Computer vision: models, learning and inference

Chapter 10

Graphical Models

Independence

- Two variables x_1 and x_2 are independent if their joint probability distribution factorizes as

  $Pr(x_1, x_2)=Pr(x_1)\ Pr(x_2)$

Conditional independence

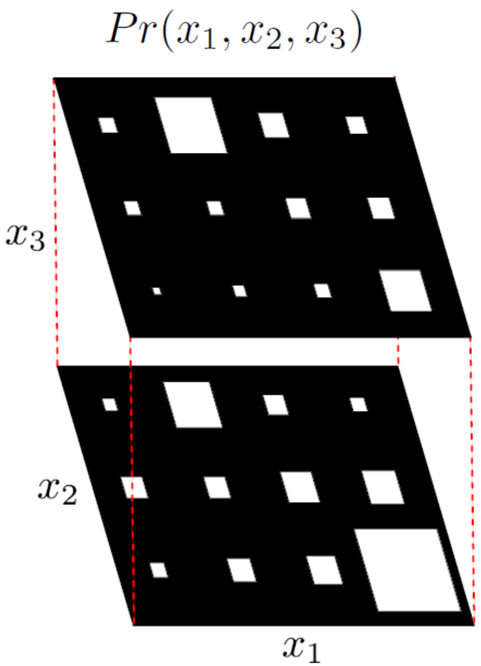
- The variable x_1 is said to be conditionally independent of x_3 given x_2 when x_1 and x_3 are independent for fixed x_2.

$$\begin{aligned} Pr(x_1|x_2,x_3) &= Pr(x_1|x_2) \\ Pr(x_3|x_1,x_2) &= Pr(x_3|x_2) \end{aligned}$$

- When this is true the joint density factorizes in a certain way and is hence redundant.

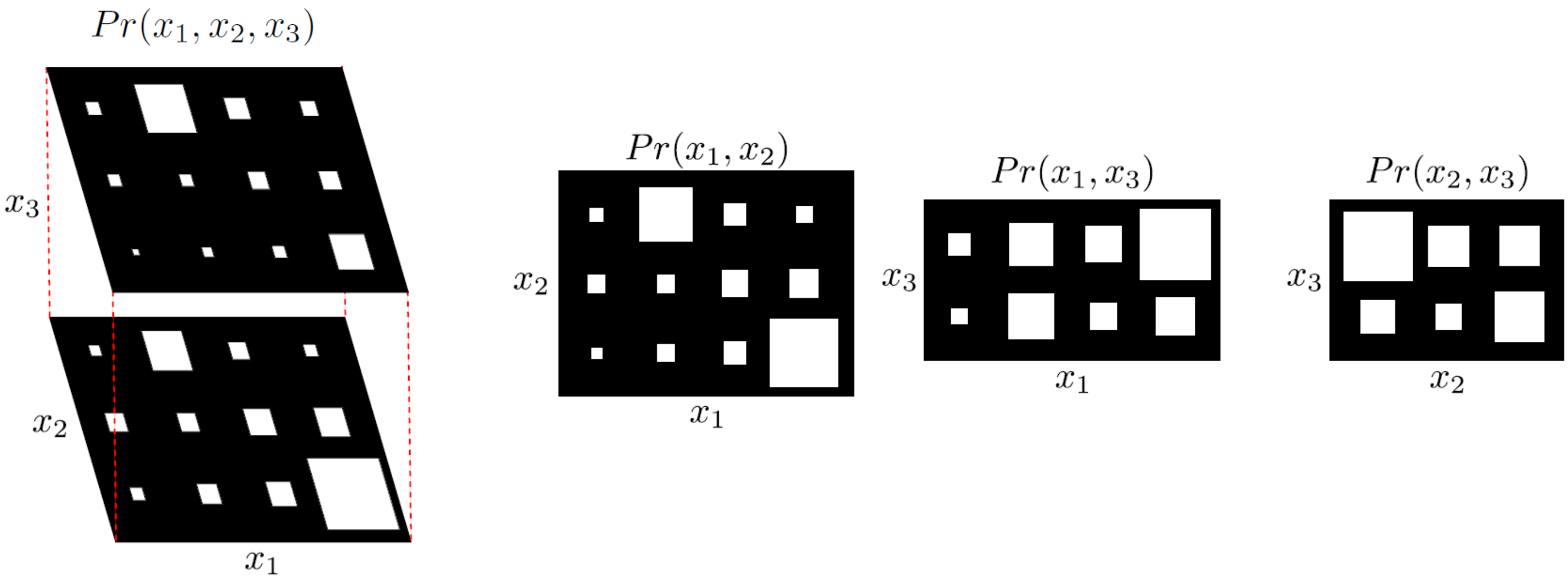
$$\begin{aligned} Pr(x_1,x_2,x_3) &= Pr(x_3|x_2,x_1)Pr(x_2|x_1)Pr(x_1) \\ &= Pr(x_3|x_2)Pr(x_2|x_1)Pr(x_1). \end{aligned}$$

Conditional independence



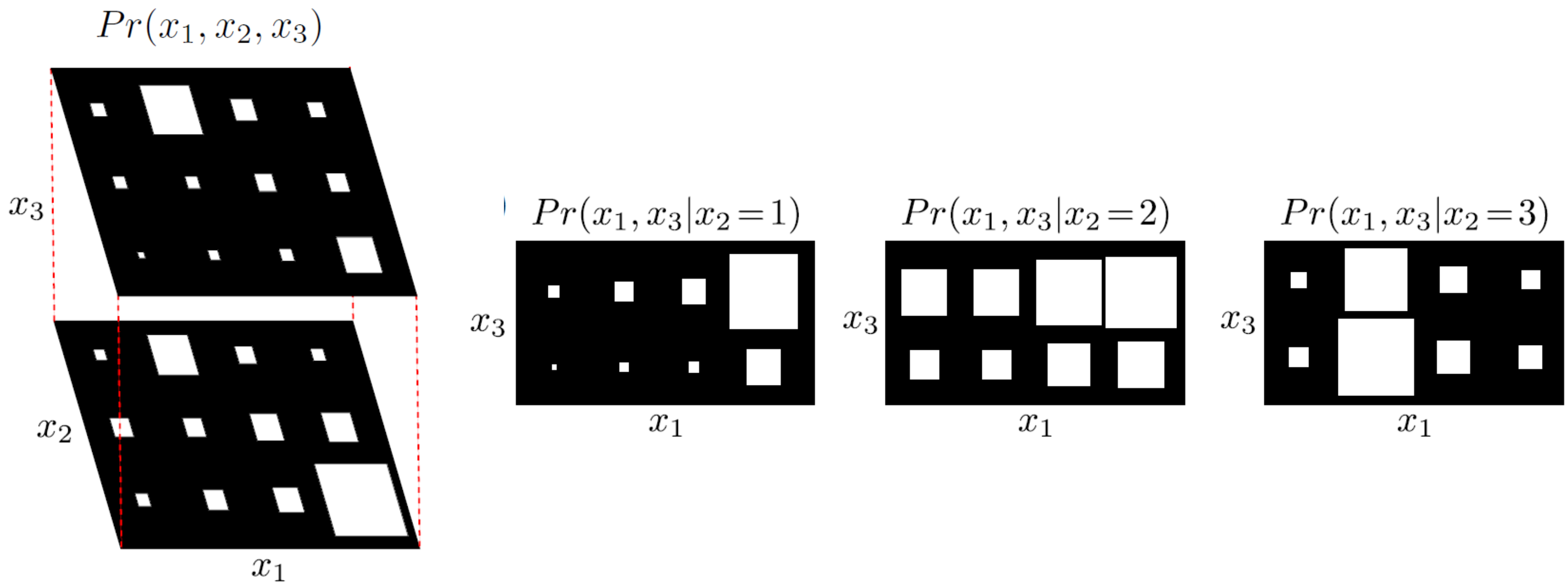
- Consider joint pdf of three discrete variables x_1, x_2, x_3

Conditional independence



- Consider joint pdf of three discrete variables x_1, x_2, x_3
  - The three marginal distributions show that no pair of variables is independent

Conditional independence



- Consider joint pdf of three discrete variables x_1, x_2, x_3
  - The three marginal distributions show that no pair of variables is independent
  - But x_1 is independent of x_2 given x_3

Graphical models

- A graphical model is a graph based representation that makes both factorization and conditional independence relations easy to establish

- Two important types:
  - Directed graphical model or Bayesian network
  - Undirected graphical model or Markov network

Directed graphical models

- Directed graphical model represents probability distribution that factorizes as a product of conditional probability distributions

$$Pr(x_{1...N}) = \prod_{n=1}^{N} Pr(x_n | x_{\mathrm{pa}[n]})$$

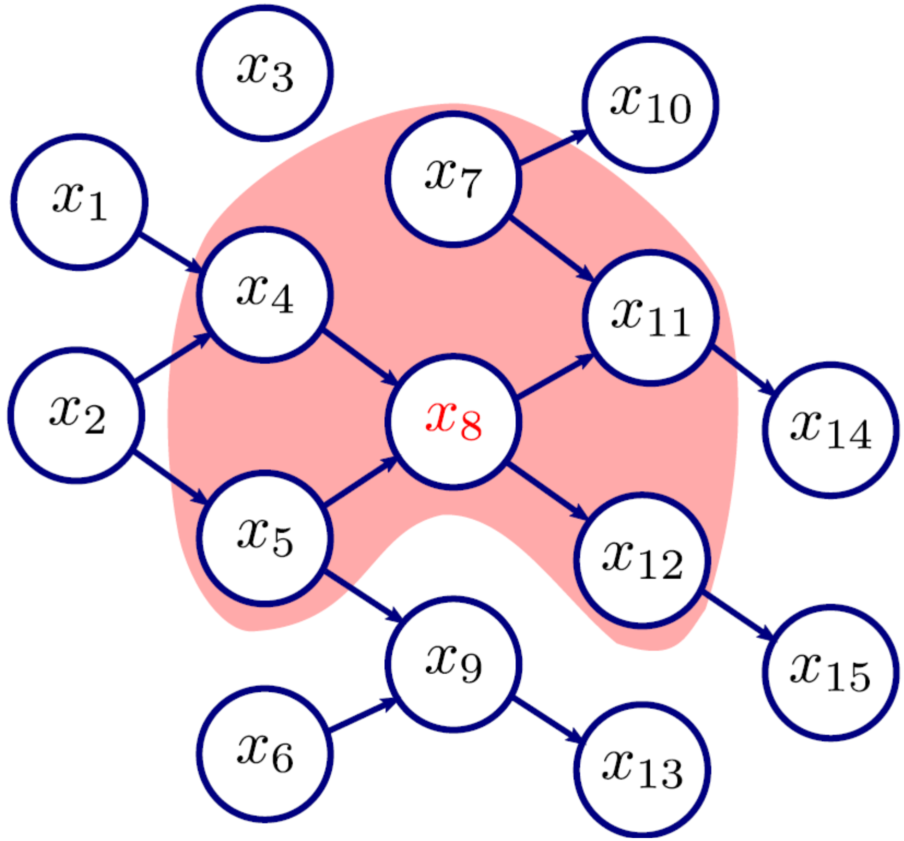
where pa[n] denotes the parents of node n

Directed graphical models

- To visualize graphical model from factorization
  - add one node per random variable and draw arrow to each variable from each of its parents.

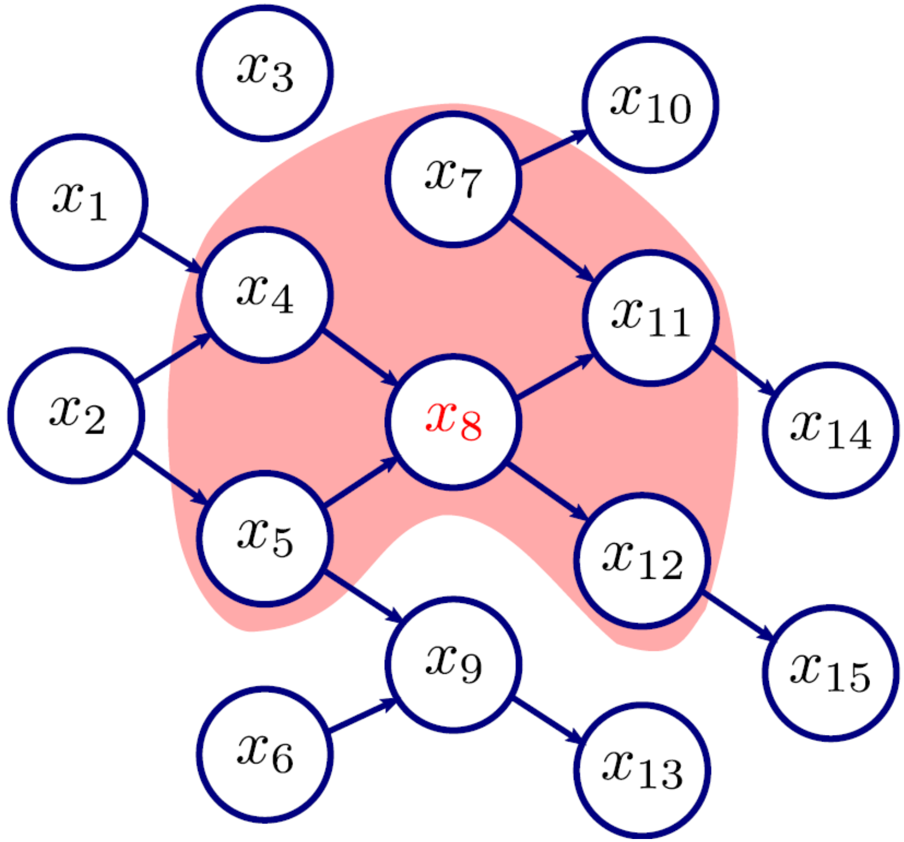
- To extract factorization from graphical model
  - Add one term per node in the graph $Pr(x_n | x_{pa[n]})$
  - If no parents then just add $Pr(x_n)$

Example 1



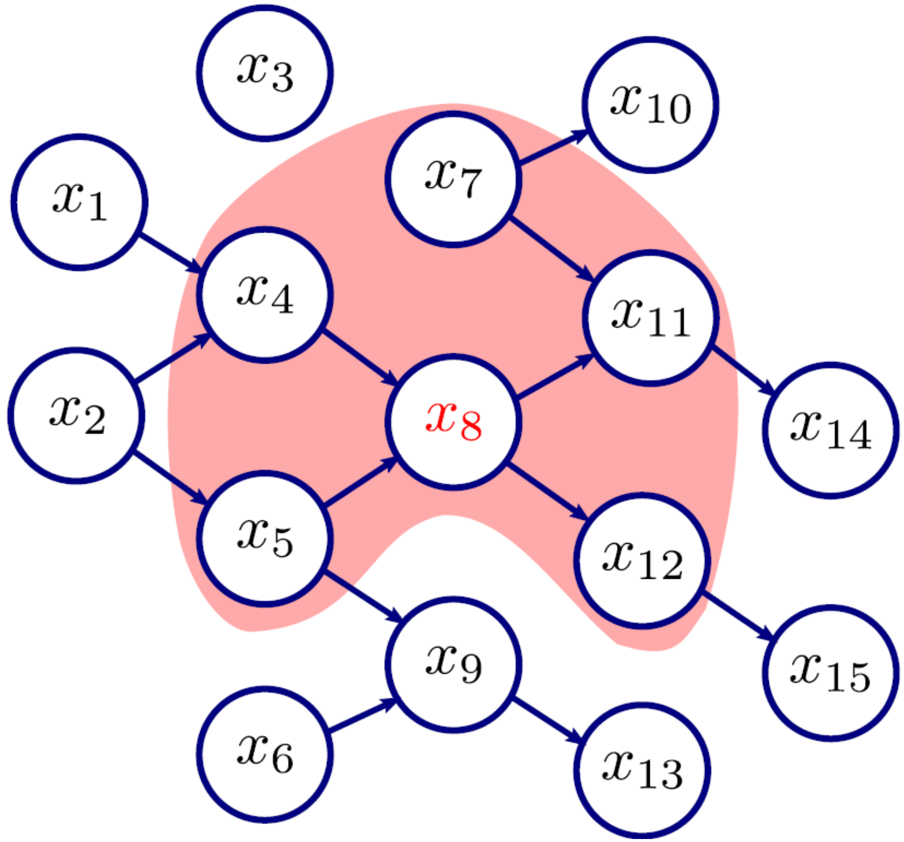
$$
\begin{aligned}
Pr(x_1 \ldots x_{15}) \;=\; & Pr(x_1)Pr(x_2)Pr(x_3)Pr(x_4|x_1,x_2)Pr(x_5|x_2)Pr(x_6) \\
& Pr(x_7)Pr(x_8|x_4,x_5)Pr(x_9|x_5,x_6)Pr(x_{10}|x_7)Pr(x_{11}|x_7,x_8) \\
& Pr(x_{12}|x_8)Pr(x_{13}|x_9)Pr(x_{14}|x_{11})Pr(x_{15}|x_{12}).
\end{aligned}
$$

Example 1



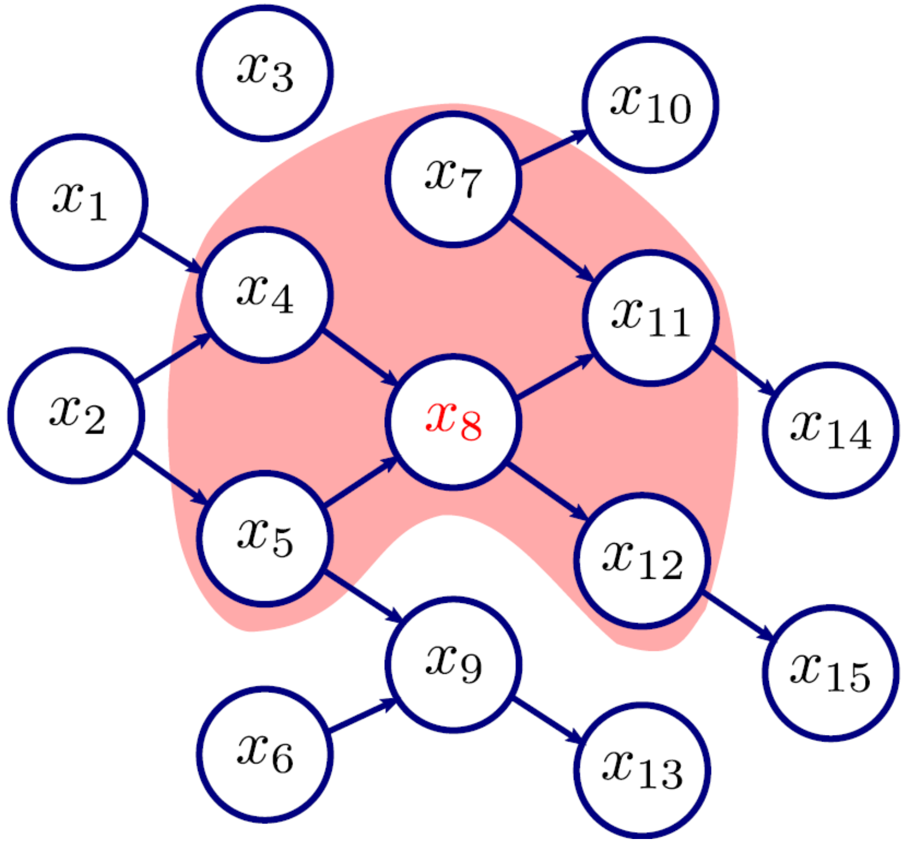
= Markov Blanket of variable x_8 – Parents, children and parents of children

Example 1



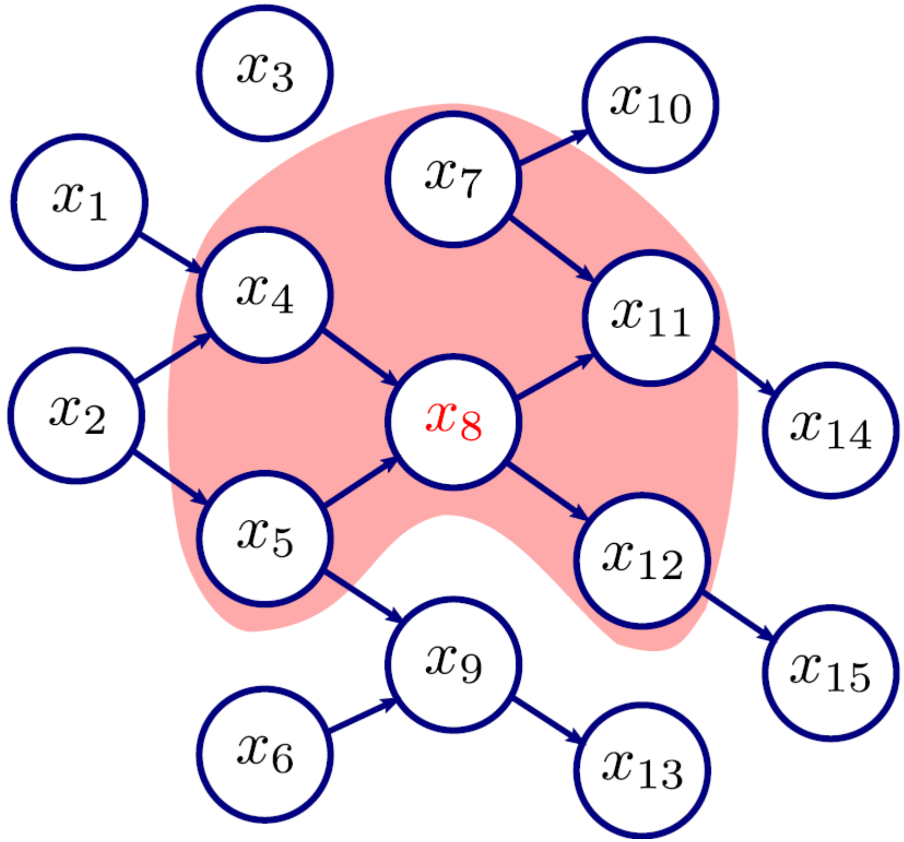
If there is no route between two variables and they share no ancestors, they are independent.

Example 1



A variable is conditionally independent of all others, given its Markov Blanket

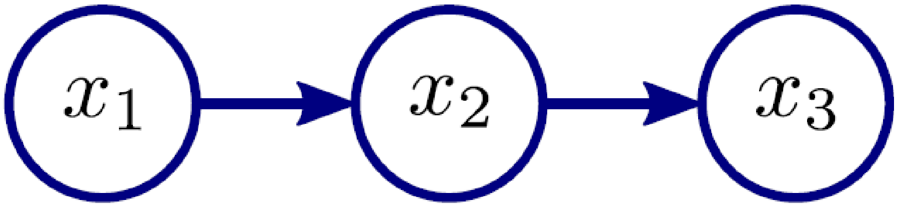
Example 1



General rule:

The variables in set $\mathcal{A}$ are conditionally independent of those in set $\mathcal{B}$ given set $\mathcal{C}$ if all routes from $\mathcal{A}$ to $\mathcal{B}$ are blocked. A route is blocked at a node if (i) this node is in $\mathcal{C}$ and the arrows meet head to tail or tail to tail or (ii) neither this node nor any of its descendants are in $\mathcal{C}$ and the arrows meet head to head.

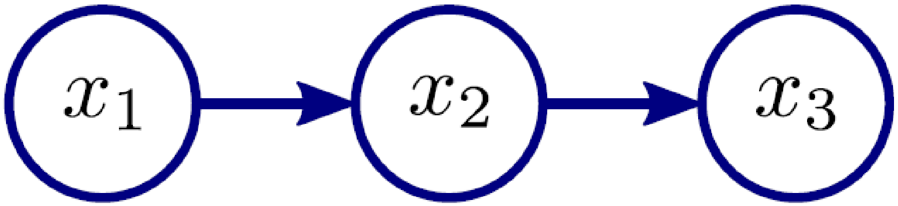
Example 2



The joint pdf of this graphical model factorizes as:

$$Pr(x_1, x_2, x_3) = Pr(x_1)Pr(x_2|x_1)Pr(x_3|x_2)$$

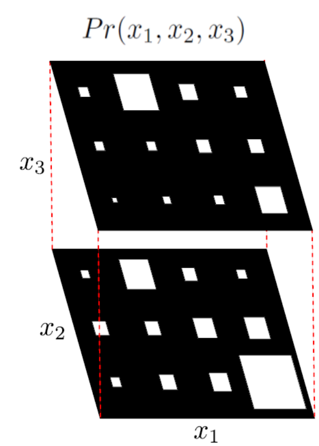
Example 2



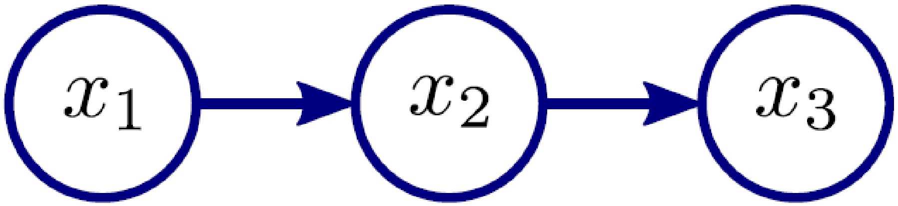
The joint pdf of this graphical model factorizes as:

$$Pr(x_1, x_2, x_3) = Pr(x_1)Pr(x_2|x_1)Pr(x_3|x_2)$$

It describes the original example:



Example 2

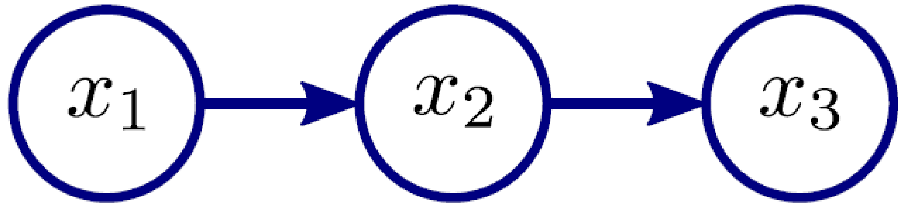


Here the arrows meet head to tail at x_2, and so x_1 is conditionally independent of x_3 given x_2.

General rule:

The variables in set $\mathcal{A}$ are conditionally independent of those in set $\mathcal{B}$ given set $\mathcal{C}$ if all routes from $\mathcal{A}$ to $\mathcal{B}$ are blocked. A route is blocked at a node if (i) this node is in $\mathcal{C}$ and the arrows meet head to tail or tail to tail or (ii) neither this node nor any of its descendants are in $\mathcal{C}$ and the arrows meet head to head.

Example 2

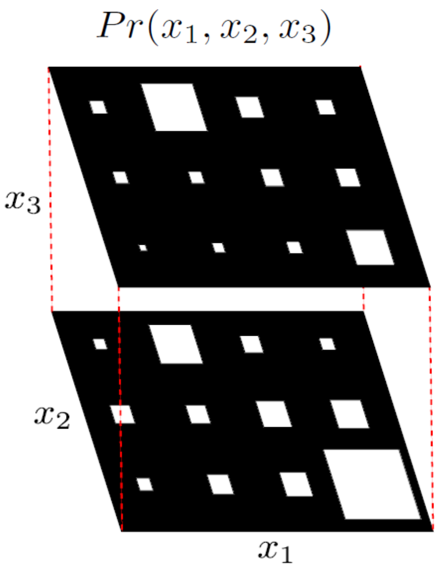


Algebraic proof:

$$
\begin{aligned}
Pr(x_1|x_2,x_3) &= \frac{Pr(x_1,x_2,x_3)}{Pr(x_2,x_3)} \\
&= \frac{Pr(x_1)Pr(x_2|x_1)Pr(x_3|x_2)}{\int Pr(x_1)Pr(x_2|x_1)Pr(x_3|x_2)dx_1} \\
&= \frac{Pr(x_1)Pr(x_2|x_1)}{\int Pr(x_1)Pr(x_2|x_1)dx_1},
\end{aligned}
$$

No dependence on x_3 implies that x_1 is conditionally independent of x_3 given x_2.

Redundancy



Conditional independence can be thought of as redundancy in the full distribution

$$Pr(x_1, x_2, x_3) = Pr(x_1)Pr(x_2|x_1)Pr(x_3|x_2)$$

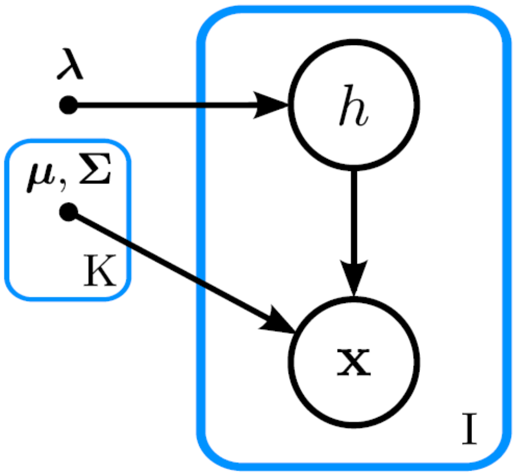
4 + 3 x 4 + 2 x 3

= 22 entries

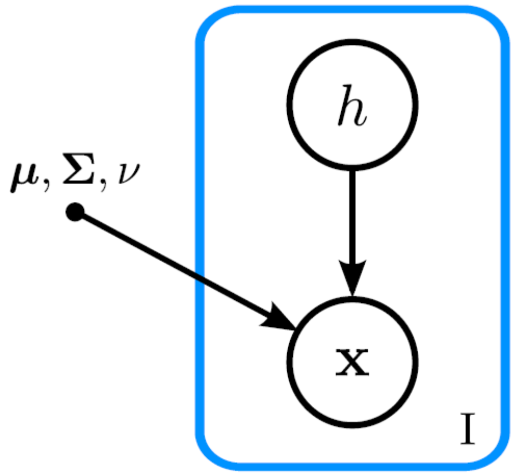
4 x 3 x 2 = 24 entries

Redundancy here only very small, but with larger models can be very significant.

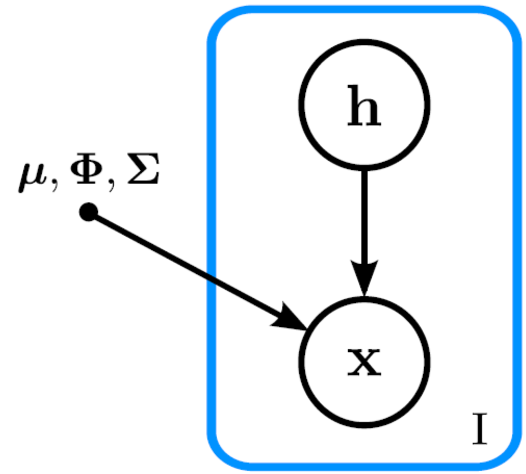
Example 3



Mixture of Gaussians



t-distribution



Factor analyzer

Blue boxes = Plates. Interpretation: repeat contents of box number of times in bottom right corner.
Bullet = variables which are not treated as uncertain

Undirected graphical models

Probability distribution factorizes as:

$$Pr(x_{1\ldots N}) = \frac{1}{Z}\prod_{c=1}^{C}\phi_c[x_{1\ldots N}]$$

Partition function (normalization constant)

Product over C functions

Potential function (returns non-negative number)

Undirected graphical models

Probability distribution factorizes as:

$$Pr(x_{1\ldots N}) = \frac{1}{Z}\prod_{c=1}^{C}\phi_c[x_{1\ldots N}]$$

Partition function

(normalization constant)

$$Z = \sum_{x_1}\sum_{x_2}\cdots\sum_{x_N}\prod_{c=1}^{C}\phi_c[x_{1\ldots N}]$$

For large systems, intractable to compute

Alternative form

$$Pr(x_{1\ldots N}) = \frac{1}{Z}\prod_{c=1}^{C}\phi_c[x_{1\ldots N}]$$

Can be written as Gibbs Distribution:

$$Pr(x_{1\ldots N}) = \frac{1}{Z}\exp\left[-\sum_{c=1}^{C}\psi_c[x_{1\ldots N}]\right]$$

where

$$\psi_c[x_{1\ldots N}] = -\log[\phi_c[x_{1\ldots N}]]$$

Cost function
(positive or negative)

Cliques

Better to write undirected model as

$$Pr(x_{1\ldots N}) = \frac{1}{Z}\prod_{c=1}^{C}\phi_c[\mathcal{S}_c]$$

Product over cliques

Clique
Subset of variables

$$\mathcal{S}_c \subset \{x_n\}_{n=1}^{N}$$

Undirected graphical models

- To visualize graphical model from factorization
  - Sketch one node per random variable
  - For every clique, sketch connection from every node to every other

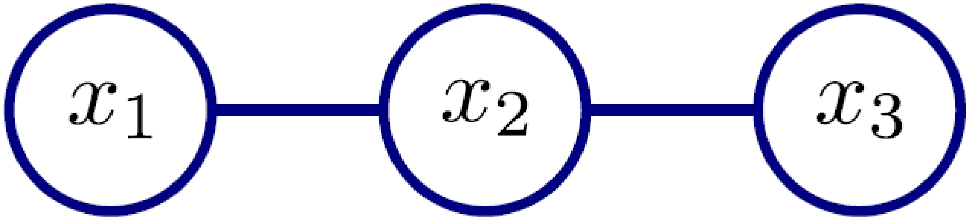
- To extract factorization from graphical model
  - Add one term to factorization per maximal clique (fully connected subset of nodes where it is not possible to add another node and remain fully connected)

Conditional independence

- Much simpler than for directed models:

One set of nodes is conditionally independent of another given a third if the third set separates them (i.e. Blocks any path from the first node to the second)

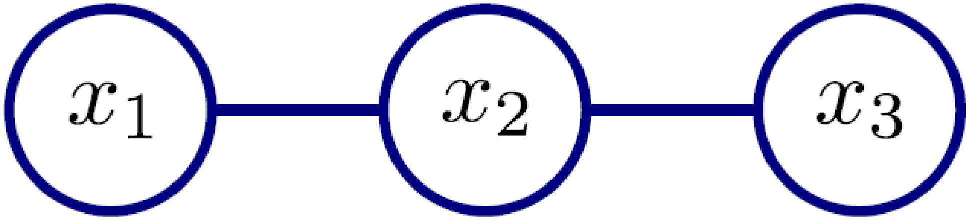
Example 1



Represents factorization:

$$Pr(x_1, x_2, x_3) = \frac{1}{Z}\phi_1[x_1, x_2]\phi_2[x_2, x_3]$$

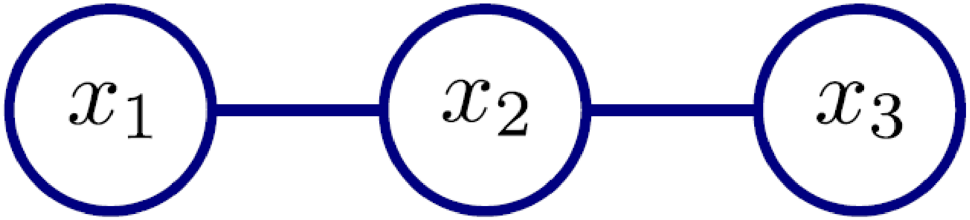
Example 1



By inspection of graphical model:

x_1 is conditionally independent of x_3 given x_2, as the route from x_1 to x_3 is blocked by x_2.

Example 1

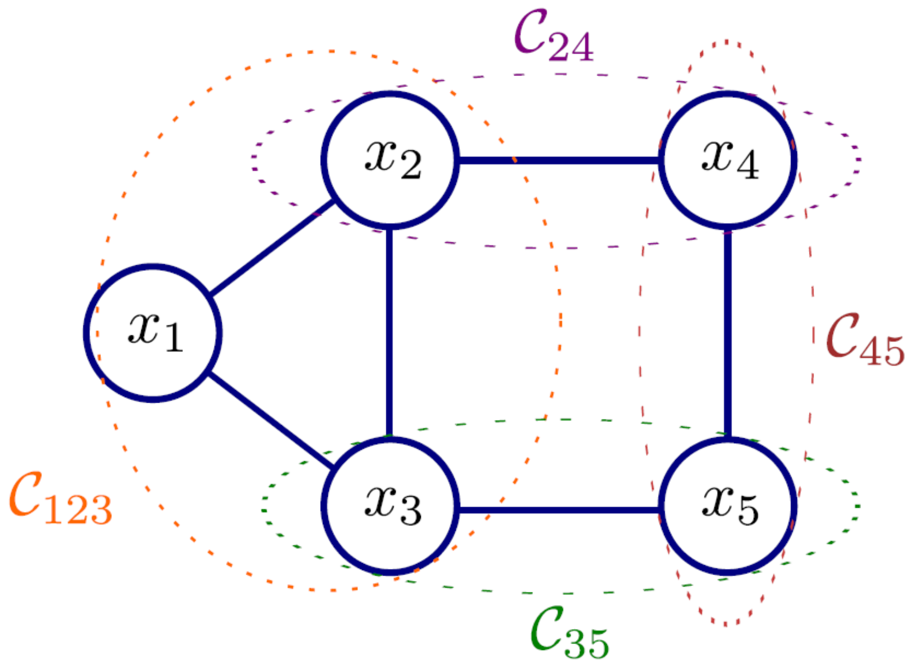


Algebraically:

$$
\begin{aligned}
Pr(x_1|x_2,x_3) &= \frac{Pr(x_1,x_2,x_3)}{Pr(x_2,x_3)} \\
&= \frac{\frac{1}{Z}\phi_1[x_1,x_2]\phi_2[x_2,x_3]}{\int \frac{1}{Z}\phi_1[x_1,x_2]\phi_2[x_2,x_3]dx_1} \\
&= \frac{\phi_1[x_1,x_2]}{\int \phi_1[x_1,x_2]dx_1}.
\end{aligned}
$$

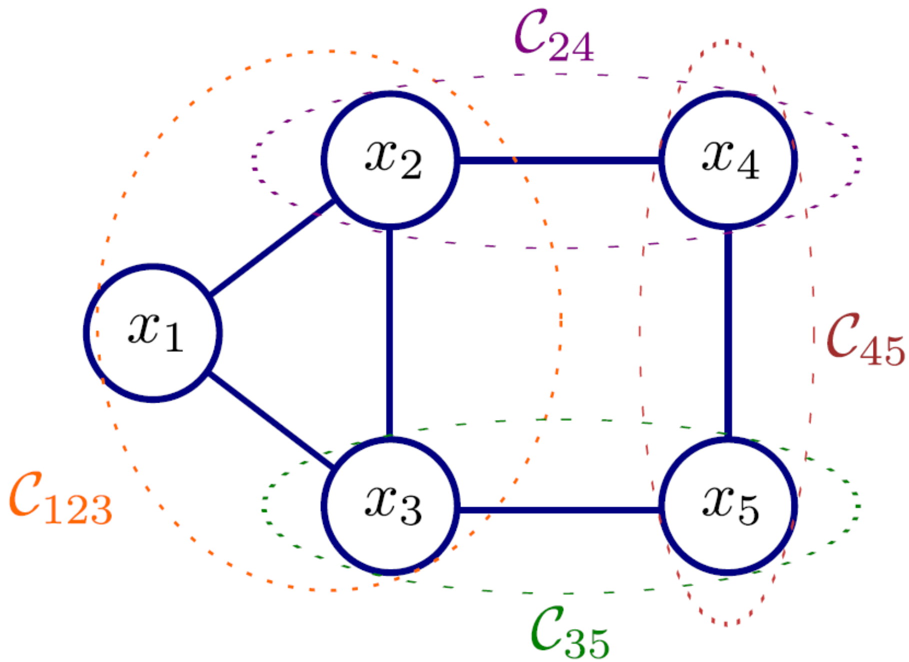
No dependence on x_3 implies that x_1 is conditionally independent of x_3 given x_2.

Example 2



- Variables x_1 and x_2 form a clique (both connected to each other)
- But not a maximal clique, as we can add x_3 and it is connected to both

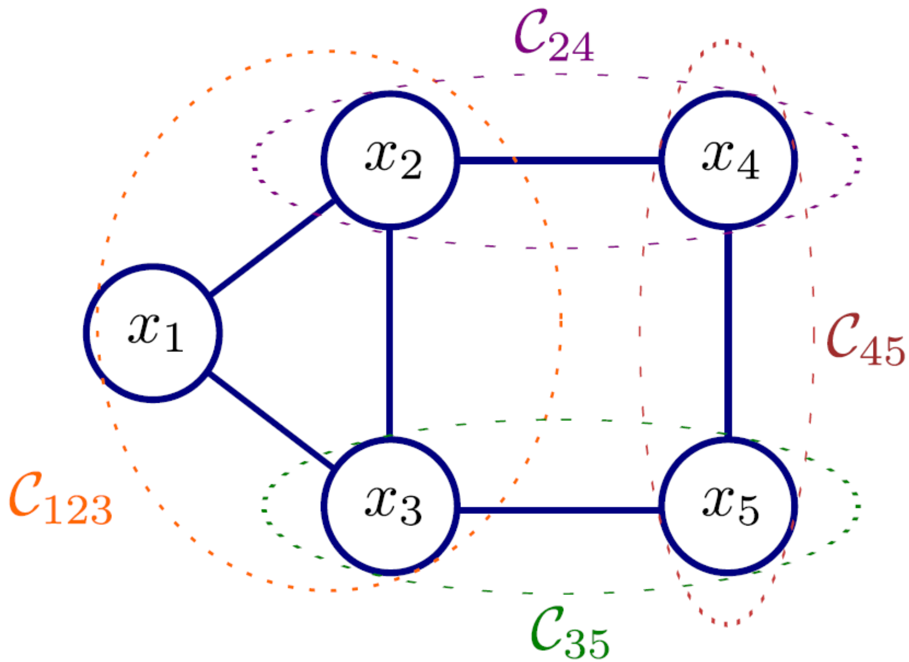
Example 2



Graphical model implies factorization:

$$Pr(x_{1\ldots5}) = \frac{1}{Z}\phi_1[x_1, x_2, x_3]\phi_2[x_2, x_4], \phi_3[x_3, x_5]\phi_4[x_4, x_5]$$

Example 2



$$Pr(x_{1\ldots5}) = \frac{1}{Z}\phi_1[x_1, x_2, x_3]\phi_2[x_2, x_4], \phi_3[x_3, x_5]\phi_4[x_4, x_5]$$

Or could be....

$$Pr(x_{1\ldots5}) = \frac{1}{Z}\left(\phi_1[x_1, x_2]\phi_2[x_2, x_3]\phi_3[x_1, x_3]\right)\phi_4[x_2, x_4], \phi_5[x_3, x_5]\phi_6[x_4, x_5]$$

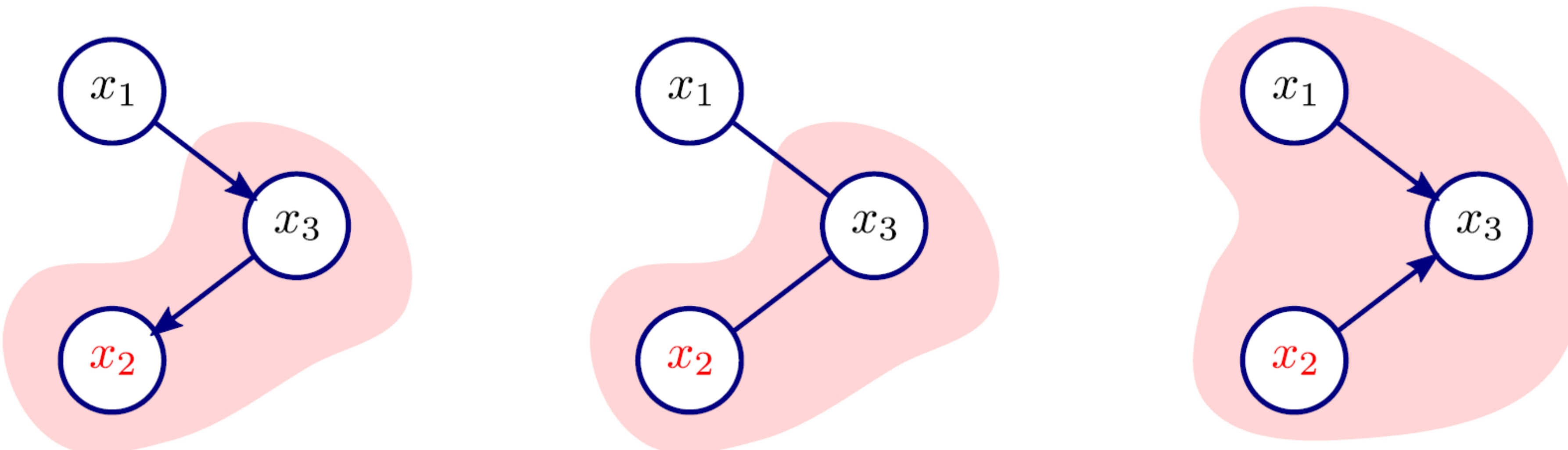
... but this is less general

Comparing directed and undirected models

Executive summary:

- Some conditional independence patterns can be represented as both directed and undirected
- Some can be represented only by directed
- Some can be represented only by undirected
- Some can be represented by neither

Comparing directed and undirected models

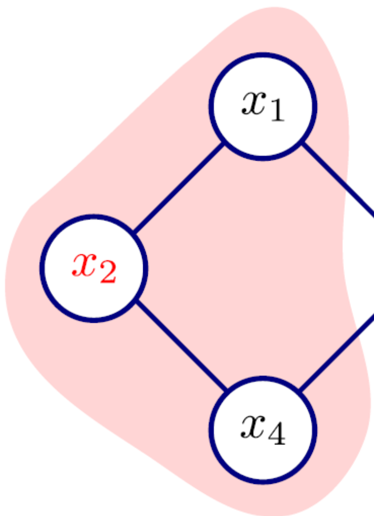


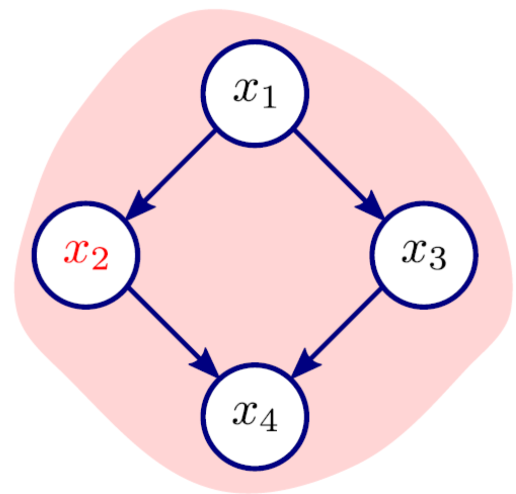
$x_2 \perp\!\!\!\perp x_1 | x_3$ $\quad$ $x_2 \perp\!\!\!\perp x_1 | x_3$ $\quad$ $x_2 \perp\!\!\!\perp x_1$

These models represent same independence / conditional independence relations

There is no undirected model that can describe these relations

Comparing directed and undirected models





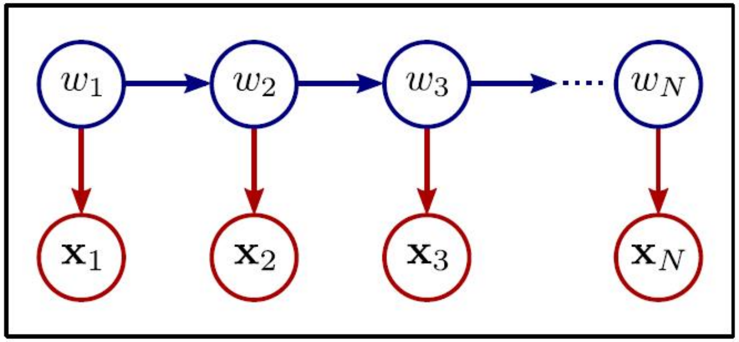
$$x_1 \perp\!\!\!\perp x_4 | x_2, x_3$$
$$x_2 \perp\!\!\!\perp x_3 | x_1, x_4$$

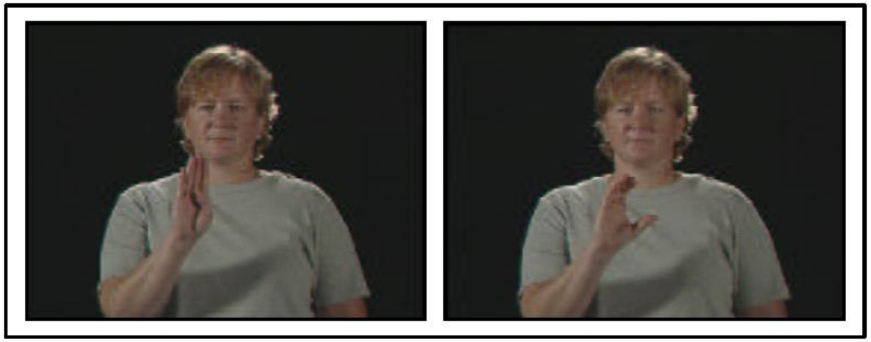
$$x_1 \perp\!\!\!\perp x_4 | x_2, x_3$$
$$x_2 \perp\!\!\!\perp x_3 | x_1$$

There is no directed model that can describe these relations

Closest example, but not the same

Graphical models in computer vision

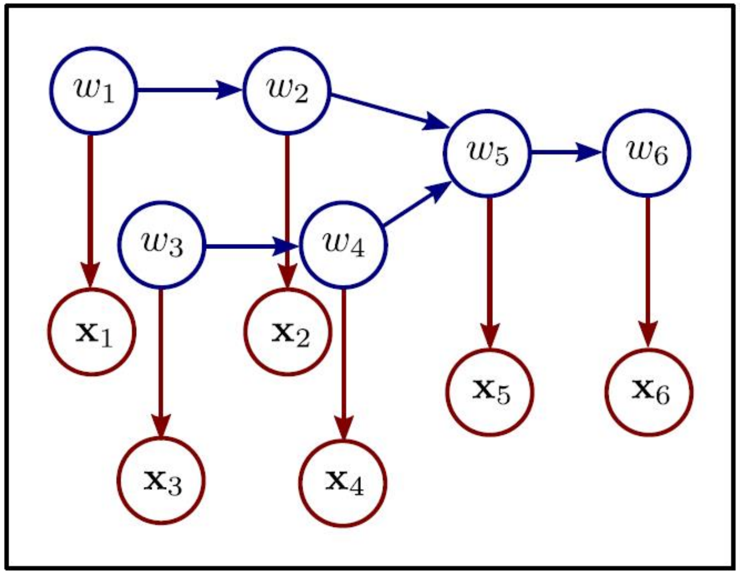




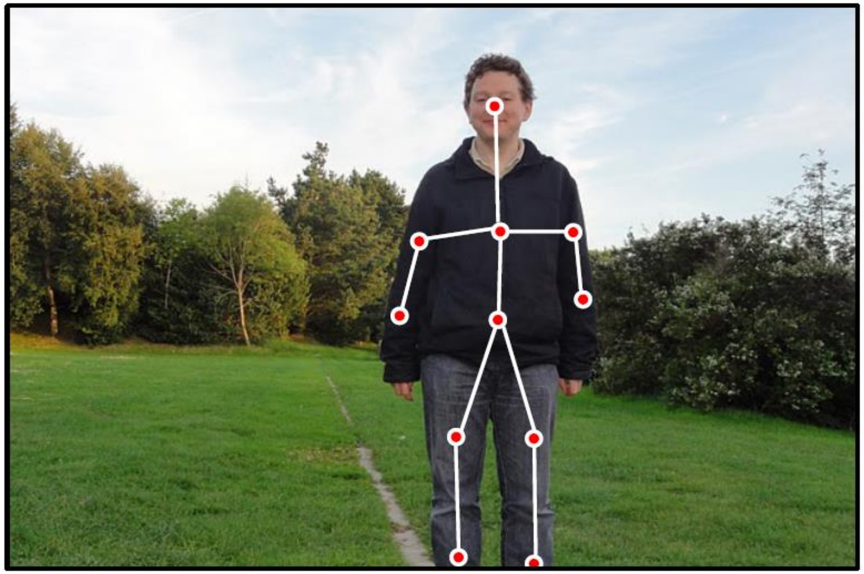
Chain model
(hidden Markov model)

Interpreting sign
language sequences

Graphical models in computer vision



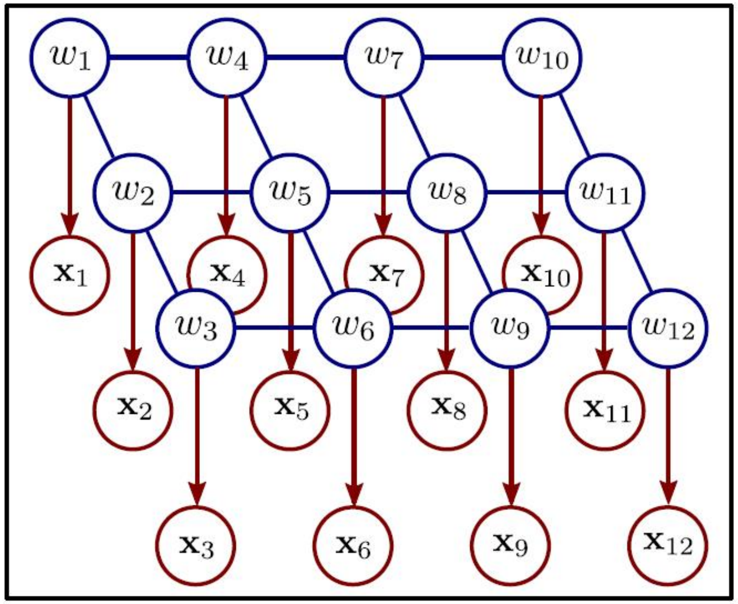
Tree model



Parsing the human body

Note direction of links, indicating that we're building a probability distribution over the data, i.e. generative models: $Pr(\mathbf{x}|\mathbf{w})$

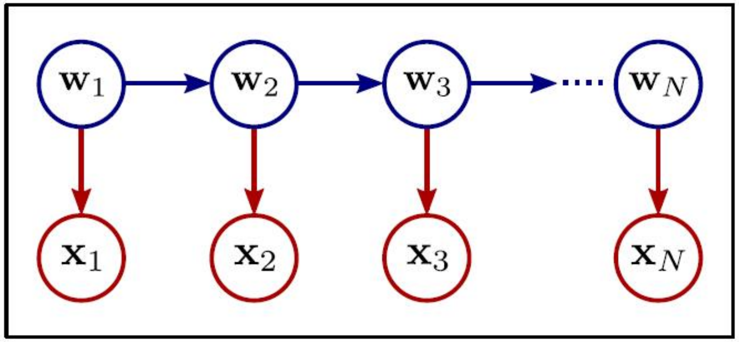
Graphical models in computer vision



Grid model
Markov random field
(blue nodes)

Semantic
segmentation

Graphical models in computer vision



Chain model
Kalman filter

Tracking contours

Inference in models with many unknowns

- Ideally we would compute full posterior distribution $Pr(w_{1\ldots N}|\mathbf{x}_{1\ldots N})$.
- But for most models this is a very large discrete distribution – intractable to compute
- Other solutions:
  - Find MAP solution
  - Find marginal posterior distributions
  - Maximum marginals
  - Sampling posterior

Finding MAP solution

$$\begin{aligned}\hat{w}_{1\ldots N} &= \underset{w_{1\ldots N}}{\operatorname{argmax}}\left[Pr(w_{1\ldots N}|\mathbf{x}_{1\ldots N})\right]\\ &= \underset{w_{1\ldots N}}{\operatorname{argmax}}\left[Pr(\mathbf{x}_{1\ldots N}|w_{1\ldots N})Pr(w_{1\ldots N})\right]\end{aligned}$$

- Still difficult to compute – must search through very large number of states to find the best one.

Marginal posterior distributions

$$Pr(w_n|\mathbf{x}_{1\ldots N}) = \int\int Pr(w_{1\ldots N}|\mathbf{x}_{1\ldots N})dw_{1\ldots n-1}dw_{n+1\ldots N}$$

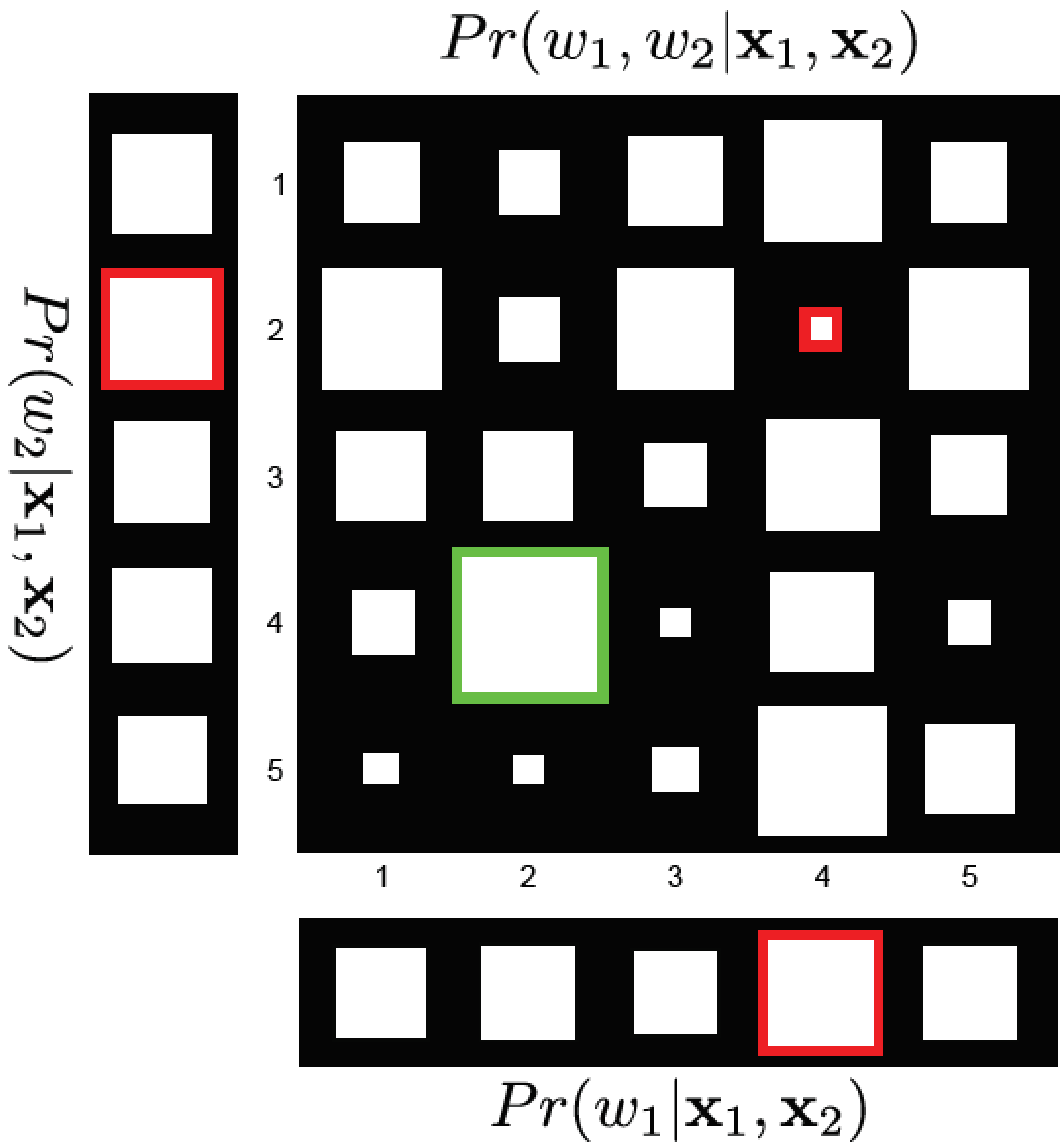
- Compute one distribution for each variable w_n.
- Obviously cannot be computed by computing full distribution and explicitly marginalizing.
- Must use algorithms that exploit conditional independence!

Maximum marginals

$$\hat{w}_n = \underset{w_n}{\operatorname{argmax}} \left[Pr(w_n | \mathbf{x}_{1\ldots N}) \right]$$

- Maximum of marginal posterior distribution for each variable w_n.
- May have probability zero; the states can be individually probable, but never co-occur.

Maximum marginals



Sampling the posterior

- Draw samples from posterior $Pr(w_{1...N}|\mathbf{x}_{1...N})$.
  - use samples as representation of distribution
  - select sample with highest prob. as point sample
  - compute empirical max-marginals
    - Look at marginal statistics of samples

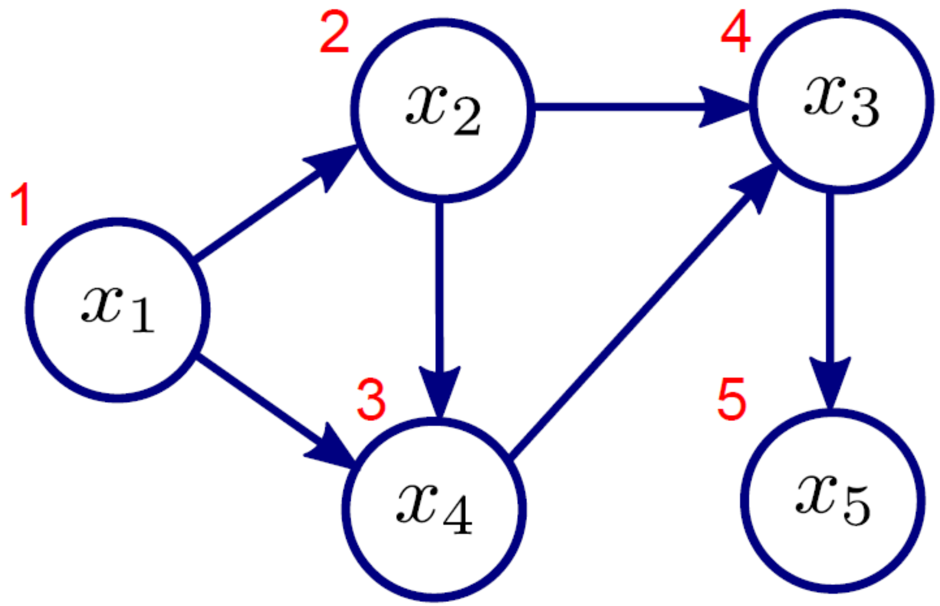
Drawing samples - directed

$$Pr(x_{1\ldots N}) = \prod_{n=1}^{I} Pr(x_n | x_{\mathrm{pa}[n]})$$

To sample from directed model, use ancestral sampling

- work through graphical model, sampling one variable at a time.
- Always sample parents before sampling variable
- Condition on previously sampled values

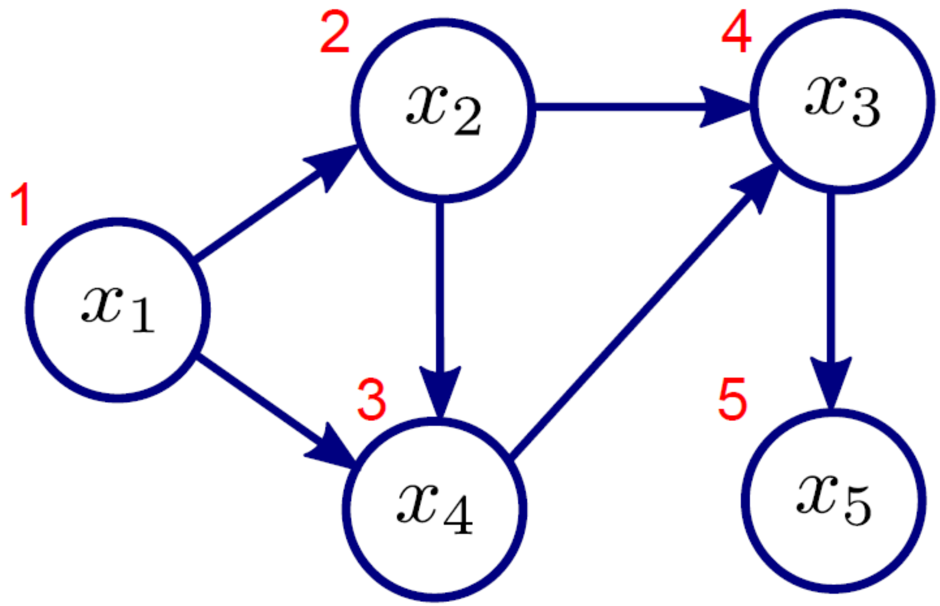
Ancestral sampling example



$$Pr(x_1, x_2, x_3, x_4, x_5) =$$

$$Pr(x_1)Pr(x_2|x_1)Pr(x_3|x_4, x_2)Pr(x_4|x_2, x_1)Pr(x_5|x_3)$$

Ancestral sampling example



To generate one sample:

1. Sample x_1^* from $Pr(x_1)$
2. Sample x_2^* from $Pr(x_2| x_1^*)$
3. Sample x_4^* from $Pr(x_4| x_1^*, x_2^*)$
4. Sample x_3^* from $Pr(x_3| x_2^*,x_4^*)$
5. Sample x_5^* from $Pr(x_5| x_3^*)$

Drawing samples - undirected

- Can't use ancestral sampling as no sense of parents / children and don't have conditional probability distributions
- Instead us Markov chain Monte Carlo method
  - Generate series of samples (chain)
  - Each depends on previous sample (Markov)
  - Generation stochastic (Monte Carlo)
- Example MCMC method = Gibbs sampling

Gibbs sampling

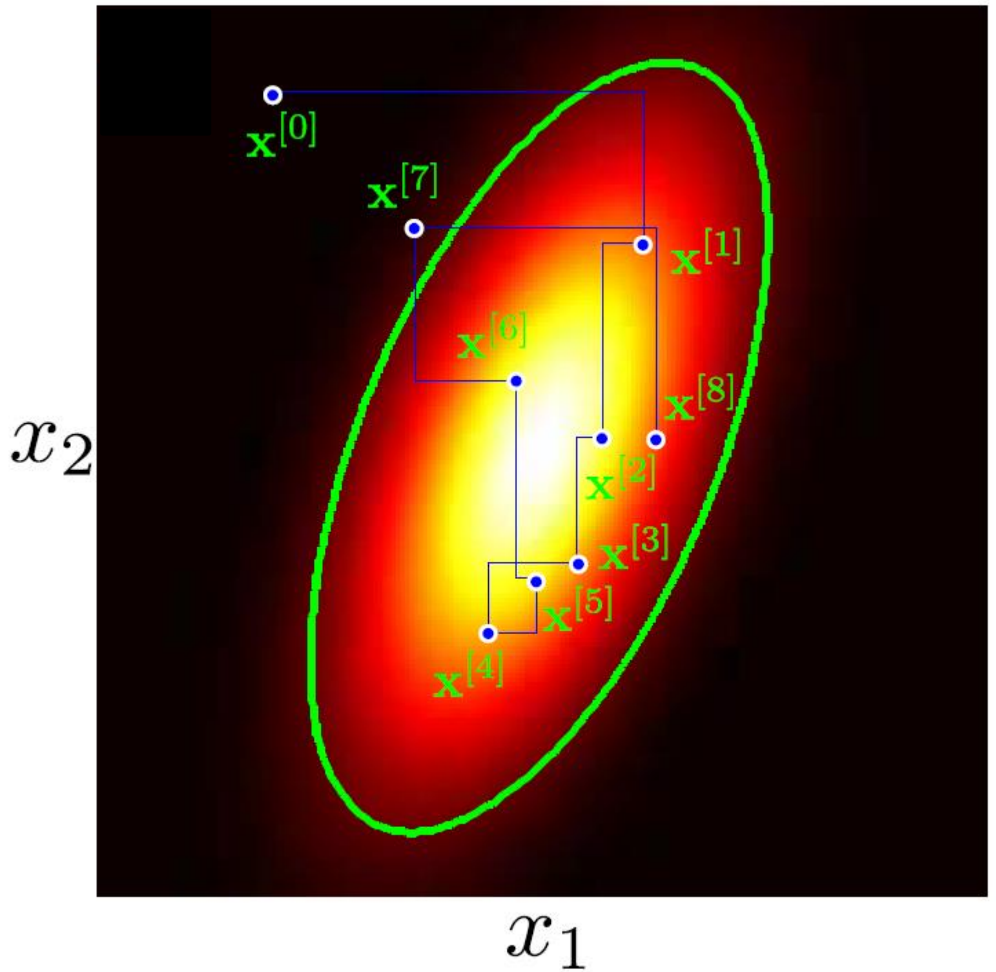
To generate new sample **x** in the chain

- Sample each dimension in any order
- To update n^{th} dimension x_n
  - Fix other N-1 dimensions
  - Draw from conditional distribution $Pr(x_n | x_{1...N\backslash n})$

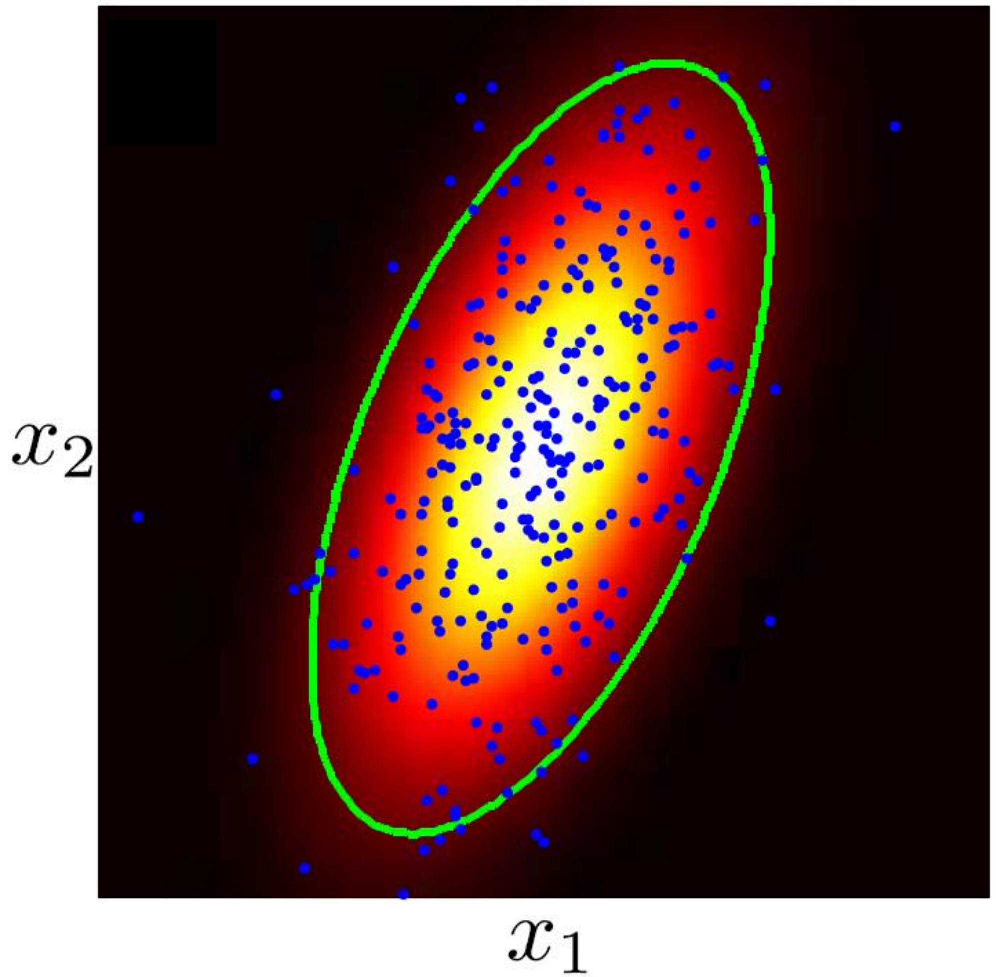
Get samples by selecting from chain

- Needs burn-in period
- Choose samples spaced apart, so not correlated

Gibbs sampling example: bi-variate normal distribution



Gibbs sampling example: bi-variate normal distribution



Learning in directed models

$$Pr(x_{1\ldots N}) = \prod_{n=1}^{I} Pr(x_n | x_{\text{pa}[n]})$$

Use standard ML formulation

$$\begin{aligned}\hat{\boldsymbol{\theta}} &= \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \left[\prod_{i=1}^{I} \prod_{n=1}^{N} Pr(x_{i,n} | x_{i,\text{pa}[n]}, \boldsymbol{\theta}) \right] \\ &= \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \left[\sum_{i=1}^{I} \sum_{n=1}^{N} \log[Pr(x_{i,n} | x_{i,\text{pa}[n]}, \boldsymbol{\theta})] \right]\end{aligned}$$

where $x_{i,n}$ is the n^{th} dimension of the i^{th} training example.

Learning in undirected models

Write in form of Gibbs distribution

$$Pr(\mathbf{x}) = \frac{1}{Z}\exp\left[-\sum_{c=1}^{C}\psi_c[x_{1\ldots N},\boldsymbol{\theta}]\right]$$

Maximum likelihood formulation

$$\begin{aligned}\hat{\boldsymbol{\theta}} &= \underset{\boldsymbol{\theta}}{\arg\max}\,\frac{1}{Z(\boldsymbol{\theta})^I}\exp\left[-\sum_{i=1}^{I}\sum_{c=1}^{C}\psi_c(\mathbf{x}_i,\boldsymbol{\theta})\right]\\ &= \underset{\boldsymbol{\theta}}{\arg\max}\,-I\log[Z(\boldsymbol{\theta})]-\sum_{i=1}^{I}\sum_{c=1}^{C}\psi_c(\mathbf{x}_i,\boldsymbol{\theta})\end{aligned}$$

Learning in undirected models

$$\begin{aligned}\frac{\partial L}{\partial \boldsymbol{\theta}} &= -I\frac{\partial \log[Z(\boldsymbol{\theta})]}{\partial \boldsymbol{\theta}} - \sum_{i=1}^{I}\sum_{c=1}^{C}\frac{\partial \psi_c(\mathbf{x}_i,\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\\ &= -I\frac{\partial \log\left[\sum_{\mathbf{x}_i}\exp\left[-\sum_{c=1}^{C}\psi_c(\mathbf{x}_i,\boldsymbol{\theta})\right]\right]}{\partial \boldsymbol{\theta}} - \sum_{i=1}^{I}\sum_{c=1}^{C}\frac{\partial \psi_c(\mathbf{x}_i,\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\end{aligned}$$

PROBLEM: To compute first term, we must sum over all possible states. This is intractable

Contrastive divergence

Some algebraic manipulation

$$
\begin{aligned}
\frac{\partial \log[Z(\boldsymbol{\theta})]}{\partial \boldsymbol{\theta}} &= \frac{1}{Z(\boldsymbol{\theta})}\frac{\partial Z(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \\
&= \frac{1}{Z(\boldsymbol{\theta})}\frac{\partial \sum_{\mathbf{x}} f[\mathbf{x},\boldsymbol{\theta}]}{\partial \boldsymbol{\theta}} \\
&= \frac{1}{Z(\boldsymbol{\theta})}\sum_{\mathbf{x}}\frac{\partial f[\mathbf{x},\boldsymbol{\theta}]}{\partial \boldsymbol{\theta}} \\
&= \frac{1}{Z(\boldsymbol{\theta})}\sum_{\mathbf{x}} f[\mathbf{x},\boldsymbol{\theta}]\frac{\partial \log[f[\mathbf{x},\boldsymbol{\theta}]]}{\partial \boldsymbol{\theta}} \\
&= \sum_{\mathbf{x}} Pr(\mathbf{x})\frac{\partial \log[f[\mathbf{x},\boldsymbol{\theta}]]}{\partial \boldsymbol{\theta}}.
\end{aligned}
$$

Contrastive divergence

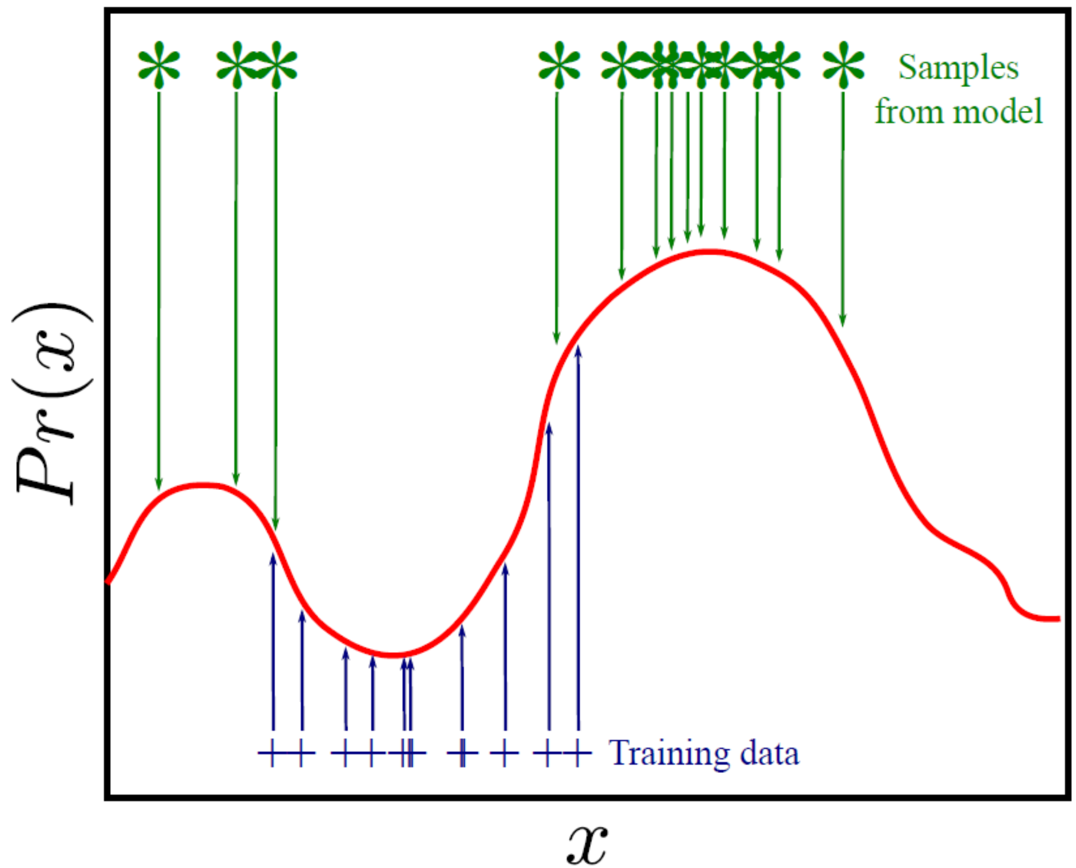
Now approximate:

$$\frac{\partial \log[Z(\boldsymbol{\theta})]}{\partial \boldsymbol{\theta}} = \sum_{\mathbf{x}} Pr(\mathbf{x}) \frac{\partial \log[f[\mathbf{x}, \boldsymbol{\theta}]]}{\partial \boldsymbol{\theta}}$$

$$\approx \frac{1}{J} \sum_{j=1}^{J} \frac{\partial \log[f[\mathbf{x}_j^*, \boldsymbol{\theta}]]}{\partial \boldsymbol{\theta}}$$

Where x_j^* is one of J samples from the distribution.
Can be computed using Gibbs sampling. In practice, it is possible to run MCMC for just 1 iteration and still OK.

Contrastive divergence



$$\frac{\partial \log[Pr(\mathbf{x})]}{\partial \boldsymbol{\theta}} = -\frac{\partial \log[Z(\boldsymbol{\theta})]}{\partial \boldsymbol{\theta}} + \frac{\partial \log[f[\mathbf{x}, \boldsymbol{\theta}]]}{\partial \boldsymbol{\theta}}$$

Conclusions

Can characterize joint distributions as

- Graphical models
- Sets of conditional independence relations
- Factorizations

Two types of graphical model, represent different but overlapping subsets of possible conditional independence relations

- Directed (learning easy, sampling easy)
- Undirected (learning hard, sampling hard)