Markov Random Fields (MRFs)

So far:

1) pdfs: $p(x|y,z)$ (but not multivariate $p(\underline{x}|y,z)$)

2) directed graphical models (Bayes Nets)

Are there other ways to represent complex distributions? (which is to say knowledge)

Bayes Nets (or DAGs) are one way to factor (and therefore simplify) a complex joint distribution.

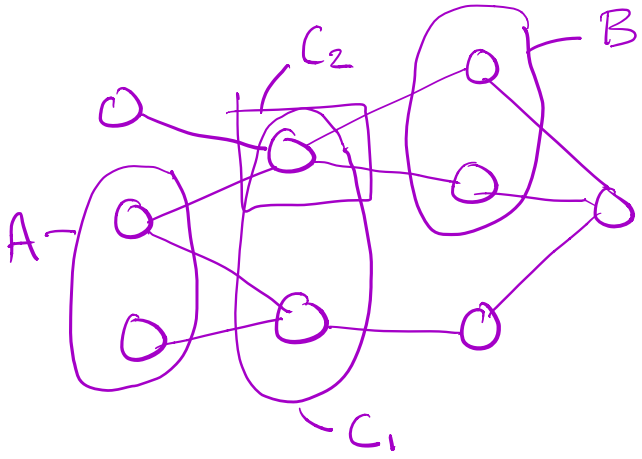
$$p(x_{1:N}) = \prod_{i=1}^{N} p(x_i | pa(x_i))$$

$$p(x_i | pa(x_i)) = p(x_i) \text{ if } pa(x_i) = \emptyset$$

Are there other ways? Yes.

Undirected graphical models.

Consider:



Note: nodes still represent variables, but now there is no implied causality, i.e. no arrows.

What are the independence properties?

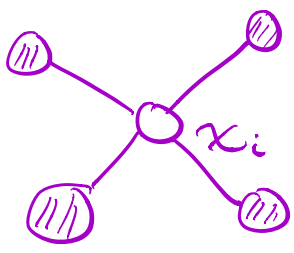
$$A \perp\!\!\!\perp B \mid C ?$$

Much simpler than DAGs:
If all paths from A to B pass though C_1, then C "blocks" A and B and $A \perp\!\!\!\perp B \mid C$.

In the graph above:

$$A \perp\!\!\!\perp B \mid C_1 \quad \text{but} \quad A \not\perp\!\!\!\perp B \mid C_2$$

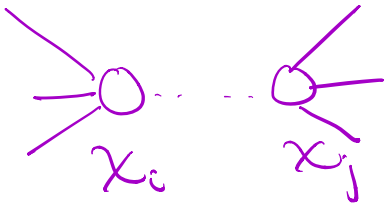
There is no "explaining away" phenomenon.
Markov blanket is simpler too:



x_i is conditionally indep. of all other nodes given only its neighbors.

What factorization model to use?

What does a link mean? (Or the absense of one?)



If there is no connection between nodes x_i and x_j then

$x_i \perp\!\!\!\perp x_j \mid$ rest of graph.

Why? All other paths are blocked.

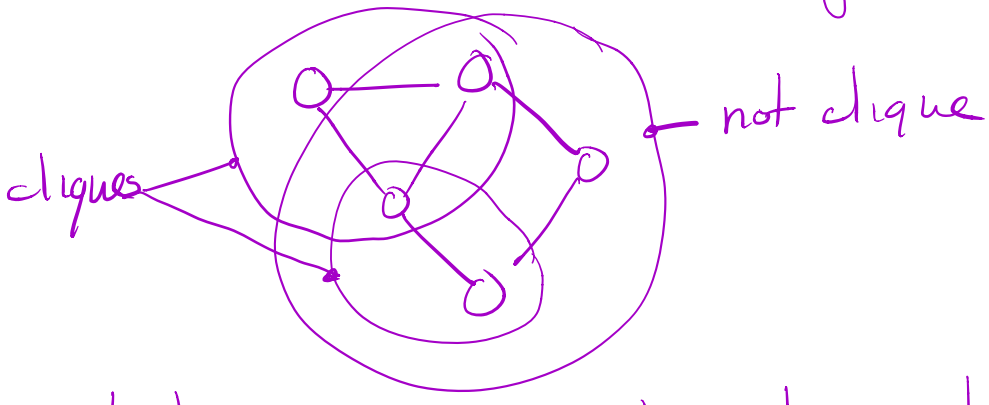
$$p(x_i, x_j \mid x_k) = p(x_i \mid x_k)\, p(x_j \mid x_k) \quad k \neq i, j$$

Defining the joint probability.

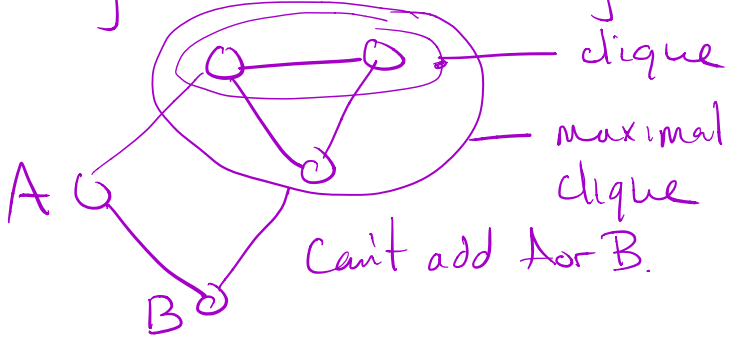
Depends on the graph cliques.

Cliques of a graph: a subset of nodes s.t. $\exists$ a link between all pairs in the subset.

The set of nodes in a clique is fully connected.



Maximal clique: a clique where it is not possible to add any more nodes w/o breaking the clique.



MRFs define joint probability in terms a maximal cliques.

Let C denote a clique. ψ = "potential function"

$\underline{x}_c$ = set of vars in C.

$$p(\underline{x}_{1:N}) = \frac{1}{Z} \prod_c \psi_c(\underline{x}_c), \quad \psi_c(\underline{x}_c) > 0$$

This factorizes the joint pdf into smaller pdfs defined on cliques.

Z = normalization const. to make $p(\underline{x}_{1:N})$ a valid, i.e. normalized, pdf.

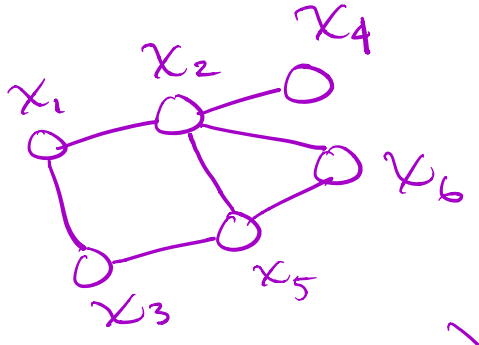
Also called a "partition" function. (from physics)

$$Z = \sum_{\underline{x}} \prod_c \psi_c(\underline{x}_c) \quad \text{(discrete } \underline{x}\text{)}$$

$$= \int_{\underline{x}_c} \prod_c \psi_c(\underline{x}_c) \quad \text{(continuous } \underline{x}\text{)}$$

can be comp. expensive

Example:



Let $\psi_{i,j,\ldots} = \psi(x_i, x_j, \ldots)$ then

$$p(\underline{x}) = \psi_{1,2}\psi_{1,3}\psi_{2,4}\psi_{2,5}\psi_{2,5,6}\psi_{3,5}$$

Note that for local cond. prob.

$$p(\underline{x}_a | \underline{x}_b) = \frac{\cancel{\frac{1}{Z}} p(\underline{x}_a, \underline{x}_b)}{\cancel{\frac{1}{Z}} p(\underline{x}_b)}$$

The Z's cancel, so don't need to compute it.

Hammersly-Clifford Theorem

Relates factorization to conditional independence.

#1) An undirected GM G is an MRF if two nodes are conditionally indep. whenever they are separated by evidence nodes:

$$p(x_i | x_{G \setminus i}) = p(x_i | x_{N_i})$$

$G \setminus i$ = all nodes in graph G except i

N_i = neighboring nodes of x_i

#2) A pdf $p(x)$ on an undirected GM G is a Gibbs distribution if it can be factored into positive functions defined on cliques that cover all nodes and edges of G:

$$p(x) = \frac{1}{Z} \prod_{c \in C_G} \psi_c(x_c)$$

C_G = all (maximal) cliques of G

Hammersly-Clifford theorem says def. #1 $\Leftrightarrow$ def #2.

How do we define $\psi(\underline{x}_c)$?

We need $\psi_c(\underline{x}_c) > 0$, so it's convenient to write:

$$\psi_c(\underline{x}_c) = \exp\left[-E(\underline{x}_c)\right]$$

↳ "Energy function"

This is also called a Boltzman distribution (from statistical mechanics)

Total energy is the sum of the energies of each clique.

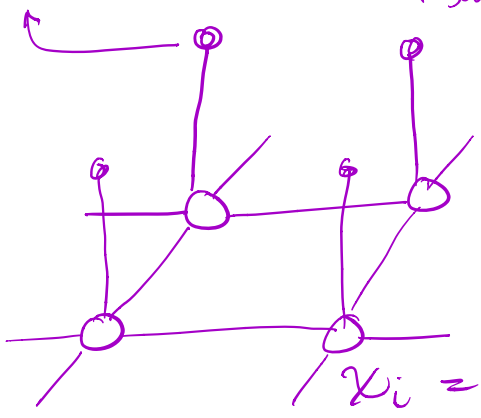
Potential functions do not have a specific probabilistic interpretation (although they can since $p(\underline{x}_c) > 0$)

Example: Image denoising with binary pixels

[This is more about MRFs than it is about image denoising — we'll learn better ways later.]

$y_i \in [-1, +1]$ = noisy pixels (observed)

Assume 10% pixels are flipped.



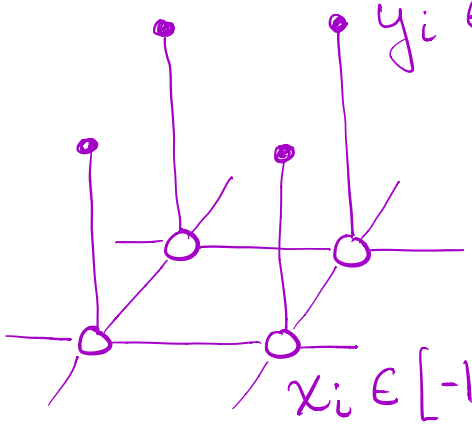
x_i = true pixel value (hidden)

How do we express knowledge of visual structure in the model?

Idea: pixels are correlated. Noise is not.

⇒ links specify degree of correlation.

How do we specify the energy function?



$y_i \in [-1,+1]$ = noisy pixels (observed)

$x_i \in [-1,+1]$ = true pixel value (hidden)

The energy fn E should describe image structure:

low energy = good = values are consistent with image structure

Most basic assumption: pixels are correlated

$\Rightarrow x_i \& x_j$ should be correlated

Also $x_i \& y_i$ should be correlated.
(or the same if there is no noise)

Can use $-\beta x_i x_j$ and $-\eta x_i y_i$

same $\Rightarrow$ lower energy
different $\Rightarrow$ higher energy

β, η const. > 0

Also need to model tendency of pixels to be on or off:

$h x_i$ $\quad h = \text{const} > 0$ $\quad$ "bias" $\quad h = 0 \Rightarrow -1$ and $+1$ have equal prob.

We will change x_i to minimize energy and denoise the image

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

$$p(\underline{x}, \underline{y}) = \frac{1}{Z} \exp\left[-E(\underline{x}, \underline{y})\right]$$

We are given y_i (observed pixels) so $p(\underline{x} | \underline{y})$ is defined implicitly.

How do we solve for the optimal x_i?

That is, $\underline{x} = \arg\max_{\underline{x}} p(\underline{x}, \underline{y})$.

Idea: use Monte Carlo (or Markov Chain Monte Carlo = MCMC)

Start with an initial solution $\underline{x}^0$.

Heratively (and randomly) change x_i to minimize $E(\underline{x}, \underline{y})$.

This is coordinate-wise gradient descent (or ascent on $p(\underline{x}, \underline{y})$

1) $x_i = y_i \ \forall i$ this is the initial soln. $\underline{x}^0$

2) for each x_i (or at random)

calc $E | x_i = +1$ all other vars are fixed.

$E | x_i = -1$

3) change x_i to state with lower energy.
(This can be computed efficiently.)

4) repeat until converged. (to local minimum of E)

→ Could also accept changes probabilistically, i.e. accept some increases, but decrease this probability over time. This is simulated annealing. Can converge to global min.