Bayesian networks: Modeling

CS194-10 Fall 2011 Lecture 21

Outline

- ♢ Overview of Bayes nets
- ♢ Syntax and semantics
- ♢ Examples
- ♢ Compact conditional distributions

Learning with complex probability models

Learning cannot succeed without imposing some prior structure on the hypothesis space (by *constraint* or by *preference*)

Generative models $P(\mathbf{X} \mid \theta)$ support MLE, MAP, and Bayesian learning for domains that (approximately) reflect the assumptions underlying the model:

- ◇ Naive Bayes—conditional independence of attributes given class value
- ◇ Mixture models—domain has a flat, discrete category structure
- ◇ All i.i.d. models—model doesn't change over time
- ◇ Etc.

Would like to express arbitrarily complex and flexible prior knowledge:

- ◇ Some attributes depend on others
- ◇ Categories have hierarchical structure;
  objects may be mixtures of several categories
- ◇ Observations at time t may depend on earlier observations
- ◇ Etc.

Bayesian networks

A simple, graphical notation for conditional independence assertions among a predefined set of random variables X_j, $j = 1, \ldots, D$ and hence for compact specification of arbitrary joint distributions

Syntax:

a set of nodes, one per variable

a directed, acyclic graph (link $\approx$ "directly influences")

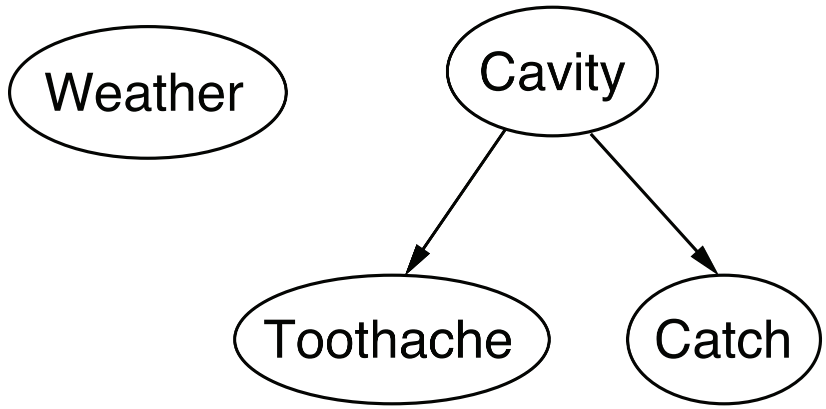
a set of parameters for each node given its parents:

$\theta(X_j | Parents(X_j))$

In the simplest case, parameters consist of a conditional probability table (CPT) giving the distribution over X_j for each combination of parent values

Example

Topology of network encodes conditional independence assertions:



$Weather$ is independent of the other variables

$Toothache$ and $Catch$ are conditionally independent given $Cavity$

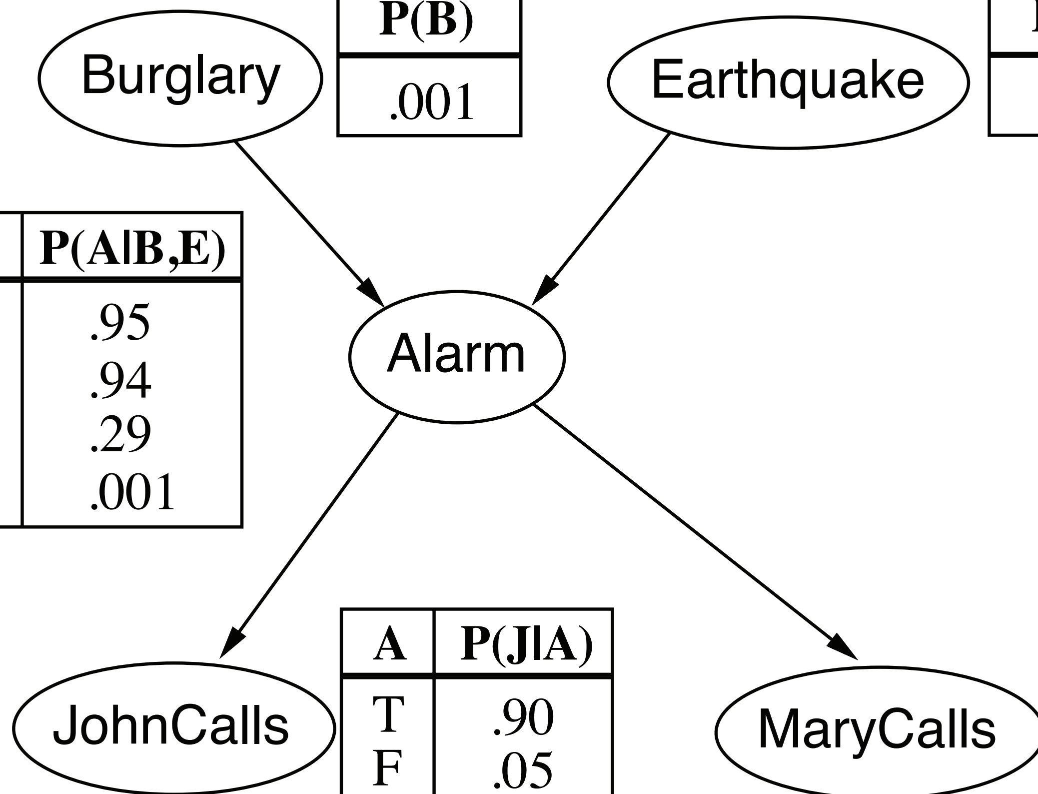
Example

I'm at work, neighbor John calls to say my alarm is ringing, but neighbor Mary doesn't call. Sometimes it's set off by minor earthquakes. Is there a burglar?

Variables: $Burglar$, $Earthquake$, $Alarm$, $JohnCalls$, $MaryCalls$
Network topology reflects "causal" knowledge:

- A burglar can set the alarm off
- An earthquake can set the alarm off
- The alarm can cause Mary to call
- The alarm can cause John to call

Example contd.



P(B)
.001

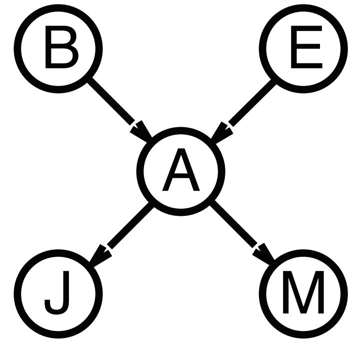
P(E)
.002

B	E	P(A|B,E)
T	T	.95
T	F	.94
F	T	.29
F	F	.001

A	P(J|A)
T	.90
F	.05

A	P(M|A)
T	.70
F	.01

Compactness



A CPT for Boolean X_j with L Boolean parents has 2^L rows for the combinations of parent values

Each row requires one parameter p for $X_j = true$ (the parameter for $X_j = false$ is just $1 - p$)

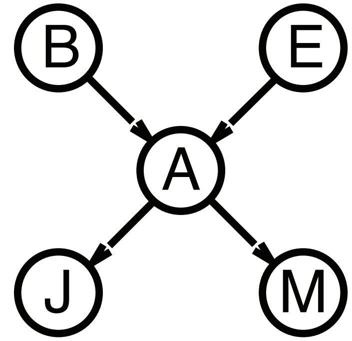
If each variable has no more than L parents, the complete network requires $O(D \cdot 2^L)$ parameters

I.e., grows linearly with D, vs. $O(2^D)$ for the full joint distribution

For burglary net, $1 + 1 + 4 + 2 + 2 = 10$ parameters (vs. $2^5 - 1 = 31$)

Global semantics

Global semantics defines the full joint distribution as the product of the local conditional distributions:



$$P(x_1, \ldots, x_D) = \prod_{j=1}^{D} \theta(x_j | parents(X_j)) \ .$$

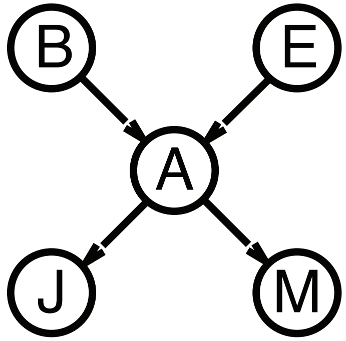
e.g., $P(j \wedge m \wedge a \wedge \neg b \wedge \neg e)$

$=$

Global semantics

Global semantics defines the full joint distribution
as the product of the local conditional distributions:

$$P(x_1, \ldots, x_D) = \prod_{j=1}^{D} \theta(x_j | parents(X_j)) \ .$$

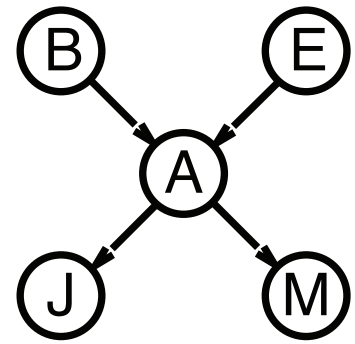


e.g., $P(j \wedge m \wedge a \wedge \neg b \wedge \neg e)$

$$\begin{aligned} &= \theta(j|a)\theta(m|a)\theta(a|\neg b, \neg e)\theta(\neg b)\theta(\neg e) \\ &= 0.9 \times 0.7 \times 0.001 \times 0.999 \times 0.998 \\ &\approx 0.00063 \end{aligned}$$

Global semantics

Global semantics defines the full joint distribution as the product of the local conditional distributions:



$$P(x_1, \ldots, x_D) = \prod_{j=1}^{D} \theta(x_j | parents(X_j)) \ .$$

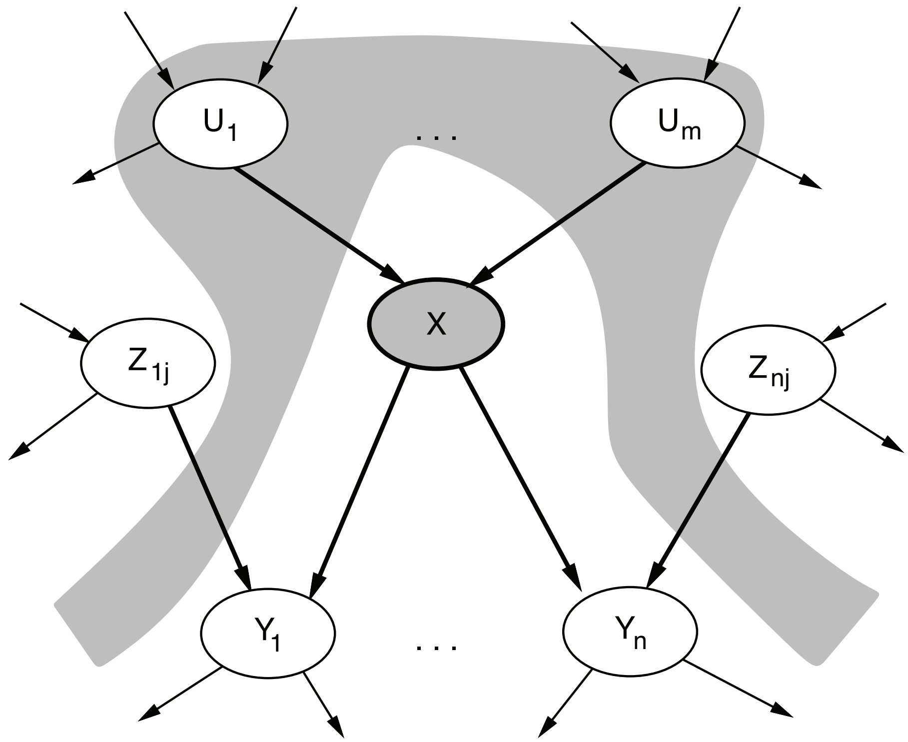
e.g., $P(j \wedge m \wedge a \wedge \neg b \wedge \neg e)$

$$\begin{aligned} &= \theta(j|a)\theta(m|a)\theta(a|\neg b, \neg e)\theta(\neg b)\theta(\neg e) \\ &= 0.9 \times 0.7 \times 0.001 \times 0.999 \times 0.998 \\ &\approx 0.00063 \end{aligned}$$

Theorem: $\theta(X_j | Parents(X_j)) = \mathbf{P}(X_j | Parents(X_j))$

Local semantics

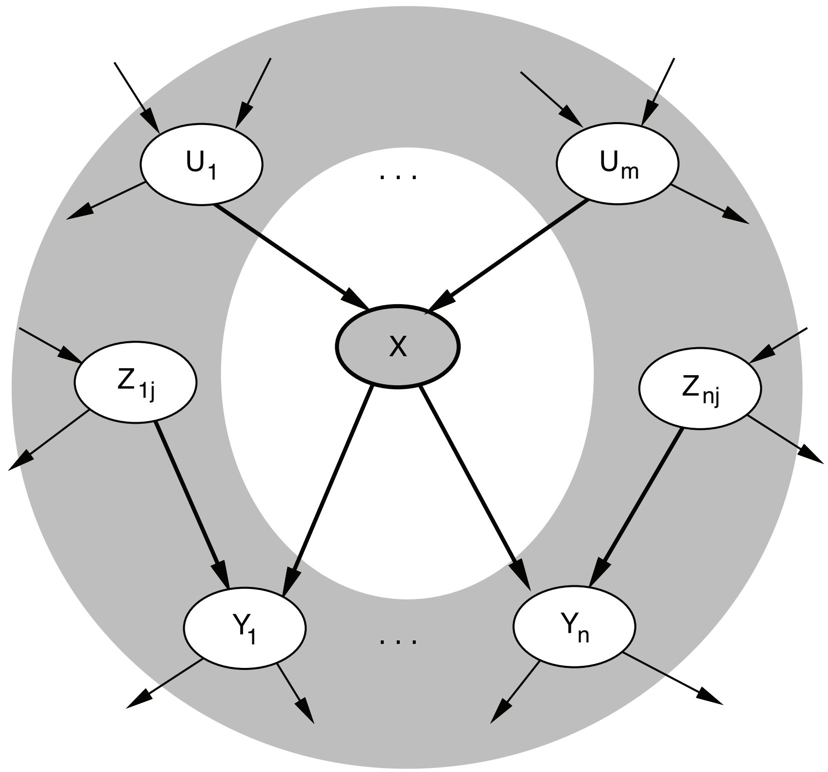
Local semantics: each node is conditionally independent of its nondescendants given its parents



Theorem: Local semantics $\Leftrightarrow$ global semantics

Markov blanket

Each node is conditionally independent of all others given its Markov blanket: parents + children + children's parents



Constructing Bayesian networks

Need a method such that a series of locally testable assertions of conditional independence guarantees the required global semantics

1. Choose an ordering of variables $X_1, \ldots, X_D$
2. For $j = 1$ to D
   add X_j to the network
   select parents from $X_1, \ldots, X_{j-1}$ such that
   $\mathbf{P}(X_j | Parents(X_j)) = \mathbf{P}(X_j | X_1,\, \ldots,\, X_{j-1})$
   i.e., X_j is conditionally independent of other variables given parents

This choice of parents guarantees the global semantics:

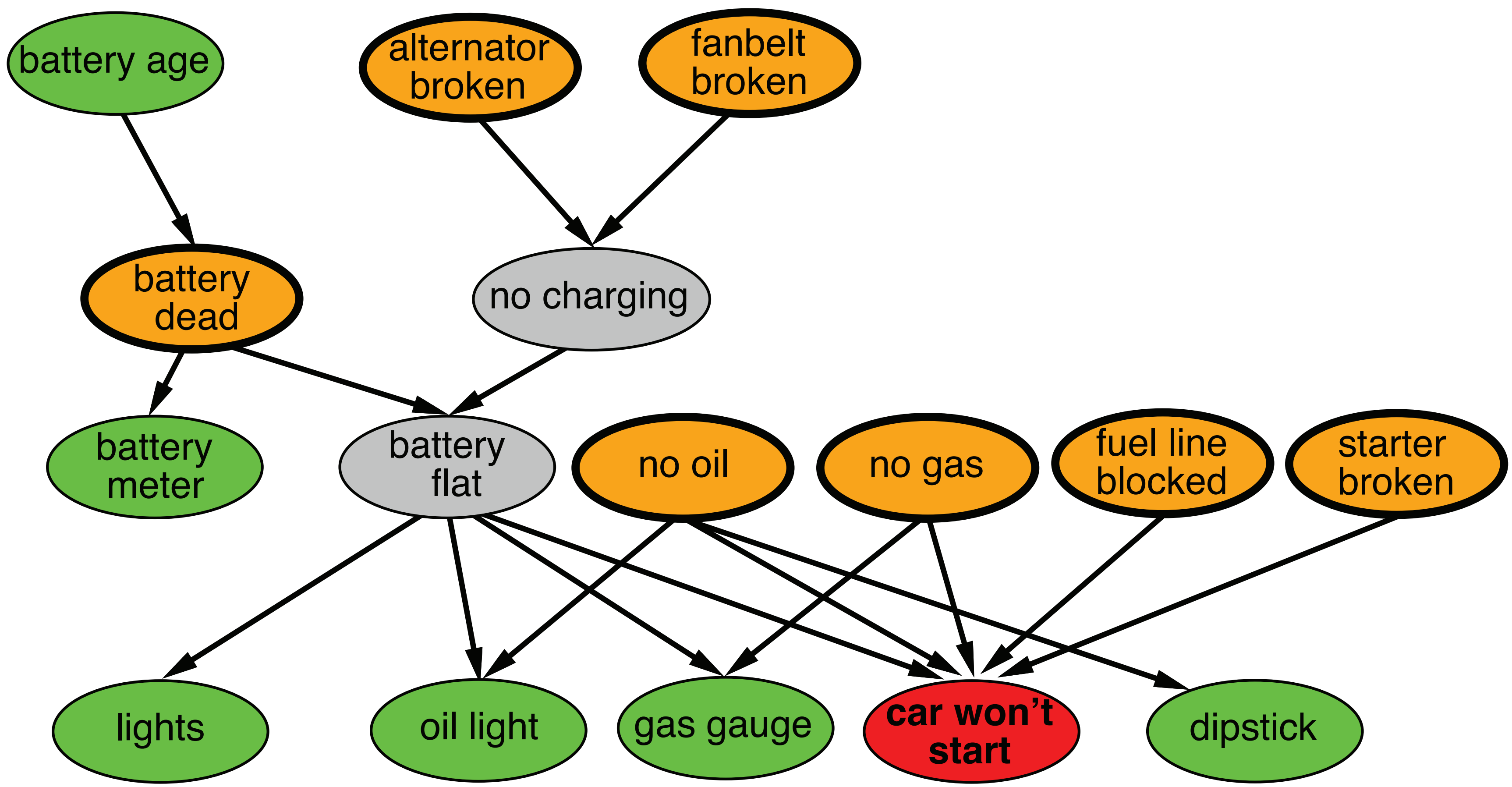
$$\begin{aligned} \mathbf{P}(X_1, \ldots, X_D) &= \prod_{j=1}^{D} \mathbf{P}(X_j | X_1,\, \ldots,\, X_{j-1}) \quad \text{(chain rule)} \\ &= \prod_{j=1}^{D} \mathbf{P}(X_j | Parents(X_j)) \quad \text{(by construction)} \end{aligned}$$

Example: Car diagnosis

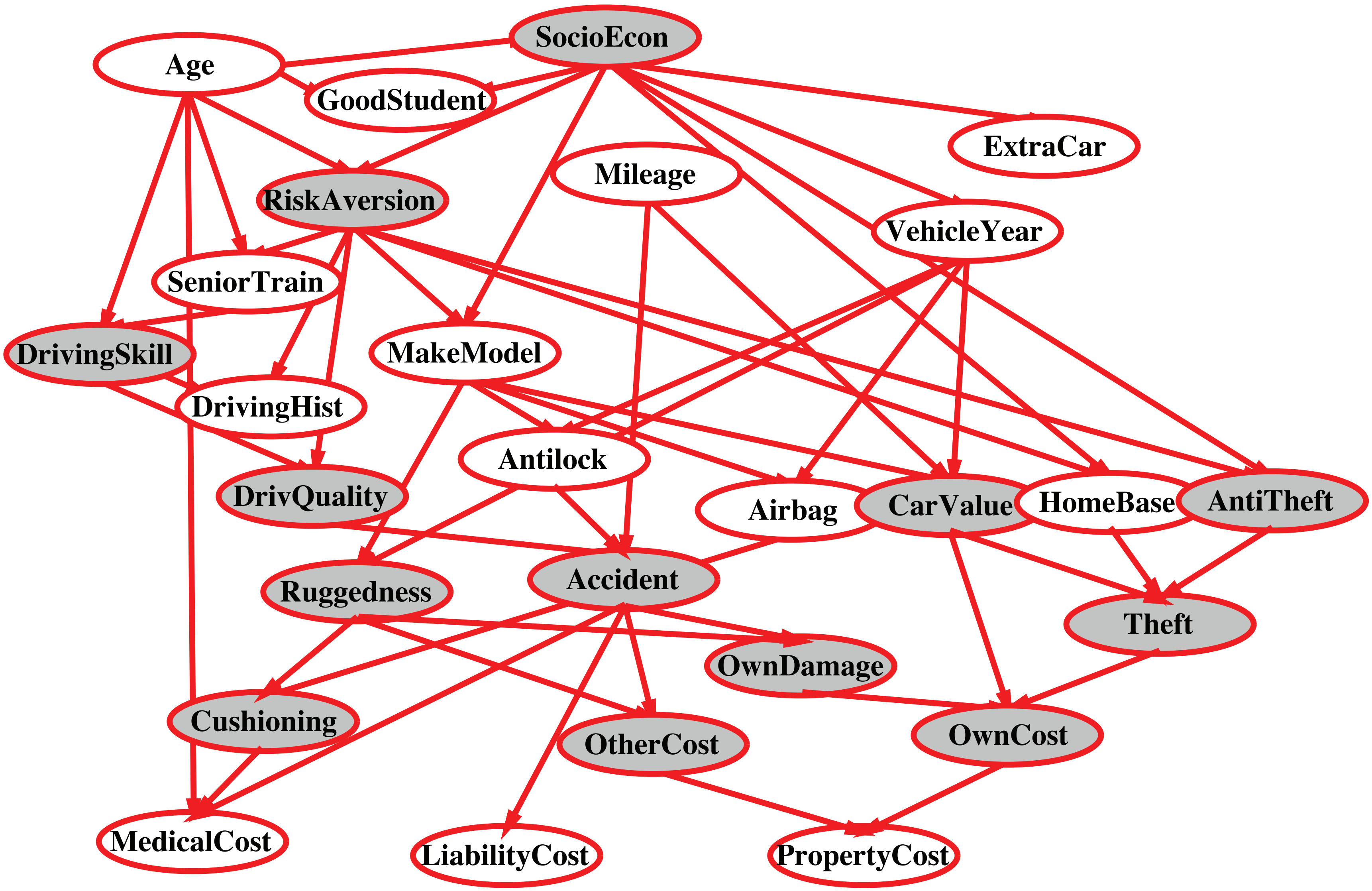
Initial evidence: car won't start

Testable variables (green), "broken, so fix it" variables (orange)

Hidden variables (gray) ensure sparse structure, reduce parameters



Example: Car insurance



Compact conditional distributions

CPT grows exponentially with number of parents
CPT becomes infinite with continuous-valued parent or child

Solution: canonical distributions that are defined compactly

Deterministic nodes are the simplest case:
$X = f(Parents(X))$ for some function f

E.g., Boolean functions
$NorthAmerican \Leftrightarrow Canadian \lor US \lor Mexican$

E.g., numerical relationships among continuous variables

$$\frac{\partial LakeLevel}{\partial t} = \text{inflow + precipitation - outflow - evaporation}$$

Compact conditional distributions contd.

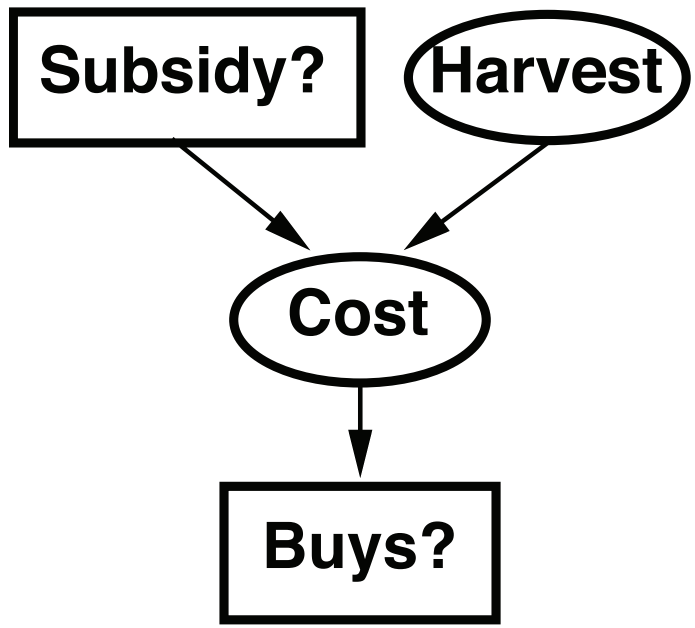
Noisy-OR distributions model multiple noninteracting causes
1) Parents $U_1 \ldots U_L$ include all causes (can add leak node)
2) Independent failure probability q_ℓ for each cause alone
$\Rightarrow P(X|U_1 \ldots U_M, \neg U_{M+1} \ldots \neg U_L) = 1 - \prod_{\ell=1}^{M} q_\ell$

$Cold$	Flu	$Malaria$	$P(Fever)$	$P(\neg Fever)$
F	F	F	$\mathbf{0.0}$	1.0
F	F	T	0.9	$\mathbf{0.1}$
F	T	F	0.8	$\mathbf{0.2}$
F	T	T	0.98	$0.02 = 0.2 \times 0.1$
T	F	F	0.4	$\mathbf{0.6}$
T	F	T	0.94	$0.06 = 0.6 \times 0.1$
T	T	F	0.88	$0.12 = 0.6 \times 0.2$
T	T	T	0.988	$0.012 = 0.6 \times 0.2 \times 0.1$

Number of parameters **linear** in number of parents

Hybrid (discrete+continuous) networks

Discrete (*Subsidy?* and *Buys?*); continuous (*Harvest* and *Cost*)



Option 1: discretization—possibly large errors, large CPTs
Option 2: finitely parameterized canonical families

1) Continuous variable, discrete+continuous parents (e.g., *Cost*)
2) Discrete variable, continuous parents (e.g., *Buys?*)

Continuous child variables

Need one conditional density function for child variable given continuous parents, for each possible assignment to discrete parents

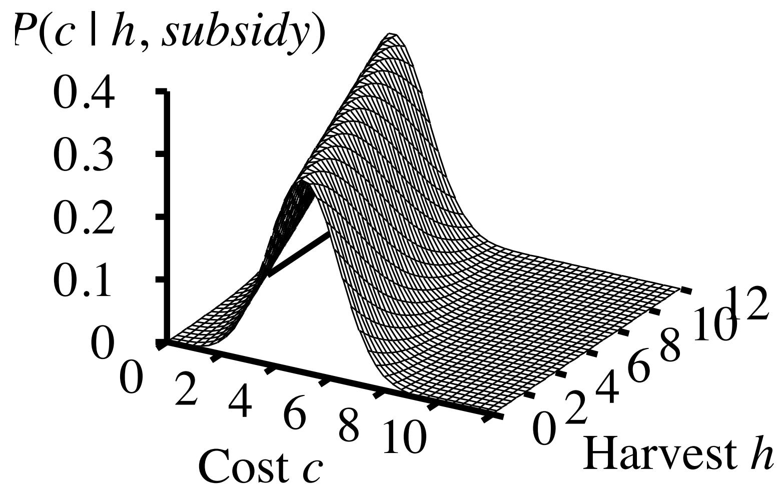
Most common is the linear Gaussian model, e.g.,:

$$
\begin{aligned}
& P(Cost\!=\!c|Harvest\!=\!h, Subsidy?\!=\!true) \\
&= N(a_t h + b_t, \sigma_t)(c) \\
&= \frac{1}{\sigma_t\sqrt{2\pi}} exp\left(-\frac{1}{2}\left(\frac{c-(a_t h + b_t)}{\sigma_t}\right)^2\right)
\end{aligned}
$$

Mean $Cost$ varies linearly with $Harvest$, variance is fixed

Linear variation is unreasonable over the full range
but works OK if the **likely** range of $Harvest$ is narrow

Continuous child variables



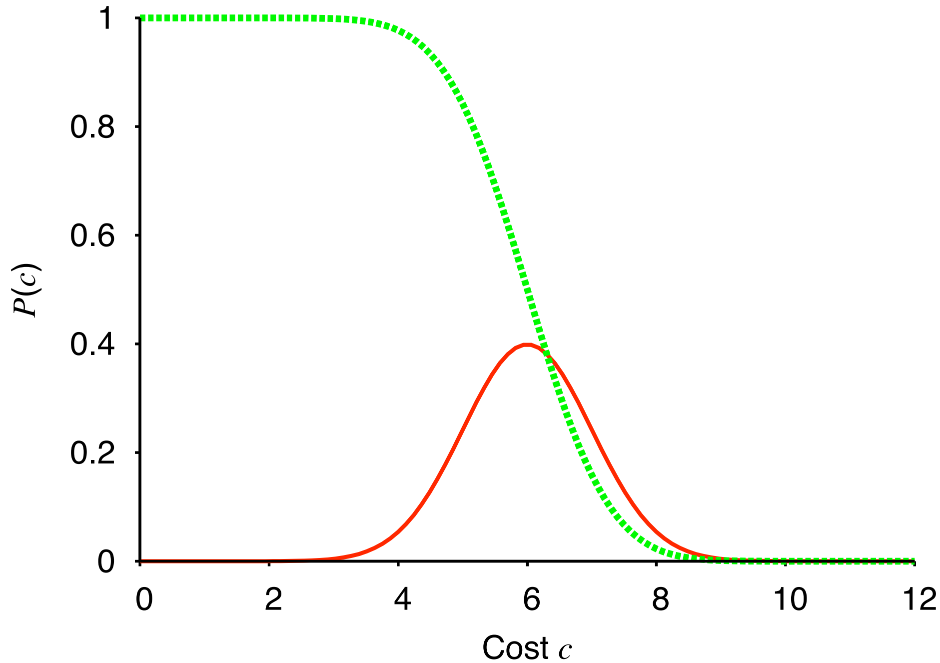
All-continuous network with LG distributions

⇒ full joint distribution is a multivariate Gaussian

Discrete+continuous LG network is a conditional Gaussian network i.e., a multivariate Gaussian over all continuous variables for each combination of discrete variable values

Discrete variable w/ continuous parents

Probability of $Buys?$ given $Cost$ should be a "soft" threshold:



Probit distribution uses integral of Gaussian:

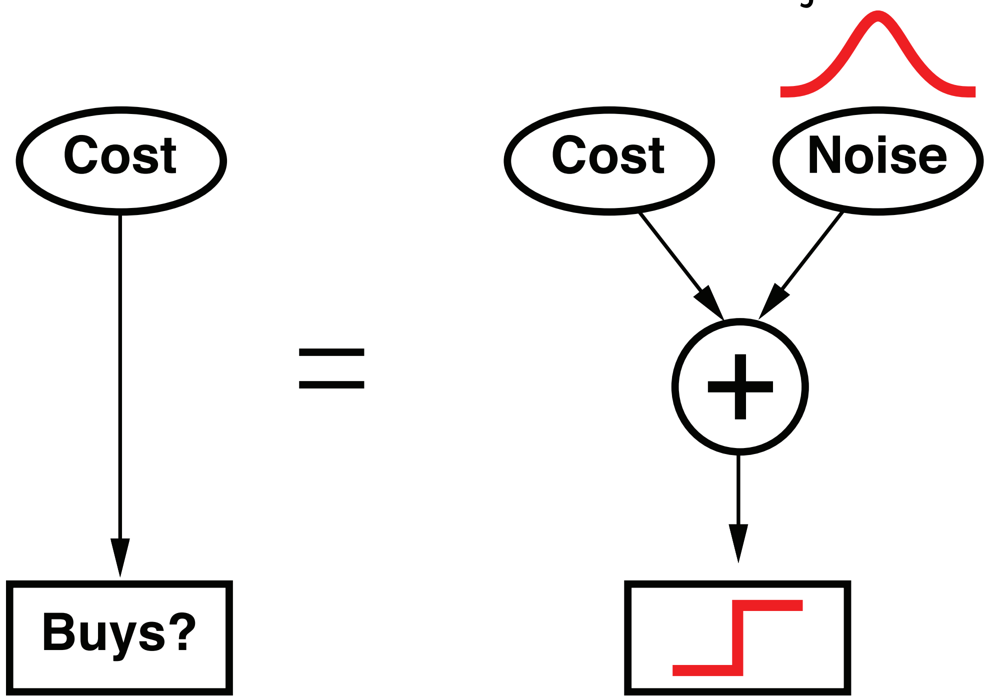
$\Phi(x) = \int_{-\infty}^{x} N(0,1)(x)dx$

$P(Buys? = true \mid Cost = c) = \Phi((-c + \mu)/\sigma)$

Why the probit?

1. It's sort of the right shape

2. Can view as hard threshold whose location is subject to noise

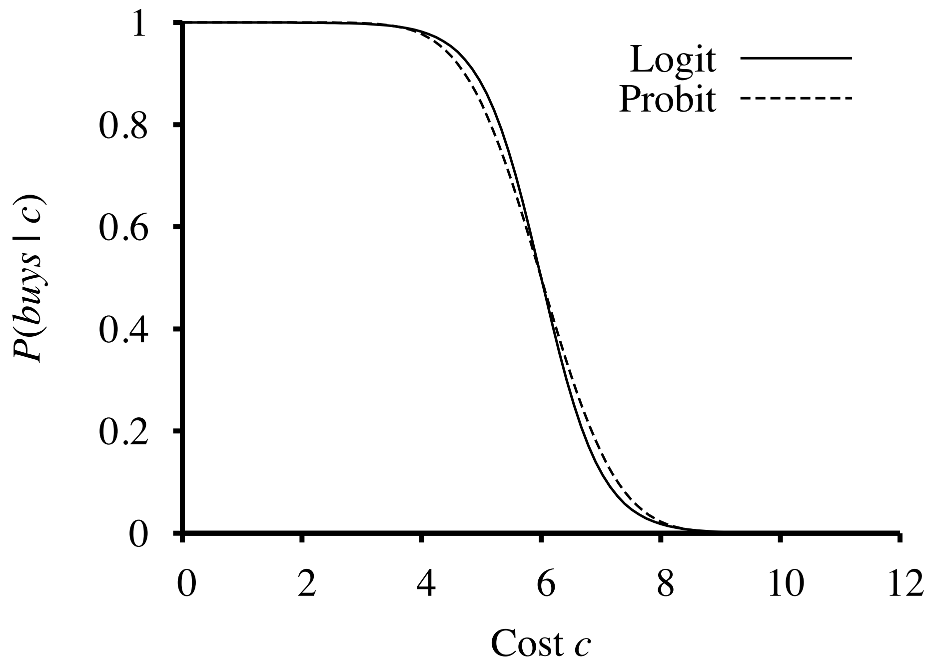


Discrete variable contd.

Sigmoid (or logit) distribution also used in neural networks:

$$P(Buys? = true \mid Cost = c) = \frac{1}{1 + exp(-2\frac{-c+\mu}{\sigma})}$$

Sigmoid has similar shape to probit but much longer tails:



Summary (representation)

Bayes nets provide a natural representation for (causally induced) conditional independence

Topology + CPTs = compact representation of joint distribution
$\Rightarrow$ fast learning from few examples

Generally easy for (non)experts to construct

Canonical distributions (e.g., noisy-OR, linear Gaussian)
$\Rightarrow$ compact representation of CPTs
$\Rightarrow$ faster learning from fewer examples