Part I

C. M. Bishop

PATTERN RECOGNITION AND MACHINE LEARNING

CHAPTER 8: GRAPHICAL MODELS

Probabilistic Graphical Models

- Graphical representation of a probabilistic model
- Each variable corresponds to a node in the graph
- Links in the graph denote probabilistic relations between variables

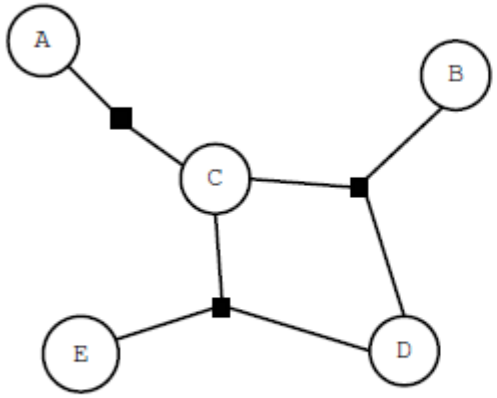
Why do we need graphical models?

- Graphs are an intuitive way of representing and visualising the relationships between many variables. (Examples: family trees, electric circuit diagrams, neural networks)

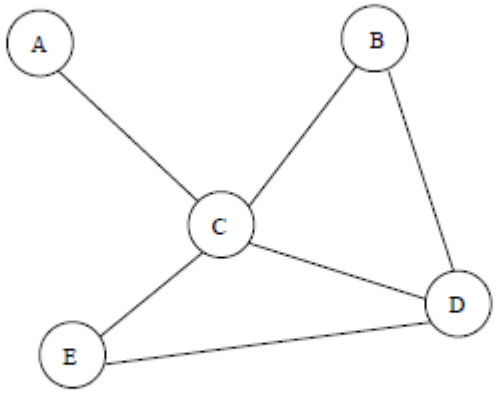
- A graph allows us to abstract out the conditional independence relationships between the variables from the details of their parametric forms. Thus we can ask questions like: “Is A dependent on B given that we know the value of C ?” just by looking at the graph.

- Graphical models allow us to define general message-passing algorithms that implement Bayesian inference efficiently. Thus we can answer queries like “What is P(A|C = c)?” without enumerating all settings of all variables in the model.

Three kinds of Graphical Models

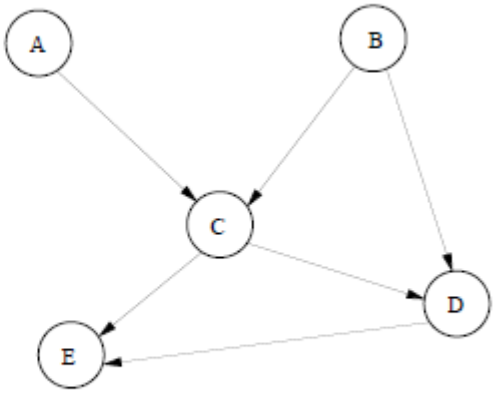


factor graph



undirected graph

or Markov random fields



directed graph

or Bayesian Networks

convenient for solving inference problems

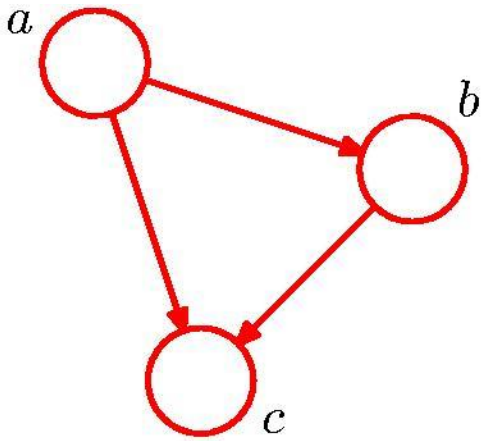
useful to express soft constraints between variables

useful to express causal relationships between variables

image by Zoubin Ghahramani

Bayesian Networks

Directed Acyclic Graph (DAG)



$$p(a,b,c) = p(c|a,b)p(a,b) = p(c|a,b)p(b|a)p(a)$$

Note: the left-hand side is symmetrical w/r to the variables whereas the right-hand side is not.

Bayesian Networks

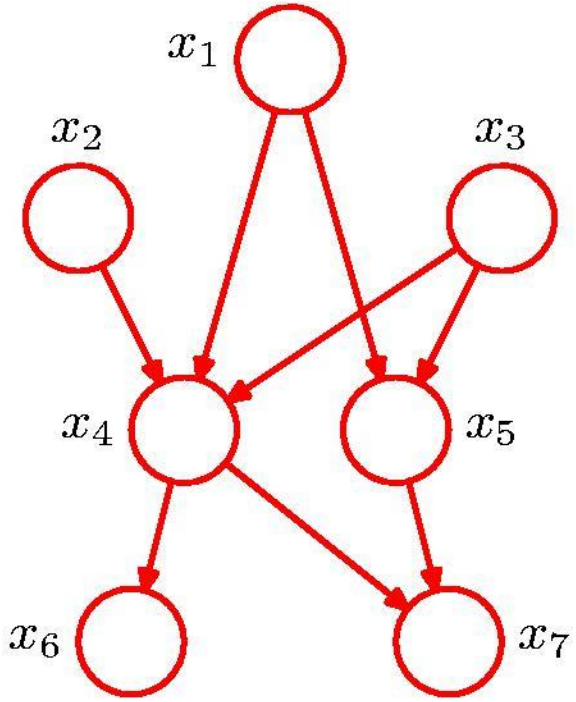
Generalization to K variables:

$$p(x_1, \ldots, x_K) = p(x_K|x_1, \ldots, x_{K-1}) \ldots p(x_2|x_1)p(x_1)$$

- The associated graph is fully connected.
- The absence of links conveys important information.

Bayesian Networks

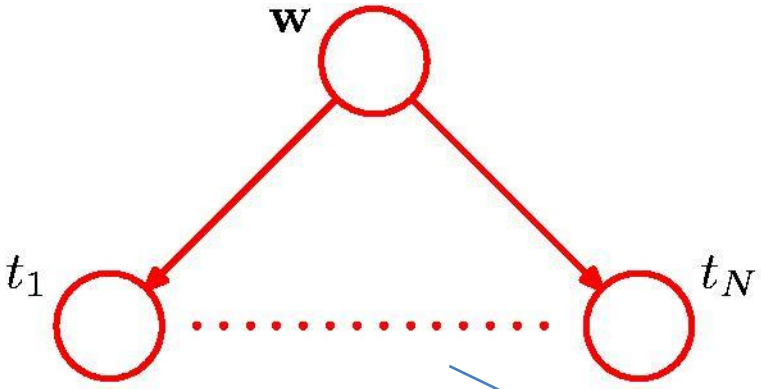
$$p(x_1, \ldots, x_7) = p(x_1)p(x_2)p(x_3)p(x_4|x_1, x_2, x_3)$$
$$p(x_5|x_1, x_3)p(x_6|x_4)p(x_7|x_4, x_5)$$



General Factorization

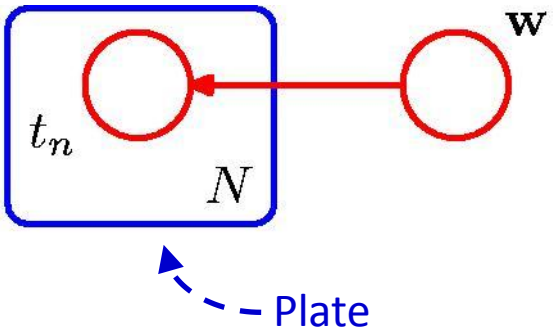
$$p(\mathbf{x}) = \prod_{k=1}^{K} p(x_k|\mathrm{pa}_k)$$

Some Notations(1)



$$p(\mathrm{t}, \mathrm{w}) = p(\mathrm{w}) \prod_{n=1}^{N} p(t_n \mid \mathrm{w})$$

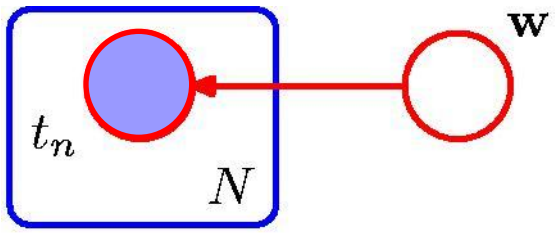
More compact representation



Some Notations(2)

Condition on data

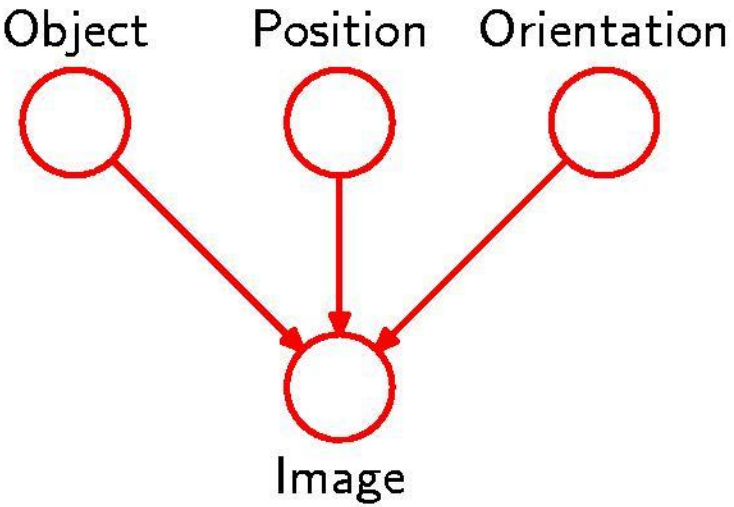
$$p(\mathbf{w}|\mathbf{t}) \propto p(\mathbf{w}) \prod_{n=1}^{N} p(t_n|\mathbf{w})$$



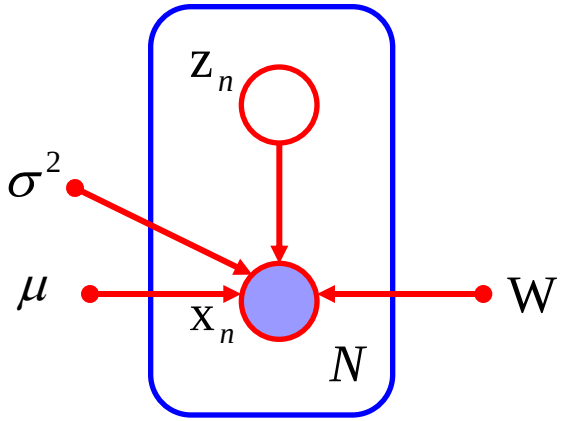
Shaded to indicate that the r.v. is set to its observed value

Generative Models

Causal process for generating images



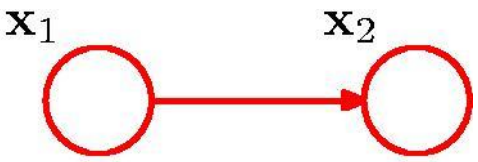
PPCA



Discrete Variables

Denote the probability of observing both $x_{1k} = 1$ and $x_{2l} = 1$ by μ_{kl}, $\sum_k \sum_l \mu_{kl} = 1$

General joint distribution: $K^2 - 1$ parameters



K states K states

$$p(\mathbf{x}_1, \mathbf{x}_2 | \boldsymbol{\mu}) = \prod_{k=1}^{K} \prod_{l=1}^{K} \mu_{kl}^{x_{1k} x_{2l}}$$

$$p(\mathrm{x}_1, \mathrm{x}_2) = p(\mathrm{x}_2 \mid \mathrm{x}_1) p(\mathrm{x}_1)$$

K^2-1 parameters

K -1 parameters

K -1 parameters

Discrete Variables

Independent joint distribution: $2(K-1)$ parameters

$$\hat{p}(\mathbf{x}_1, \mathbf{x}_2|\boldsymbol{\mu}) = \prod_{k=1}^{K} \mu_{1k}^{x_{1k}} \prod_{l=1}^{K} \mu_{2l}^{x_{2l}}$$

Discrete Variables

General joint distribution over M variables:

$K^M - 1$ parameters

M-node Markov chain: $\overbrace{K-1}^{p(\mathrm{x}_1)} + (M-1)\overbrace{K(K-1)}^{p(\mathrm{x}_i \mid \mathrm{x}_{i-1})\, p(\mathrm{x}_{i-1}),\, i=2,\ldots,M}$ parameters

Reduce the number of parameters by dropping link in the graph, at the expense of having a restricted class of distributions.

Conditional Independence

a is independent of b given c

$$p(a|b,c) = p(a|c)$$

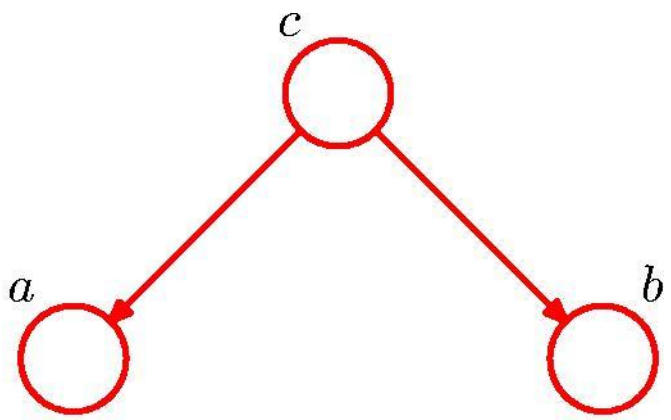
Equivalently

$$\begin{aligned} p(a,b|c) &= p(a|b,c)p(b|c) \\ &= p(a|c)p(b|c) \end{aligned}$$

Notation

$$a \perp\!\!\!\perp b \mid c$$

Conditional Independence: Example 1



tail-to-tail

$$p(a, b, c) = p(a|c)p(b|c)p(c)$$

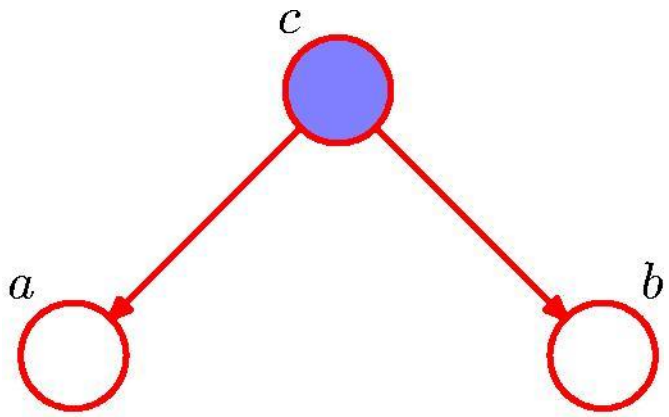
marginalize w.r.t c

$$p(a, b) = \sum_c p(a|c)p(b|c)p(c)$$

$$a \not\perp\!\!\!\perp b \mid \emptyset$$

as it doesn't factorize into $p(a)p(b)$ in general

Conditional Independence: Example 1



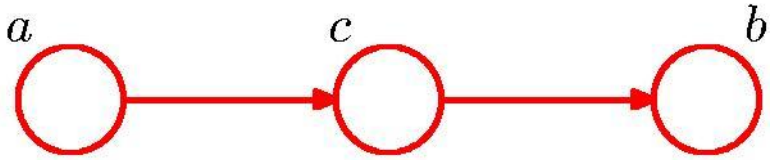
tail-to-tail

$$p(a,b|c) = \frac{p(a,b,c)}{p(c)}$$
$$= p(a|c)p(b|c)$$

$$a \perp\!\!\!\perp b \mid c$$

Conditional Independence: Example 2

head-to-tail



$$p(a, b, c) = p(a)p(c|a)p(b|c)$$

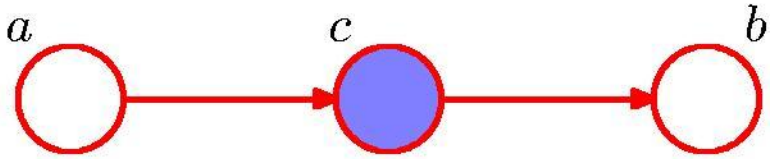
marginalize w.r.t c

$$p(a, b) = p(a) \sum_c p(c|a)p(b|c) = p(a)p(b|a)$$

$$a \not\perp b \mid \emptyset$$

Conditional Independence: Example 2

head-to-tail

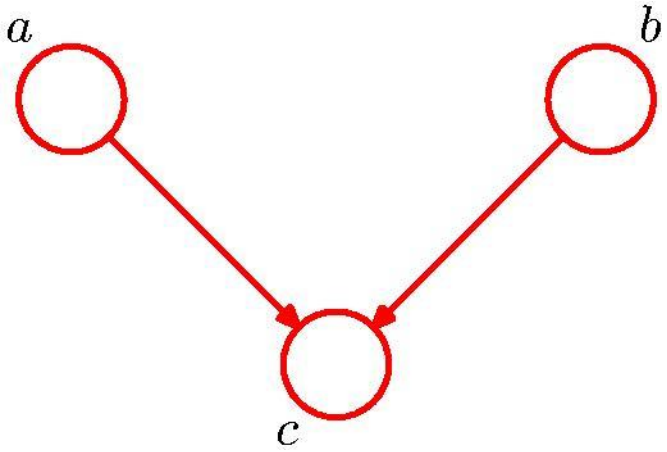


$$
\begin{aligned}
p(a, b|c) &= \frac{p(a, b, c)}{p(c)} \\
&= \frac{p(a)p(c|a)p(b|c)}{p(c)} \\
&= p(a|c)p(b|c)
\end{aligned}
$$

$$a \perp\!\!\!\perp b \mid c$$

Conditional Independence: Example 3

head-to-head



$$p(a, b, c) = p(a)p(b)p(c|a, b)$$

marginalize w.r.t c

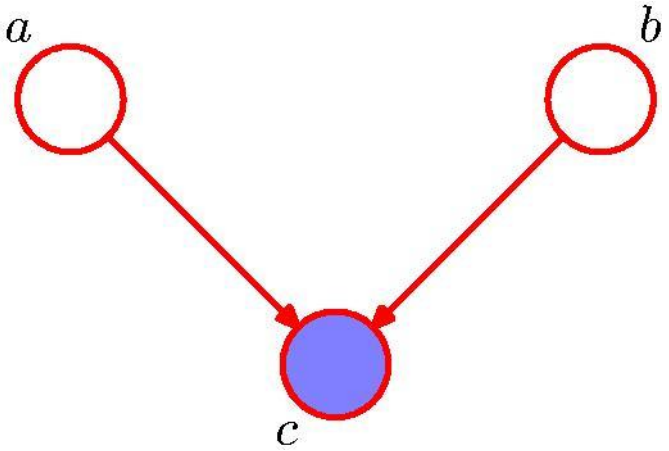
$$p(a, b) = p(a)p(b)$$

$$a \perp\!\!\!\perp b \mid \emptyset$$

Note: this is the opposite of Example 1 and 2, with c unobserved.

Conditional Independence: Example 3

head-to-head



$$p(a,b|c) = \frac{p(a,b,c)}{p(c)}$$

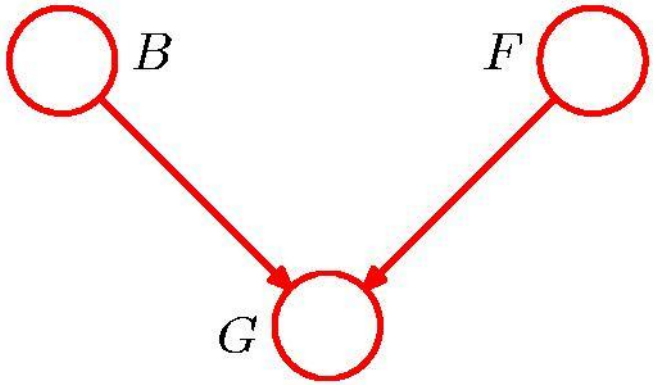
$$= \frac{p(a)p(b)p(c|a,b)}{p(c)}$$

$$a \not\perp b \mid c$$

Note: this is the opposite of Example 1 and 2, with c observed.

“Am I out of fuel?”

$$
\begin{aligned}
p(G=1|B=1,F=1) &= 0.8 \\
p(G=1|B=1,F=0) &= 0.2 \\
p(G=1|B=0,F=1) &= 0.2 \\
p(G=1|B=0,F=0) &= 0.1
\end{aligned}
$$



$$
\begin{aligned}
p(B=1) &= 0.9 \\
p(F=1) &= 0.9
\end{aligned}
$$

and hence

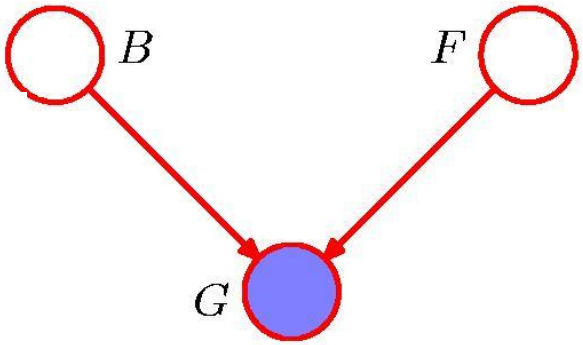
$$
p(F=0) = 0.1
$$

B = Battery (0=flat, 1=fully charged)
F = Fuel Tank (0=empty, 1=full)
G = Fuel Gauge Reading (0=empty, 1=full)

“Am I out of fuel?”

$$p(G=0) = \sum_{B\in\{0,1\}} \sum_{F\in\{0,1\}} p(G=0|B,F)p(B)p(F) = 0.315$$

$$p(G=0|F=0) = \sum_{B\in\{0,1\}} p(G=0|B,F=0)p(B) = 0.81$$

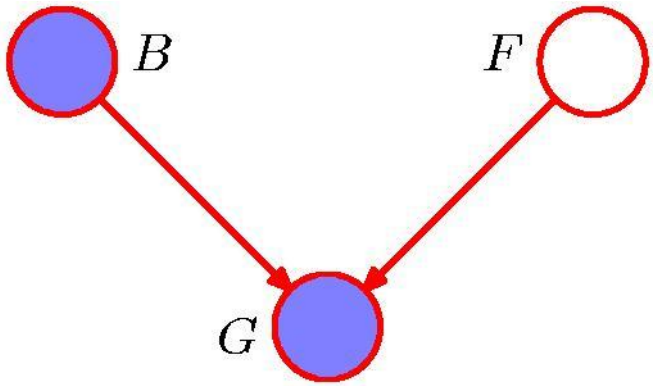


$$\begin{aligned} p(F=0|G=0) &= \frac{p(G=0|F=0)p(F=0)}{p(G=0)} \\ &\simeq 0.257 \end{aligned}$$

Probability of an empty tank increased by observing $G = 0$.

“Am I out of fuel?”

the state of the fuel tank and that of the battery have become dependent on each other as a result of observing the reading on the fuel gauge.



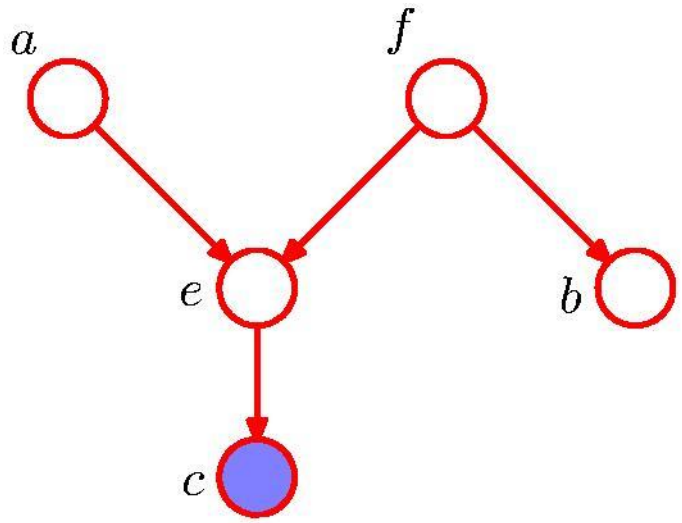
$$p(F=0|G=0,B=0) = \frac{p(G=0|B=0,F=0)p(F=0)}{\sum_{F\in\{0,1\}} p(G=0|B=0,F)p(F)}$$
$$\simeq 0.111$$

Probability of an empty tank reduced by observing $B=0$.
This referred to as “explaining away”.

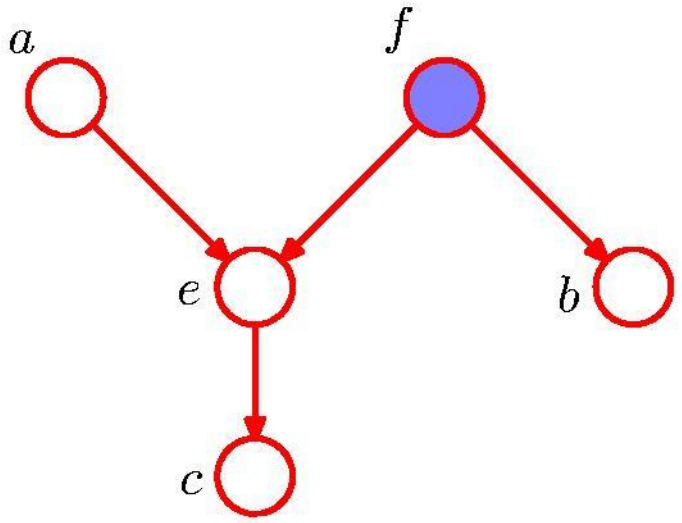
D-separation

- A, B, and C are non-intersecting subsets of nodes in a directed graph.
- A path from A to B is blocked if it contains a node such that either
  a) the arrows on the path meet either head-to-tail or tail-to-tail at the node, and the node is in the set C, or
  b) the arrows meet head-to-head at the node, and neither the node, nor any of its descendants, are in the set C.
- If all paths from A to B are blocked, A is said to be d-separated from B by C.
- If A is d-separated from B by C, the joint distribution over all variables in the graph satisfies $A \perp\!\!\!\perp B \mid C$.

D-separation: Example

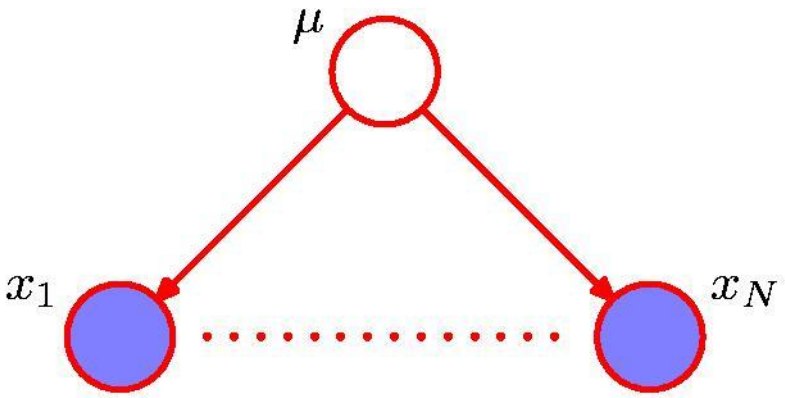


$a \not\perp\!\!\!\perp b \mid c$



$a \perp\!\!\!\perp b \mid f$

D-separation: I.I.D. Data

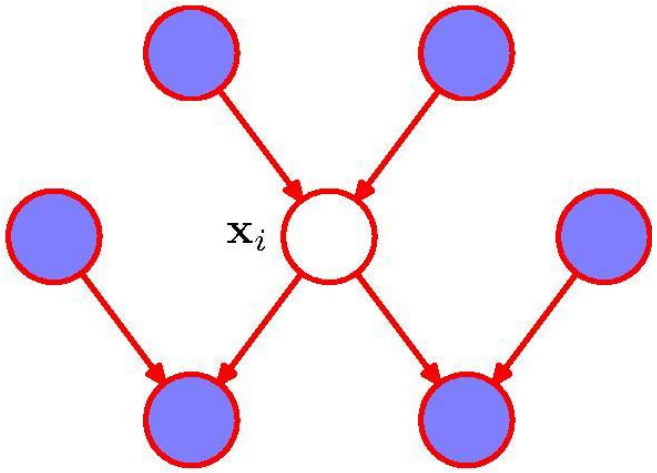


$$p(\mathcal{D}|\mu) = \prod_{n=1}^{N} p(x_n|\mu)$$

$$p(\mathcal{D}) = \int_{-\infty}^{\infty} p(\mathcal{D}|\mu)p(\mu)\,\mathrm{d}\mu \neq \prod_{n=1}^{N} p(x_n)$$

The Markov Blanket

due to ‘explaining away’ phenomenon



$$p(\mathbf{x}_i|\mathbf{x}_{\{j\neq i\}}) = \frac{p(\mathbf{x}_1,\ldots,\mathbf{x}_M)}{\int p(\mathbf{x}_1,\ldots,\mathbf{x}_M)\,\mathrm{d}\mathbf{x}_i}$$

$$= \frac{\prod_k p(\mathbf{x}_k|\mathrm{pa}_k)}{\int \prod_k p(\mathbf{x}_k|\mathrm{pa}_k)\,\mathrm{d}\mathbf{x}_i}$$

Markov Blanket (remaining factors)

- Parents and children of x_i
- Co-parents: parents of children of x_i

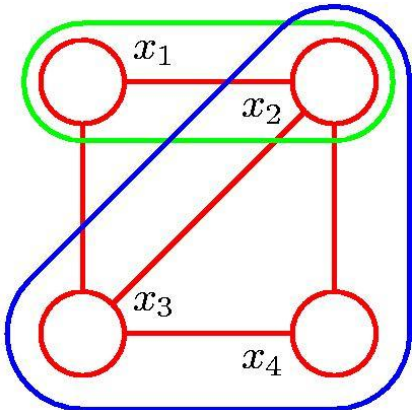
Factorization Properties

- Consider two nodes x_i and x_j that are not connected by a link then these are conditionally independent given all other nodes
- As there is no direct path between the nodes
- All other paths are blocked by nodes that are observed

$$p(x_i, x_j | \mathbf{x}_{\setminus\{i,j\}}) = p(x_i | \mathbf{x}_{\setminus\{i,j\}}) p(x_j | \mathbf{x}_{\setminus\{i,j\}})$$

Cliques and Maximal Cliques

Clique



Maximal Clique

- *Clique*: A set of fully connected nodes
- *Maximal Clique*: clique in which it is not possible to include any other nodes without it ceasing to be a clique
- Joint distribution can thus be factored it terms of maximal cliques
- Functions defined on maximal cliques includes the subsets of maximal cliques

Thus, if {x1, x2, x3} is a maximal clique and we define an arbitrary function over this clique, then including another factor defined over a subset of these variables would be redundant.

Joint Distribution

Notations: C – max. clique, $\mathbf{x}_C$ - the set of variables in that clique.

$$p(\mathbf{x}) = \frac{1}{Z} \prod_C \psi_C(\mathbf{x}_C)$$

where $\psi_C(\mathbf{x}_C)$ is the potential over clique C, which is a non-negative function which measures "compatibility" between settings of the variables.

$$Z = \sum_{\mathbf{x}} \prod_C \psi_C(\mathbf{x}_C)$$

is the partition function (used for normalization).

note: M K-state variables $\rightarrow$ K^M terms in Z.

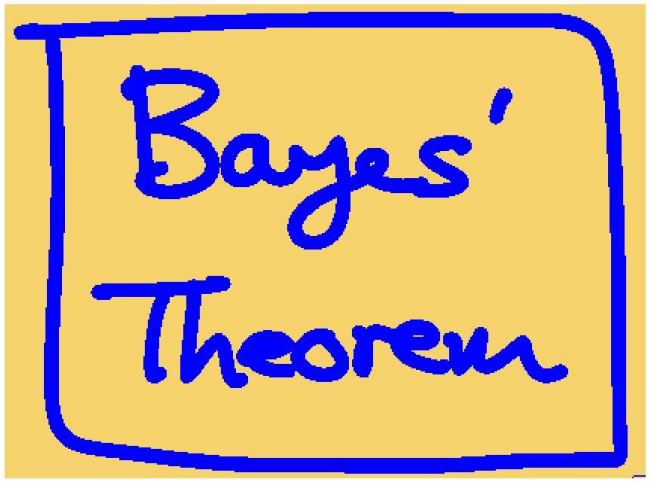
Hammersley and Clifford Theorem

- $\mathcal{UI}$: the set of distributions that are consistent with the set of conditional independence statements read from the graph using graph separation.
- $\mathcal{UF}$: the set of distributions that can be expressed as a factorization described with respect to the maximal cliques.
- The Hammersley-Clifford theorem states that the sets $\mathcal{UI}$ and $\mathcal{UF}$ are identical if $\psi_C(\mathbf{x}_C)$ is strictly positive.
- In such case

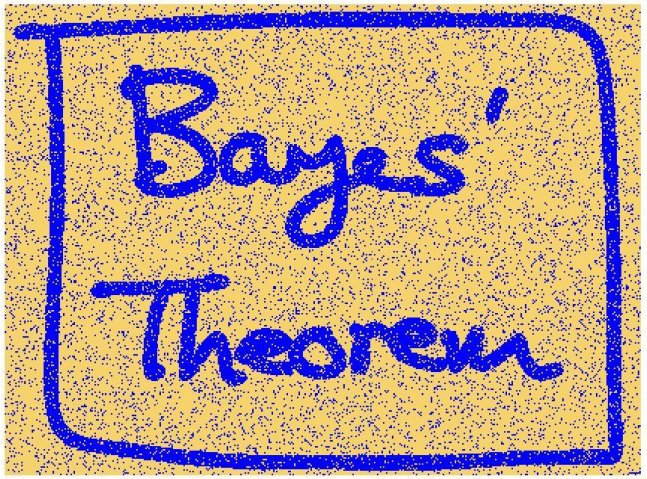
$$\psi_C(\mathbf{x}_C) = \exp\{-E(\mathbf{x}_C)\}$$

where $E(\mathbf{x}_C)$ is called an energy function, and the exponential representation is called the Boltzmann distribution.

Illustration: Image De-Noising using MRF



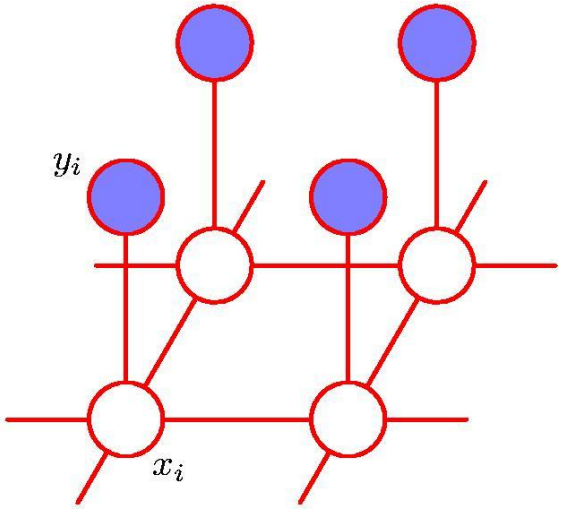
Original Image



Noisy Image

- Noisy Image: $y_i \in \{-1, +1\}$ where i runs over all the pixels
- Unknown Noise Free Image: $x_i \in \{-1, +1\}$
- Noisy image is obtained by randomly flipping the sign of pixels in the Noise free image with some small probability.
- The Goal: Given Noisy image recover Noise Free Image

Markov Model

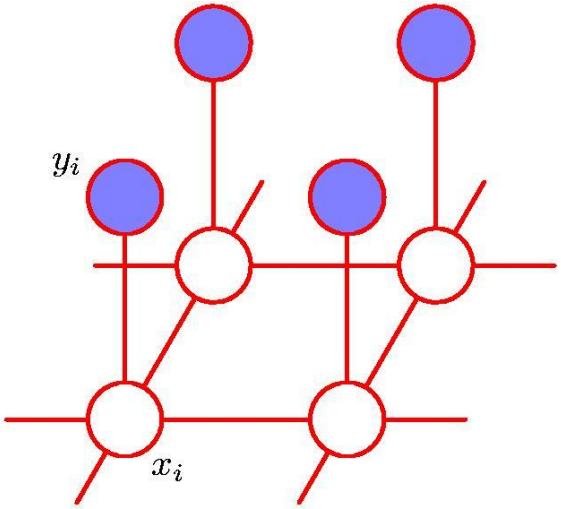


- strong correlation between x_i and y_i.
- neighbouring pixels x_i and x_j in an image are strongly correlated.

Two types of cliques

- $-\eta x_i\, y_i$: giving a lower energy when x_i and y_i have the same sign and a higher energy when they have the opposite sign
- $-\beta x_i\, x_j$: the energy is lower when the neighboring pixels have the same sign than when they have the opposite sign

Markov Model



biasing the model towards pixel values that have one particular sign in preference to the other.

$$E(\mathbf{x}, \mathbf{y}) = h \sum_i x_i - \beta \sum_{\{i,j\}} x_i x_j - \eta \sum_i x_i y_i$$

The joint distribution

$$p(\mathbf{x}, \mathbf{y}) = \frac{1}{Z} \exp\{-E(\mathbf{x}, \mathbf{y})\}$$

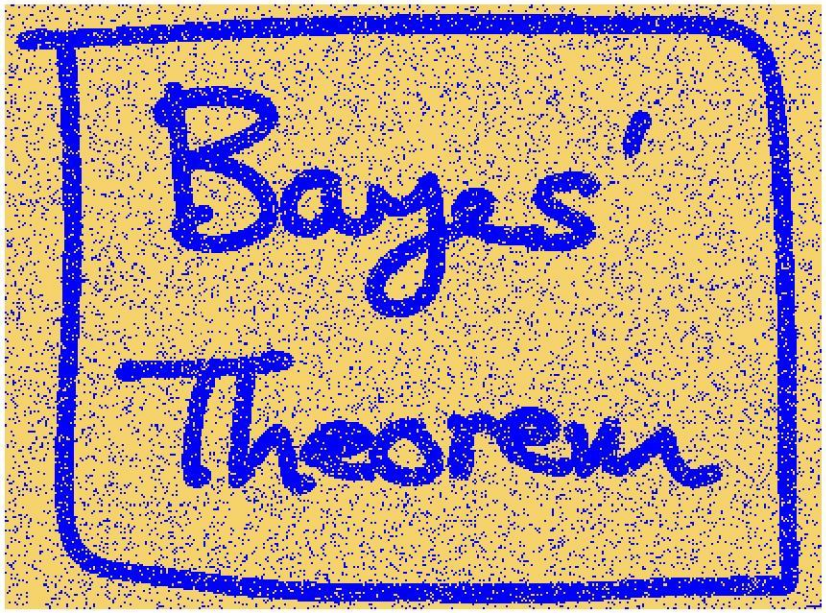
Fixing **y** as observed values implicitly defines $p(\mathbf{x}|\,\mathbf{y})$

To obtain the image **x** having a high probability

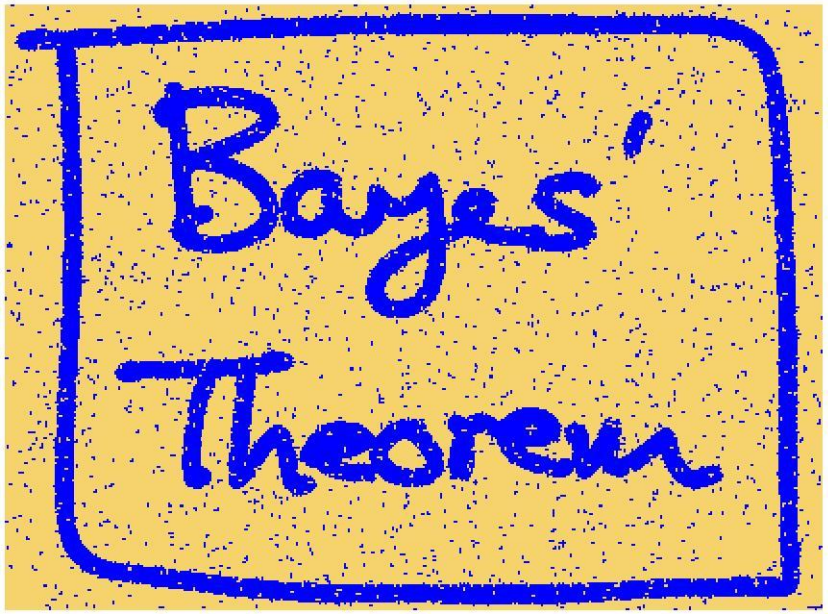
ICM - iterated conditional modes

- Initialize the variables $x_i = y_i$ for all i
- For x_j evaluate the total energy for the two possible states $x_j = +1$ and $x_j = -1$ with other node variables fixed
- set x_j to whichever state has the lower energy
- Repeat the update for another site, and so on, until some suitable stopping criterion is satisfied

Illustration: Image De-Noising (3)

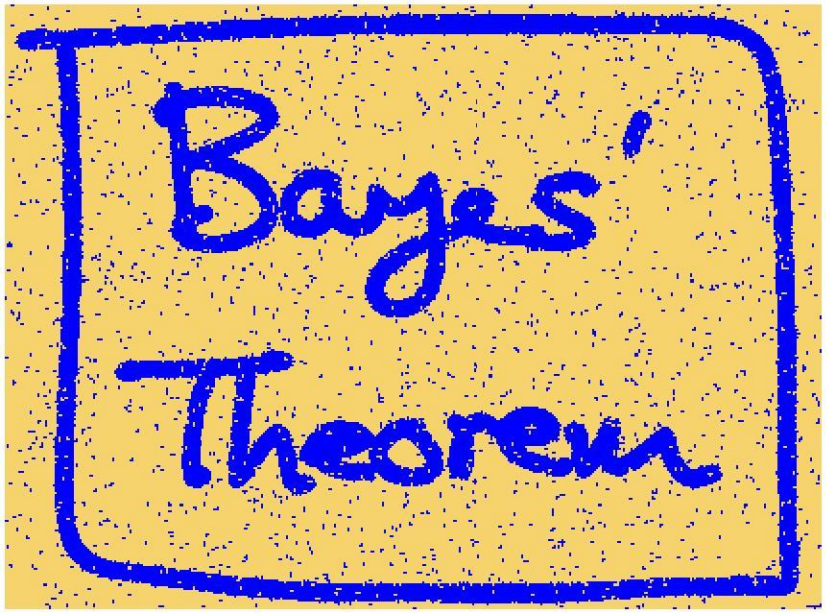


Noisy Image

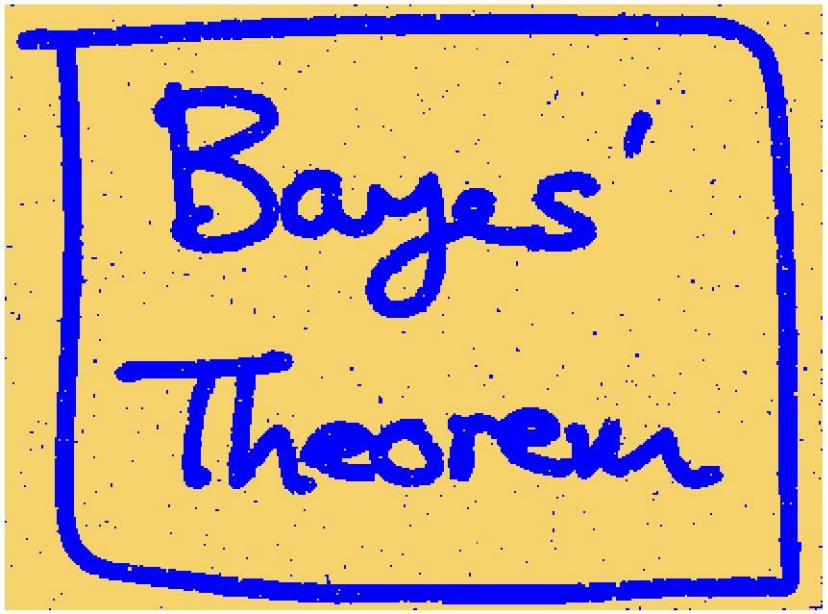


Restored Image (ICM)

Illustration: Image De-Noising (4)



Restored Image (ICM)



Restored Image (Graph cuts)