Probabilistic Graphical Models

Frequentist Estimation,
Bayesian Networks

Frequentist Estimation

θ is an unknown number

Each θ defines a different probabilistic model for your data

Some models explain your data better than others

$P(x_1, \ldots, x_n; \theta)$ is the likelihood function (not a conditional probability since θ is not random

Estimation: Find $\hat{\theta}$

The estimator is a random variable (why?)

Frequentist Estimation

Given $X = x$, the maximum likelihood estimate (MLE) will be a function of x.

Notation: $\hat{\theta} = \delta(\mathrm{X}_1, \dots, \dots X_{\mathrm{n}})$

Potentially confusing notation: Sometimes $\hat{\theta}$ is used for both the estimator and the estimate.

- Note: The MLE is required to be in the parameter space Ω.
- Often it is easier to maximize the log-likelihood $L(\theta) = \log(f(x \mid \theta)$

Maximum Likelihood Estimation

- Let $X \sim \text{Binomial}(\theta)$. Find the maximum likelihood estimator of θ. Say we observe $X = 3$, what is the maximum likelihood estimate of θ ?
- Let $X_1, \dots, X_n$ be i.i.d. $N(\mu, \sigma^2)$.

- Find the MLE of μ when σ^2 is known.
- Find the MLE of μ and σ^2 (both unknown).

- Let $X_1, \dots, X_n$ be i.i.d. Uniform $[0, \theta]$, where $\theta > 0$. Find $\hat{\theta}$
- Let $X_1, \dots, X_n$ be i.i.d. Uniform $[\theta, \theta + 1]$. Find $\hat{\theta}$

Quantifying Uncertainty

- Suppose $X = (X_1, \dots, X_n)$ is a random sample from $f(x \mid \theta)$.
- A function $r(X_1, \dots, X_n)$ is a statistic (and a random variable).
- A sampling distribution: the distribution of a statistic (given θ)
- Estimator $\hat{\theta}$ is a statistic
- Can use the sampling distributions to compare different estimators and quantify uncertainty
- Can be used to estimate number of samples we need to limit bias
- Leads to definitions of new distributions, e.g., χ_m^2 and t_m.

Quantifying Uncertainty

Let $X_1, \ldots, X_n$ be a random sample from a $\mathcal{N}(\mu, \sigma^2)$ with unknown μ, σ^2.
The sample mean and the sample variance are defined as

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i, \qquad \hat{\sigma}^2_{MLE} = \frac{1}{n}\sum_{i=1}^{n} (X_i - \bar{X}_n)^2, \quad S_n = \left(\frac{\sum (X_i - \bar{X}_n)^2}{n-1}\right)^{1/2}$$

If you know μ but not σ^2

$$\frac{n\hat{\sigma}^2_{MLE}}{\sigma^2} \sim \chi^2_n$$

If you do not know μ or σ^2, then

$$n^{1/2}(\bar{X}_n - \mu)/S_n \sim t_{n-1}$$

Confidence interval

You can compute

$$P\left(\bar{X}_n - \frac{cS_n}{n^{1/2}} < \mu < \bar{X}_n + \frac{cS_n}{n^{1/2}}\right) \geq \gamma$$

- γ – Confidence Interval for μ.
- μ is not random, the interval is.
- Interpretation: γ is the frequency we expect the random interval to include the true value, if we repeat the experiment multiple times

Properties of an Estimator

An estimator $\hat{\theta} = g(X_1, \ldots, X_n)$ is a function of random variables $X_1, \ldots, X_n$ and therefore has a distribution. The distribution of $\hat{\theta}$ is called sampling distribution.

Unbiased estimator
An estimator is unbiased if $E(\hat{\theta}) = \theta$. $E(\hat{\theta}) - \theta$ is called the bias of the estimator.

Consistent estimator
An estimator is consistent if $\widehat{\theta_n} \xrightarrow{p} \theta$.

Example
Bernoulli MLE $\hat{\theta}_{MLE} = \sum_{i=1}^{n} x_i$. Is it unbiased? Is it consistent?

Mean Squared Error of an Estimator

Mean squared error estimator

$$\mathrm{MSE}(\hat{\theta}) = \mathbb{E}\left[(\hat{\theta} - \theta)^2\right]$$
$$= \mathrm{Var}(\hat{\theta} - \theta) + (\mathbb{E}[\hat{\theta} - \theta])^2$$
$$= \mathrm{Var}(\hat{\theta}) + \mathrm{Bias}^2(\hat{\theta})$$

For unbiased estimators, $MSE(\hat{\theta}) \; = \; Var(\hat{\theta})$

Sufficient Statistics

- A statistic: $T = r(X_1, \ldots, X_n)$
  Def: Sufficient Statistics
  Let $X_1, \ldots, X_n$ be a random sample from $f(x \mid \theta)$ and let T be a statistic.
  If the conditional distribution of
  $$X_1, \ldots, X_n \mid T = t$$
  does not depend on θ then T is called a sufficient statistic.
- The idea: Just as good to have the observed sufficient statistic as it is to have the individual observations of $X_1, \ldots, X_n$.
- Can limit our search for a good estimator to sufficient statistics

Sufficient Statistics

- Theorem: Factorization Criterion
  Let $X_1, \dots, X_n$ be a random sample form $f(x \mid \theta)$ where $\theta \in \Omega$ is unknown. A statistic $T = r(X_1, \dots, X_n)$ is a sufficient statistic for θ if and only if for all $\mathbf{x} \in \mathbb{R}^n$ and all $\theta \in \Omega$, the joint pdf/pf $f_n(\mathbf{x} \mid \theta)$ can be factored as

$$f_n(\mathbf{x} \mid \theta) = u(\mathbf{x})v(r(\mathbf{x}), \theta)$$

  where function u and v are nonnegative.

- The function u may depend on $\mathbf{x}$ but not on θ
- The function v depends on θ but depends on $\mathbf{x}$ only through the value of the statistic $r(\mathbf{x})$

Both MLEs and Bayesian estimators depend on data only through sufficient statistics.

MLE vs Bayesian Estimation

MLE
Does not always exist
Is not always appropriate
Is not always unique

Bayes
More difficult computationally
Not a single point

Probabilistic Graphical Models

Directed graphical models

- Bayes Nets
- Conditional dependence

Undirected graphical models

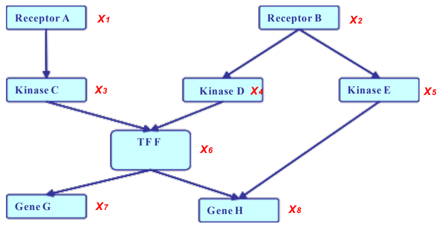
- Markov random fields (MRFs)
- Factor graphs

Two types of GMs

- Directed edges give causality relationships (Bayesian Network or Directed Graphical Model):

$$P(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8)$$

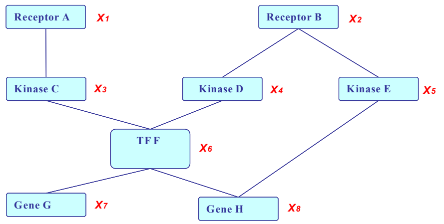
$$= P(X_1)\, P(X_2)\, P(X_3 | X_1)\, P(X_4 | X_2)\, P(X_5 | X_2)\, P(X_6 | X_3, X_4)\, P(X_7 | X_6)\, P(X_8 | X_5, X_6)$$



- Undirected edges simply give correlations between variables (Markov Random Field or Undirected Graphical model):

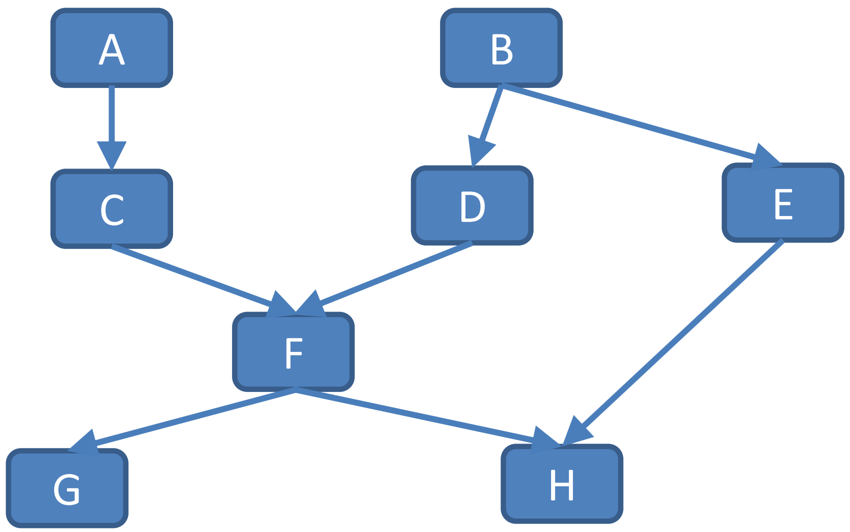
$$P(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8)$$

$$= 1/Z \exp\{E(X_1)+E(X_2)+E(X_3, X_1)+E(X_4, X_2)+E(X_5, X_2) + E(X_6, X_3, X_4)+E(X_7, X_6)+E(X_8, X_5, X_6)\}$$



Directed Graphical Models

A Directed Acyclic Graph



A joint Probability Distribution

$$P(A,B,C,D,E,F,G,H)$$

$$P(A,\dots,H) = \prod_{V\in\{A,\dots,H\}} P(V|Pa_G(V))$$

(Local) Markov Condition:
Every variable is independent of its non-descendants given its parents (in the graph)

Example: Expert systems

- Beinlich et al. 1989
- Encodes medical knowledge
- Patient monitoring system
  - Measurements:
    - Blood pressure 120/80 mmHg
    - Heart rate 80/min
    - Respiratory rate 10/min
    - …
  - Query:
    - Pr(kinked tube=true | measurements) = ?

The ALARM Monitoring System: A Case Study with two Probabilistic Inference Techniques for Belief Networks

Ingo A. Beinlich, M.D., H. J. Suermondt, R. Martin Chavez, Gregory F. Cooper, M.D., Ph.D.

Section on Medical Informatics, Stanford University School of Medicine, Stanford, California, USA

Abstract ALARM (A Logical Alarm Reduction Mechanism) is a diagnostic application used to explore probabilistic reasoning techniques in belief networks. ALARM implements an alarm message system for patient monitoring; it calculates probabilities for a differential diagnosis based on available evidence. The medical knowledge is encoded in a graphical structure connecting 8 diagnoses, 16 findings and 13 intermediate variables. Two algorithms were applied to this belief network: (1) a message-passing algorithm by Pearl for probability updating in multiply connected networks using the method of conditioning; and (2) the Lauritzen–Spiegelhalter algorithm for local probability computations on graphical structures. The characteristics of both algorithms are analyzed and their specific applications and time complexities are shown.

Introduction

The goal of the ALARM monitoring system is to provide specific text messages advising the user of possible problems. This is a diagnostic task, and we have chosen to represent the relevant knowledge in the language of a belief network *(Fig.1)*. This graphical representation [Pearl 86b] facilitates the integration of qualitative and quantitative knowledge, the assessment of multiple faults, as required by our domain, and nonmonotonic and bidirectional reasoning.

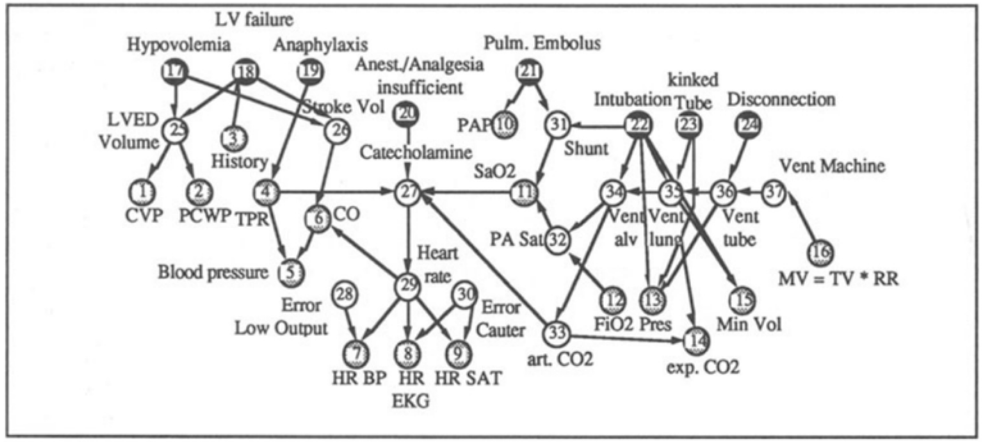
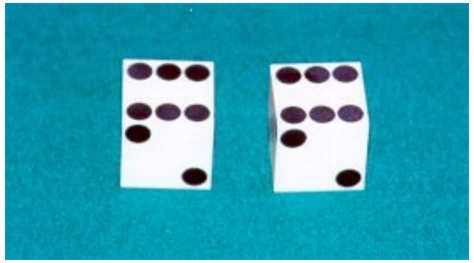


Fig. 1 The ALARM network representing causal relationships is shown with diagnostic (●), intermediate (○) and measurement (◉) nodes. CO: cardiac output, CVP: central venous pressure, LVED volume: left ventricular end-diastolic volume, LV failure: left ventricular failure, MV: minute ventilation, PA Sat: pulmonary artery oxygen saturation, PAP: pulmonary artery pressure, PCWP: pulmonary capillary wedge pressure, Pres: breathing pressure, RR: respiratory rate, TPR: total peripheral resistance, TV: tidal volume

© Eric Xing @ CMU, 2005-2020

Example: The Dishonest Casino



- A casino has two dice:
  - Fair die
    - P(1) = P(2) = P(3) = P(5) = P(6) = 1/6
  - Loaded die
    - P(1) = P(2) = P(3) = P(5) = 1/10
    - P(6) = 1/2
  - Casino player switches back-&-forth between fair and loaded die once every 20 turns

- Game:
  - You bet $1
  - You roll (always with a fair die)
  - Casino player rolls (maybe with fair die, maybe with loaded die)
  - Highest number wins $2

Puzzles regarding the dishonest casino

GIVEN: A sequence of rolls by the casino player

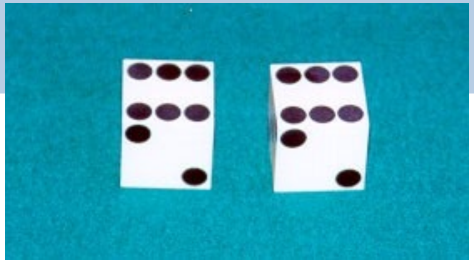
1245526462146146136136661664661636616366163616515615115146123562344

QUESTION

- How likely is this sequence, given our model of how the casino works?
  - This is the EVALUATION problem
- What portion of the sequence was generated with the fair die, and what portion with the loaded die?
  - This is the DECODING question
- How “loaded” is the loaded die? How “fair” is the fair die? How often does the casino player change from fair to loaded, and back?
  - This is the LEARNING question

© Eric Xing @ CMU, 2005-2020

Knowledge Engineering



- Picking variables
  - Observed
  - Hidden
  - Discrete
  - Continuous
- Picking structure
  - CAUSAL
  - Generative
  - Coupling
- Picking Probabilities
  - “Natural”
  - Zero probabilities
  - Orders of magnitudes
  - Relative values

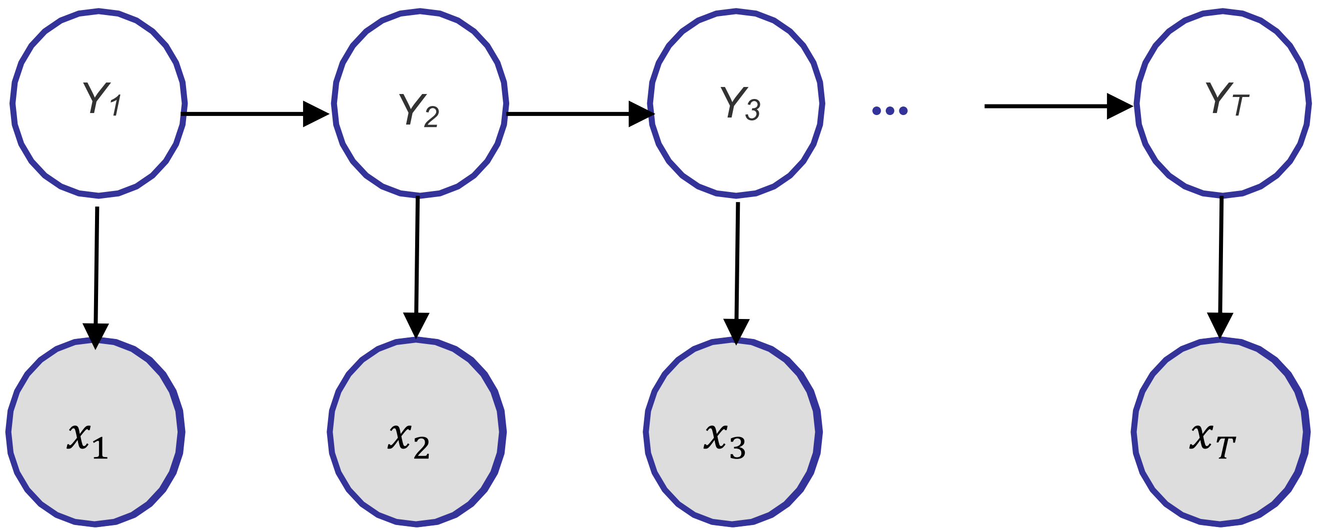
Hidden Markov Model

The underlying source:

Speech signal
genome function
dice

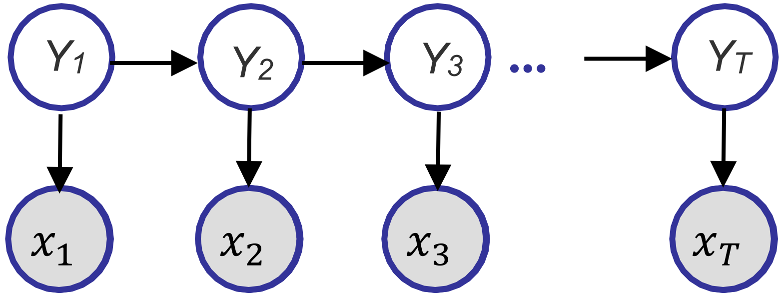
The sequence:

Phonemes
DNA sequence
sequence of rolls



© Eric Xing @ CMU, 2005-2020

Probability of a parse



- Given a sequence $\mathbf{x} = x_1 \ldots\ldots x_T$
  and a parse $\mathbf{y} = y_1, \ldots\ldots, y_T,$
- To find how likely is the parse:
  (given our HMM and the sequence)

$$p(\mathbf{x}, \mathbf{y}) = p(x_1 \ldots\ldots x_T, y_1, \ldots\ldots, y_T) \qquad \text{(Joint probability)}$$

$$= p(y_1)\, p(x_1 | y_1)\, p(y_2 | y_1)\, p(x_2 | y_2) \ldots p(y_T | y_{T-1})\, p(x_T | y_T)$$

$$= p(y_1)\, P(y_2 | y_1) \ldots p(y_T | y_{T-1}) \times p(x_1 | y_1)\, p(x_2 | y_2) \ldots p(x_T | y_T)$$

$$= p(y_1, \ldots\ldots, y_T)\, p(x_1 \ldots\ldots x_T | y_1, \ldots\ldots, y_T)$$

- Marginal probability: $p(\mathbf{x}) = \sum_{\mathbf{y}} p(\mathbf{x},\mathbf{y}) =$ $_y$
- Posterior probability: $p(\mathbf{y} | \mathbf{x}) = p(\mathbf{x},\mathbf{y}) / p(\mathbf{x})$

- We will learn how to do this efficiently (polynomial time)

© Eric Xing @ CMU, 2005-2020

Bayesian Network

- A BN is a directed graph whose nodes represent the random variables and whose edges represent direct influence of one variable on another.

- It is a data structure that provides the skeleton for representing **a joint distribution** compactly in a **factorized** way;

- It offers a compact representation for **a set of conditional independence assumptions** about a distribution;

- We can view the graph as encoding a generative sampling process executed by nature, where the value for each variable is selected by nature using a distribution that depends only on its parents. In other words, each variable is a stochastic function of its parents.

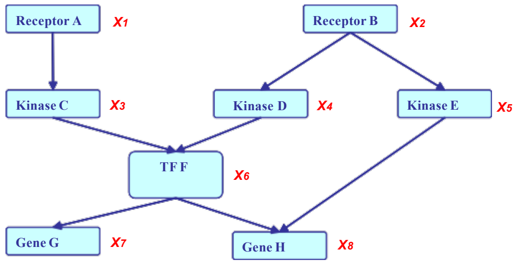
© Eric Xing @ CMU, 2005-2020

Bayesian Network: Factorization

Given a DAG, The most general form of the probability distribution that is consistent with the graph factors according to "node given its parents":

$$P(\mathbf{X}) = \prod_{i=1:d} P(X_i \mid \mathbf{X}_{\pi_i})$$

where $\mathbf{X}_{\pi_i}$ is the set of parents of X_i, d is the number of nodes (variables) in the graph.

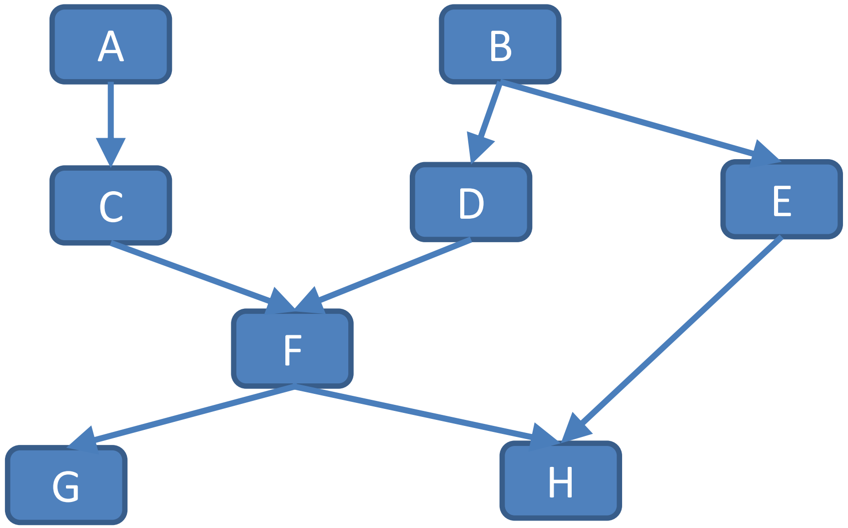


$$P(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8)$$

$$= P(X_1)\, P(X_2)\, P(X_3 | X_1)\, P(X_4 | X_2)\, P(X_5 | X_2)\, P(X_6 | X_3, X_4)\, P(X_7 | X_6)\, P(X_8 | X_5, X_6)$$

Specification of a directed GM

- There are two components to any GM:
  - the *qualitative* specification (graph)
  - the *quantitative* specificationon (jpd)

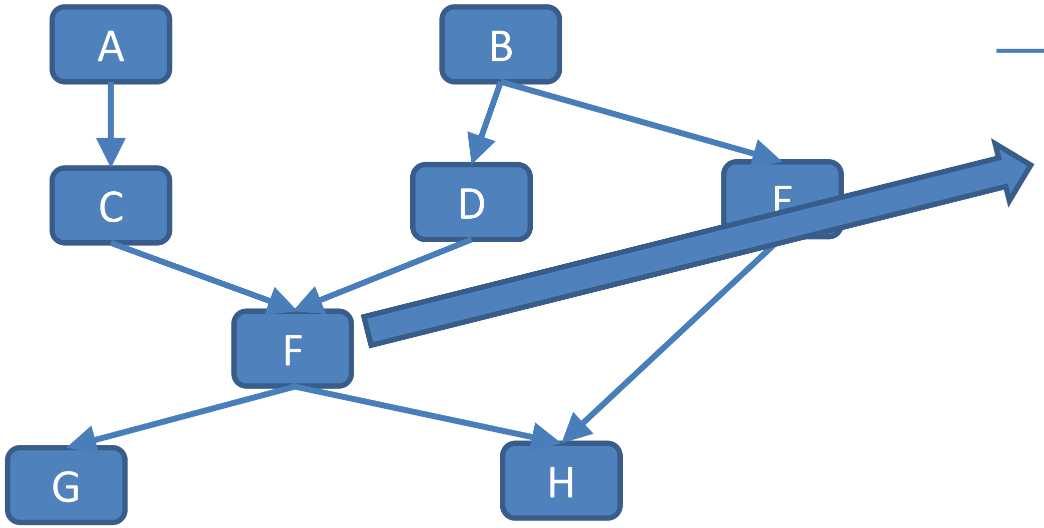


The graph dictates the factorization of the joint probability distribution

$$P(A, \ldots, H) = \prod_{V \in \{A, \ldots, H\}} P(V | Pa_G(V))$$

Specification of a directed GM

- There are two components to any GM:
  - the *qualitative* specification (graph)
  - the *quantitative* specification (jpd)



C	*D*	*P(F | C,D)*	
0	*0*	0.9	0.1
1	*0*	0.2	0.8
0	*1*	0.9	0.1
1	*1*	0.01	0.99

Qualitative Specification

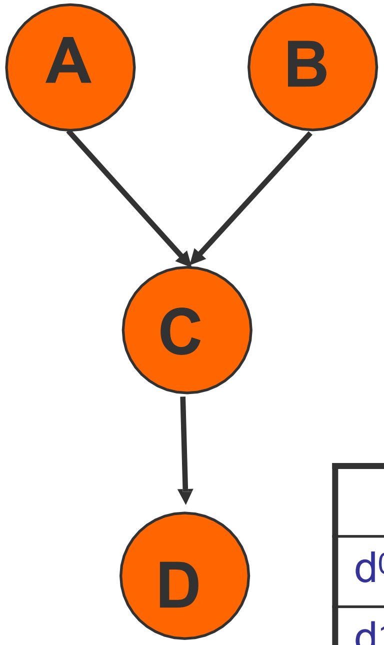
- Where does the qualitative specification come from?
  - Prior knowledge of causal relationships
  - Prior knowledge of modular relationships
  - Assessment from experts
  - Learning from data
  - We simply like a certain architecture (e.g. a layered graph)
  - …

Conditional probability tables (CPTs)

a^0	0.75
a^1	0.25

b^0	0.33
b^1	0.67

P(a,b,c.d) = P(a)P(b)P(c|a,b)P(d|c)



	a^0b^0	a^0b^1	a^1b^0	a^1b^1
c^0	0.45	1	0.9	0.7
c^1	0.55	0	0.1	0.3

	c^0	c^1
d^0	0.3	0.5
d^1	07	0.5

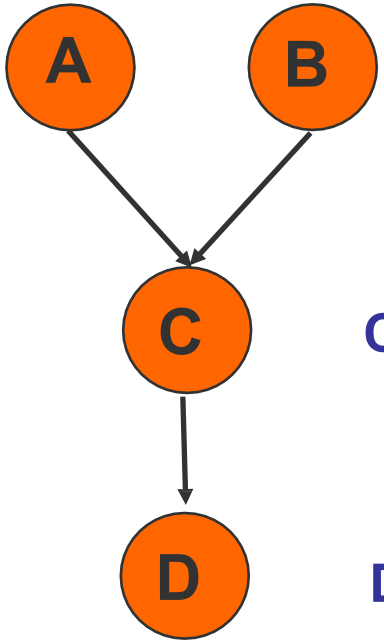
© Eric Xing @ CMU, 2005-2020

Qualitative Specification

Conditional probability density func. (CPDs)

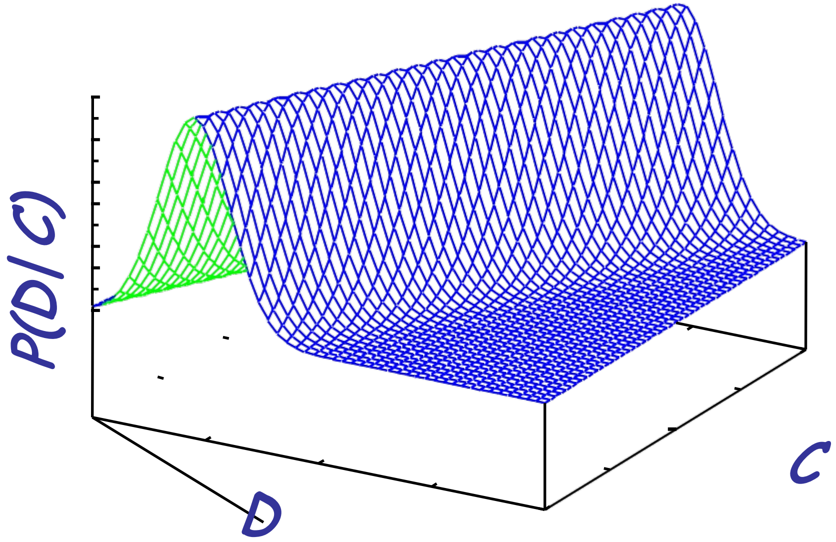
$A \sim N(\mu_a, \Sigma_a)$ $B \sim N(\mu_b, \Sigma_b)$

$$P(a,b,c.d) = P(a)P(b)P(c|a,b)P(d|c)$$



$C \sim N(A+B, \Sigma_c)$

$D \sim N(\mu_d+C, \Sigma_d)$

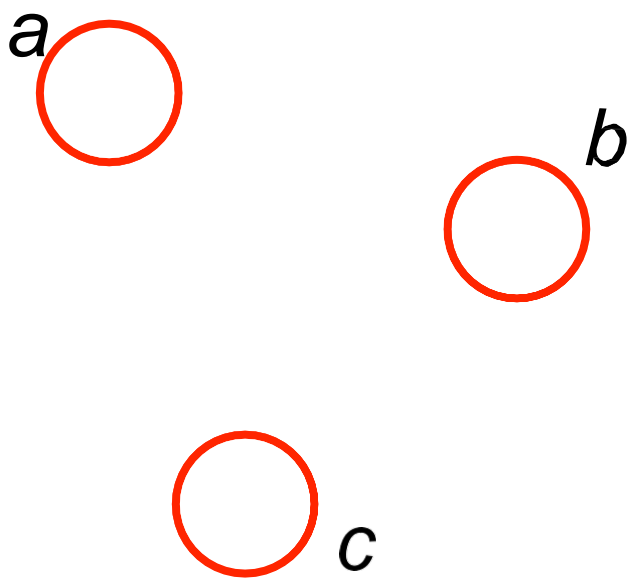


Qualitative Specification

- Where does the qualitative specification come from?
  - Prior knowledge of causal relationships
  - Prior knowledge of modular relationships
  - Assessment from experts
  - Learning from data
  - We simply like a certain architecture (e.g. a layered graph)
  - …
- Graphs imply some conditional independencies. (wherever you got them)
  - What does this mean?
  - For every distribution that factorizes according to the graph

Implied Independencies.

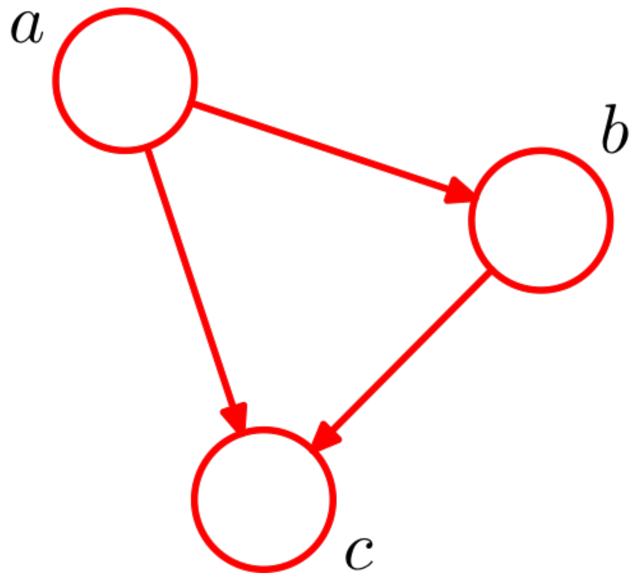
Are a and b independent ($a \perp b$ *)?*



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

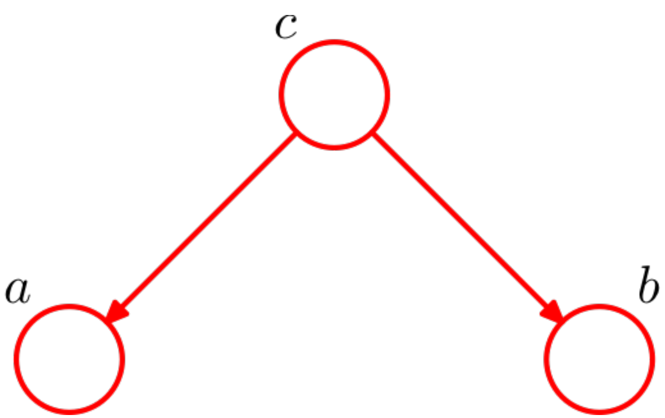
Implied Independencies

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

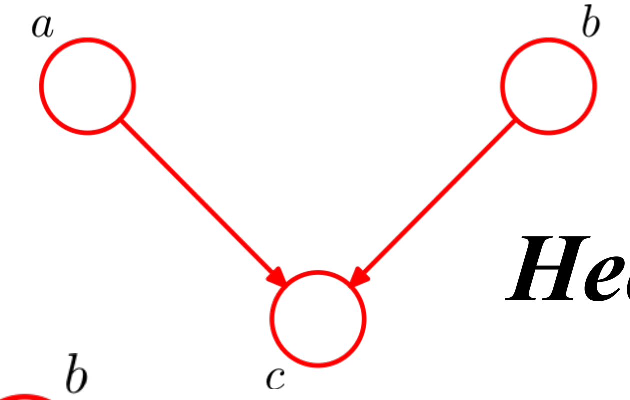


Note there are **no conditional independencies** (fully connected graph)

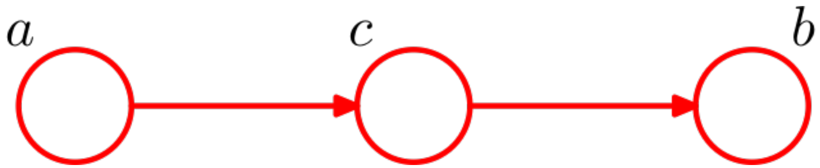
Three interesting cases



Tail-to-tail

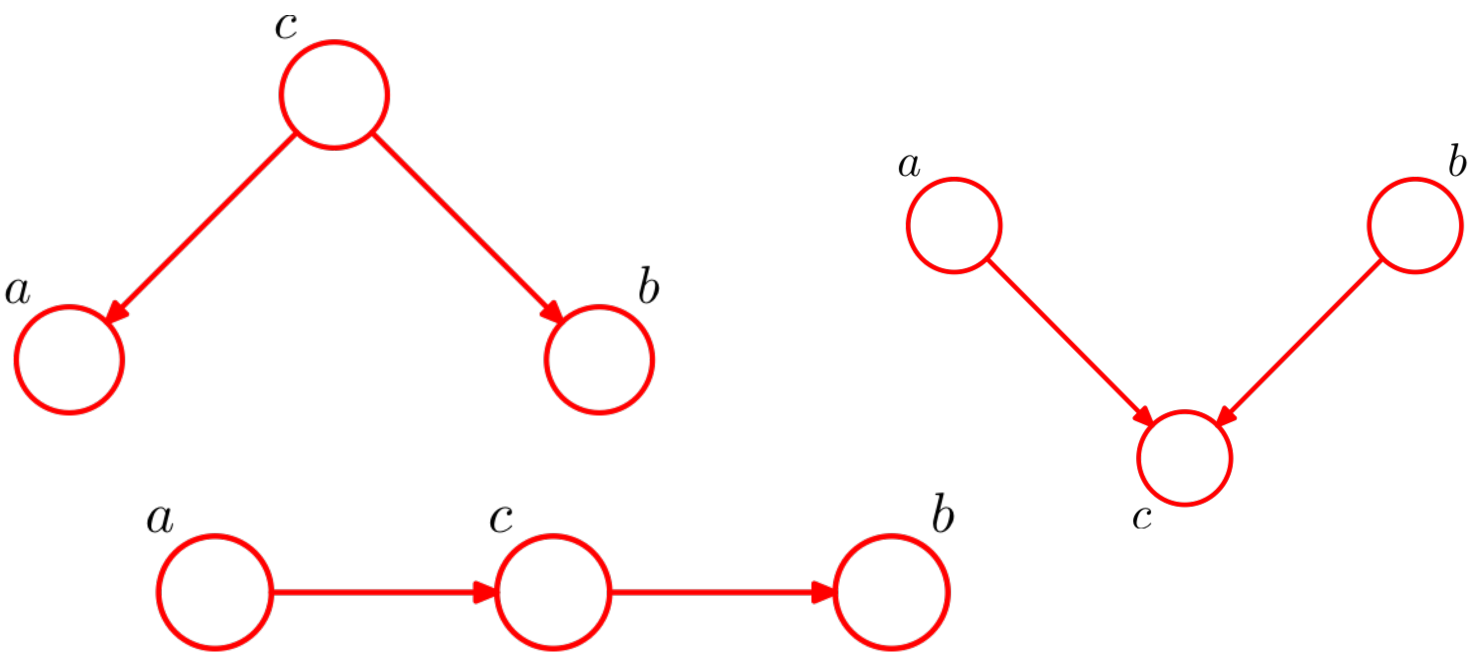


Head-to-head



Head-to-tail

Three interesting cases

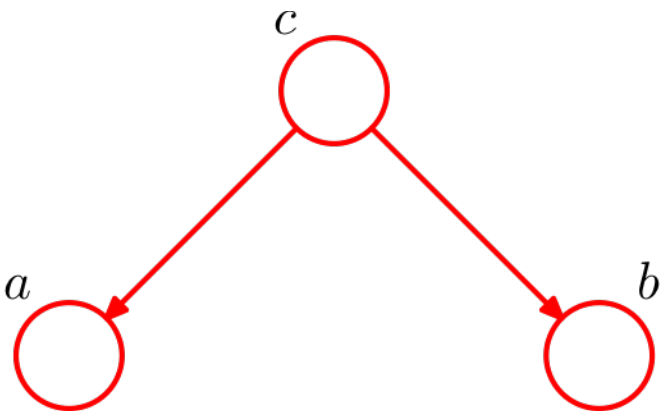


For each case, consider two questions:

1) Is $a \perp b$?

2) Is $a \perp b \mid c$? (i.e. c is observed)

Case one (tail-to-tail)

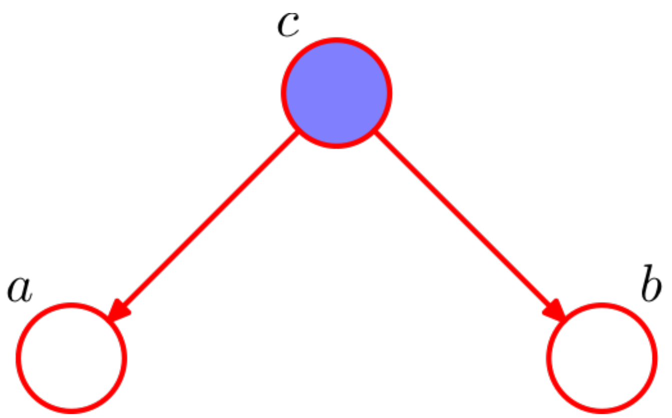


$$a \not\perp b$$

This graph represents $P(a, b, c) = P(c)P(a|c)P(b|c)$

To prove independence, we need to come up with a counter-example

Case one (tail-to-tail)



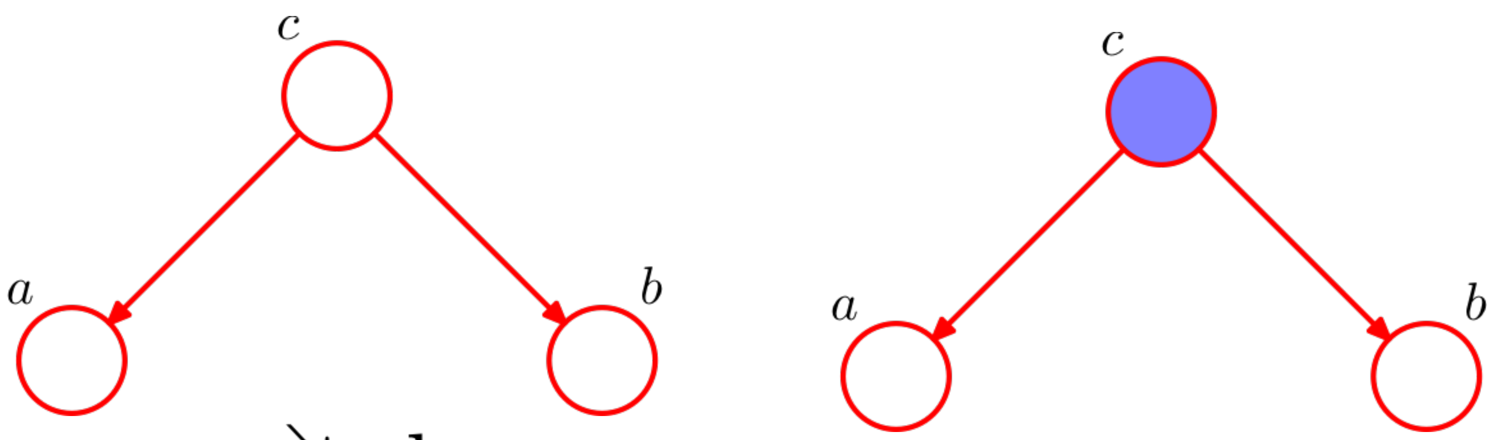
$$a \perp b \mid c$$

$p(a,b,c) = p(c)p(a|c)p(b|c)$ (what the graph represents in general)

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

This is the definition of $a \perp b|c$

Case one (tail-to-tail) summary



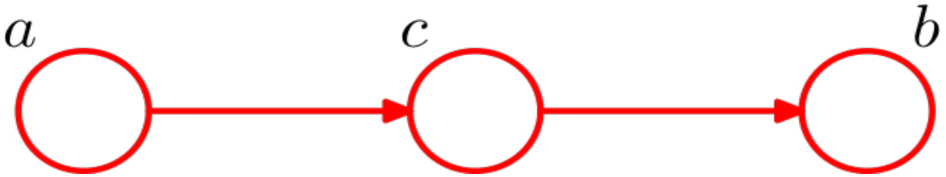
$a \not\perp b$ $\qquad$ $a \perp b \mid c$

Tail-to-tail case

With no conditioning, no independence ($\exists P$)

With conditioning, we have independence

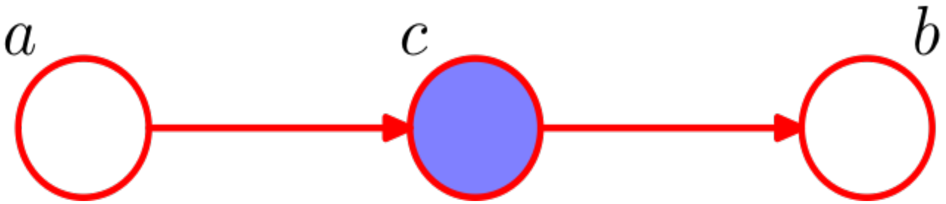
Case two (head-to-tail)



$a \not\perp b$

This graph represents $P(a, b, c) = P(a)P(a|c)P(b|c)$

Case two (head-to-tail)

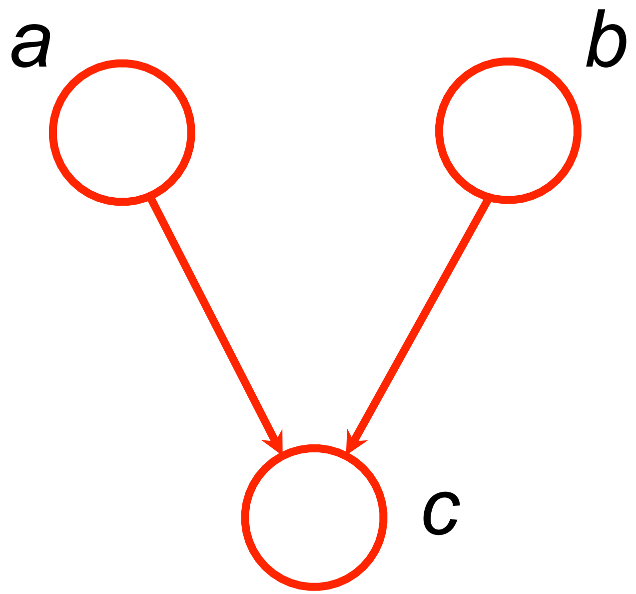


$$a \perp b \mid c$$

$$
\begin{aligned}
p(a,b \mid c) &= \frac{p(a,b,c)}{p(c)} && \text{(definition)} \\
&= \frac{p(a)\,p(c|a)\,p(b|c)}{p(c)} && \text{(from graph)} \\
&= \frac{p(a)\,p(a|c)\,p(c)\,p(b|c)}{p(a)\,p(c)} && \text{(Bayes on } p(c|a)) \\
&= p(a|c)\,p(b|c)
\end{aligned}
$$

Case three (head-to-head)

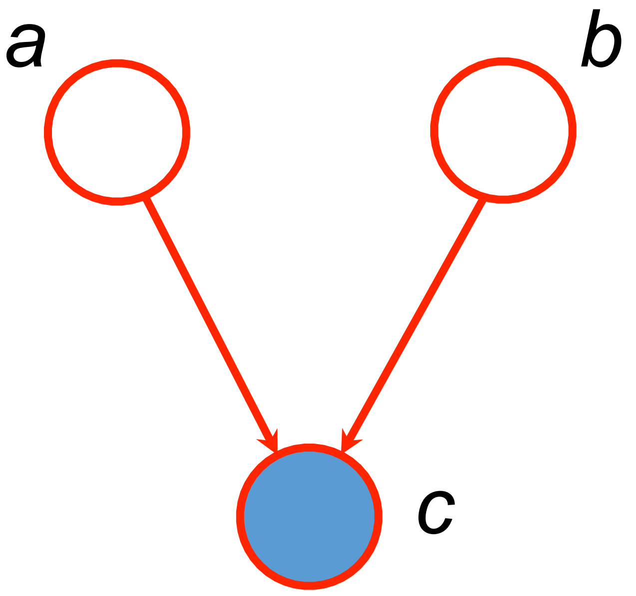
Are a and b independent ($a \perp b$)?



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

Case three (head-to-head)

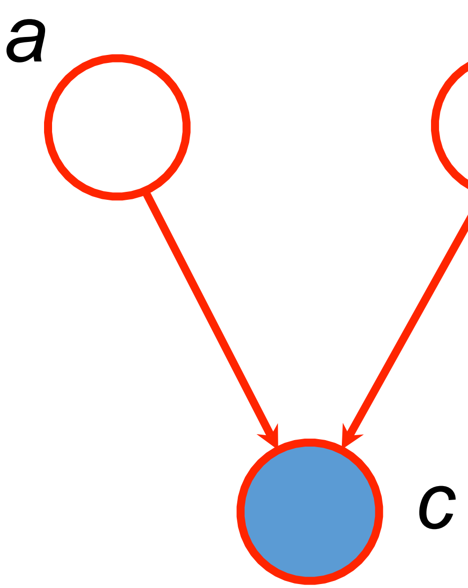
Are a and b conditionally independent ($a \perp b \mid c$ *)?*



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

Case three (head-to-head)

Are a and b conditionally independent ($a \perp b \mid c$ *)?*

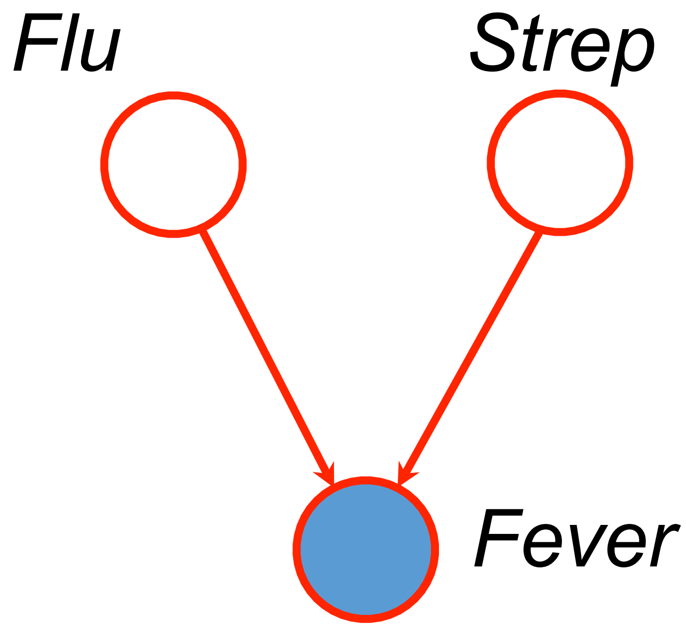


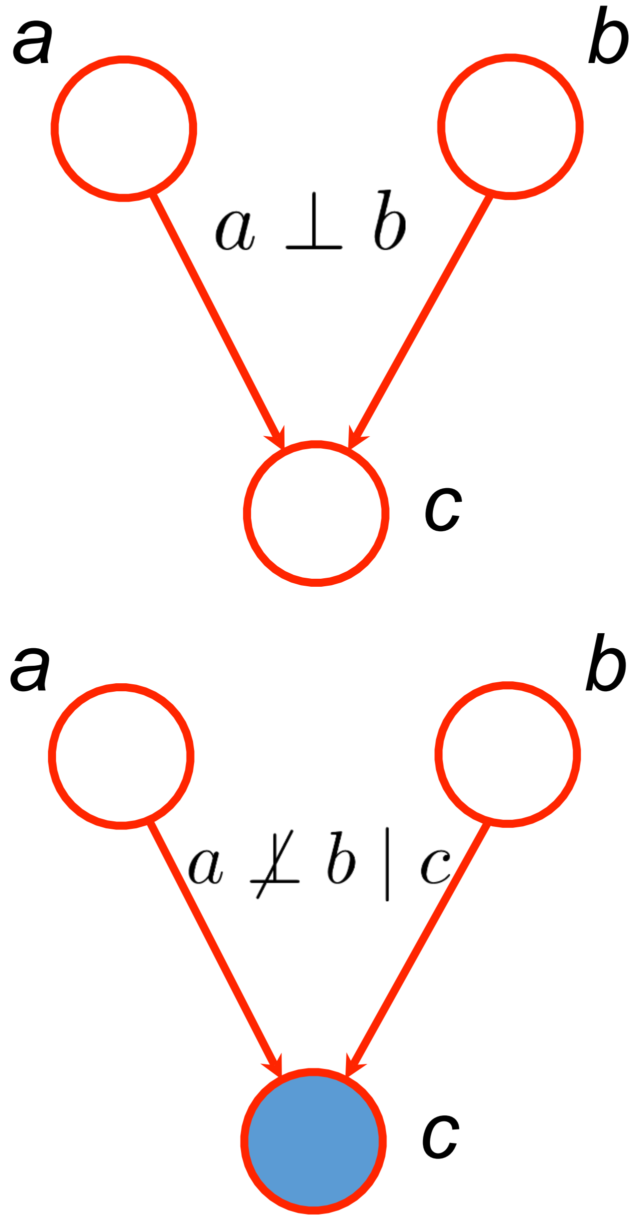
$$p(a,b|c) = \frac{p(a,b,c)}{p(c)} = \frac{p(a)p(b)p(c|a,b)}{p(c)} \neq p(a|c)p(b|c) \quad \text{(in general)}$$

Attempt at algebraic proof.

Unless the algebra reduces to something obviously false, we typically look for a counter example

Case three (head-to-head)





Phenomenon in Bayes networks known as **explaining away**

Summary

Bayesian networks: Graph (DAG)+JPD

JPD factorizes according to the factorization theorem

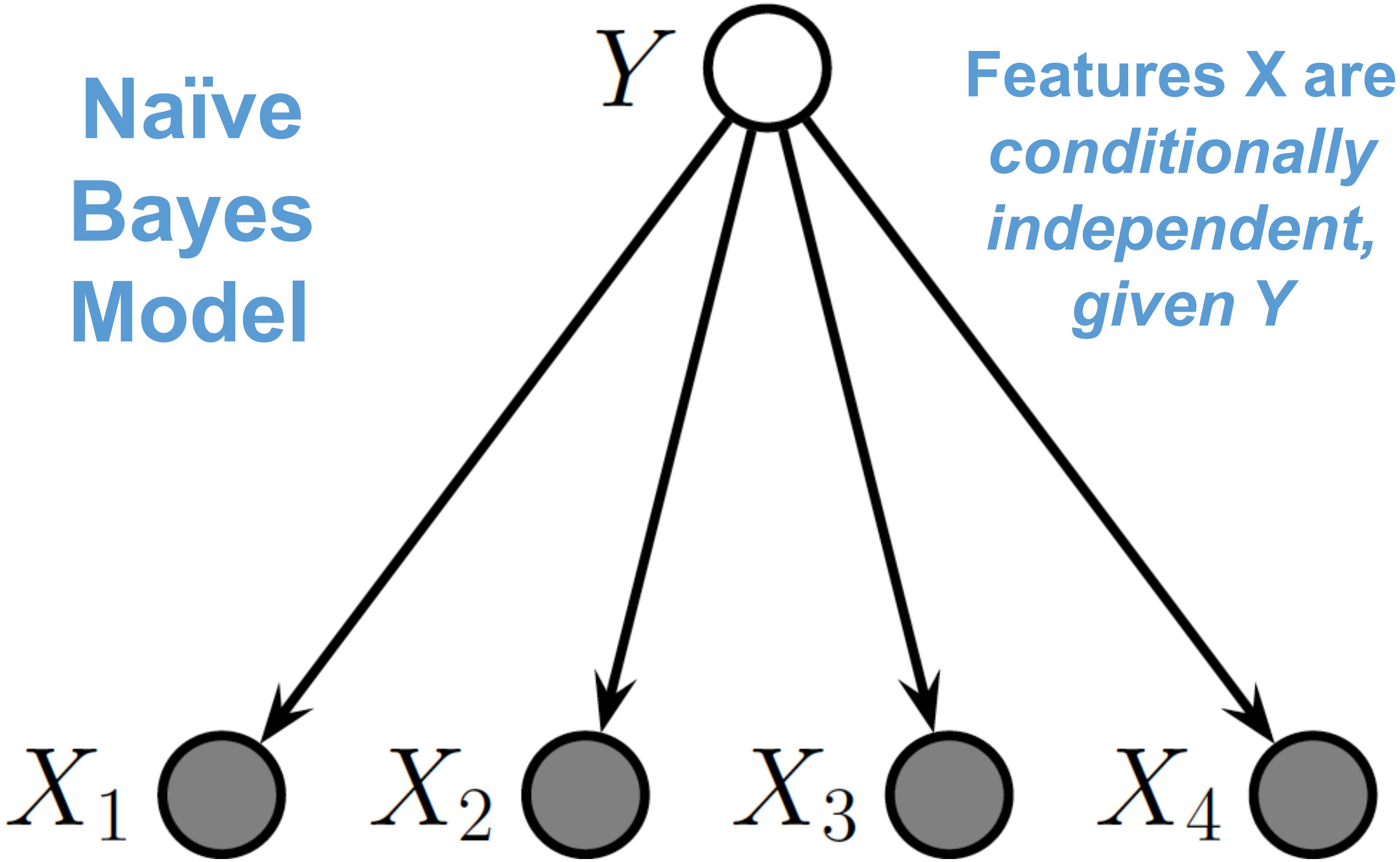
Factorization theorem implies a set of conditional independencies.

Next: A general algorithm for reading independencies from the graphs.

Shading & Plate Notation

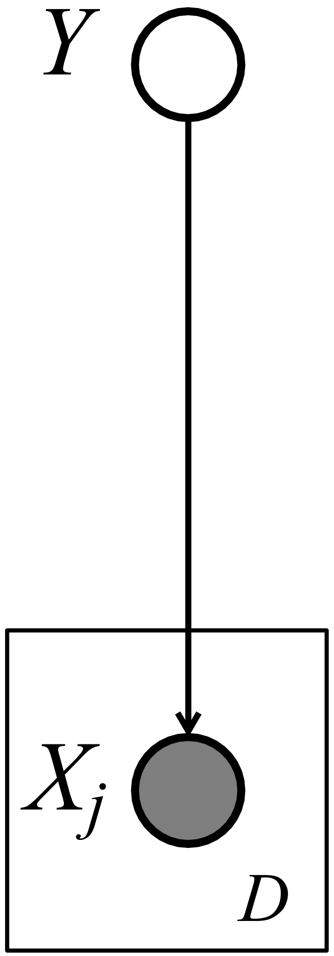
***Convention:** Shaded nodes are observed, open nodes are latent/hidden/unobserved*

Naïve Bayes Model



Features X are conditionally independent, given Y

$$p(y, \mathbf{x}) = p(y) \prod_{j=1}^{D} p(x_j | y)$$



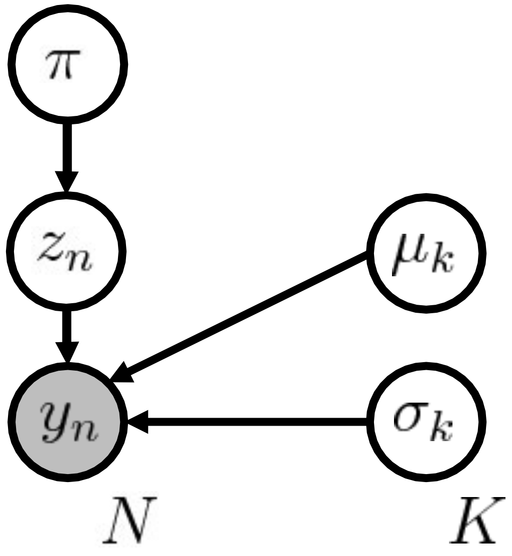
Plates denote replication of random variables

Example: Gaussian Mixture Model

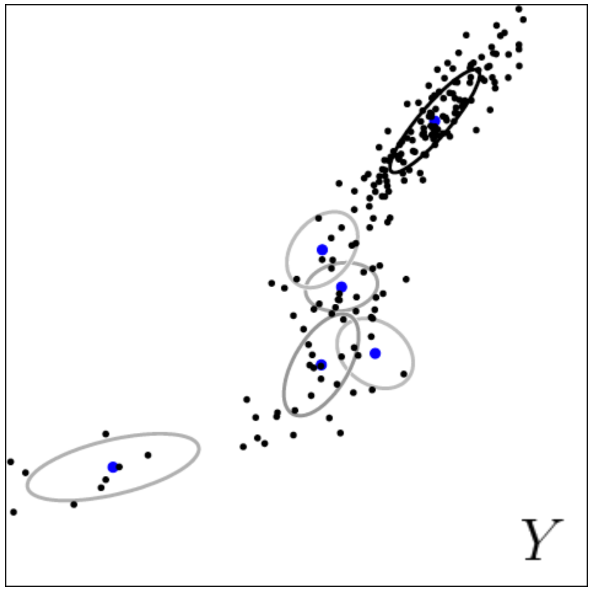
Probability Model

$$\pi \sim \text{Dirichlet}(\cdot)$$
$$\mu_k \sim \mathcal{N}(\cdot)$$
$$\sigma_k \sim \text{Inv-Gamma}(\cdot)$$
$$z_n \mid \pi \sim \text{Cat}(\pi)$$
$$y_n \mid z_n, \mu_{z_n}, \sigma_{z_n} \sim \mathcal{N}(\mu_{z_n}, \sigma_{z_n})$$

Bayes Net



Joint Sample



Sample all nodes with no parents, then children, etc., to terminals. Can sample nodes at same level in parallel.