First-order probabilistic inference

David Poole

Department of Computer Science,
University of British Columbia

IJCAI 2011 Tutorial

Overview

- Simple representation: parametrized belief networks. Means grounding.
- Inference: combine variable elimination and unification
    - One step of first-order variable elimination corresponds to many VE steps.
- Allows for new queries that depend on population size: probability that someone is guilty of a crime depends on how many other people could have done it.

Outline

1. Background
   - Probability
   - Logic
   - Relational Probabilistic Models
2. First-order Probabilistic Inference
   - Unification and Splitting
   - Lifted VE Operations
3. Conclusions

Bayesians

- Probability is a measure of belief.
- All individuals about which we have the same information should have the same probability.
  - Idea: share probability tables both initially and during inference.

Background: Belief (Bayesian) networks

- Totally order the variables of interest: $X_1, \ldots, X_n$
- $P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i|\pi_i)$
  π_i are the parents of X_i: a set of predecessors such that

$$P(X_i|\pi_i) = P(X_i|X_1, \ldots, X_{i-1})$$

Background: variable elimination

To compute the probability of a variable X given evidence $\overline{Z} = \overline{e}$:

$$P(X|\overline{Z} = \overline{e}) = \frac{P(X \wedge \overline{Z} = \overline{e})}{P(\overline{Z} = \overline{e})}$$

Suppose the other variables are $Y_1, \ldots, Y_m$:

$$\begin{aligned} & P(X \wedge \overline{Z}) \\ & = \sum_{Y_m} \cdots \sum_{Y_1} P(X_1, \ldots, X_n) \\ & = \sum_{Y_m} \cdots \sum_{Y_1} \prod_{i=1}^{n} P(X_i|\pi_i) \end{aligned}$$

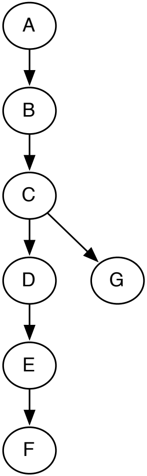
Eliminating a variable

- to compute $AB + AC$ efficiently, distribute out A: $A(B + C)$.
- to compute

$$\sum_{Y_j} \prod_{i=1}^{n} P(X_i | \pi_i)$$

  distribute out those factors that don't involve Y_j.
- Can be used for directed and undirected models.
- Closely related to nonserial dynamic programming [Bertelè & Brioschi, 1972]

Variable Elimination Example



$$
\begin{aligned}
& P(G|f) \\
& \quad \propto \quad \sum_C \sum_B \sum_D \sum_E \sum_A \\
& \qquad\qquad\qquad P(A)P(B|A)P(C|B)(D|C) \\
& \qquad\qquad\qquad P(E|D)P(f|E)P(G|C) \\
& \quad = \quad \sum_C (\sum_B (\sum_A P(A)P(B|A)) \\
& \qquad\qquad\qquad\quad P(C|B)) \\
& \qquad\qquad\quad (\sum_D P(D|C) \\
& \qquad\qquad\qquad\quad (\sum_E P(E|D)P(f|E))) \\
& \qquad\qquad\qquad\quad P(G|C)
\end{aligned}
$$

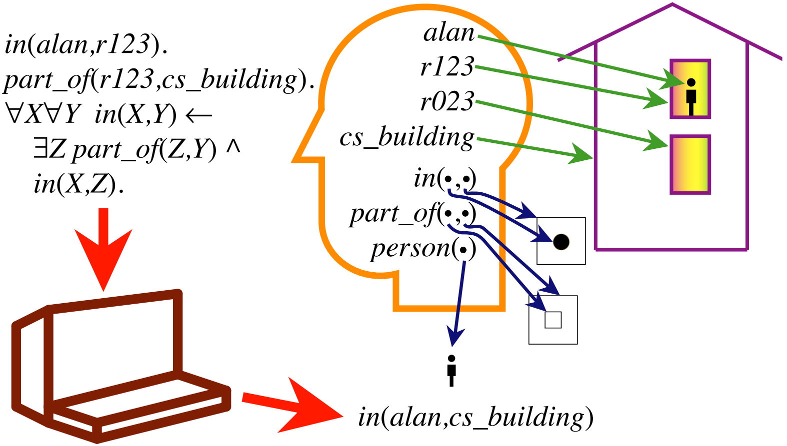
Variable Elimination: basic operations

- conditioning on observations (local to each factor)
- multiplying factors
- summing a variable from a factor

Outline

1. Background
   - Probability
   - Logic
   - Relational Probabilistic Models
2. First-order Probabilistic Inference
   - Unification and Splitting
   - Lifted VE Operations
3. Conclusions

First-order predicate calculus



Skolemization and Herbrand's Theorem

Skolemization: give a name for an object said to exist

$$\forall X \exists Y \; q(X, Y) \text{ becomes } q(X, f(X))$$

Herbrand's theorem [1930]:

- If a logical theory has a model it has a model where the domain is made of ground terms, and each term denotes itself.
- If a logical theory T is unsatisfiable, there is a finite set of ground instances of formulas of T which is unsatisfiable.

Unique Names Assumption & Negation as Failure

- Unique Names Assumption:
  — different names denote different individuals
  — different ground terms denote different individuals
- Herbrand's theorem holds even without the unique names assumption.

Theorem Proving and Unification

In 1965, Robinson showed how unification allows many ground steps with one step:

$$\underbrace{f(X,Z) \vee p(X,a) \qquad \neg p(b,Y) \vee g(Y,W)}_{f(b,Z) \vee g(a,W)}$$

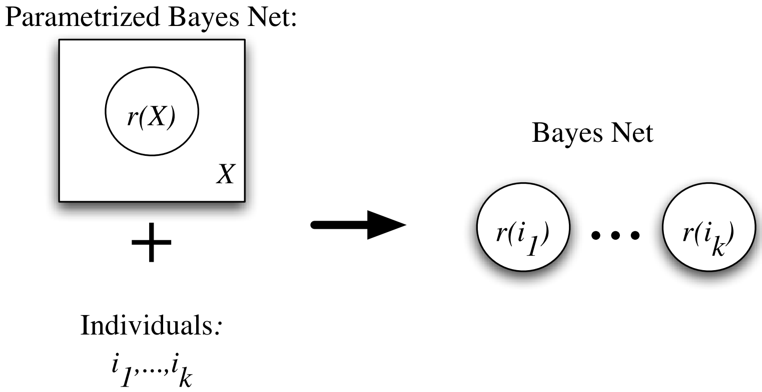
Substitution $\{X/b, Y/a\}$ is the most general unifier of $p(X,a)$ and $p(b,Y)$.

Outline

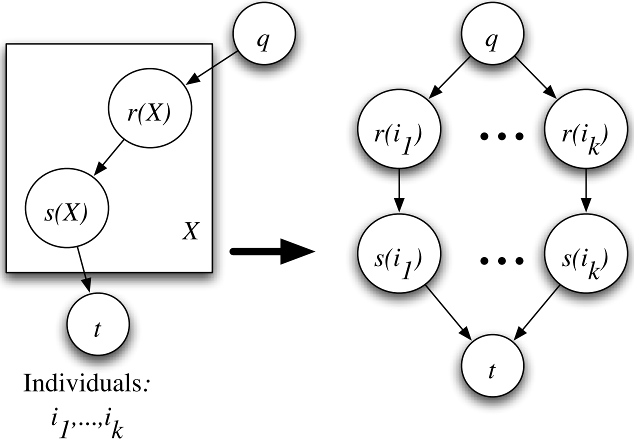
Parametrized belief networks

- Allow random variables to be parametrized. $height(X)$
- Parameters correspond to logical variables. X
- Each parameter is typed with a population. $X : person$
- Each population has a size. $|person| = 1000000$
- Parametrized belief network means its grounding. Ground instances are random variables: $height(p_1) \ldots height(p_{1000000})$
- Instances are independent (but can have common ancestors and descendents).

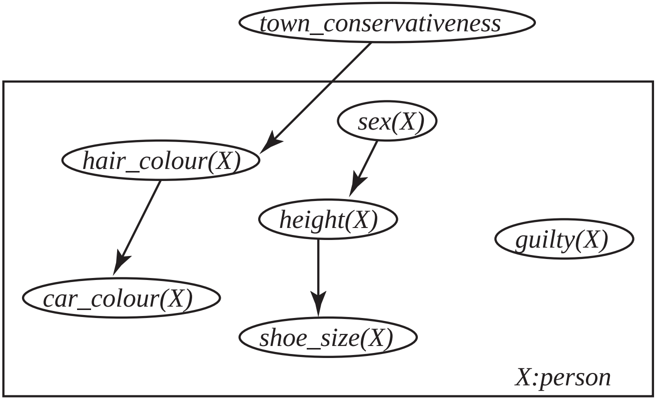
Parametrized Bayesian networks / Plates



Parametrized Bayesian networks / Plates (2)



Example parametrized belief network



$$\forall X \; P(car\_colour(X){=}pink \mid hair\_colour(X){=}pink) = 0.1$$
$$\forall X \; P(hair\_colour(X){=}pink \mid town\_conservative) = 0.001.$$

Outline

Theorem Proving and Unification (reprise)

In 1965, Robinson showed how unification allows many ground steps with one step:

$$\underbrace{f(X,Z) \vee p(X,a) \qquad \neg p(b,Y) \vee g(Y,W)}_{f(b,Z) \vee g(a,W)}$$

Substitution $\{X/b, Y/a\}$ is the most general unifier of $p(X,a)$ and $p(b,Y)$.

Variable Elimination and Unification

- Multiplying parametrized factors:

$$\underbrace{[f(X,Z),p(X,a)] \quad \times \quad [p(b,Y),g(Y,W)]}_{[f(b,Z),p(b,a),g(a,W)]}$$

  Doesn't quite work because the first parametrized factor can't be used for $X = b$ but can be used for other instances of X.

- Intuitively, we want to add the constraint $X \neq b$ to $[f(X,Z),p(X,a)]$ after the above multiplication.

Parametric Factors

A parametric factor (parfactor) is a triple $\langle C, V, t\rangle$ where

- C is a set of inequality constraints on parameters,
- V is a set of parametrized random variables
- t is a table representing a factor from the random variables to the non-negative reals.

$$\left\langle \{X \neq sue\}, \{hair\_col(X), cons\}, \begin{array}{|ll|l|} \hline hair\_col & cons & Val \\ \hline purple & yes & 0.001 \\ purple & no & 0.01 \\ & \ldots & \\ \hline \end{array} \right\rangle$$

Splitting

Instead of applying substitutions to parametric factors, we split the parametric factors on the substitution.

A **split** of $\langle C, V, t\rangle$ on $X = \gamma$, results in parametric factors:

$\langle C[X/\gamma], V[X/\gamma], t\rangle$

$\langle \{X \neq \gamma\} \cup C, V, t\rangle$ $\longleftarrow$ residual

where $V[X/\gamma]$ is V with γ substituted for X.

Splitting on a substitution

- Splitting on a substitution, means splitting on each equality in the substitution.
- Different orders of splitting give the same final result, but may give different residuals.
- Example: Split

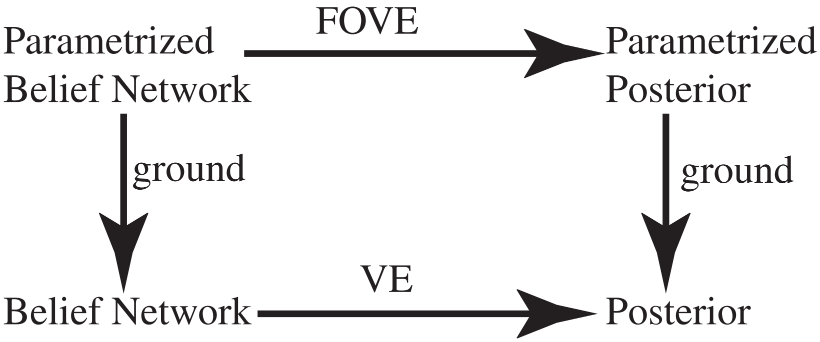
$$\langle \{\}, \{foo(X, Y, Z)\}, t_1 \rangle$$

on $\{X = Z, Y = b\}$ results in

$$\langle \{\}, \{foo(X, b, X)\}, t_1 \rangle$$
$$\langle \{X \neq Z\}, \{foo(X, Y, Z)\}, t_1 \rangle$$
$$\langle \{Y \neq b\}, \{foo(X, Y, X)\}, t_1 \rangle$$

Outline

First-order probabilistic inference



Multiplying Parametric Factors

Suppose we were to eliminate p and multiply the two parametric factors:

$$\langle\{\}, \{p(X,a), q(Y,c), s(X,Y)\}, t_1\rangle$$
$$\langle\{W \neq d\}, \{p(b,Z), q(W,T), r(W,T)\}, t_2\rangle$$

- If we grounded these, then did VE, some instances of these would be multiplied and some wouldn't.
- We unify $p(X,a)$ and $p(b,Z)$ resulting in the substitution $\theta = \{X/b, Z/a\}$.
- Unification finds the most general instances that need to be multiplied.

Splitting when Multiplying I

We are multiplying the two parametric factors:

$$\langle\{\}, \{p(X,a), q(Y,c), s(X,Y)\}, t_1\rangle \quad (1)$$
$$\langle\{W \neq d\}, \{p(b,Z), q(W,T), r(W,T)\}, t_2\rangle \quad (2)$$

We split parametric factor (1) on $\theta = \{X/b, Z/a\}$:

$$\langle\{\}, \{p(b,a), q(Y,c), s(b,Y)\}, t_1\rangle \quad (3)$$
$$\langle\{X \neq b\}, \{p(X,a), q(Y,c), s(X,Y)\}, t_1\rangle \quad (4)$$

We can split (2) on θ resulting in:

$$\langle\{W \neq d\}, \{p(b,a), q(W,T), r(W,T)\}, t_2\rangle \quad (5)$$
$$\langle\{Z \neq a, W \neq d\}, \{p(b,Z), q(W,T), r(W,T)\}, t_2\rangle \quad (6)$$

Splitting when Multiplying II

When we are multiplying:

$$\langle \{\}, \{p(b,a), q(Y,c), s(b,Y)\}, t_1 \rangle$$
$$\langle \{W \neq d\}, \{p(b,a), q(W,T), r(W,T)\}, t_2 \rangle$$

- All ground instances would need to be multiplied.
- Not all instances have the same number of variables: some will have two different q instances, and some have one.
- We need to split again on the most general unifier of $q(Y,c)$ and $q(W,T)$.

Summing out variables

- If we are not removing a parameter, we sum out as normal. E.g., summing out p:

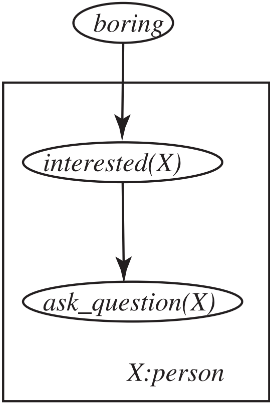
$$\langle \{\}, \{p(X), q(X)\}, t[p, q] \rangle$$

- If we are removing a parameter, we must take to the power of the effective population size. E.g., summing out p:

$$\langle \{Y \neq a\}, \{p(X, Y), q(X)\}, t[p, q] \rangle$$

- Other functions such as noisy-or, you need to take into account the population size.

Removing a parameter when summing



$|people| = 100$

observe no questions

Eliminate *interested*:

$$\langle \{\}, \{boring, interested(X)\}, t_1 \rangle$$
$$\langle \{\}, \{interested(X)\}, t_2 \rangle$$
$$\downarrow$$
$$\langle \{\}, \{boring\}, (t_1 \times t_2)^{100} \rangle$$

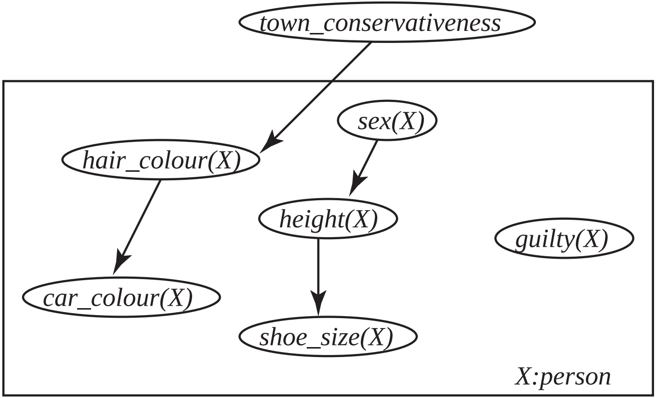
Existential Observations

Suppose we observe:

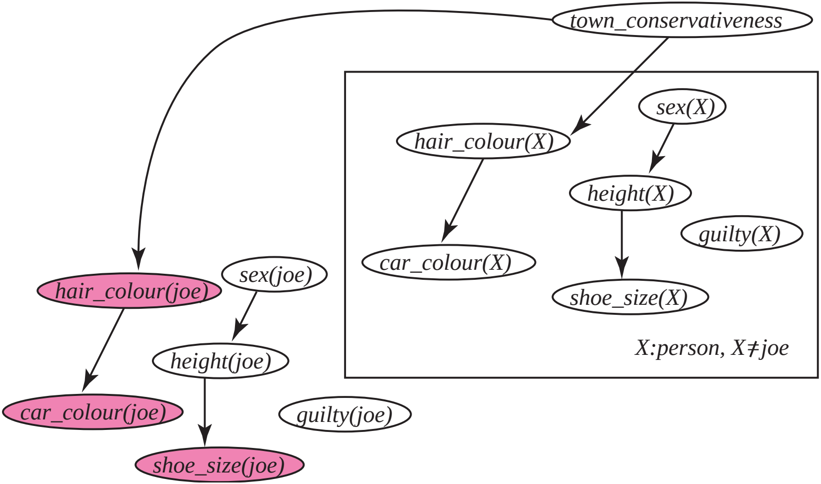
- Joe has purple hair, a purple car, and has big feet.
- A person with purple hair, a purple car, and who is very tall was seen committing a crime.

What is the probability that Joe is guilty?

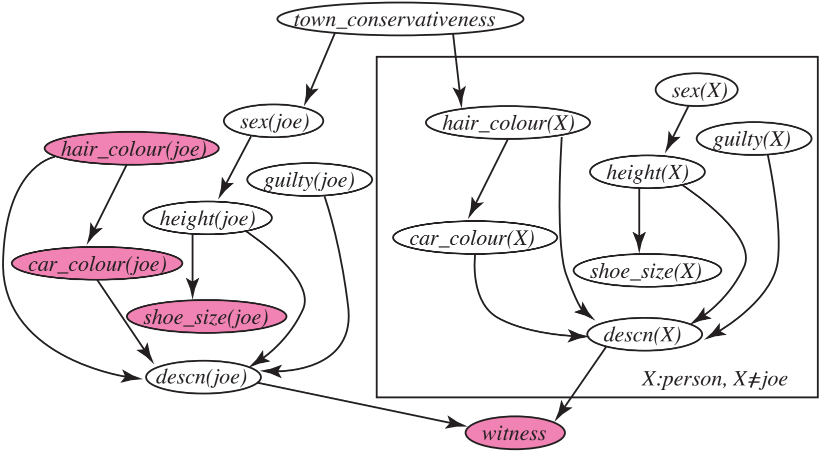
Background parametrized belief network



Observing information about Joe



Observing Joe and the crime



Last Steps

We end up with parametric Factors:

$$\langle \{\}, \{guilty(joe), descn(joe), conservativeness\}, t_1 \rangle$$
$$\langle \{X \neq joe\}, \{descn(X), conservativeness\}, t_2 \rangle$$
$$\langle \{\}, \{descn(X), witness\}, t_3 \rangle$$
$$\langle \{\}, \{conservativeness\}, t_4 \rangle$$

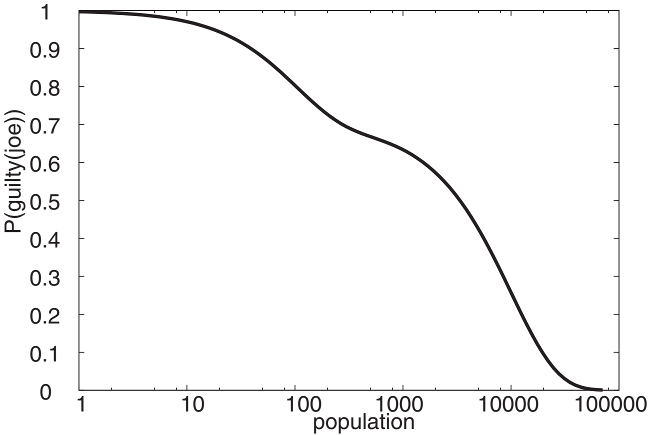
We eliminate *descn*(X):

$$\langle \{\}, \{guilty(joe), witness, conservativeness\}, t_5 \rangle$$

We sum out *conservativeness* and condition on *witness*:

$$\langle \{\}, \{guilty(joe)\}, t_6 \rangle$$

Guilty as a function of population



Outline

Conclusions

- We combine variable elimination + unification.
    - One step of first-order variable elimination corresponds to many steps in ground representation.
    - We can condition on existential and universal observations.
- Contributions of IJCAI-93 paper:
    - parametrized random variables
    - splitting to complement unification
    - parfactor representation of intermediate results
    - an algorithm for multiplying factors in a lifted manner (sometimes)