implies a factorization of the joint distribution.
- Variables are represented by nodes, and edges represent dependence.

DAG induces the following factorization of the joint distribution of random variables $x_1, x_2, \ldots, x_N$, we can write:

$$p(x_1, \ldots, x_N) = \prod_{i=1}^{N} p(x_i | x_1, \ldots, x_{i-1}) = \prod_{i=1}^{N} p(x_i | \text{parents}(x_i))$$

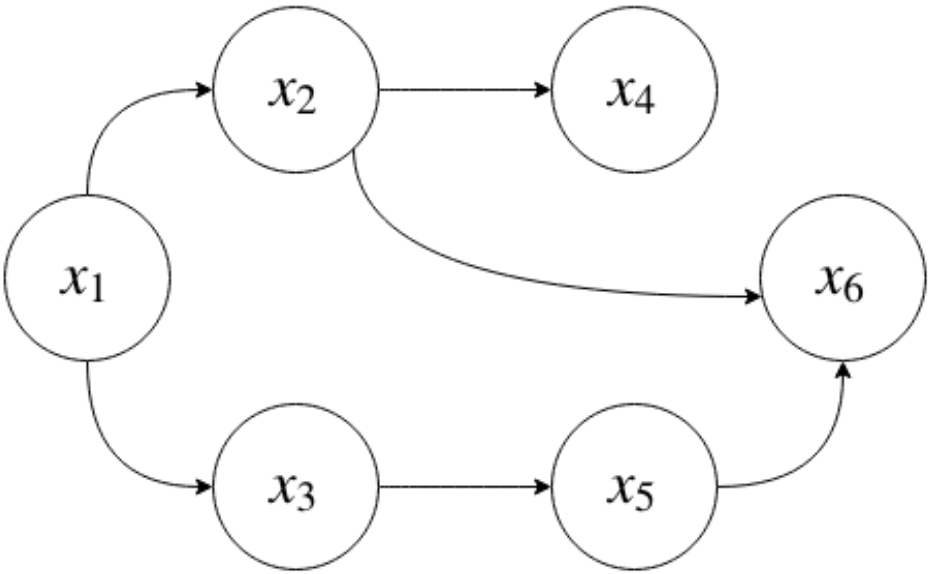
where $\text{parents}(x_i)$ is the set of nodes with edges pointing to x_i.

DAGs and Conditional Independence

In a directed acyclic graphical model (DAGs)

$$p(x_1, x_2, \ldots, x_N) = \prod_{i=1}^{N} p(x_i|\text{parents}(x_i))$$

where parents(x_i) is the set of nodes with edges pointing to x_i.

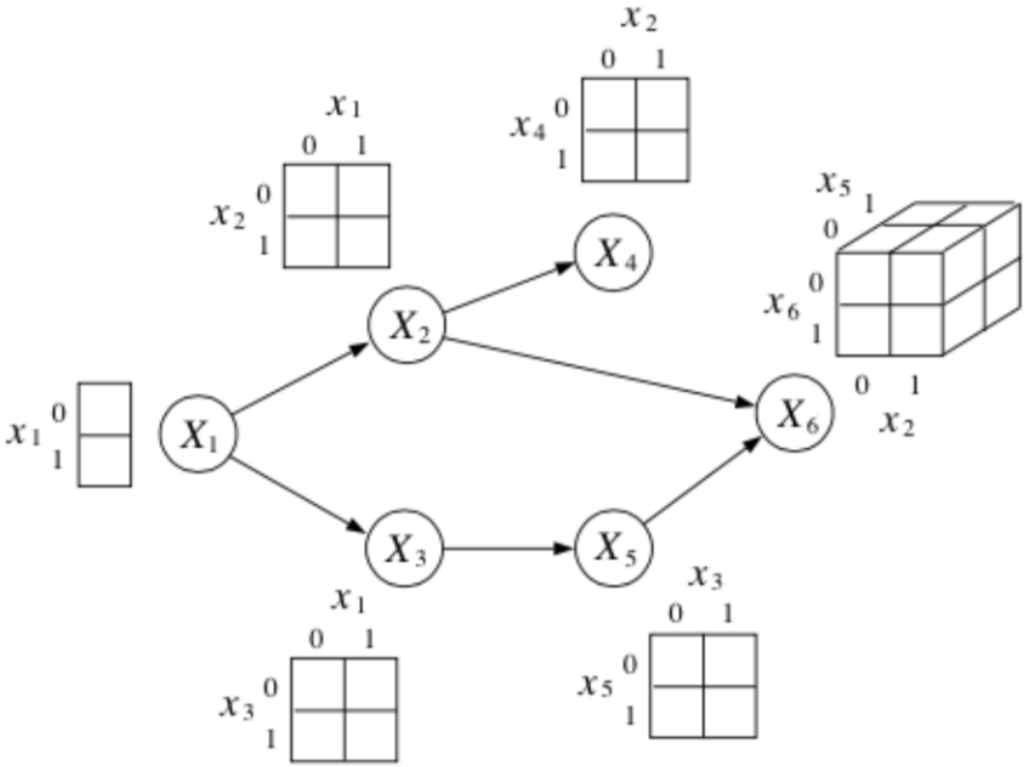


- This DAG corresponds to the following factorization of the joint distribution:

$$p(x_1, x_2, ..., x_6) = p(x_1)p(x_2|x_1)p(x_3|x_1)p(x_4|x_2)p(x_5|x_3)p(x_6|x_2, x_5)$$

Conditional Probability Tables (CPT)

Suppose each x_i is a binary random variable. How many parameters does it take to represent this joint distribution?



- For example, 2x2 CPT for the node x_4 corresponds to $p(x_4|x_2)$ requires 2 parameters.
- Each CPT with K_i parents requires 2^{K_i} parameters. In total, $\sim N2^{\max K_i}$ parameters.
- If we allow all possible dependencies (fully-connected DAG), which requires $2^N - 1$ parameters.

DAGs reduce the computational burden of making inferences by introducing conditional independencies.

Conditional Independence in DAGs

- **D-separation** (directed-separation) is a notion of connectedness in DAGs in which two (sets of) variables may or may not be connected conditioned on a third (set of) variable(s).
- D-separation implies conditional independence and vice versa.
- For a set $A \subset \{1, 2, ..., N\}$, we denote by $x_A = \{x_i \ : \ i \in A\}$. In particular, we say that

$$x_A \perp x_B \mid x_C$$

if every variable in A is d-separated from every variable in B conditioned on all the variables in C.

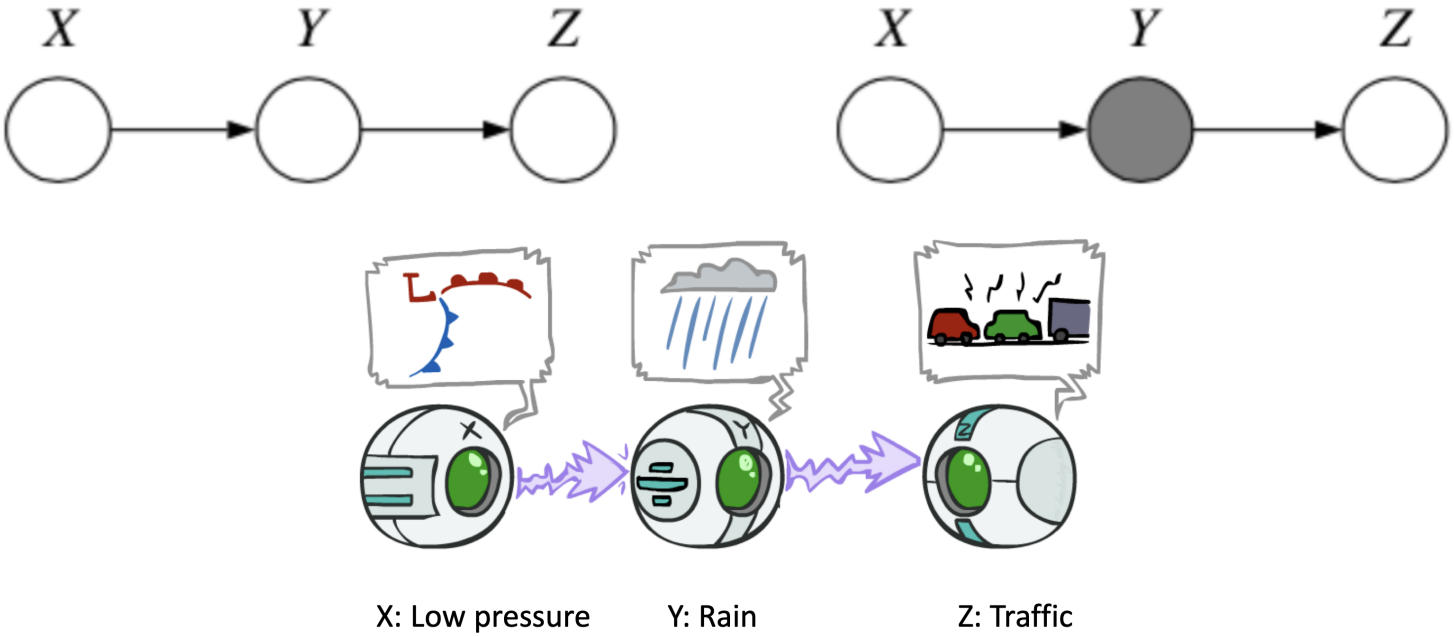
DFS Algorithm for Checking Independence

Let A, B, C be disjoint subsets of $\{1, 2, ..., N\}$.

- We 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 have d-separated end points (i.e., conditionally independent nodes), then

$$x_A \perp x_B \mid x_C$$

Causal Chain

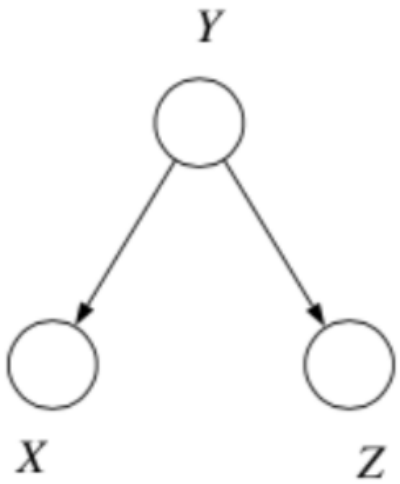


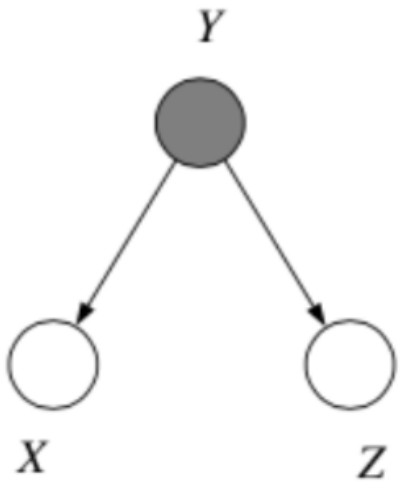
$$
\begin{aligned}
p(z|x,y) &= \frac{p(x,y,z)}{p(x,y)} \\
&= \frac{p(x)p(y|x)p(z|y)}{p(x)p(y|x)} \\
&= p(z|y) \quad X \text{ and } Z \text{ d-separated given } Y.
\end{aligned}
$$

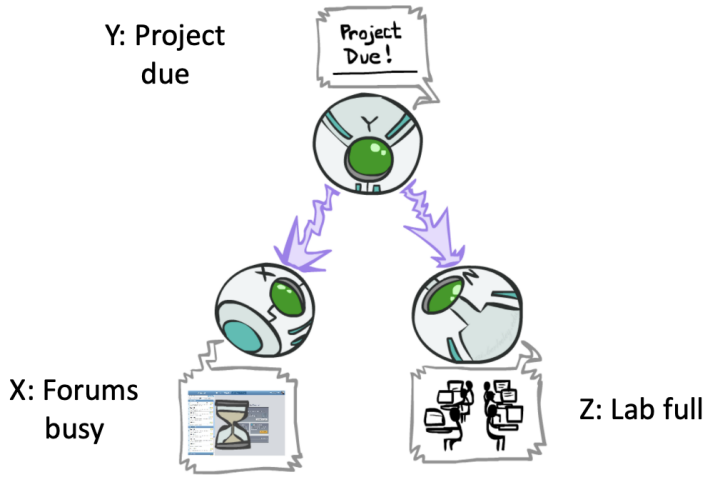
image credit Abbeel & Klein

Common Cause

Where we think of y as the "common cause" of the two independent effects x and z.





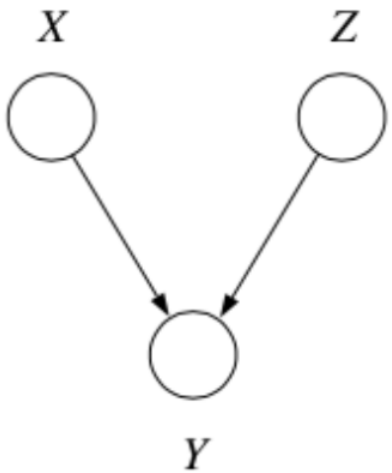


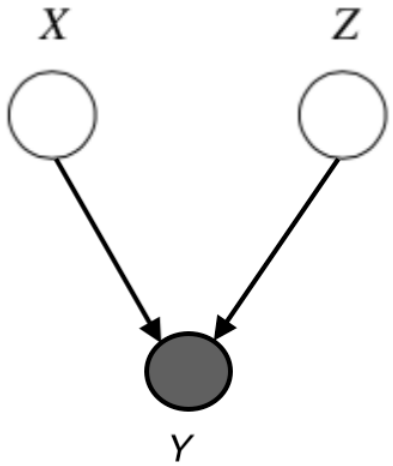
Question: When we condition on y, are x and z independent?
Answer: From the graph, we get

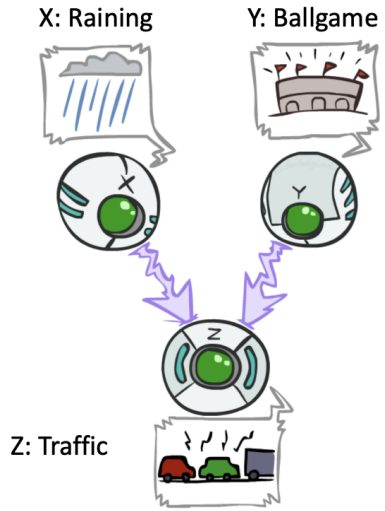
$$p(x,z|y) = \frac{p(x,y,z)}{p(y)} = \frac{p(y)p(x|y)p(z|y)}{p(y)} = p(x|y)p(z|y) \quad \text{yes!}$$

image credit Abbeel & Klein

Explaining Away (Common Effect)







Question: When we condition on y, are x and z independent?
Answer: From the graph, we get

$$p(z|x,y) = \frac{p(x)p(z)p(y|x,z)}{p(x)p(y|x)}$$
$$= \frac{p(z)p(y|x,z)}{p(y|x)} \neq p(z|y)$$

image credit Abbeel & Klein

How to check Conditional Independence?

An algorithm for determining conditional independence in a DAG can be constructed based on the rules we discussed.

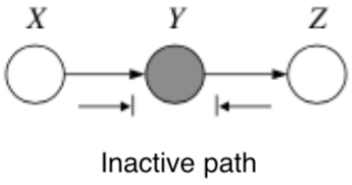
- To check if $x_A \perp x_B | x_C$ we need to check if every variable in A is d-seperated from every variable in B conditioned on all variables in C.
- In other words, given that all the nodes in x_C are "clamped", when we "wiggle" nodes x_A can we change any of the nodes in x_B?

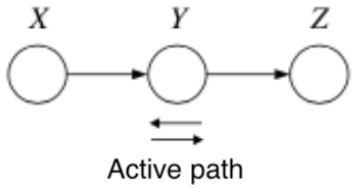
Algorithm

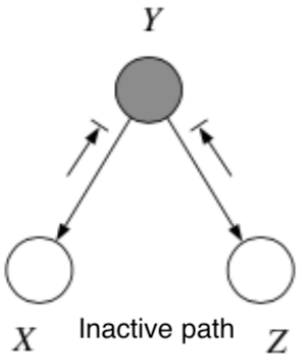
In general, the algorithm works as follows:

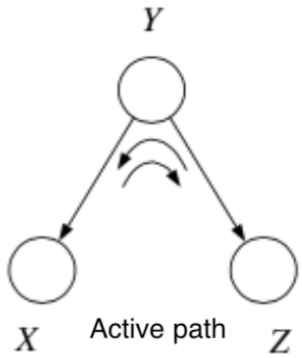
1. Shade all nodes x_C (these are observed)
2. Try to reach from node x_A to node x_B (or vice versa)
3. ... according to the rules we came up with
   - - If we can reach any of the nodes in x_B from x_A (or x_A from x_B) then $x_A \not\perp x_B | x_C$
   - - Otherwise $x_A \perp x_B | x_C$

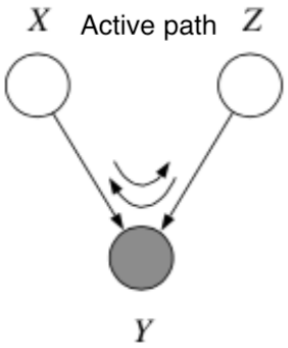
Rules for active/inactive triples

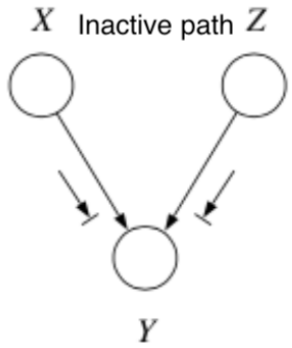




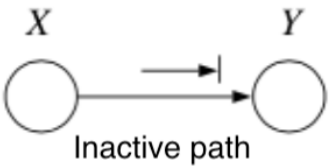


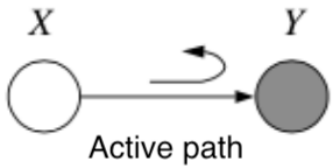


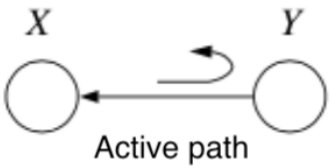


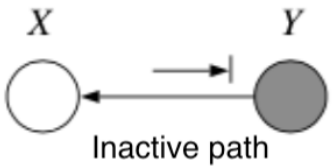


- Arrows: paths we can move along
- Arrows with bars: paths that are blocked
- Pairs are boundary cases.



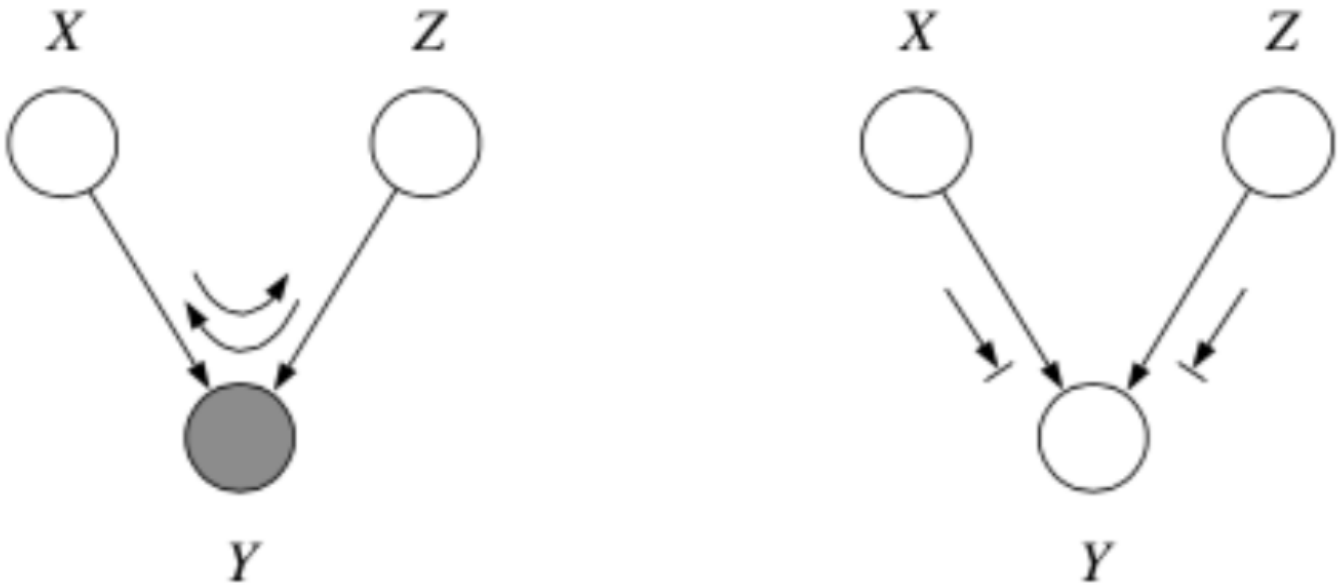






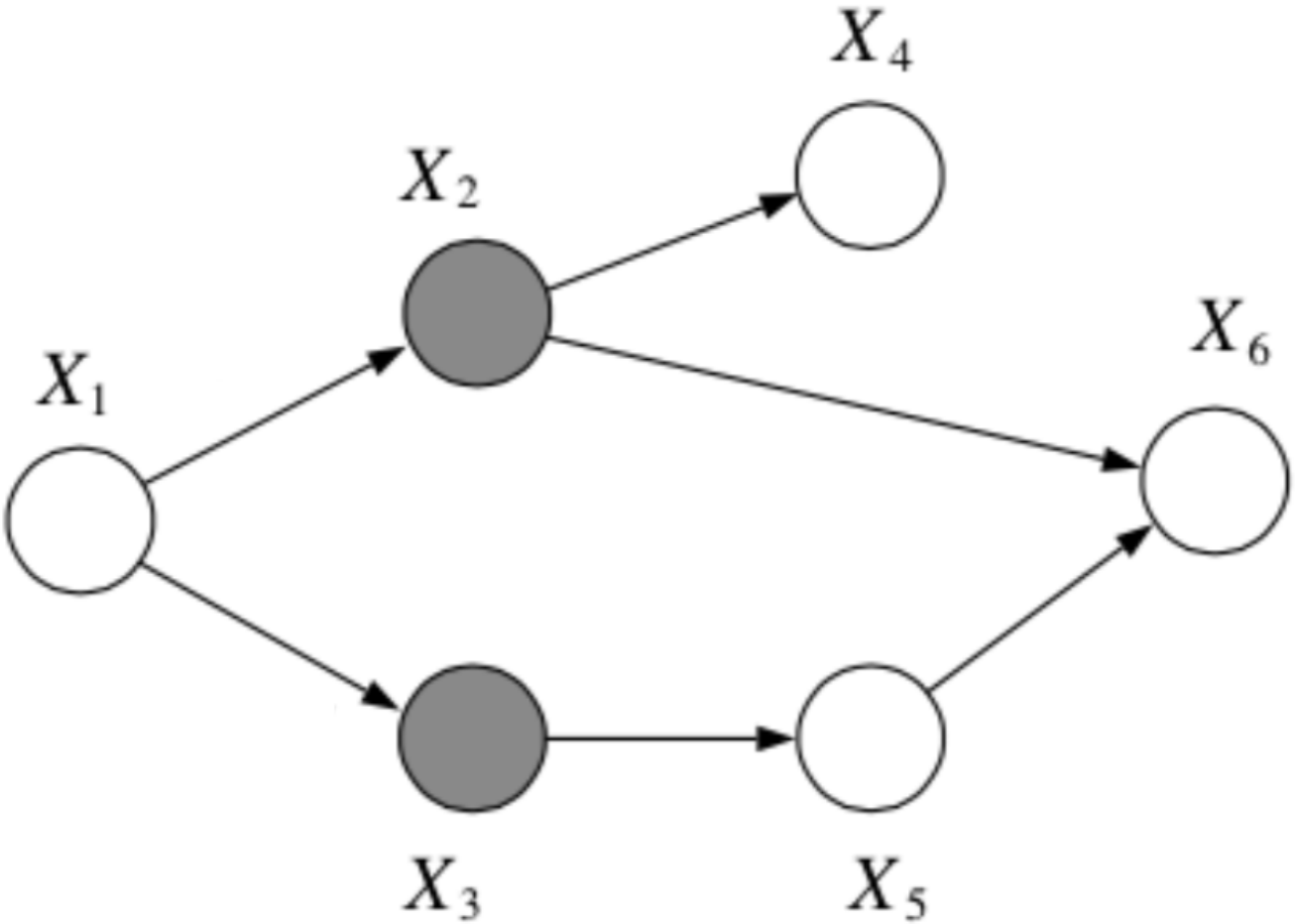
Example I: Explaining Away

If y or any of its descendants is shaded, we can travel through.



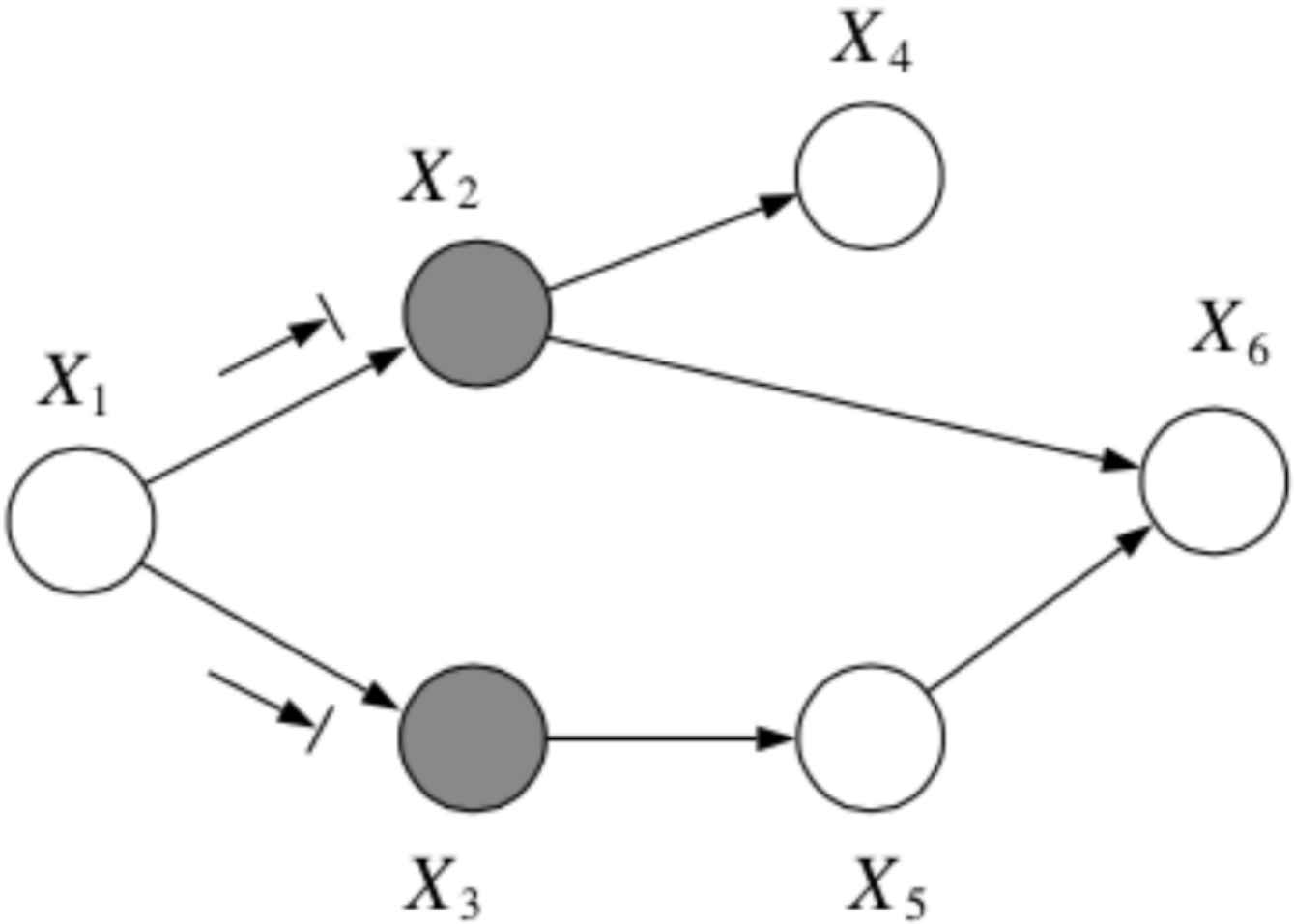
Example II: Large DAG

In the following graph, is $x_1 \perp x_6 | \{x_2, x_3\}$?



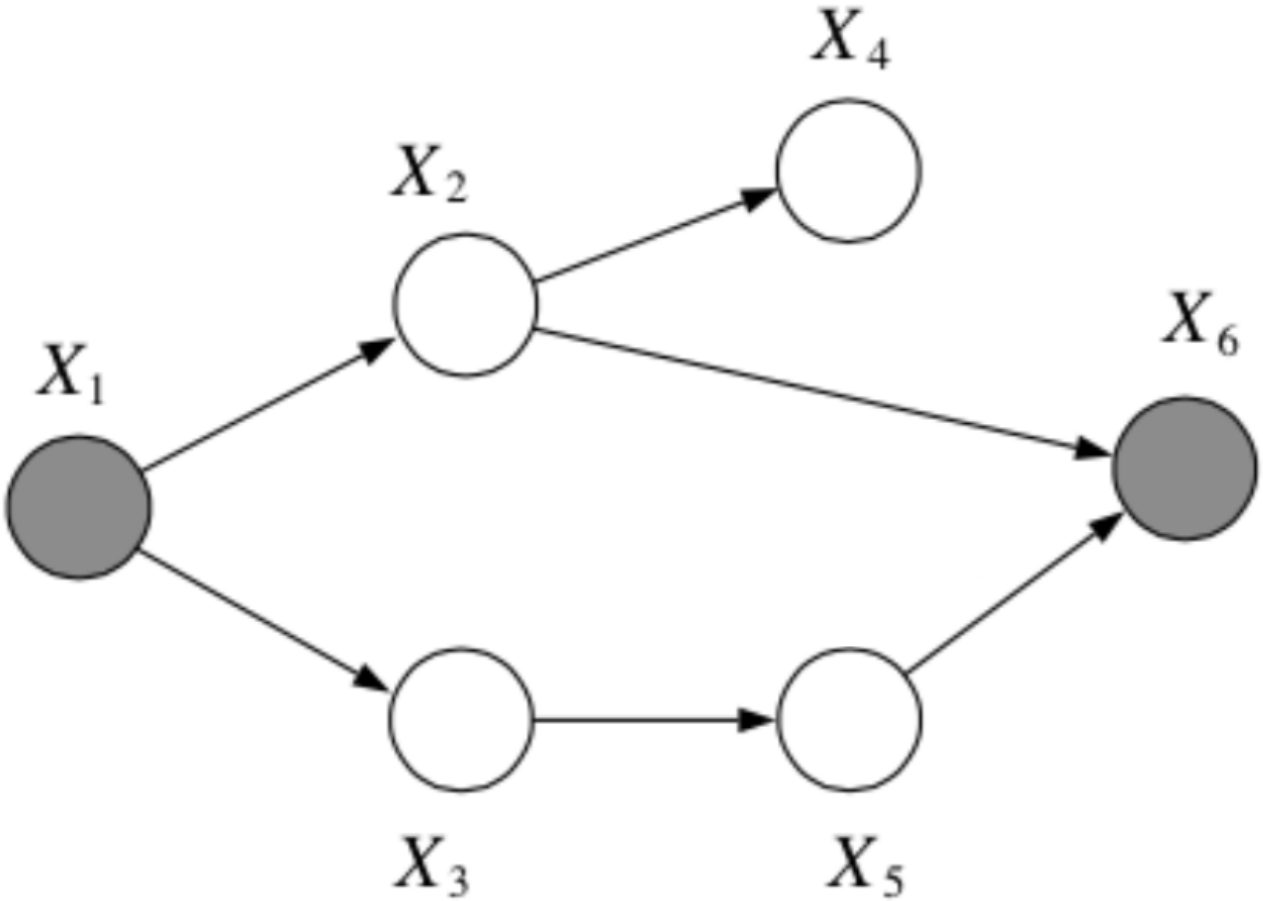
Example II: Solution

Yes.



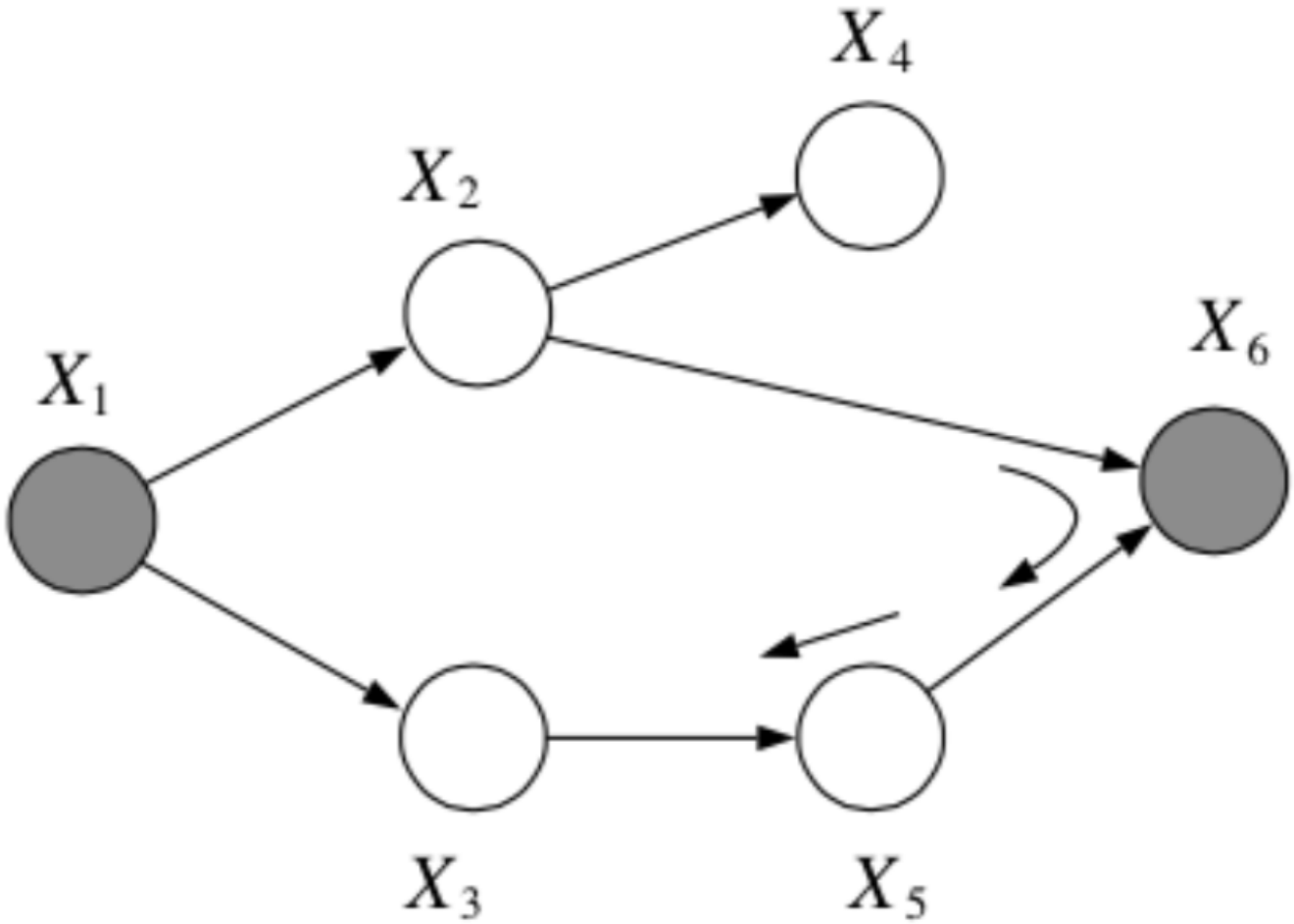
Example III:

In the following graph, is $x_2 \perp x_3 | \{x_1, x_6\}$?



Example III:

No.



Summary

- DAGs are great for encoding conditional independencies.
- They can reduce the number of parameters significantly.
- Conditional independence between two sets of variables on a DAG can be found using the Bayes ball method.
- Next lecture: Markov Random Fields.

Are DAGMs always useful?

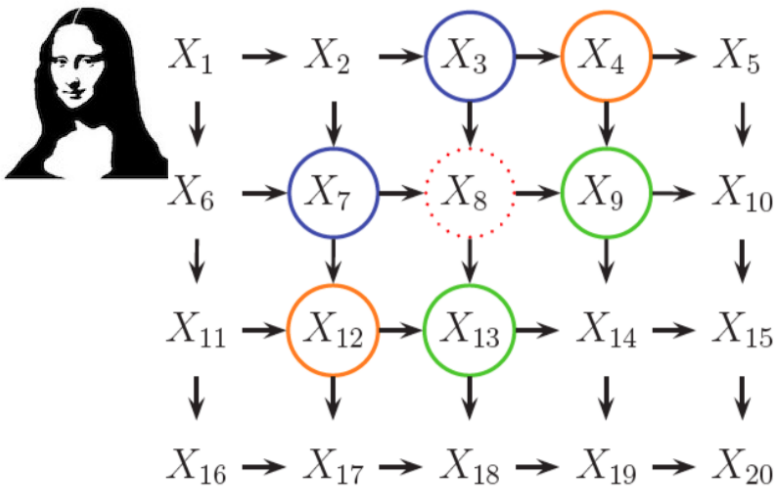


Figure : Causal MRF or a Markov mesh

- Each node is conditionally independent of its non-descendants given its parents

$$\{X_i \perp \text{non-desc}(X_i) \mid \text{parents}(X_i)\} \quad \forall i.$$

- For some problems, it is not clear how to choose the edge directions in DAGMs.

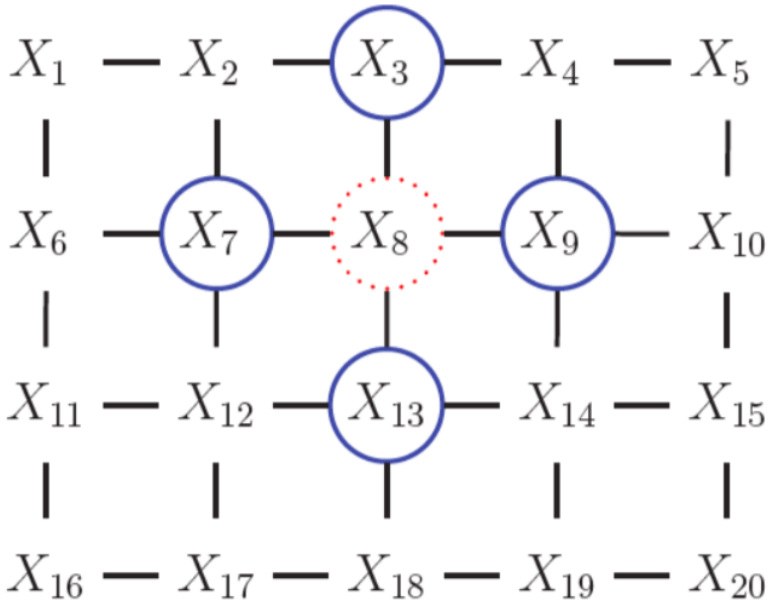
Markov blanket (mb): the set of nodes that makes X_i conditionally independent of all the other nodes.

In our example: $\text{mb}(X_8) = \{X_3, X_4, X_7, X_9, X_{12}, X_{13}\}$.

One would expect X_4 and X_{12} not to be in the Markov blanket $\text{mb}(X_8)$, especially given X_2 and X_{14} are not.

Markov Random Fields

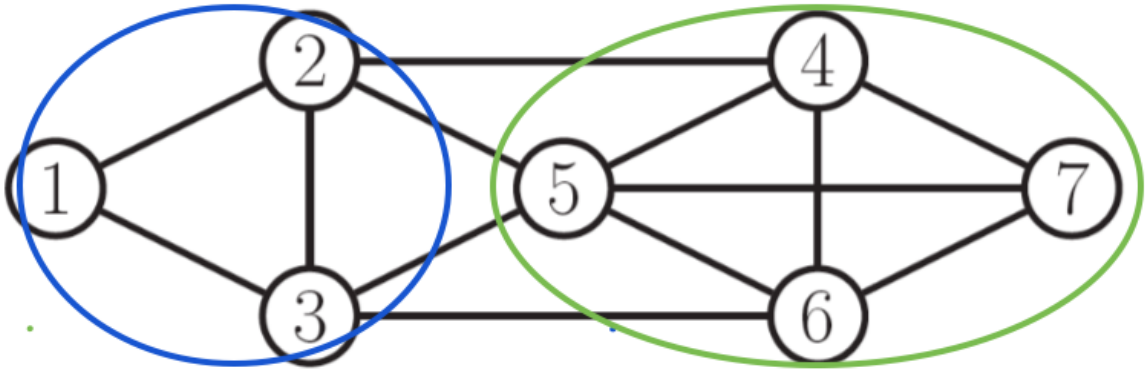
- Undirected graphical models (aka **Markov random fields (MRFs)**) are models with dependencies described by an undirected graph.
- The nodes in the graph represent random variables. However, in contrast to DAGMs, edges represent probabilistic interactions between neighbors (as opposed to conditional dependence).



Cliques

A **clique** is a subset of nodes such that every two vertices in the subset are connected by an edge.

A **maximal clique** is a clique that cannot be extended by including one more adjacent vertex.



Distributions Induced by MRFs

Let $\boldsymbol{x} = (x_1, ..., x_m)$ be the set of all random variables in our graph G.

Let $\mathcal{C}$ be the set of all maximal cliques of G.

The distribution p of X factorizes with respect to G if

$$p(\boldsymbol{x}) \propto \prod_{C \in \mathcal{C}} \psi_C(x_C)$$

for some nonnegative *potential* functions ψ_C, where $x_C = (x_i)_{i \in C}$.

The MRF on G represents the distributions that factorize wrt G.

The factored structure of the distribution makes it possible to more efficiently do the sums/integrals and is a form of dimension reduction.

Hammersley-Clifford Theorem

If $p(\boldsymbol{x}) > 0$ for all $\boldsymbol{x}$, the following statements are equivalent:

- p factorizes according to G, that is,

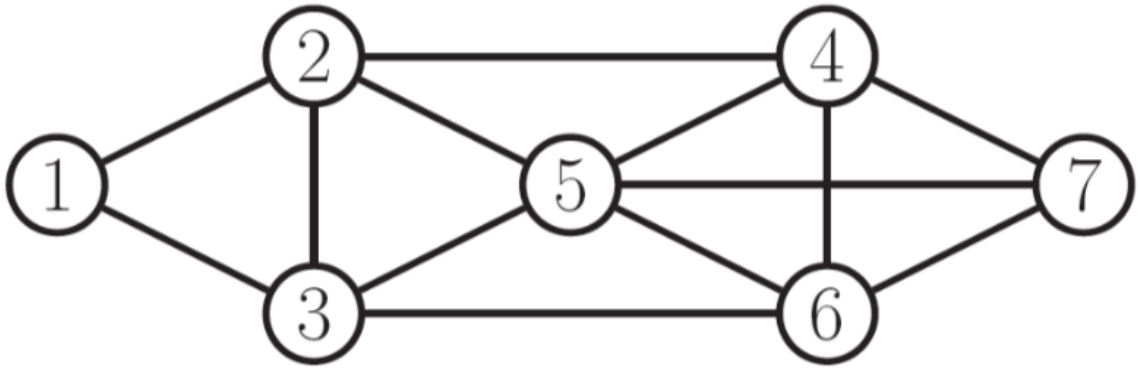
$$p(\boldsymbol{x}) \propto \prod_{C \in \mathcal{C}} \psi_C(x_C)$$

  for some nonnegative potential functions ψ_C.
- **Global Markov Properties**: $X_A \perp X_B | X_S$ if the sets A and B are *separated* by S in G (every path from A to B crosses S).

In particular,

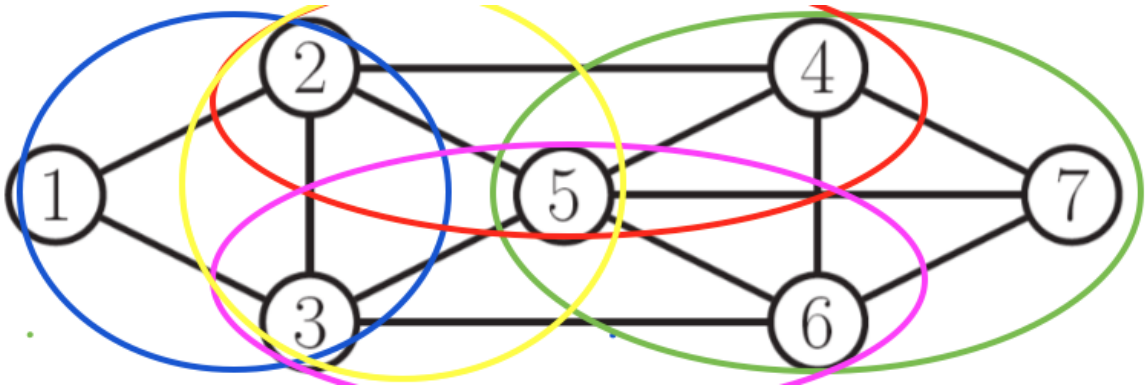
- If i, j are not connected by an edge then $X_i \perp X_j | X_{\text{rest}}$.
- The Markov blanket of X_i is given by its neighbors in G.

Example:



- How many maximal cliques are there?
- What is the underlying factorization?
- What are the induced conditional independence statements?

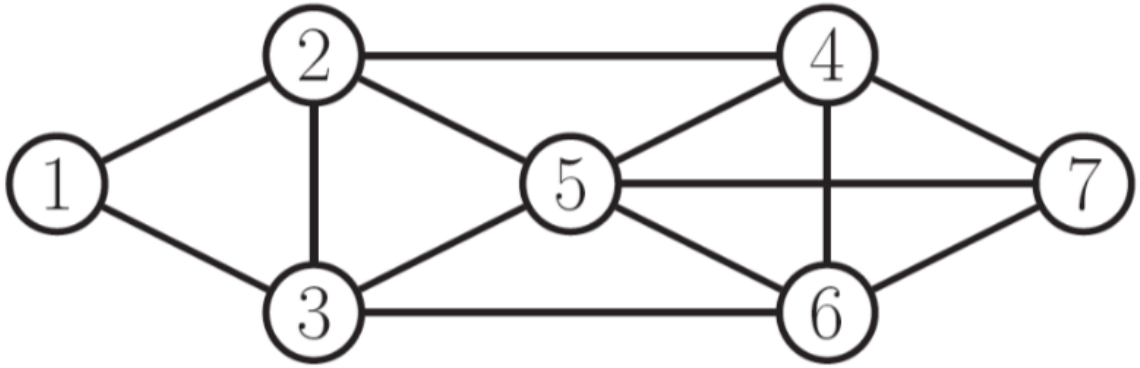
Example:



Lets see how to factorize the undirected graph of our running example:

$$p(\boldsymbol{x}) \propto \psi_{1,2,3}(x_1, x_2, x_3)\psi_{2,3,5}(x_2, x_3, x_5)\psi_{2,4,5}(x_2, x_4, x_5) \times \psi_{3,5,6}(x_3, x_5, x_6)\psi_{4,5,6,7}(x_4, x_5, x_6, x_7)$$

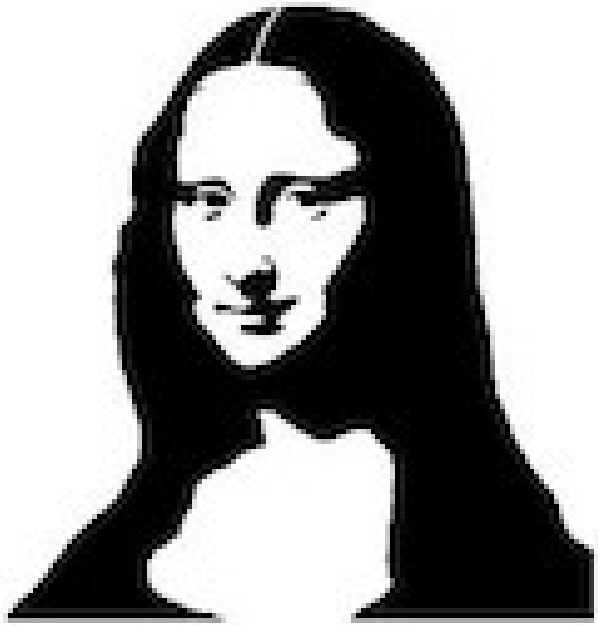
Example:

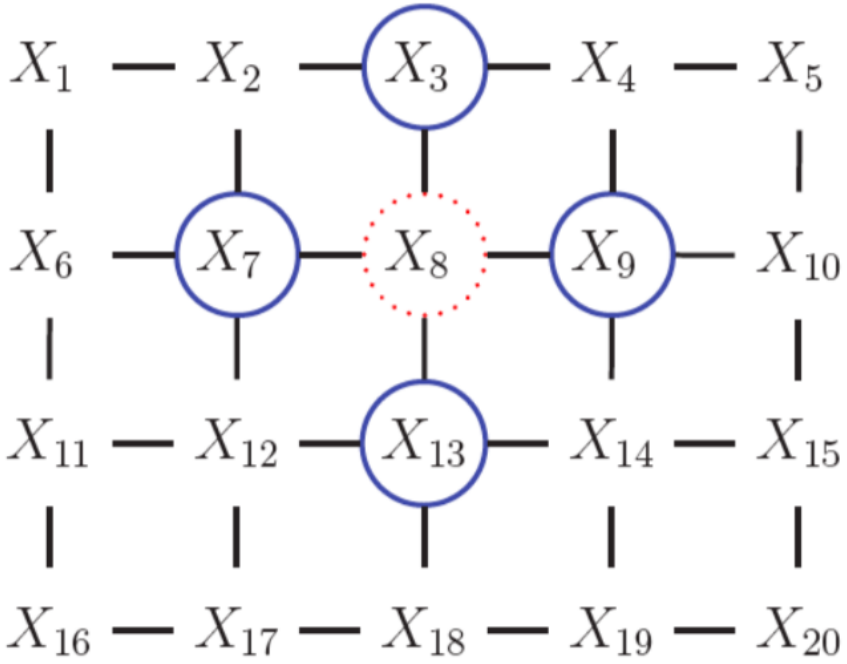


e.g. $(X_1, X_2) \perp (X_6, X_7) \mid (X_3, X_4, X_5)$

$X_1 \perp X_5 \mid (X_2, X_3)$

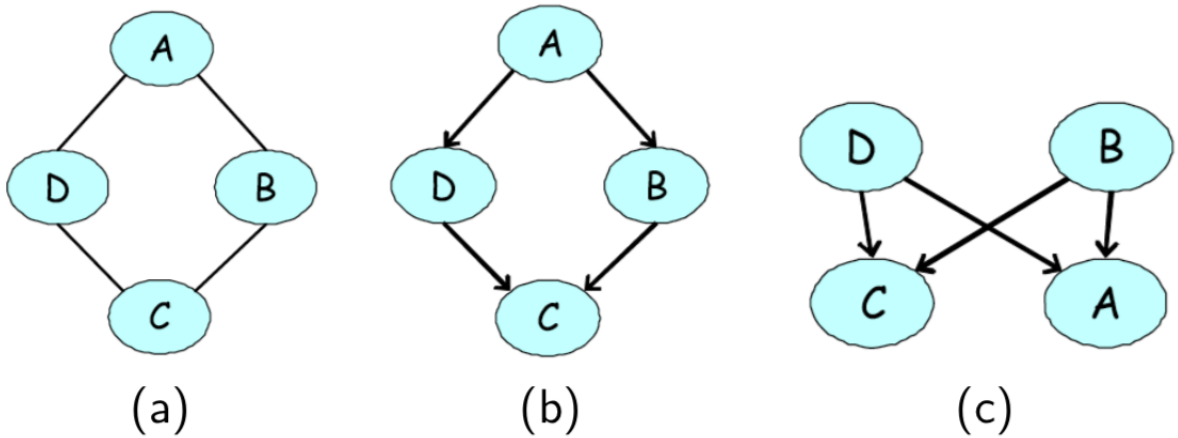
Image MRF




X_1 X_2 X_3 X_4 X_5
X_6 X_7 X_8 X_9 X_{10}
X_{11} X_{12} X_{13} X_{14} X_{15}
X_{16} X_{17} X_{18} X_{19} X_{20}

Not all MRFs can be represented as DAGMs

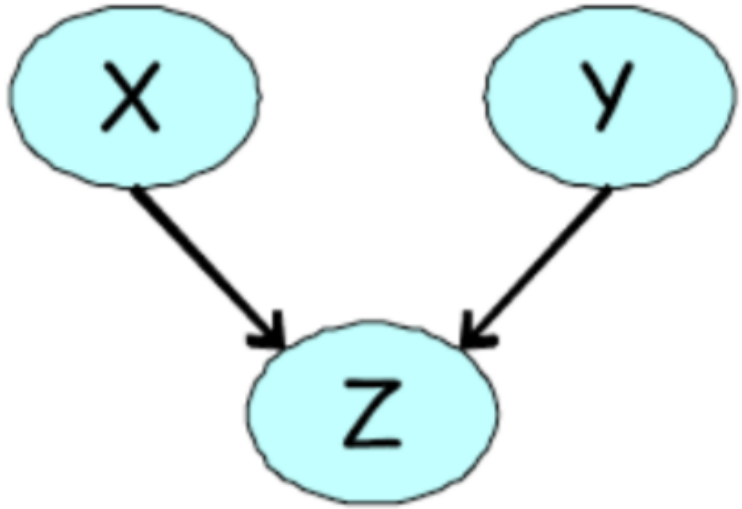
Take the following MRF for example (a) and our attempts at encoding this as a DAGM (b, c).



- Two conditional independencies in (a):
  - 1. $A \perp C | D, B$ 2. $B \perp D | A, C$
- In (b), we have the first independence, but not the second.
- In (c), we have the first independency, but not the second. Also, B and D are marginally independent.

Not all DAGMs can be represented as MRFs

Not all DAGMs can be represented as MRFs.
E.g. explaining away:



An undirected model is unable to capture the marginal independence, $X \perp Y$ that holds at the same time as $X \not\perp Y | Z$.

MRFs as Exponential Families

- Consider a parametric family of factorized distributions

$$p(\boldsymbol{x}|\theta) \;=\; \frac{1}{Z(\theta)} \prod_{C \in \mathcal{C}} \psi_C(x_C|\theta_C), \qquad \theta = (\theta_C)_{C \in \mathcal{C}}.$$

- We can write this in an exponential form:

$$p(\boldsymbol{x}|\theta) = \exp\Big\{ \sum_{C \in \mathcal{C}} \log \psi_C(x_C|\theta_C) - \underbrace{\log Z(\theta)}_{=A(\theta)} \Big\}$$

- Suppose the potentials have a log-linear form

$$\log \psi_C(x_C|\theta_C) = \theta_C^\top \, \phi_C(x_C)$$

we get the exponential family

$$p(\boldsymbol{x}|\theta) = \exp\Big\{ \sum_{C \in \mathcal{C}} \theta_C^\top \phi_C(x_C) - \underbrace{\log Z(\theta)}_{=A(\theta)} \Big\}$$

MRFs as Exponential Families

Question: When $\log \psi_C(x_C|\theta_C) = \theta_C^\top \phi_C(x_C)$?

Finite discrete case:

- If X is finite discrete then x_C takes a finite number of values and so $\log \psi_C$ takes a finite number of values.
- Take θ_C as all these possible values, and let $\phi_C(x_C)$ is a vector 1 on the entry correspond to x_C and zeros otherwise.
- Then $\log \psi_C(x_C|\theta_C) = \theta_C^\top \phi_C(x_C)$ as required.

Multivariate Gaussian case will be covered later in the lecture.

We can find the expectation of the C-th feature

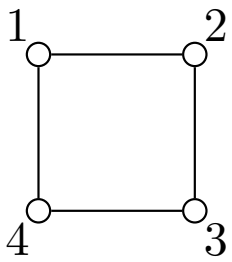
$$\frac{\partial \log Z(\theta)}{\partial \theta_C} = \mathbb{E}[\phi_C(X_C)].$$

Representing potentials

If the variables are finite discrete, we can represent the potential functions as tables of (non-negative) numbers.

e.f. consider a 4-cycle and binary random variables

$$p(x_1, x_2, x_3, x_4) = \frac{1}{Z}\psi_{1,2}(x_1, x_2)\psi_{2,3}(x_2, x_3)\psi_{3,4}(x_3, x_4)\psi_{1,4}(x_1, x_4)$$



$\psi_{1,2}(x_1, x_2)$

x_1	x_2	
0	0	30
0	1	5
1	0	1
1	1	10

$\psi_{2,3}(x_2, x_3)$

x_2	x_3	
0	0	100
0	1	1
1	0	1
1	1	100

$\psi_{3,4}(x_3, x_4)$

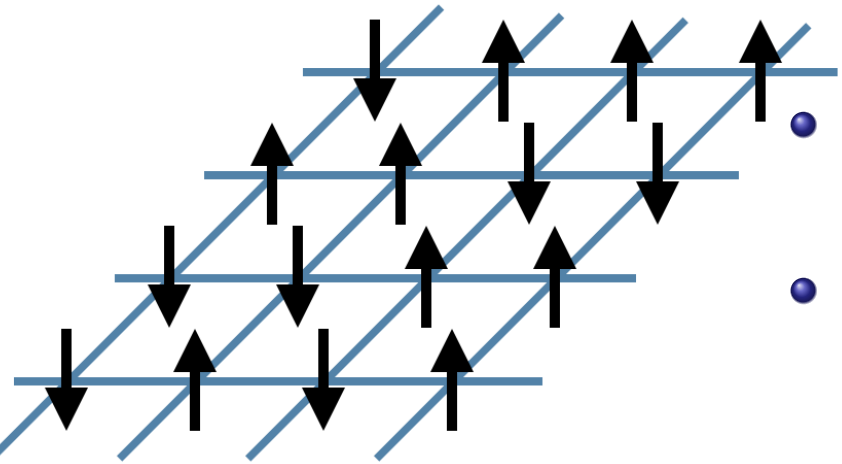
x_3	x_4	
0	0	1
0	1	100
1	0	100
1	1	1

$\psi_{1,4}(x_1, x_4)$

x_1	x_4	
0	0	100
0	1	1
1	0	1
1	1	100

These potentials are not probabilities since we ignored the normalization constant!

Example: Ising model



- The Ising model is an MRF that is used to model magnets.
- The nodes variables are spins, i.e., we use $x_s \in \{-1, +1\}$.

- Define the pairwise **clique potentials** as

$$\psi_{st}(x_s, x_t) \;=\; e^{J_{st} x_s x_t}.$$

  where J_{st} is the coupling strength between nodes s and t.
- $\psi_{st}(-1,-1) = \psi_{st}(1,1) = e^{J_{st}}; \quad \psi_{st}(-1,1) = \psi_{st}(1,-1) = e^{-J_{st}}$
- If two nodes are not connected set $J_{st} = 0$.

Ising model

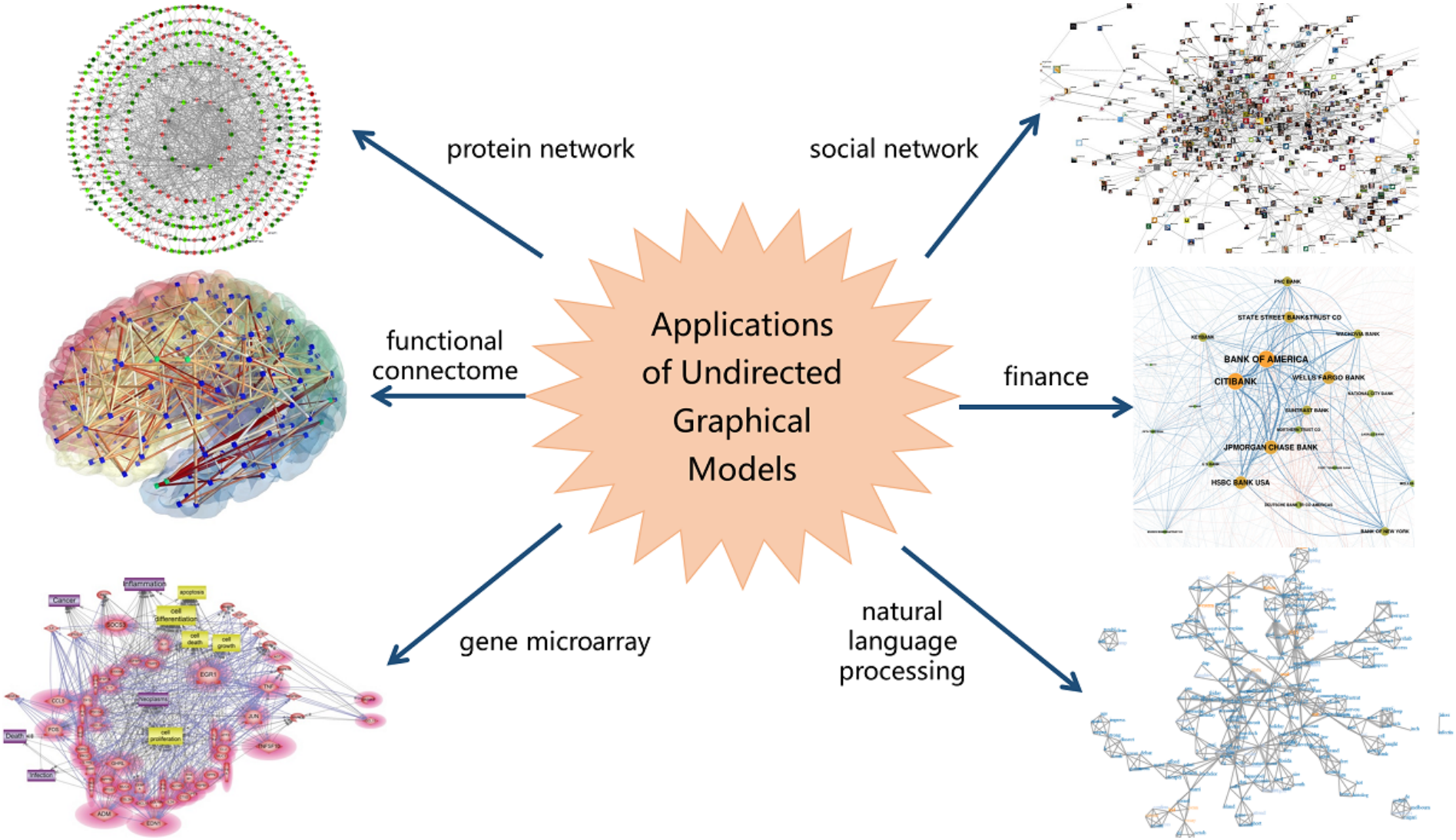
- We might want to add node potentials as well

$$\psi_s(x_s) = e^{b_s x_s}$$

- The overall distribution becomes

$$p(\boldsymbol{x}) \propto \prod_{s\sim t} \psi_{st}(x_s, x_t) \prod_s \psi_s(x_s) = \exp\Big\{ \sum_{s\sim t} J_{st} x_s x_t + \sum_s b_s x_s \Big\}.$$

- If $J_{st} > 0$ the model promotes same spins on neighboring spins.
- Hammersley-Clifford theorem: $J_{ij} = 0$ then $X_i \perp X_j | X_{\text{rest}}$.


protein network
social network
functional connectome
Applications of Undirected Graphical Models
finance
gene microarray
natural language processing
STATE STREET BANK&TRUST CO
BANK OF AMERICA
CITIBANK
WELLS FARGO BANK
JPMORGAN CHASE BANK
HSBC BANK USA

Multivariate Gaussian distribution

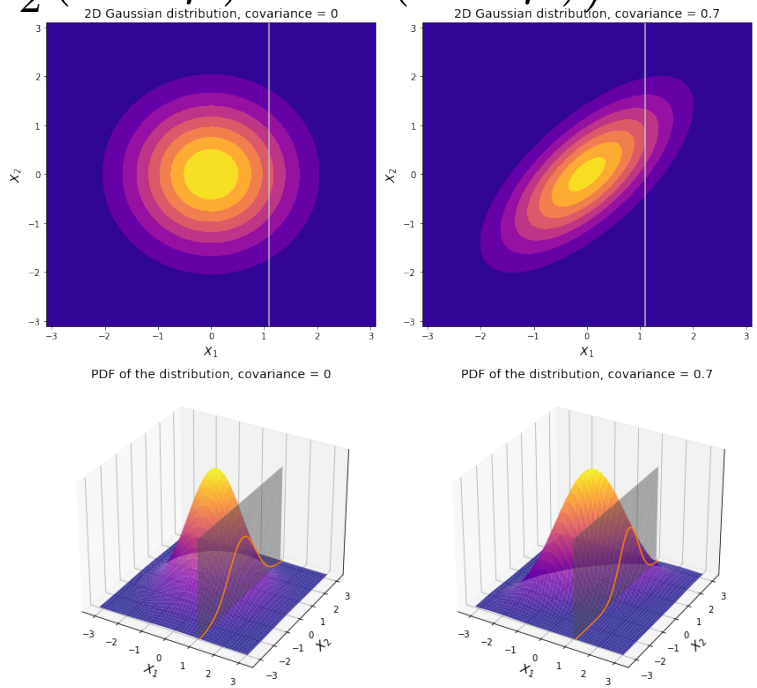
Univariate Gaussian: $f(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{1}{2\sigma^2}(x-\mu)^2)$.

Recall: Multivariate normal distribution, $X = (X_1,\ldots,X_m)$:

Let $\mu \in \mathbb{R}^m$ and Σ symmetric positive definite $m \times m$ matrix. We write $X \sim N_m(\mu,\Sigma)$ if the density of the vector X is

$$f(\boldsymbol{x};\mu,\Sigma) = \frac{1}{(2\pi)^{m/2}}(\det\Sigma)^{-1/2}\exp\left(-\frac{1}{2}(\boldsymbol{x}-\mu)^T\Sigma^{-1}(\boldsymbol{x}-\mu)\right).$$

Positive definite: $\forall \boldsymbol{u} \neq \mathbf{0} \quad \boldsymbol{u}^\top\Sigma\boldsymbol{u} > 0$.



Moments:

- mean vector: $\mathbb{E}X = \mu$,
- covariance: $\mathrm{var}(X) = \Sigma$.

Recall: Marginal and conditional distributions

Split X into two blocks $X = (X_A, X_B)$. Denote

$$\mu = (\mu_A, \mu_B) \qquad \text{and} \qquad \Sigma = \begin{bmatrix} \Sigma_{AA} & \Sigma_{AB} \\ \Sigma_{BA} & \Sigma_{BB} \end{bmatrix}.$$

Marginal distribution

$X_A \sim N(\mu_A, \Sigma_{AA})$

Conditional distribution

$X_A | X_B = x_B \sim N\left(\mu_A + \Sigma_{AB}\Sigma_{BB}^{-1}(x_B - \mu_B), \Sigma_{AA} - \Sigma_{AB}\Sigma_{BB}^{-1}\Sigma_{BA}\right)$

- Note that the conditional covariance is constant.

Some other properties

Linear transformations:

$A \in \mathbb{R}^{m \times p}$ for $m \leq p$ and $X \sim N_p(\mu, \Sigma)$ then $AX \sim N_m(A\mu, A\Sigma A^T)$.

Conditional independence:

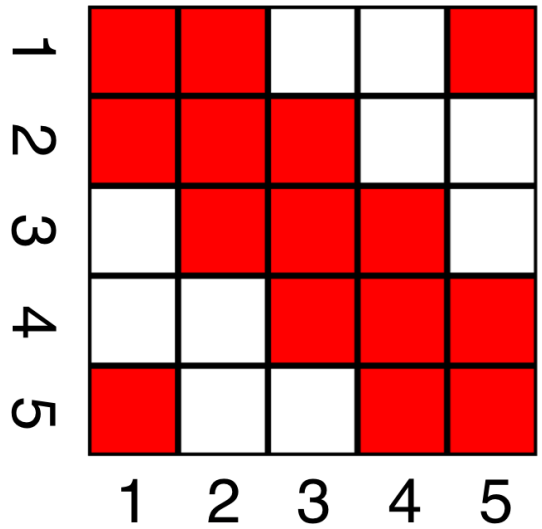
- $X_i \perp X_j$ if and only if $\Sigma_{ij} = 0$.
- $X_i \perp X_j | X_C$ if and only if $\Sigma_{ij} - \Sigma_{i,C}\Sigma_{C,C}^{-1}\Sigma_{C,j} = 0$
- Let $R = V \setminus \{i, j\}$. The following are equivalent:
  - $X_i \perp X_j | X_R$
  - $\Sigma_{ij} - \Sigma_{i,R}\Sigma_{R,R}^{-1}\Sigma_{R,j} = 0$
  - $(\Sigma^{-1})_{ij} = 0$

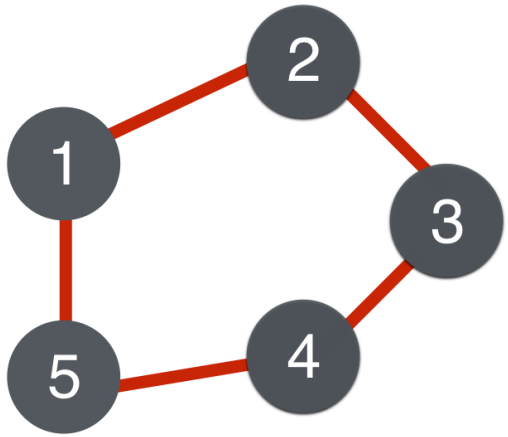
Gaussian Graphical models

Denote $K = \Sigma^{-1}$ then

$$p(\boldsymbol{x}|\mu, \Sigma) \propto \prod_s e^{-\frac{1}{2}K_{ss}(x_s-\mu_s)^2} \prod_{s<t} e^{-K_{st}(x_s-\mu_s)(x_t-\mu_t)}.$$

Important interpretation: $K_{ij} = 0$ if and only if $X_i \perp X_j | X_{\text{rest}}$.





Show that this is an exponential family.

Summary

Graphical models:

- Directed graphical models
- Undirected graphical models
- .... and the conditional independence they induce
- Next lecture: exact inference.