Graphical Models in Computer Vision

Andreas Geiger

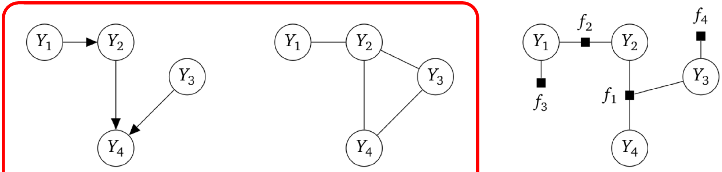
Max Planck Institute for Intelligent Systems
Perceiving Systems

April 18, 2016

MAX-PLANCK-GESELLSCHAFT

Todays topic

- Recap
  - Probability Theory
- Graphical Models
  - Directed Models
  - Undirected Models
  - Filter View



(a) Bayesian Network **(b)** Markov Random Field **(c)** Factor Graph

Recap

Probabilities: Discrete Case

- Joint Probability

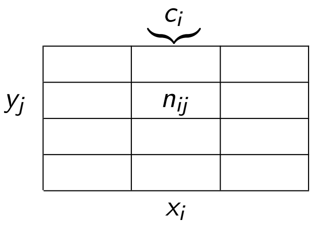
$$p(x_i, y_j) = \frac{n_{ij}}{N}$$

- Marginal Probability

$$p(x_i) = \frac{c_i}{N}$$

- Conditional Probability

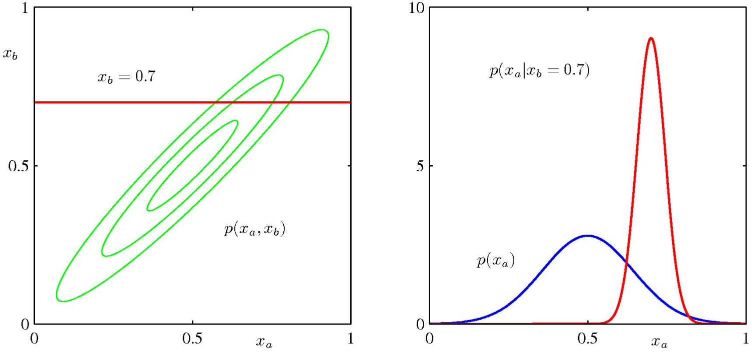
$$p(y_j \mid x_i) = \frac{n_{ij}}{c_i}$$



$$c_i = \sum_j n_{ij}$$

$$N = \sum_{ij} n_{ij}$$

Probabilities: Continuous Case



joint, marginal, conditional probability

The Rules of Probability

- Sum rule

$$p(X) = \sum_{y \in \mathcal{Y}} p(X, Y = y)$$

- Product Rule

$$p(X, Y) = p(Y|X)p(X)$$

- Bayes Theorem

$$p(Y|X) = \frac{p(X|Y)p(Y)}{p(X)}$$

Probability Variables – Notation

- Two random variables X and Y

Independence

X and Y are independent if

$$p(X, Y) = p(X)p(Y)$$

- Provided $p(X) \neq 0, p(Y) \neq 0$ this is equivalent with

$$p(X \mid Y) = p(X) \Leftrightarrow p(Y \mid X) = p(Y)$$

Probability Variables – Notation

- Sets of random variables $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$

Conditional independence

$\mathcal{X}$ and $\mathcal{Y}$ are independent provided we know the state of $\mathcal{Z}$ if

$$p(\mathcal{X}, \mathcal{Y} \mid \mathcal{Z}) = p(\mathcal{X} \mid \mathcal{Z}) p(\mathcal{Y} \mid \mathcal{Z})$$

for all states of $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$. They are conditional independent given $\mathcal{Z}$

- For conditional independence we write

$$\mathcal{X} \perp\!\!\!\perp \mathcal{Y} \mid \mathcal{Z}$$

- And thus we write for (unconditional) independence

$$\mathcal{X} \perp\!\!\!\perp \mathcal{Y} \mid \emptyset \quad \text{or shorter} \quad \mathcal{X} \perp\!\!\!\perp \mathcal{Y}$$

Probability Variables – Notation

- Similarly we write

$$\mathcal{X} \top\!\!\!\top \mathcal{Y} \mid \mathcal{Z}$$

  for conditionally dependent sets of random variables

- and

$$\mathcal{X} \top\!\!\!\top \mathcal{Y} \mid \emptyset \quad \text{or shorter} \quad \mathcal{X} \top\!\!\!\top \mathcal{Y}$$

  for unconditional dependent random variables

Dependent or not?

- A is independent of B ($A \perp\!\!\!\perp B$)
- B is independent of C ($B \perp\!\!\!\perp C$)
- C and A are ... ?

- Consider this distribution

$$p(A, B, C) = p(B)p(A, C)$$

- $a \perp\!\!\!\perp b$ and $b \perp\!\!\!\perp c$ because:

$$\begin{aligned} p(A, B) &= p(B) \sum_c p(A, C) \\ p(C, B) &= p(B) \sum_a p(A, C) \end{aligned}$$

- So A and C may or may not be independent

- And the other direction? $\{A \top\!\!\!\top B, B \top\!\!\!\top C\} \Rightarrow A \top\!\!\!\top C$?

Belief Networks

An example

- Mr. Holmes leaves his house
  - He sees that the lawn in front of his house is wet
  - This can have two reasons: he left the sprinkler turned on or it rained during the night.
  - Without any further information the probability of both events increases
- Now he also observes that his neighbour's lawn is wet
  - This lowers the probability that he left his sprinkler on. This event is *explained away*

Example continued

- Let's formalize:
- There are several random variables
  - $R \in \{0,1\}$, $R = 1$ means it has been **R**aining
  - $S \in \{0,1\}$, $S = 1$ means the **S**prinkler was left on
  - $N \in \{0,1\}$, $N = 1$ means **N**eighbours lawn is wet
  - $H \in \{0,1\}$, $H = 1$ means **H**olmes lawn is wet
- How many states to be specified?

$$p(R,S,N,H) = \underbrace{p(H \mid R,S,N)}_{2^3=8}\,\underbrace{p(N \mid R,S)}_{2^2=4}\,\underbrace{p(R \mid S)}_{2}\,\underbrace{p(S)}_{1}$$

- $8 + 4 + 2 + 1 = 15$ numbers needed to specify all probabilities
- In general $2^n - 1$ for binary states only

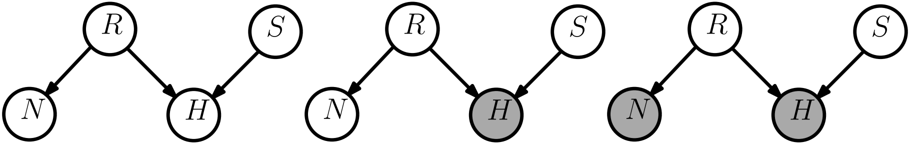
Example – Conditional Independence

- As a modeler of this problem we have prior knowledge of the dependencies / independencies
- $p(H \mid R, S, N) = p(H \mid R, S)$
- $p(N \mid R, S) = p(N \mid R)$
- $p(R \mid S) = p(R)$
- In effect our model becomes

$$p(R, S, N, H) = \underbrace{p(H \mid R, S)}_{4} \underbrace{p(N \mid R)}_{2} \underbrace{p(R)}_{1} \underbrace{p(S)}_{1}$$

- How many states? 8
- **H**olmes grass, **N**eighbours grass, **R**ain, **S**prinkler

This example as a Belief Network



- This is called a directed graphical model or belief network
  - observing the wet grass
  - observing the neighbours wet grass

Example – Inference

- The most pressing question is: was the sprinkler on?
- in other words what is $p(S = 1 \mid H = 1)$?
- First we need to specify the eight states (conditional probability table) (CPT)

$$
\begin{gathered}
p(R = 1) = 0.2, p(S = 1) = 0.1 \\
p(N = 1 \mid R = 0) = 0.2, \quad p(N = 1 \mid R = 1) = 1 \\
p(H = 1 \mid R = 0, S = 0) = 0, \quad p(H = 1 \mid R = 0, S = 1) = 0.9 \\
p(H = 1 \mid R = 1, S = 0) = 1, \quad p(H = 1 \mid R = 1, S = 1) = 1
\end{gathered}
$$

- $p(S = 1 \mid H = 1) = \ldots = 0.3382$
- $p(S = 1 \mid H = 1, N = 1) = \ldots = 0.1604$ (explained away)

Belief Networks

Belief network

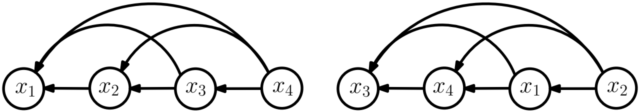
A belief network is a distribution of the form

$$p(x_1, \ldots, x_D) = \prod_{i=1}^{D} p(x_i \mid pa(x_i)),$$

where $pa(x)$ denotes the parental variables of x

- No cycles allowed! Why?
- $\Rightarrow$ Directed uncyclic graph (DAG)

Different factorizations



- Two factorizations of four variables:

$$
\begin{aligned}
p(x_1, x_2, x_3, x_4) &= p(x_1 \mid x_2, x_3, x_4)p(x_2 \mid x_3, x_4)p(x_3 \mid x_4)p(x_4) \\
p(x_1, x_2, x_3, x_4) &= p(x_3 \mid x_1, x_2, x_4)p(x_4 \mid x_1, x_2)p(x_1 \mid x_2)p(x_2)
\end{aligned}
$$

- Any distribution can be written in such a cascade form as a belief network (using Bayes' theorem)
- With independence assumptions the factorization is important

Conditional Independence

Conditional Independence

- Structure of the DAG corresponds to a set of conditional independence assumptions
  - which parents are sufficient (are the causes) to specify the CPT
  - to complete need to specify all $p(x \mid pa(x))$
- This does **not** mean non-parental variables have no influence:

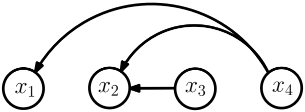
$$p(x_1 \mid x_2)p(x_2 \mid x_3)p(x_3)$$

with DAG $x_1 \leftarrow x_2 \leftarrow x_3$ does **not** imply (Exercise)

$$p(x_2 \mid x_1, x_3) = p(x_2 \mid x_3)$$

Conditional Independence

- Important task:
  - given graph, read of conditional independence statements
- Question:
  - are x_1 and x_2 conditional independent given x_4?
  - and what about $x_1 \perp\!\!\!\perp x_2 \mid x_3$?



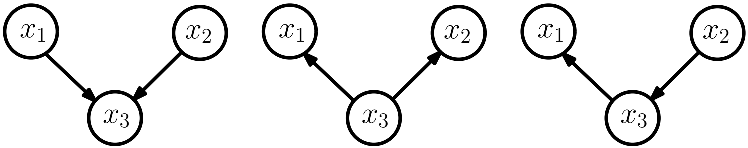
- how to automate?

Collisions

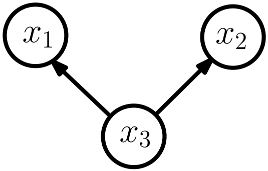
Collision

Given a path from node a to b, a collider is a node c for which there are two nodes a, b in the path pointing *towards* c. $(a \to c \leftarrow b)$

- Let's check these for colliders:



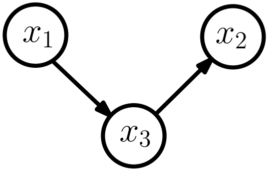
Collider and conditional independence



- x_3 a collider ? no
- $x_1 \perp\!\!\!\perp x_2 \mid x_3$? yes

$$
\begin{aligned}
p(x_1, x_2 \mid x_3) &= p(x_1, x_2, x_3)/p(x_3) \\
&= p(x_1 \mid x_3)p(x_2 \mid x_3)p(x_3)/p(x_3) \\
&= p(x_2 \mid x_3)p(x_1 \mid x_3)
\end{aligned}
$$

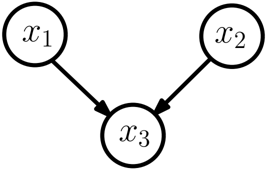
Collider and conditional independence



- x_3 a collider ? no
- $x_1 \perp\!\!\!\perp x_2 \mid x_3$? yes

$$\begin{aligned} p(x_1, x_2 \mid x_3) &= p(x_1, x_2, x_3)/p(x_3) \\ &= p(x_2 \mid x_3)p(x_3 \mid x_1)p(x_1)/p(x_3) \\ &= p(x_2 \mid x_3)p(x_1, x_3)/p(x_3) \\ &= p(x_2 \mid x_3)p(x_1 \mid x_3) \end{aligned}$$

Collider and conditional independence



- x_3 a collider ? yes
- $x_1 \perp\!\!\!\perp x_2 \mid x_3$? no! (explaining away)

$$\begin{aligned} p(x_1, x_2 \mid x_3) &= p(x_1, x_2, x_3)/p(x_3) \\ &= p(x_1)p(x_2)\underbrace{p(x_3 \mid x_1, x_2)/p(x_3)}_{\neq 1 \text{ in general}} \end{aligned}$$

- $x_1 \perp\!\!\!\perp x_2$? yes

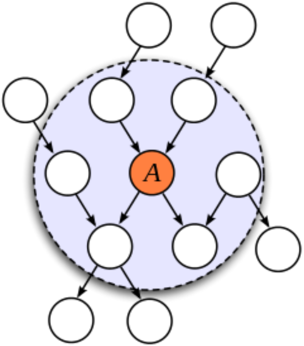
$$p(x_1, x_2) = \sum_{x_3} p(x_3 \mid x_1, x_2)p(x_1)p(x_2) = p(x_1)p(x_2)$$

Determining Conditional Independence I

Special case: The distribution of A conditioned on all other variables depends only on the variables in the "Markov blanket".

The Markov blanket comprises:

- Parents
- Children
- Parents of children



Determining Conditional Independence II

- There is a general algorithm to check for conditional independence $\mathcal{X} \perp\!\!\!\perp \mathcal{Y} \mid \mathcal{Z}$ in any belief network, called "D-separation":

D-separation

For every $x \in \mathcal{X}, y \in \mathcal{Y}$ check every path U between x and y.
A path is blocked if there is a node w on U such that either:

1. w is a collider and neither w nor any descendant is in $\mathcal{Z}$
2. w is not a collider on U and w is in $\mathcal{Z}$

If all such paths are blocked then $\mathcal{X}$ and $\mathcal{Y}$ are d-separated by $\mathcal{Z}$

- But, as always in life, there is alternatives ...

Determining Conditional Independence III

- Given $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$ how to determine whether $\mathcal{X} \perp\!\!\!\perp \mathcal{Y} \mid \mathcal{Z}$?
- The first two steps apply to directed graphs only
- Let $\mathcal{D} = \{\mathcal{X} \cup \mathcal{Y} \cup \mathcal{Z}\}$

1. Build the Ancestral Graph
    - Remove all nodes that are $\notin \mathcal{D}$ and not an ancestor of a node in $\mathcal{D}$
    - Also remove all edges in or out of such nodes
2. Moralisation
    - Connect parents with common child
    - Remove directions
3. Separation
    - Remove links neighbouring $\mathcal{Z}$
    - If no path links a node in $\mathcal{X}$ to a node in $\mathcal{Y} \Rightarrow \mathcal{X} \perp\!\!\!\perp \mathcal{Y} \mid \mathcal{Z}$

Markov equivalence

Markov equivalence

Two graphs are Markov equivalent if they represent the same set of conditional independence statements.
(holds for directed and undirected graphs)

Skeleton

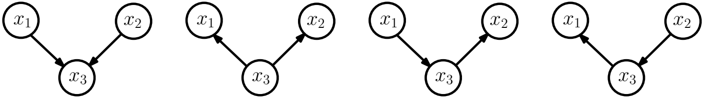
Graph resulting when removing all arrows of edges

Immorality

Parents of a child with no connection

- Markov equivalent $\Leftrightarrow$ Same skeleton and same set of immoralities

Three variable graphs revisited



(a) (b) (c) (d)

- All have the same skeleton
- (b,c,d) have no immoralities
- (a) has immorality (x_1, x_2) and is thus not equivalent

Markov Networks

Markov Networks

- So far:
  - Factorization with each factor a (conditional) probability distribution
  - Normalization as a by-product
- Alternative:

$$p(a, b, c) = \frac{1}{Z}\phi(a, b)\phi(b, c)$$

- Here Z normalization constant or partition function

$$Z = \sum_{a,b,c} \phi(a, b)\phi(b, c)$$

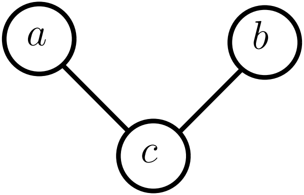
Definitions

Potential

A potential $\phi(x)$ is a non-negative function of the variable x. A joint potential $\phi(x_1, \ldots, x_D)$ is a non-negative function of the set of variables.

- Distribution (as in belief networks) is a special choice

Example



$$p(a, b, c) = \frac{1}{Z}\phi_{ac}(a, c)\phi_{bc}(b, c)$$

Markov Network

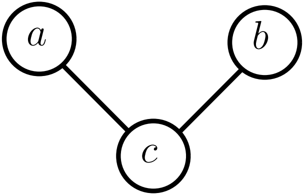
Markov Network

For a set of variables $\mathcal{X} = \{x_1, \ldots, x_D\}$ a Markov network is defined as a product of potentials over the maximal cliques $\mathcal{X}_c$ of the graph $\mathcal{G}$

$$p(x_1, \ldots, x_D) = \frac{1}{Z} \prod_{c=1}^{C} \phi_c(\mathcal{X}_c)$$

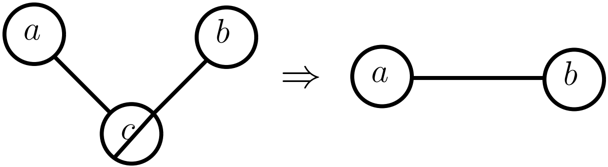
- Special case: cliques of size 2 – pairwise Markov network
- If all potentials are strictly positive this is called a Gibbs distribution

Properties of Markov Networks



$$p(a, b, c) = \frac{1}{Z}\phi_{ac}(a, c)\phi_{bc}(b, c)$$

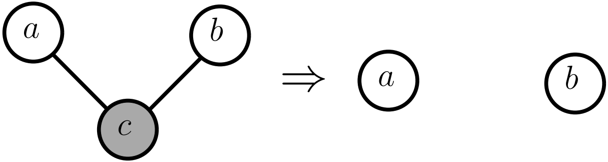
Properties of Markov Networks



- Marginalizing over c makes a and b "graphically" dependent
- Check

$$p(a, b) = \sum_c \frac{1}{Z} \phi_1(a, c)\phi_2(b, c)$$

Properties of Markov Networks



- Conditioning on c makes a and b independent (whiteboard)

$$p(a, b \mid c) = p(a \mid c)p(b \mid c)$$

- This is opposite to the directed version $a \rightarrow c \leftarrow b$ where conditioning *introduced* dependency

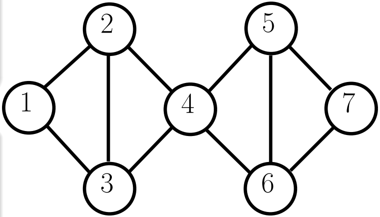
Global Markov Property

Separation

A subset $\mathcal{S}$ separates $\mathcal{A}$ from $\mathcal{B}$ if every path from a member of $\mathcal{A}$ to any member of $\mathcal{B}$ passes through $\mathcal{S}$.

Global Markov Property

For disjoint sets of variables $(\mathcal{A}, \mathcal{B}, \mathcal{S})$ where $\mathcal{S}$ separates $\mathcal{A}$ from $\mathcal{B}$, then $\mathcal{A} \perp\!\!\!\perp \mathcal{B} \mid \mathcal{S}$



Local Markov Property

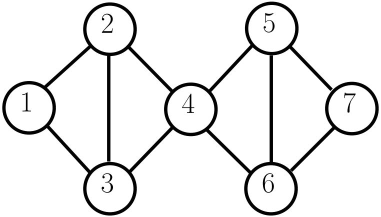
- For positive potentials, the so-called local Markov property holds

Local Markov Property

$$p(x \mid \mathcal{X} \setminus \{x\}) = p(x \mid ne(x))$$

- The set of neighboring nodes $ne(x)$ is called the Markov blanket
- This also holds for set of variables $\Rightarrow$ Simple independence check by separation (third step in algorithm III discussed for directed models)

Local Markov Property – Example



- $p(x_4 \mid x_1, x_2, x_3, x_4, x_5, x_6, x_7) = p(x_4 \mid x_2, x_3, x_5, x_6)$
- in other words $x_4 \perp\!\!\!\perp \{x_1, x_7\} \mid \{x_2, x_3, x_5, x_6\}$
- and others

Markov Random Field (MRF)

Markov Random Field

A MRF is defined by a set of distributions $p(x_i \mid ne(x_i))$. A distribution is a *Markov Random Field* with respect to an undirected graph G if

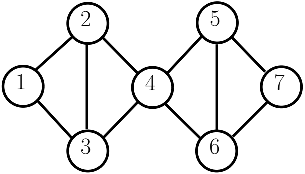
$$p(x_i \mid x_{\setminus i}) = p(x_i \mid ne(x_i))$$

- Not every set of conditional distributions $p(x_i \mid x_{\setminus i})$ yields a valid joint distribution (exercise)

Finding the factorization

- An undirected graph G specifies a set of conditional independence statements
- Question: What is the most general factorization F (form of the distribution) that satisfies these independences?

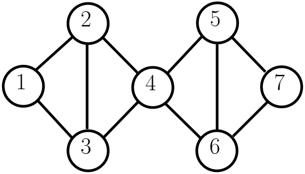
Finding the factorization



- Eliminate variable one by one
- Let's start with x_1

$$p(x_1, \ldots, x_7) = p(x_1 \mid x_2, x_3) p(x_2, \ldots, x_7)$$

Finding the factorization



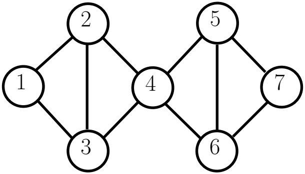
- Graph specifies:

$$\begin{aligned} & p(x_1, x_2, x_3 \mid x_4 \ldots, x_7) &= p(x_1, x_2, x_3 \mid x_4) \\ \Rightarrow \quad & p(x_2, x_3 \mid x_4, \ldots x_7) &= p(x_2, x_3 \mid x_4) \end{aligned}$$

- Hence

$$p(x_1, \ldots, x_7) = p(x_1 \mid x_2, x_3) p(x_2, x_3, \mid x_4) p(x_4, x_5, x_6, x_7)$$

Finding the factorization



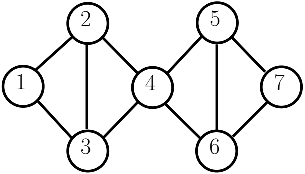
- We continue to find

$$
\begin{aligned}
p(x_1, \ldots, x_7) \quad = \quad & p(x_1 \mid x_2, x_3) p(x_2, x_3 \mid x_4) \\
& p(x_4 \mid x_5, x_6) p(x_5, x_6 \mid x_7) p(x_7)
\end{aligned}
$$

- A factorization into clique potentials (maximal cliques)

$$
p(x_1, \ldots, x_7) = \frac{1}{Z} \phi(x_1, x_2, x_3) \phi(x_2, x_3, x_4) \phi(x_4, x_5, x_6) \phi(x_5, x_6, x_7)
$$

Finding the factorization



- Markov conditions of graph $G \Rightarrow$ factorization F into clique potentials
- And conversely: $F \Rightarrow G$

Hammersley-Clifford Theorem

Hammersely-Clifford

This factorization property $G \Leftrightarrow F$ holds for any undirected graph provided that the potentials are positive

- Thus also loopy ones: $x_1 - x_2 - x_3 - x_4 - x_1$
- Theorem says, distribution is of the form

$$p(x_1, x_2, x_3, x_4) = \frac{1}{Z}\phi_{12}(x_1, x_2)\phi_{23}(x_2, x_3)\phi_{34}(x_3, x_4)\phi_{41}(x_4, x_1)$$

Directed vs. Undirected

Bayes or Markov?

- So which one is better? Directed or Undirected ?
- Both directed and undirected graphical models imply sets of conditional independences
- Which one models more distributions? Or are they the same?
- First introduce "canonical" representation

Relationship directed – undirected models: maps

D Map

A graph is said to be a D map (dependency map) of a distribution if every conditional independence statement satisfied by the distribution is reflected in the graph

- A completely disconnected graph contains all possible independence statements for its variables
- $\Rightarrow$ it is a trivial D map for any distribution

Relationship directed – undirected models: maps

I Map

A graph is said to be a **I map** (independence map) of a distribution if every conditional independence implied by the graph is satisfied by the distribution

- A fully connected graph implies no independence statements
- $\Rightarrow$ it is a trivial I map for any distribution

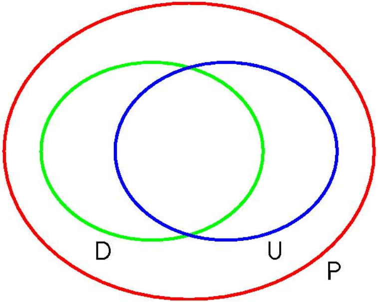
Relationship directed – undirected models: maps

Perfect Map

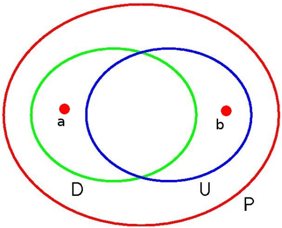
If every conditional independence property of the distribution is reflected in the graph, **and vice versa**, then the graph is said to be a perfect map for that distribution.

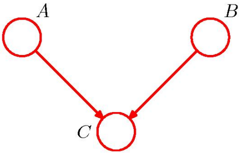
- A perfect map: Both I map and a D map of the distribution

Relationship directed – undirected GM

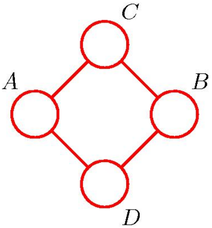


- P – set of all distributions for a given set of variables
- Distributions that can be represented as a perfect map
  - using undirected graph – U
  - using a directed graph – D





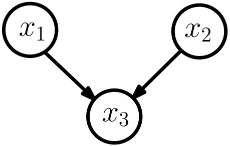
(a)



(b)

- Middle: conditional independence properties cannot be expressed using an undirected graph over the same three variables
- Right: conditional independence properties cannot be expressed using a directed graph over the same four variables

Filter View of a Graphical Model

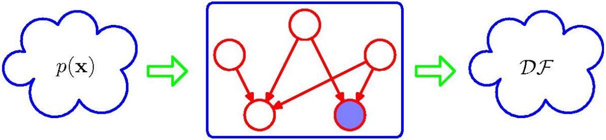


- This graph has one conditional independence statement only: $x_1 \perp\!\!\!\perp x_2$
- The following distribution satisfies this:
  $p_1(x_3 = 1 \mid x_1, x_2) = (x_1 - x_2)^2,\ p_1(x_1 = 1) = 0.3,\ p_1(x_2 = 1) = 0.4$
- **BUT:** More conditional independencies are possible:

$$p_2(x_3 = 1 \mid x_1, x_2) = 0.5,\ p_2(x_1 = 1) = 0.3,\ p_2(x_2 = 1) = 0.4$$

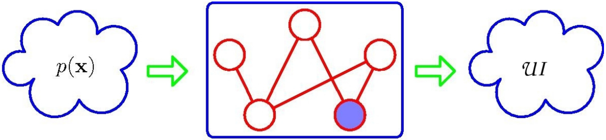
p_2 satisfies $\{x_1 \perp\!\!\!\perp x_2, x_1 \perp\!\!\!\perp x_3, x_2 \perp\!\!\!\perp x_3\}$

Filter View of a Graphical Model



- Belief network implies a list of conditional independences
- Regard as filter:
  - Only distributions $\mathcal{DF}$ that satisfy all conditional independences are allowed to pass
  - All distributions satisfying the d-separation theorem pass
- One graph describes a whole family of probability distributions
- Extremes:
  - Fully connected, no constraints, all p pass
  - no connections, only product of marginals may pass

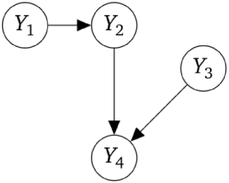
Filter View of a Graphical Model



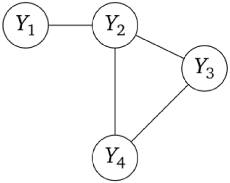
- Let $\mathcal{UI}$ denote the distributions that can pass
  - those that satisfy all conditional independence statements
  - those which can be read from the graph using graph separation
- Let $\mathcal{UF}$ denote the distributions with factorization over cliques
- Hammersley-Clifford says: $\mathcal{UI} = \mathcal{UF}$

Next Time ...

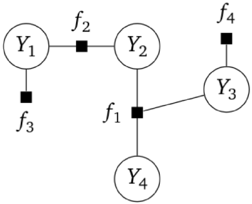
- ... we will meet our last friend:



(a) Bayesian Network



(b) Markov Random Field



(c) Factor Graph

Questions?