Introduction to Probabilistic Graphical Models

Sargur Srihari
srihari@cedar.buffalo.edu

Topics

1. What are probabilistic graphical models (PGMs)
2. Use of PGMs Engineering and AI
3. Directionality in graphs
4. Bayesian Networks
5. Generative Models and Sampling
6. Using PGMs with fully Bayesian Models
7. Complexity Issues

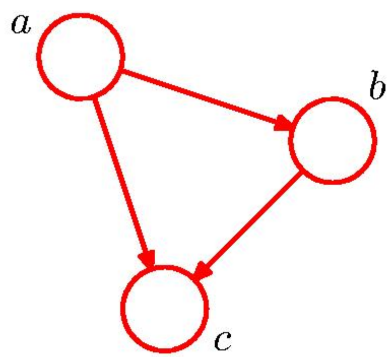
What are Graphical Models?

- They are *diagrammatic representations* of probability distributions
  - marriage between probability theory and graph theory

- Also called *probabilistic graphical models*
- They augment analysis instead of using pure algebra

What is a Graph?

- Consists of nodes (also called vertices) and links (also called edges or arcs)



- In a probabilistic graphical model
  - each node represents a random variable (or group of random variables)
  - Links express probabilistic relationships between variables

Graphical Models in Engineering

- Natural tool for handling Uncertainty and Complexity
  - which occur throughout applied mathematics and engineering
- Fundamental to the idea of a graphical model is the notion of modularity
  - a complex system is built by combining simpler parts.

Why are Graphical Models useful in Engineering?

- Probability theory provides the glue whereby
  - the parts are combined, ensuring that the system as a whole is consistent
  - providing ways to interface models to data.
- Graph theoretic side provides:
  - Intuitively appealing interface
    - by which humans can model highly-interacting sets of variables
  - Data structure
    - that lends itself naturally to designing efficient general-purpose algorithms

Graphical models: Unifying Framework

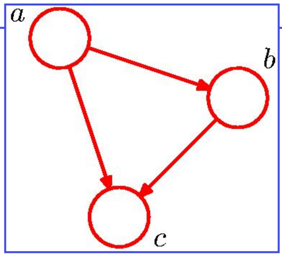
- View classical multivariate probabilistic systems as instances of a common underlying formalism
  - mixture models, factor analysis, hidden Markov models, Kalman filters and Ising models
  - Encountered in systems engineering, information theory, pattern recognition and statistical mechanics
- Advantages of View:
  - Specialized techniques in one field can be transferred between communities and exploited
  - Provides natural framework for designing new systems

Role of Graphical Models in ML

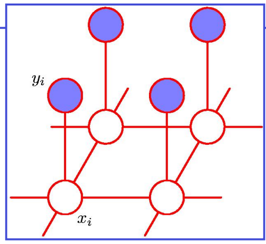
1. Simple way to visualize
   - structure of probabilistic model
2. Insights into properties of model
   - Conditional independence properties by inspecting graph
3. Complex computations
   - required to perform inference and learning expressed as graphical manipulations

Graph Directionality

- Directed graphical models
  - directionality associated with arrows
- Bayesian networks
  - Express causal relationships between random variables
- More popular in AI and statistics



- Undirected graphical models
  - links without arrows
- Markov random fields
  - Better suited to express soft constraints between variables
- More popular in Vision and physics



Independence and Inference

- MRFs have simple definition of independence
  - Two sets of nodes are conditionally independent given a third set C if all nodes in A and B are connected through nodes in C
- BN independence is more complex
  - Involves direction of arcs
- Inference problems
  - Convenient to convert both to *factor graph* representation

Bayesian Networks

- Directed graphs
  - used to describe probability distributions
- Consider Joint distribution
  - of three variables a,b,c
- Powerful aspect of graphical models
  - Not necessary to state whether they are discrete or continuous
  - A specific graph
    - can make probabilistic statements about a broad class of distributions
- Bayesian Network is not necessarily Bayesian statistics

Joint and Conditional Distributions

- The necessary probability theory can be expressed in terms of two simple equations
  - Sum Rule
    - probability of a variable is obtained by marginalizing or summing out other variables

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

  - Product Rule
    - joint probability expressed in terms of conditionals

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

All probabilistic inference and learning amounts to repeated application of sum and product rule

From Joint Distribution to Graphical Model

- Consider Joint distribution $p(a,b,c)$
  - By product rule $p(a,b,c)=p(c|a,b)p(a,b)$
  - Again by product rule we get

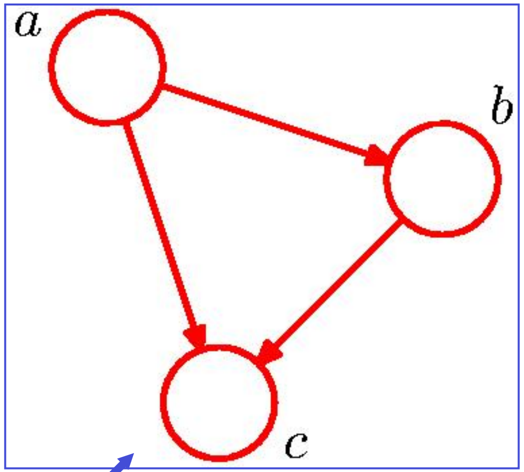
$$\boxed{p(a,b,c)=p(c|a,b)p(b|a)p(a)}$$

- This decomposition holds
  - for any choice of the joint distribution

Directed Graphical Model

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

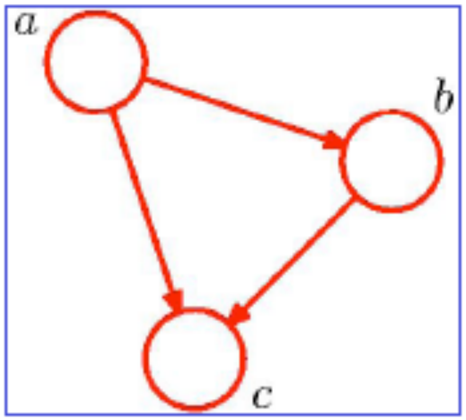
- Now represent rhs by graphical model
  - Introduce a node for each random variable
  - Associate each node with conditional distribution on rhs
    - For each conditional distribution add links (arrow): for $p(c|a,b)$ links from a and b to c
- Different ordering of variables would give a different graph



TERMINOLOGY

- Node a is *parent* of node b
- Node b is *child* of node a
- *No distinction between node and variable*

From Graph to Distribution



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

- Graph says:
  - Distribution is a product of three terms
  - There are three known distributions whose product defines the joint distribution
  - Each distribution could be specified by a table

Ex: Conditional Probabilities

- x_1= Box: red/blue
- x_2= Fruit: a/o

$p(x_1)$

$p(x_1\text{=red})= 4/10$

$p(x_1\text{=blue})= 6/10$

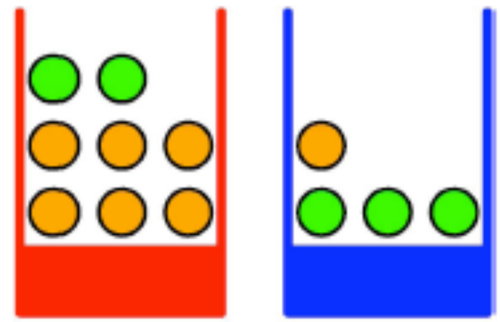
$p(x_2|x_1)$

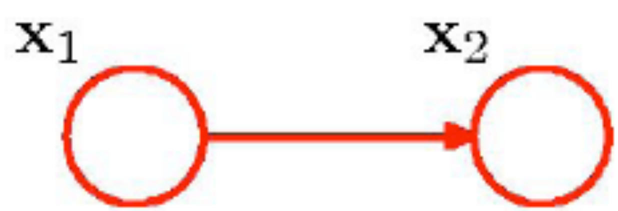
$p(x_2\text{=a}|x_1\text{=red})= 1/4$

$p(x_2\text{=o}|x_1\text{=red})= 3/4$

$p(x_1\text{=a}|x_1\text{=blue})= 3/4$

$p(x_1\text{=o}|x_1\text{=blue})= 1/4$





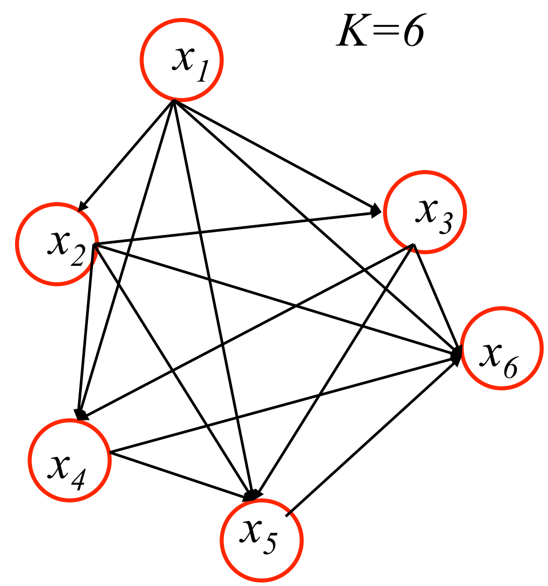
$p(x_1,x_2)=p(x_1)p(x_2|x_1)$

Joint distribution of K variables

- Repeated application of product rule gives

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

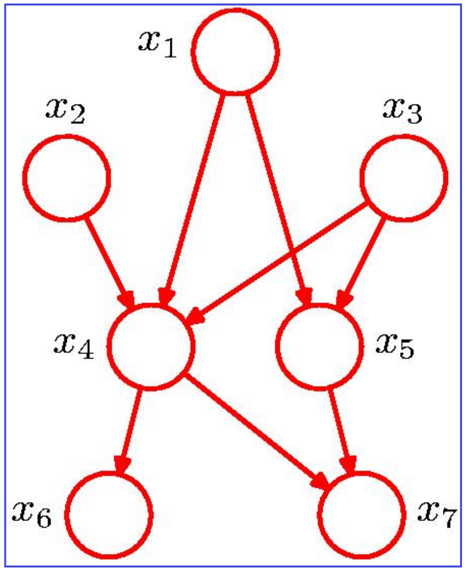
- Represent by K nodes,
  - one for each conditional distribution
- Each node has
  - incoming nodes from all lower-numbered nodes
- These are fully connected
  - Link between every pair of nodes



Not a fully connected graph

- Absence of links carry interesting information
  - Not a fully connected graph with *7* variables
    - No link from x_1 to x_2 or from x_3 to x_7
  - Joint distribution given by

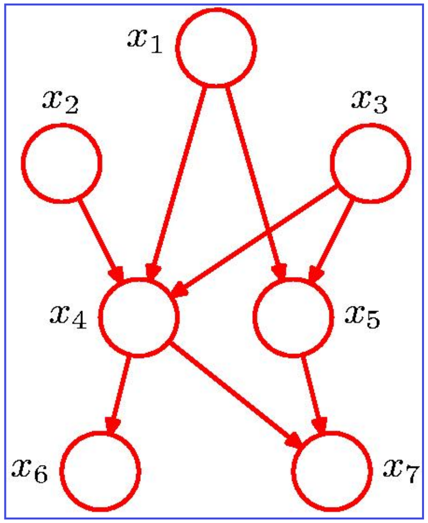
$$p(x_1)\, p(x_2)\, p(x_3) p(x_4|x_1,x_2,x_3) p(x_5|x_1, x_3)\, p(x_6|x_4)\, p(x_7|x_4,x_5)$$



Directed acyclic graph over seven variables

Graph to Probability Distribution

- Joint distribution defined by a graph is given by a product
- Product terms are conditional distributions of each node conditioned on variables corresponding to parents of that node in the graph



$$p(\mathrm{x}) = \prod_{k=1}^{K} p(x_k \mid pa_k)$$

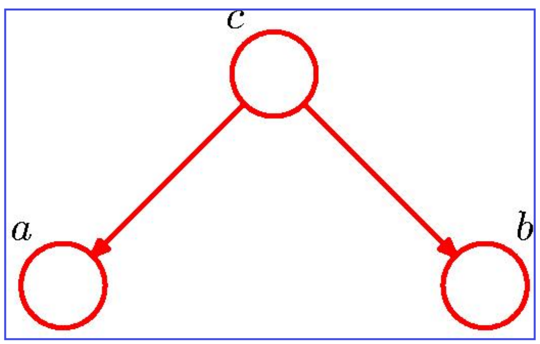
  – where $x=(x_1,..,x_K)$ and pa_k denotes set of parents of a_k

- By inspection, joint distribution given by

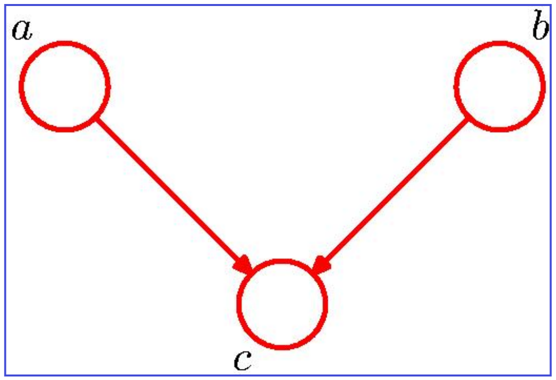
$p(x_1)\ p(x_2)\ p(x_3)p(x_4|x_1,x_2,x_3)p(x_5|x_1,x_3)p(x_6|x_4)\ p(x_7|x_4,x_5)$

Examples of graphs and distributions

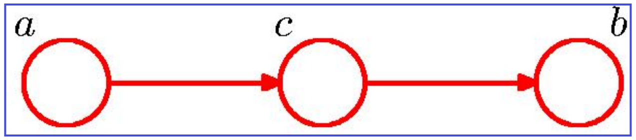
- Each graph has three nodes
- Applying parent rule we get the factors for each graph



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



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

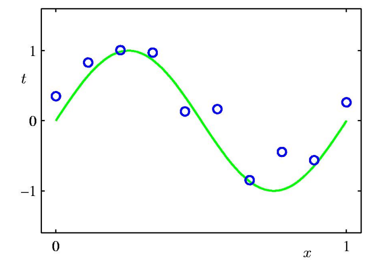


$p(a,b,c)=p(a)p(c|a)p(b|c)$

Example: Bayesian Polynomial Regression

- Illustration of directed graph to describe probability distributions
- The polynomial regression problem

$$y(x,\mathrm{w}) = w_0 + w_1 x + w_2 x^2 + .. + w_M x^M$$



- Probabilistic formulation with Random variables:
  - Vector of polynomial coefficients w
  - Observed data $\mathrm{t} = (t_1,..,t_N)^T$ where $t_n = y(x_n, \mathrm{w}) + e$ is a noise corrupted version of x_n
- Focusing on random variables, joint distribution is given by

From product rule $p(\mathrm{t},\mathrm{w}) = p(\mathrm{w})p(\mathrm{t}|\mathrm{w})$

$$p(t,\mathrm{w}) = \underbrace{p(\mathrm{w})}_{\text{prior}} \underbrace{\prod_{n=1}^{N} p(t_n \mid \mathrm{w})}_{\text{Conditional distributions}}$$

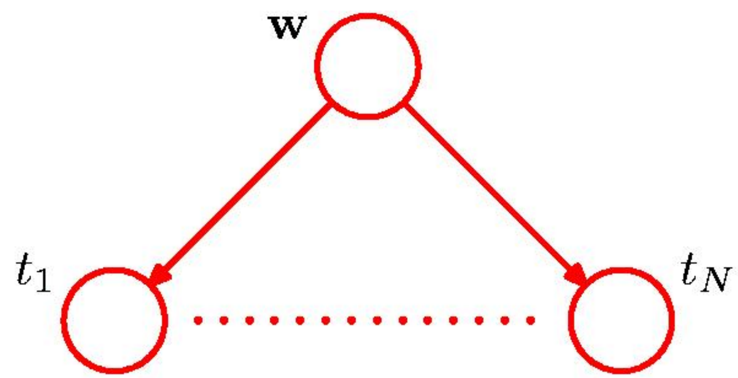
Since samples are independent

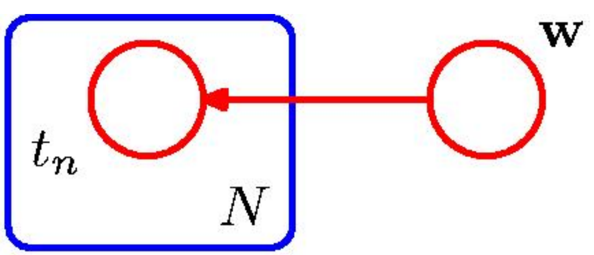
Graph for Polynomial Regression

- Joint distribution is given by

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

- It is to be represented by a graphical model
  - Observing rhs introduce a node for each variable
  - Introduce associations
    - A link for each conditional distribution
- Equivalent *plate* representation





Parameters in Bayesian Poly Regression

- Random variables:
  - Vector of polynomial coefficients w
  - Observed data $\mathrm{t} = (t_1,..,t_N)^T$ where t_n is a noise corrupted version of x_n
- Additionally model contains:
  - Input data $\mathrm{x} = (x_1,..,x_N)^T$
  - Noise variance σ^2
  - Hyper-parameter α, which is the precision (inverse variance) of Gaussian prior

Parameters of model rather than random variables

- Focusing on random variables, joint distribution is given by

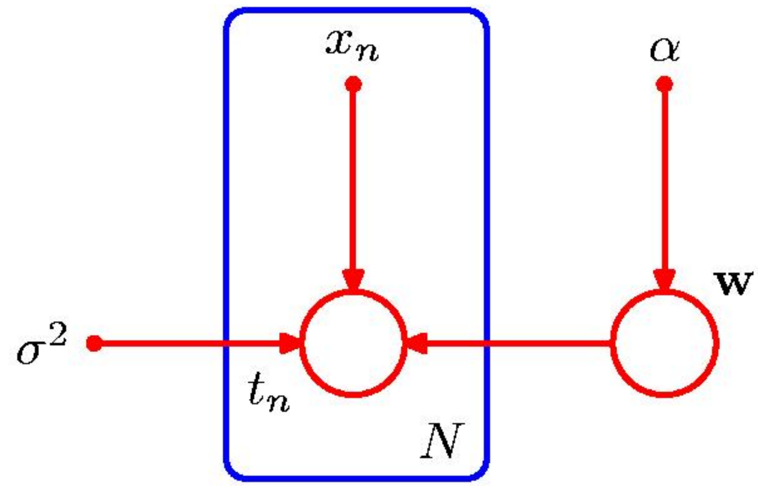
$$p(t, \mathrm{w}) = \underbrace{p(\mathrm{w})}_{\substack{\text{Gaussian prior} \\ \text{with precision } \alpha}} \prod_{n=1}^{N} \underbrace{p(t_n \mid \mathrm{w})}_{\substack{\text{Conditional distributions} \\ \text{with variance } \sigma^2}}$$

With Deterministic Parameters

- Sometimes useful to explicitly show parameters of model
- Same model with deterministic parameters

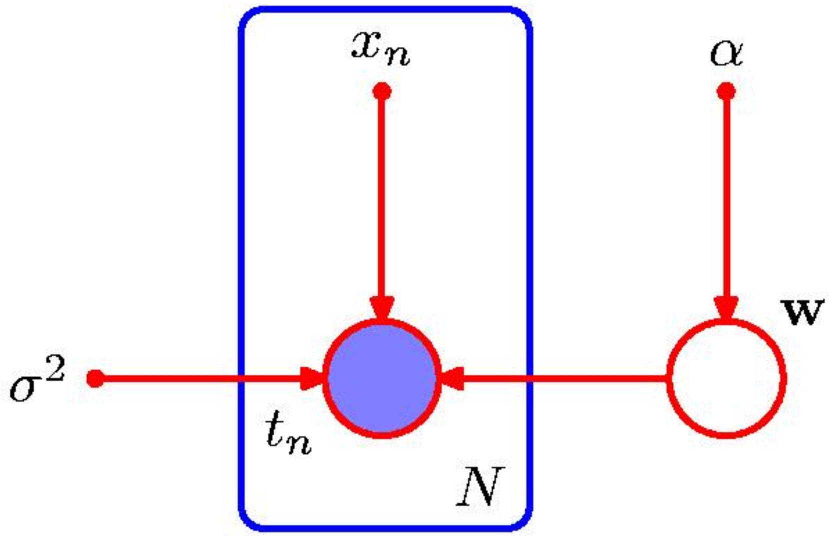
$$p(\mathrm{t},\mathrm{w}|\mathrm{x},\alpha,\sigma^2) = p(\mathrm{w}|\alpha)\prod_{n=1}^{N} p(t_n \mid \mathrm{w}, x_n, \sigma^2)$$

- Random variables are denoted by open circles
- Deterministic parameters by smaller (tiny) solid circles



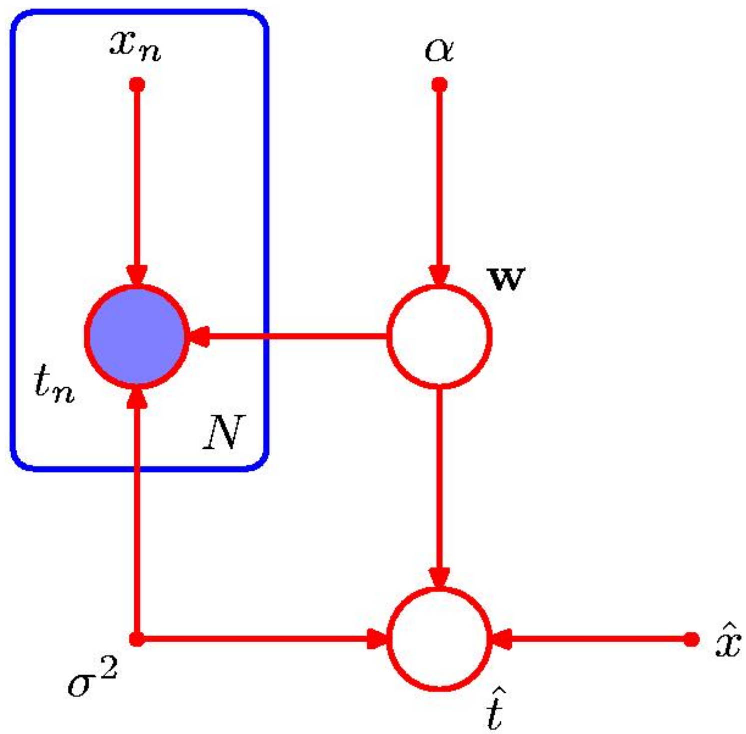
Observed and Latent Variables

- In ML some random variables are set to specific observed values
  - Observed variable nodes are shaded
- Variable $\mathbf{w}$ is not observed
  - It is *latent* or *hidden*
- Nodes t_n are shaded
  - to indicate corresponding variables are set to observed values



Including predictions

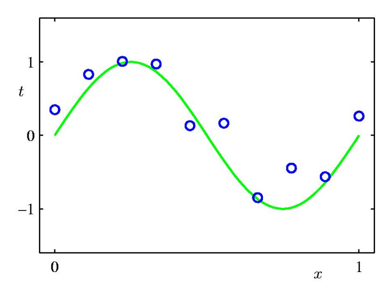
- Polynomial Regression model including input and predicted values
- Need to write out longer probability expression including
  - new input value $\hat{x}$ and model prediction $t^{\wedge}$
- Joint distribution of all variables in this model



$$p(\hat{t},\mathrm{t},\mathrm{w}\,|\,\hat{\mathrm{x}},\mathrm{x},\alpha,\sigma^2)=p(\mathrm{w}\,|\,\alpha)\prod_{n=1}^{N}p(t_n\,|\,\mathrm{w},x_n,\sigma^2)p(\hat{t}\,|\,\hat{x},\mathrm{w},\sigma^2)$$

- Predictive distribution

$$p(\hat{t}\mid\hat{x},\mathrm{x},\alpha,\sigma^2)\ \alpha\int p(\hat{t},\mathrm{t},w\mid\hat{x},\mathrm{x},\alpha,\sigma^2)\ d\mathrm{w}$$



Generative Models

- Many situations where we need to draw samples from a probability distribution
- Many methods of sampling exist
- *Ancestral Sampling* is relevant to graphical models
  - Given a graphical model we can specify how samples can be drawn from the joint or marginal distributions
    - Conditional can be drawn by setting the parent value to known value

Specifying Sampling Problem

- Joint distribution $p(x_1,..,x_K)$ over K variables which factorizes according to

$$p(\mathrm{x}) = \prod_{k=1}^{K} p(x_k \mid pa_k)$$

- Assume each node has a higher numbered node than its parents
- Goal is to draw a sample $\hat{x}_1, \hat{x}_2,..,\hat{x}_K$
  - from this distribution

Ancestral Sampling

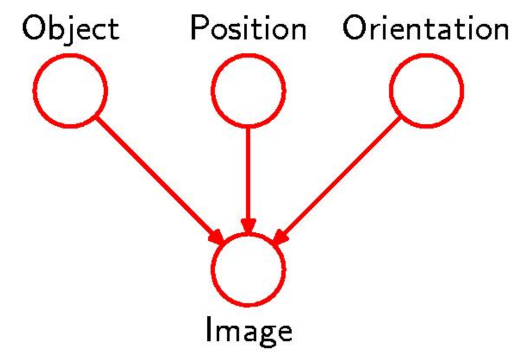
- Start with lowest numbered node
- Draw a sample from the distribution $p(x_1)$ which we call $\hat{x}_1$
- Work through each of the nodes in order
  - For node n we draw a sample from conditional distribution $p(x_n|pa_n)$
  - Where parent variables are set to their sampled values
- Once final variable x_K is sampled
  - Achieved objective of obtaining a single sample from joint distribution
- To sample from marginal distribution
  - Sample from full distribution and discard unnecessary values
  - E.g., to draw from distribution $p(x_2,x_4)$ simply sample from full distribution, retain values x_2,x_4^ and discard remaining values $\{\hat{x}_{j\neq 2,4}\}$

Numbering in Practical Applications

- Higher numbered variables
  - Correspond to terminal nodes of graph
  - Represent observations
- Lower numbered nodes
  - Correspond to latent variables
  - Latent variables allow
    - complicated distribution over observed variables to be represented in terms of simpler (typically exponential family) distributions

Causal Model Example

- Models can express processes by which the observed data arose
- Object recognition task
  - Observed data point is an image of one of the objects
  - Latent variables are position and orientation of object
- To find posterior distribution over object
  - Integrate over all positions and orientations
- The graphical model captures *causal* process by which data is generated



Generative Model

- Since data can be generated from causal model it is a generative model
- Ancestral sampling could be used to generate synthetic data
  - Mimics creation of observed data
  - Gives rise to "fantasy" data whose probability distribution is the same as model
  - Synthetic observation is informative in understanding form of distribution represented by model

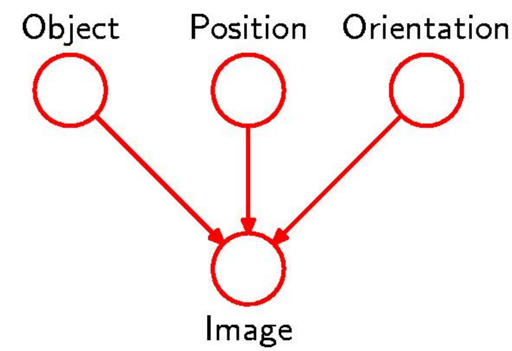
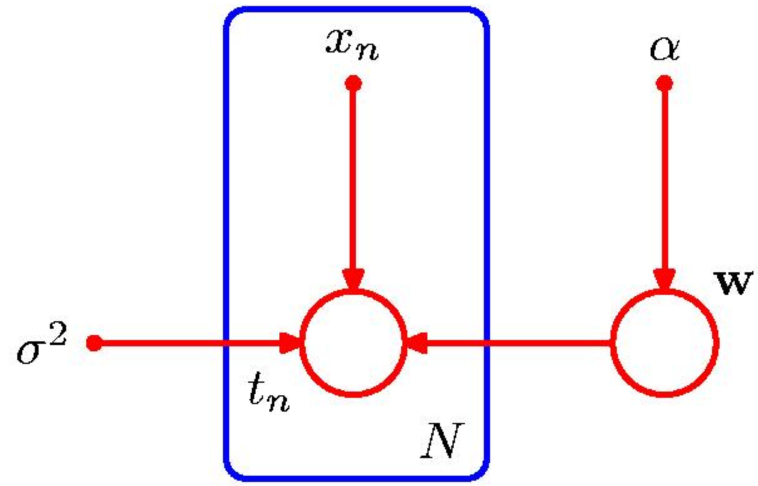


Image pixel intensities are dependent on identity of object as well as its pose

But..Polynomial Regression Model is not generative

- No probability distribution associated with input variable x
- Not possible to generate synthetic data points from this model
- Can make it generative by introducing a prior distribution $p(x)$



Discrete Variables

- When constructing more complex probability distributions from simpler (exponential) distributions graphical models are useful
- Graphical models have nice properties when each parent-child pair are conjugate
- Two cases of interest:
  - Both correspond to discrete variables
  - Both correspond to Gaussian variables

Discrete Case

- Probability distribution $p(\mathrm{x}\,|\,\mu)$ for a single discrete variable x having K states

- Using 1 of K representation
  - For $K=6$ when $x_3=1$ then x represented as $\mathrm{x}=(0,0,1,0,0,0)^T$
    Note that $\sum_{k=1}^{K} x_k = 1$
- If probability of $x_k=1$ is given by parameter μ_k

$$p(\mathrm{x}\,|\,\mu) = \prod_{k=1}^{K} \mu_k^{x_k} \quad \text{where } \mu = (\mu_1, .., \mu_{\mathrm{K}})^{\mathrm{T}}$$

- The distribution is normalized: $\sum_{k=1}^{K} \mu_k = 1$
  - There are K-1 independent values for μ_k needed to define distribution

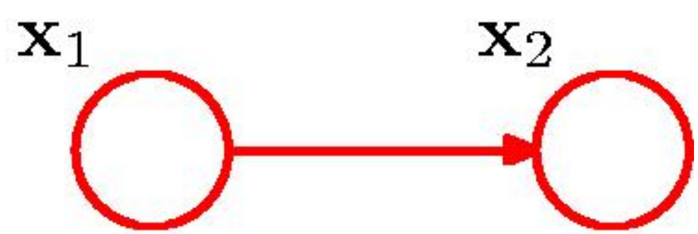
Two Discrete Variables

x_1:	x_{11}					x_{1K}
x_2:	x_{21}					x_{2K}

- x_1 and x_2 each with K states each
- Denote probability of both $x_{1k}=1$ and $x_{2l}=1$ by μ_{kl}
- where x_{1k} denotes k^{th} component of x_1
- Joint distribution is $$\boxed{p(\mathrm{x}_1,\mathrm{x}_2 \mid \mu) = \prod_{k=1}^{K}\prod_{l=1}^{K} \mu_{kl}^{x_{1k}x_{2l}}}$$
- Since parameters are subject to constraint $\sum_k \sum_l \mu_{kl} = 1$
- There are K^2-1 parameters
- For arbitrary distribution over M variables there are K^M-1 parameters

Graphical Models for Two Discrete Variables

- Joint distribution $p(\mathrm{x}_1,\mathrm{x}_2)$
- Using product rule
  - is factored as $p(\mathrm{x}_2|\mathrm{x}_1)p(\mathrm{x}_1)$



- Has two node graph
- Marginal distribution $p(\mathrm{x}_1)$ has K-1 parameters
- Conditional distribution $p(\mathrm{x}_2|\mathrm{x}_1)$ also requires K-1 parameters for each of K values of x_1
- Total number of parameters is $(K\text{-}1)+K(K\text{-}1) = K^2\text{-}1$
  - As before

Two Independent Discrete Variables

- x_1 and x_2 are independent
  - Has graphical model

- Each variable described by a separate multinomial distribution
  - Total no of parameters is $2(K-1)$
  - For M variables no of parameters is $M(K-1)$
- Reduced number of parameters by dropping links in graph
  - Grows linearly with no of variables

Fully connected has high complexity

- General case of M discrete variables $x_1,.., x_M$
- If BN is fully connected
  - Completely general distribution with K^M-1 parameters
- If there are no links
  - Joint distribution factorizes into product of marginals
  - Total no of parameters is $M(K-1)$
- Graphs of intermediate levels of connectivity
  - More general distribution than fully factorized ones
  - Require fewer parameters than general joint distribution
  - Example: chain of nodes

Special Case: Chain of Nodes

- Marginal distribution $p(\mathrm{x}_1)$ requires K-1 parameters
- Each of the M-1 conditional distributions $p(\mathrm{x}_i|\mathrm{x}_{i-1})$, for $i=2,..,M$ requires $K(K$-$1)$ parameters
- Total parameter count is

  $K\text{-}1+(M\text{-}1)K(K\text{-}1)$

  Which is quadratic in K
- Grows linearly (not exponentially) with length of chain

Alternative: Sharing Parameters

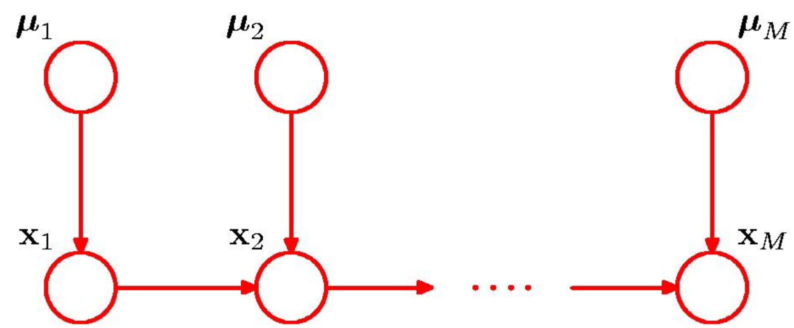
- Reduce parameters by *sharing* or *tying* parameters

- In above,
  - all conditional distributions $p(x_i|x_{i-1})$, for $i=2,..,M$
  - share same set of $K(K-1)$ parameters governing distribution of x_1
- Total of K^2-1 parameters needed to specify distribution

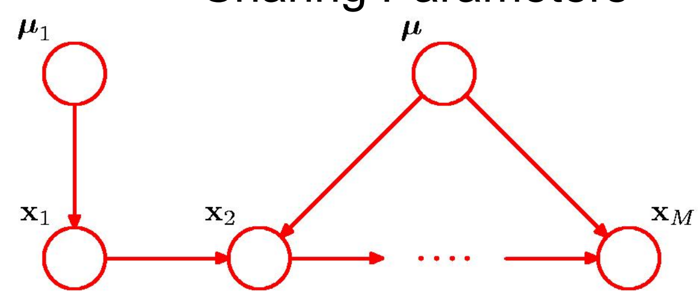
Conversion into Bayesian Model

- Given graph over discrete variables
- We can turn it into a Bayesian model by introducing Dirichlet priors for parameters
- Each node acquires an additional parent for each discrete node
- Tie the parameters governing conditional distributions $p(x_i|x_{i-1})$

Chain of Nodes
With priors



Sharing Parameters



Binomial: Beta Prior

- Bernoulli:

$$p(x=1|\mu)=\mu$$

- Likelihood of Bernoulli with $D=\{x_1,..x_N\}$

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

- Binomial:

$$Bin(m|N,\mu)=\binom{N}{m}\mu^m(1-\mu)^{N-m}$$

– Conjugate Prior:

$$Beta(\mu|a,b)=\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1}$$

Multinomial: Dirichlet Prior

- Generalized Bernoulli (*1*-of-*K*) x=*(0,0,1,0,0,0)^T K=6*

$$p(\mathrm{x}\,|\,\mu)=\prod_{k=1}^{K}\mu_k^{x_k} \text{ where } \mu=(\mu_1,..,\mu_K)^T$$

K=2 is Bernoulli

- Multinomial

K=2 is Binomial

$$Mult(m_1 m_2..m_K\,|\,\mu,N)=\binom{N}{m_1 m_2..m_k}\prod_{k=1}^{K}\mu_k^{m_k}$$

$$\binom{N}{m_1 m_2..m_k}=\frac{N!}{m_1!m_2!..m_k!}$$

  - Where the normalization coefficient is the no of ways of partitioning *N* objects into *K* groups of size $m_1, m_2..m_k$

- Conjugate prior distribution for parameters m_k

$$p(\mu\,|\,\alpha)\,\alpha\prod_{k=1}^{K}\mu_k^{\alpha_k-1} \text{ where } 0\le\mu_k\le 1 \text{ and } \sum_k \mu_k=1$$

- Normalized form is

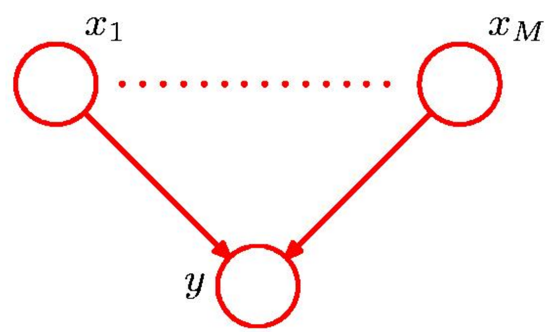
$$Dir(\mu\,|\,\alpha)=\frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)...\Gamma(\alpha_k)}\prod_{k=1}^{K}\mu_k^{\alpha_k-1} \text{ where } \alpha_0=\sum_{k=1}^{K}\alpha_k$$

Controlling Number of parameters in models: Parameterized Conditional Distributions

- Control exponential growth of parameters in models of discrete variables
- Use parameterized models for conditional distributions instead of complete tables of conditional probability values

Parameterized Conditional Distributions

- Consider graph with binary variables
- Each parent variable x_i governed by single parameter m_i representing probability $p(x_i=1)$
- M parameters in total for parent nodes
- Conditional distribution $p(y|x_1,..,x_M)$ requires 2^M parameters
  - Representing probability $p(y=1)$ for each of the 2^M settings of parent variables
  - 000000000 to 111111111



Conditional distribution using logistic sigmoid

- Parsimonious form of conditional distribution
- Logistic sigmoid acting on linear combination of parent variables

$$p(y=1 \mid x_1,..,x_M) = \sigma\left(w_0 + \sum_{i=1}^{M} w_i x_i\right) = \sigma\left(\mathrm{w}^T \mathrm{x}\right)$$

  – where $s(a) = (1+exp(-a))^{-1}$ is the logistic sigmoid
  – $\mathrm{x}=(x_0,x_1,..,x_M)^T$ is vector of parent states
- No of parameters grows linearly with M
- Analogous to choice of a restrictive form of covariance matrix in multivariate Gaussian

Linear Gaussian Models

- Expressing multivariate Gaussian as a directed graph
- Corresponding to linear Gaussian model over component variables
  - Mean of a conditional distribution is a linear function of the conditioning variable
- Allows expressing interesting structure of distribution
  - General Gaussian case and diagonal covariance case represent opposite extremes

Graph with continuous random variables

- Arbitrary acyclic graph over D variables
- Node i represents a single continuous random variable x_i having Gaussian distribution
- Mean of distribution is a linear combination of states of its parent nodes pa_i of node i

$$p(x_i \mid pa_i) = \mathrm{N}\left(x_i \,\middle|\, \sum_{j \in pa_i} w_{ij} x_j + b_i, v_i \right)$$

- Where
  - w_{ij} and b_i are parameters governing the mean
  - v_i is the variance of the conditional distribution for x_i

Joint Distribution

- Log of joint distribution

$$\ln p(\mathrm{x}) = \sum_{i=1}^{D} \ln p(x_i \mid pa_i)$$

$$= -\sum_{i=1}^{D} \frac{1}{2v_i}\left(x_i - \sum_{j \in pa_i} w_{ij} x_j - b_i \right)^2 + \underbrace{const}_{\text{Terms independent of } \mathrm{x}}$$

- Where $\mathrm{x}=(x_1,..,x_D)^T$
- This is a quadratic function of x
  - Hence joint distribution $p(\mathrm{x})$ is a multivariate Gaussian

Mean and Covariance of Joint Distribution

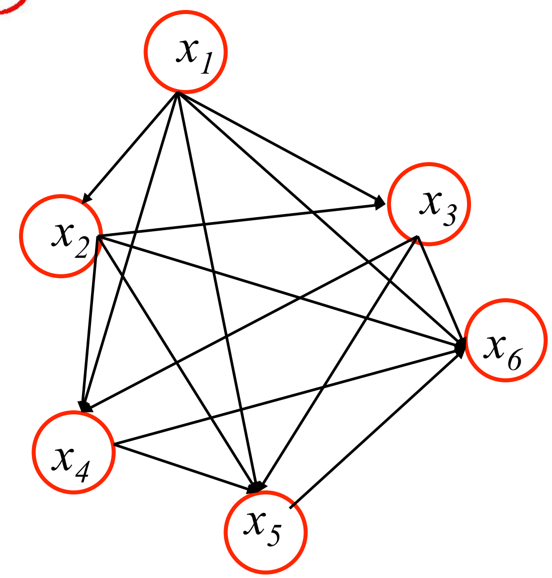
- Recursive Formulation
- Since each variable x_i has,
  - conditional on the states of its parents,
  - a Gaussian distribution, we can write $x_i = \sum_{j \in pa_i} w_{ij} + b_i + \sqrt{v_i}\varepsilon_i$
  - where e_i is
    - a zero mean, unit variance Gaussian random variable
    - satisfying $\mathbf{E}[e_i]=0$ and $\mathbf{E}[e_i e_j]=I_{ij}$
    - and I_{ij} is the i,j element of the identity matrix
- Taking expectation $\boxed{\mathbf{E}[x_i] = \sum_{j \in pa_i} w_{ij}\mathbf{E}[x_j] + b_i}$
- Thus we can find components of $\mathbf{E}[\mathrm{x}]=(\mathbf{E}[\mathrm{x}_1],..\mathbf{E}[\mathrm{x}_D])^T$
  - by starting at lowest numbered node and working recursively through the graph
- Similarly elements of covariance matrix

$$\boxed{\mathrm{cov}[x_i, x_j] = \sum_{j \in pa_j} w_{jk}\,\mathrm{cov}[x_i, x_k] + I_{ij} v_j}$$

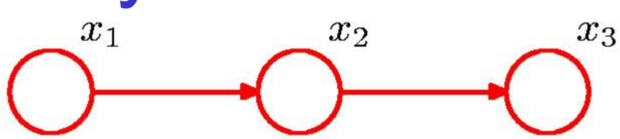
Three cases for no. of parameters

- No links in the graph

  - *2D* parameters
- Fully connected graph



  - $D(D+1)/2$ parameters
- Graphs with intermediate level of complexity
  - Chain



Extreme Case with no links

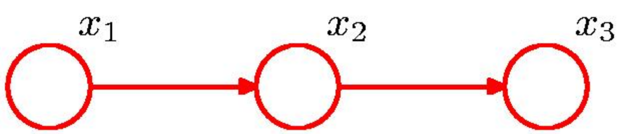
- D isolated nodes
  - There are no parameters w_{ij}
  - Only D parameters b_i and
  - D parameters v_i
- Mean of $p(x)$ given by $(b_1,..,b_D)^T$
- Covariance matrix is diagonal of form $diag(v_1,..,v_D)$
- Joint distribution has total of $2D$ parameters
- Represents set of D independent univariate Gaussian distributions

Extreme case with all links

- Fully connected graph
- Each node has all lower numbered nodes as parents
- Matrix w_{ij} has i-1 entries on the ith row and hence is a lower triangular matrix (with no entries on leading diagonal)
- Total no of parameters w_{ij} is to take D^2 no of elements in D x D matrix, subtracting D to account for diagonal and divide by 2 to account for elements only below diagonal

Graph with intermediate complexity

- Link missing between variables x_1 and x_3
- Mean and covariance of joint distribution are



$$\mu = \left(b_1,\ b_2 + w_{21}b_1,\ b_3 + w_{32}b_2 + w_{32}w_{21}b_1\right)^T$$

$$\Sigma = \begin{pmatrix} v_1 & w_{21}v_1 & w_{32}w_{21}v_1 \\ w_{21}v_1 & v_2 + w_{21}^2 v_1 & w_{32}(v_2 + w_{21}^2 v_1) \\ w_{32}w_{21}v_1 & w_{32}(v_2 + w_{21}^2 v_1) & w_{32}^2(v_2 + w_{21}^2 v_1) \end{pmatrix}$$

Extension to multivariate Gaussian variables

- Nodes in the graph represent multivariate Gaussian variables
- Write conditional distribution for node i in the form

$$p(\mathrm{x}_i \mid pa_i) = \mathrm{N}\left(\mathrm{x}_i \middle| \sum_{j \in pa_i} \mathrm{W}_{ij}\mathrm{x}_j + \mathrm{b}_i, \sum v_i \right)$$

- where W_{ij} is a matrix (non-square if x_i and x_j have different dimensionalities)

Summary

1. PGMs allow visualizing probabilistic models
   – Joint distributions are directed/undirected PGMs
2. PGMs can be used to generate samples
   – Ancestral sampling with directed PGMs is simple
3. PGMs are useful for Bayesian statistics
   – Discrete variable PGM represented using Dirichlet priors
4. Parameter explosion controlled by tying parameters
5. Multivariate Gaussian expressed as PGM
   – Graph is a linear Gaussian model over components