Bayesian AI

Bayesian Artificial Intelligence
Introduction
IEEE Computational Intelligence Society
IEEE Computer Society

Kevin Korb
Clayton School of IT
Monash University

kbkorb@gmail.com

Contents

Abstract

Bayesian networks are the basis for a new generation of probabilistic expert systems, which allow for exact (and approximate) modelling of physical, biological and social systems operating under uncertainty. Bayesian networks are also an important representational tool for data mining, in causal discovery. Applications range across the sciences, industries and government organizations. At Monash University, Bayesian AI has been used for graphical expert systems for medical diagnosis and prognosis, in meteorological predication, environmental management, intelligent tutoring systems, epidemiology, poker and other applications. Norsys (www.norsys.com) list hundreds of additional applications of Bayesian AI. This talk will explain the basics of the technology, illustrate them with example Bayesian networks, and discuss the growth in the use of Bayesian networks in recent years. The technology is not only mature, but is becoming more widely accepted in major projects.

The Certainty of Uncertainty

Sources of uncertainty:

World laws. Laws governing world (events) may be stochastic.

- Long tradition of ignoring this possibility
- This is hardly plausible in view of the probabilistic theories of: genetics, economics, physics, etc.

Inaccessibility. Operation of hidden variables makes relations btw observed variables stochastic.

Measurement error. Uncertainty of measurement translates into uncertain relations btw measured variables.

Bayes' Theorem

Discovered by Rev Thomas Bayes; published posthumously in 1763
Forward Inference: $P(e|h)$ – e.g., what is the probability of heads given a fair coin?
Bayes' Inverse Inference Rule:

$$P(h|e) = \frac{P(e|h)P(h)}{P(e)}$$

$$\text{posterior} = (\text{likelihood} \times \text{prior})\alpha$$

- Forward inference tells us likelihoods
- Finding priors is the main problem in applying Bayes' Rule

Bayes' Theorem

For 30 years Bayes' Rule has NOT been used in AI

- Not because it was thought undesirable and not due to lack of priors, but
- Because: it was (thought) infeasible
  - ⇒ requires full joint probability
  - ⇒ computation is exponential in number of possible states

Bayesian Reasoning for Humans (BRH)

First: it's important

Cancer Problem

> *You have been referred to a specialty clinic. You are told that one in one hundred appearing at the clinic have cancer X and that the tests are positive given cancer X 80% of the time and they also are positive 10% of the time in people without cancer X.*

What is the probability that you have cancer X?

1. $\approx$ 99%
2. $\approx$ 80%
3. 50-50
4. $\approx$ 30%
5. $\approx$ 7%

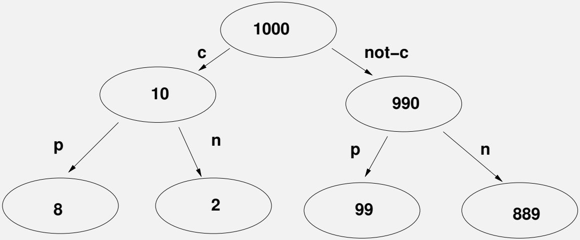
BRH: Bayes' Theorem

Second: it's "hard"

$$
\begin{aligned}
P(c|p) &= \frac{P(p|c)P(c)}{P(p|c)P(c) + P(p|\neg c)P(\neg c)} \\
&= \frac{.8 \times .01}{.8 \times .01 + .1 \times .99} \\
&= \frac{.008}{.008 + .099} \\
&= \frac{.008}{.107} \approx .075
\end{aligned}
$$

BRH: Frequency Formats

Third: it's easy — multiply!



Classification tree for breast cancer

$$P(c|p) = \frac{P(c,p)}{P(p)} = \frac{8}{8+99}$$

$\Rightarrow$ Even easier: use Bayesian networks!

Cancer X

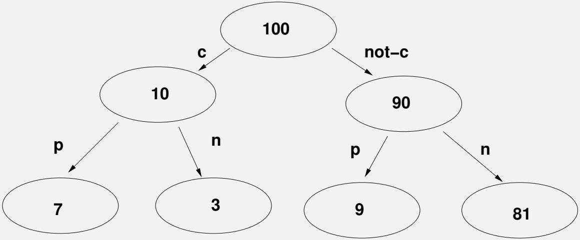
You may have cancer X. You are in a suspect category where the rate is 10%. A test Y is positive given cancer 70% of the time and it is also positive 10% of the time without cancer.

What is the probability that you have cancer X?

1. $\approx$ 90%
2. $\approx$ 70%
3. $\approx$ 60%
4. 50-50
5. $\approx$ 40%
6. $\approx$ 10%

OK, what's the probability given a positive test?

Cancer X: Frequency Formats

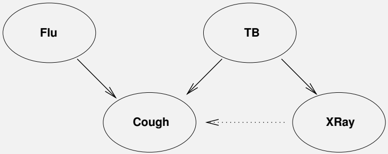


$$P(c|p) = \frac{P(c,p)}{P(p)} = \frac{7}{7+9} \approx 0.44$$

AI History: Idiot Bayes

An attempt to simplify probabilistic reasoning in 1960s medical diagnostic programs. Assumed:

- Diseases marginally independent
  - E.g., Flu and TB independent
- Symptoms independent given disease
  - E.g., Sneezing & Cough given Flu (!?)
- Diseases *remain* independent given symptoms
  - E.g., $P(Flu|Cough, \neg TB) = P(Flu|Cough)$
  - This is obviously wrong!
  - Indeed, if $P(TB \vee Flu) = 1$, $P(Flu|Cough, \neg TB) = 1$



Duda's Prospector

— Duda, Hart, Nilsson (1974)
First major success for "probabilities" in expert systems.

- Elicited marginal and conditional probabilities from experts
- Update rules were simple and fast:
  - $P(A, B) = \min(P(A), P(B))$
  - $P(A \vee B) = \max(P(A), P(B))$

Duda's Prospector

Problems?

- Update rules are stupid.
  Suppose rain and shine are equally likely. Then we get:
  - $P(rain, shine) = min(P(rain), P(shine)) = 0.5$
  - $P(rain \vee shine) = \ max(P(rain), P(shine)) = 0.5$
- Probabilities were *overspecified*
  - If you elicit $P(A), P(A|B), P(B)$ you have two different ways of computing $P(A)$:
    1. $P(A)$
    2. $P(A) = P(A|B)P(B) + P(A|\neg B)P(\neg B)$
  - Leading to inconsistencies

$\Rightarrow$ Not necessarily bad, but requires resolution!

Mycin's Certainty Factors

Supposedly, a big improvement. Used in various expert systems through the 1980s.

Certainty Factors: $CF(h, e) \in [-1, 1]$

Update Rule: Belief in conclusion =
certainty factor $\times$ belief in premises.

Mycin's Certainty Factors

The devil was in the details; for complex updates

- Let $CF(h, e_1) = x$; $CF(h, e_2) = y$

Then

$$CF(h, e_1 \wedge e_2) = \begin{cases} x + y - xy & \text{if } x, y \geq 0 \\ \frac{x+y}{1-\min(|x|,|y|)} & \text{if } xy \leq 0 \\ x + y + xy & \text{if } x, y \leq 0 \end{cases}$$

However, David Heckerman (1986) proved CF calculus is equivalent to probability theory IF evidential statements are independent.

E.g., coughing and sneezing are independent of each other!

Physical Probability

Frequentism: anti-subjective

- Aristotle: The probable is what usu happens
- John Venn (1866): $P(h) = F(h)$, relative to a reference class and in the "long run"
- Richard von Mises (1919)/Church (1940): prob identified with the limit frequency in a (hypothetical) sequence, which is invariant under prior computable selections of subsequences.
  - Prob of rain tomorrow = 0.5 means...

Corresponds better with physical experiment: probabilities don't seem to budge because of subjective opinions!

Subjective Probability

Still, there seems to be a role for subjective opinion determining betting behavior
Suppose the world is **deterministic**: all physical probabilities are 0 or 1.

- It *still* makes sense to say that prob of rain tomorrow is 0.5!

Needn't rely on Laplace's principle of indifference.
Instead, for example, use David Lewis's

Principal Principle:

$$P(h|Ch(h) = r) = r$$

I.e., theory is a source of probability

Subjective Probability

Other possible sources of prior probability:

- Observed frequencies
  - Reichenbach's "straight rule"
- Evolved bias
- Even max entropy

Bayesian Networks

Next time we will look at the new technology of Bayesian nets…

> *Note that Bayesian nets are usable regardless of your interpretation of probability.*

Causal Graphical Models

- First systematic use of graphs for reasoning

  Wigmore (1913) charts for legal reasoning

- First systematic use of specifically causal graphs

  Sewall Wright (1921) for analysing population genetics

- Simon-Blalock method for parameterization
- Structural equation models (SEMs)
- Algorithms for Bayesian network modeling

  Pearl (1988), Neapolitan (1990)

Reichenbach's Common Cause Principle

Common Cause Principle

Reichenbach (1956): When there is a enduring correlation between two types of events, there is an underlying causal explanation.

Or:

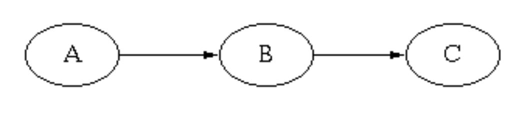
Where there's smoke, there's fire

Or:

$$\frac{1}{\text{Statistician's Mantra}}$$

Conditional Independence: Causal Chains

Causal chains give rise to conditional independence:

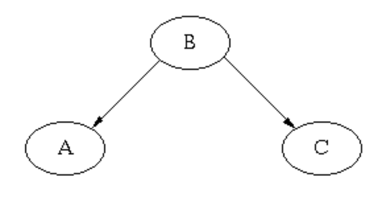


$$P(C|A \wedge B) = P(C|B) \equiv A \perp\!\!\!\perp C|B$$

E.g., a sexually transmitted disease

Conditional Independence: Common Causes

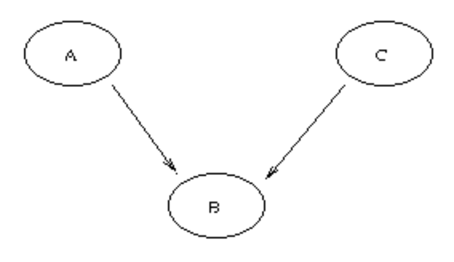
Common causes (ancestors) also give rise to conditional independence:



$$P(C|A \wedge B) = P(C|B) \equiv A \perp\!\!\!\perp C|B$$

Conditional Dependence: Common Effects

Common effects (or their descendants) give rise to conditional *dependence:*



$$P(A|C \wedge B) \neq P(A)P(C) \equiv A \not\perp\!\!\!\perp C|B$$

E.g., inheriting a recessive trait from both parents; explaining away.

Causality and Probability

Dependency signature

Note that the conditional dependency structures are exact opposite btw chains/common ancestry and "collisions".

- Margin dependence: marginal independence
- Conditional independence: conditional dependence

This is key for causal discovery.

Bayesian Networks

Definition (Bayesian Network)

A graph where:

1. The nodes are random variables.
2. Directed arcs (arrows) connect pairs of nodes.
3. Each node has a conditional probability table that *quantifies* the effects of its parents.
4. It is a directed acyclic graph (DAG), i.e. no directed cycles.

Pearl's Alarm Example

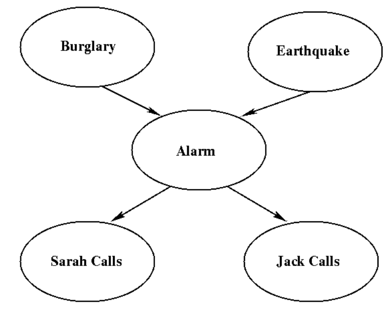
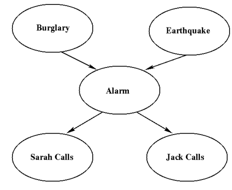


Figure: Pearl's Alarm Example

Factorization

Any joint probability distribution can be factorized using any total order. E.g.,

$$
\begin{aligned}
&P(B,E,A,S,J) \\
&\quad = \frac{P(B,E,A,S,J)}{P(J)}P(J) \\
&\quad = P(B,E,A,S|J)P(J) \\
&\quad = \ldots \\
&\quad = P(B|E,A,S,J)P(E|A,S,J)P(A|S,J)P(S|J)P(J)
\end{aligned}
$$

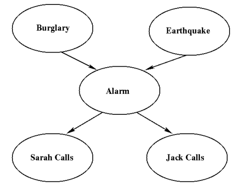


Factorization

The advantage of graphical models is that we have a grahical criterion for systematically simplifying this computation, yielding:

$$P(B, E, A, S, J) = P(S|A)P(J|A)P(A|B, E)P(B)P(E)$$

NB: Note that the order is no longer arbitrary!



The Markov condition

In order to justify the simplification, we will have to invoke (and justify) the Markov condition:

Definition (Markov Condition)

There are no direct dependencies in the system being modeled which are not explicitly shown via arcs.

Equivalently,

Definition (Markov Condition)

Every variable is independent of its non-descendants given a known state for its parents.

The Markov condition

The Markov condition is not automatically true; you have to *make* it true.

> *When it's false, there's a missing arc somewhere. The model is wrong, so go find the right model.*

Minimally, this is the right position to adopt by default; caveats below…

Inference

Given the above, a large variety of "efficient" algorithms are available for probabilistic inference — i.e., Bayesian inference conditioning upon observations

- exact
- or approximate (complex nets)

Efficiency depends upon network complexity (esp arc density)

- worst case exponential (NP-hard; Cooper, 1990)

Compactness and Node Ordering

Compactness of BN depends upon how the net is constructued, in particular upon the underling node order

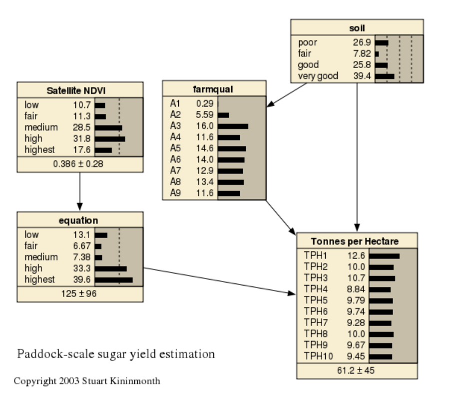
- When constructing a BN, it's best to add nodes in their natural causal order, root causes through to leaves.
- Other orderings tend to produce denser networks

Sugar Cane

Using a small number of multiscale variables, including farm quality, soil classes and Landsat ETM based normalised difference vegetation index, a Bayesian belief network was constructed. The inferential capacity of the model was used to generate the expected yield for each paddock based on assessments 5 months prior to harvesting.

The power of the Bayesian belief network to display the model uncertainty and to direct further research into data collection is a compelling reason to utilise this technique.

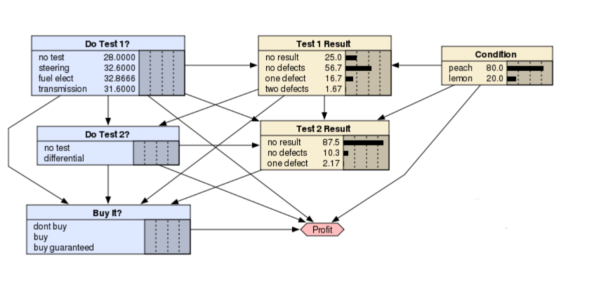
Sugar Cane


soil
poor 26.9
fair 7.82
good 25.8
very good 39.4
Satellite NDVI
low 10.7
fair 11.3
medium 28.5
high 31.8
highest 17.6
0.386 ± 0.28
farmqual
A1 0.29
A2 5.59
A3 16.0
A4 11.6
A5 14.6
A6 14.0
A7 12.9
A8 13.4
A9 11.6
equation
low 13.1
fair 6.67
medium 7.38
high 33.3
highest 39.6
125 ± 96
Tonnes per Hectare
TPH1 12.6
TPH2 10.0
TPH3 10.7
TPH4 8.84
TPH5 9.79
TPH6 9.74
TPH7 9.28
TPH8 10.0
TPH9 9.67
TPH10 9.45
61.2 ± 45
Paddock-scale sugar yield estimation
Copyright 2003 Stuart Kininmonth

Car Buyer

This is an example influence diagram for Joe, who has to decide whether to buy a certain used car which may be a 'peach' or a 'lemon'. He has the option of doing some tests beforehand, and of buying it with a guarantee or not. This is the classic example of an influence diagram derived from a decision problem with a very asymmetric decision tree, since if Joe decides not to test then the test results have no meaning, etc.

Car Buyer


Do Test 1?
no test 28.0000
steering 32.6000
fuel elect 32.8666
transmission 31.6000
Test 1 Result
no result 25.0
no defects 56.7
one defect 16.7
two defects 1.67
Condition
peach 80.0
lemon 20.0
Do Test 2?
no test
differential
Test 2 Result
no result 87.5
no defects 10.3
one defect 2.17
Buy It?
dont buy
buy
buy guaranteed
Profit

Causal Ordering

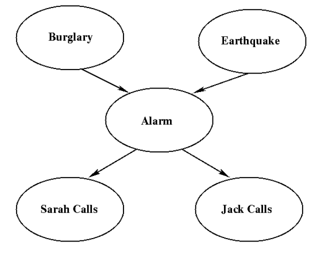
Why does variable order affect network density?

Because

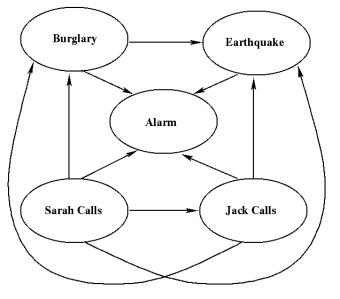
- Using the causal order allows direct representation of conditional independencies
- Violating causal order requires new arcs to re-establish conditional independencies

Causal Ordering

Using $\langle B, E, A, S, J \rangle$



Using $\langle S, J, B, E, A \rangle$



Causal Semantics

It's clear that many BNs are *not* causal networks

- E.g., the last Alarm network above

But it's also clear that many others *are* causal networks.

Furthermore, it's clear that causal nets have advantages:

- They are more intuitive
  - easier to elicit
  - possible to explain
- They are more compact and efficient
- They can be machine learned
- Interventions can be reasoned about

Bayesian Networks: Summary

BNs:

- Perform Bayesian updating on any available evidence
- Do so efficiently when possible
  - given the Markov condition
  - given low arc densities, using d-separations
- Causal models are advantageous: more understandable, more compact

Question: Can causal models really be machine learned?

Extensions to Bayesian Networks

Decision networks:
For decision making.

Dynamic Bayesian networks:
For reasoning about changes over time

Making Decisions

- Bayesian networks can be extended to support decision making.
- **Preferences** between different outcomes of various plans.
    - Utility theory
- **Decision theory** = Utility theory + Probability theory.

Expected Utility

Definition (Expected Utility)

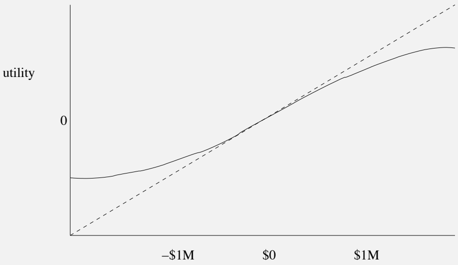
$$EU(A|E) = \sum_i P(O_i|E, A) \times U(O_i|A)$$

- E = available evidence,
- A = an action
- O_i = possible outcome state
- U = utility

Utility

How are utility functions constructed?

- Often utility is equated with money
  - Money in the future should be discounted compared to money in the present
  - And even discounted money is rarely equal to utility



Type of Nodes

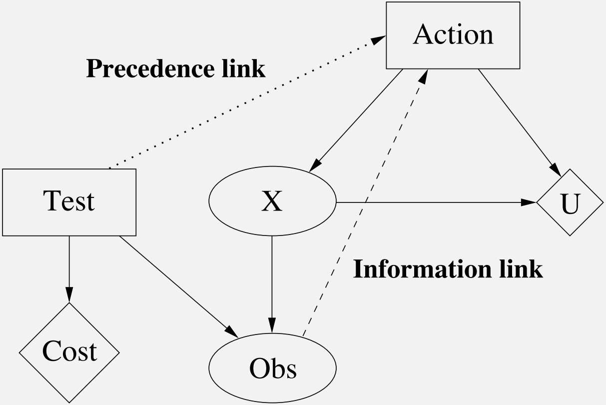
Chance nodes: (ovals)
: Represent random variables (same as Bayesian networks). Has an associated CPT. Parents can be decision nodes and other chance nodes.

Decision nodes: (rectangles)
: Represent points where the decision maker has a choice of actions.

Utility nodes: (diamonds)
: Represent the agent's utility function (also called **value nodes** in the literature). Parents are variables describing the outcome state that directly affect utility. Has an associated table representing multi-attribute utility function.

Sequential Decision Making

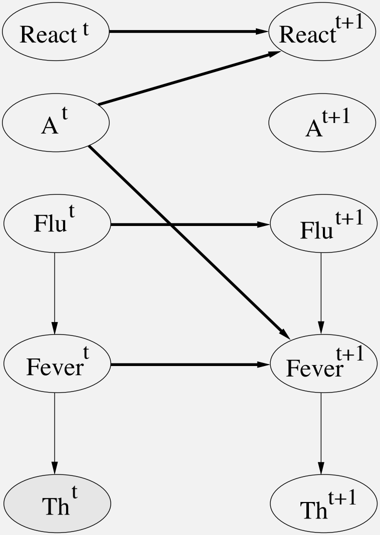
- Precedence links used to show temporal ordering.
- Network for a test-action decision sequence



Dynamic Belief Networks

- One node for each variable for each time step.
- **Intra-slice** arcs $Flu^T \longrightarrow Fever^T$
- **Inter-slice (temporal)** arcs
  1. $Flu^T \longrightarrow Flu^{T+1}$
  2. $Aspirin^T \longrightarrow Fever^{T+1}$

Fever DBN



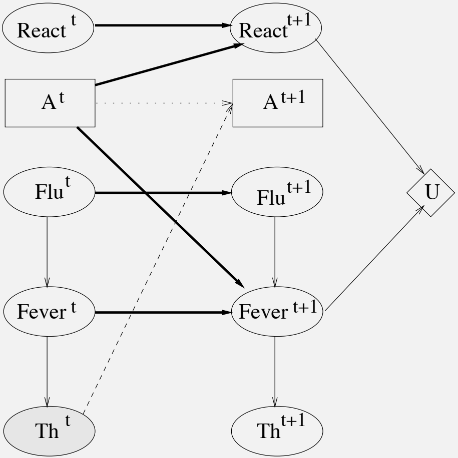
DBN reasoning

- Can calculate distributions for S_{t+1} and further: **probabilistic projection**.
- Reasoning can be done using standard BN updating algorithms
- This type of DBN gets very large, very quickly.
- Usually only keep two time slices of the network.

Dynamic Decision Network

- Similarly, Decision Networks can be extended to include temporal aspects.

Fever DDN



Extensions: Summary

- BNs can be extended with decision nodes and utility nodes to support decision making: *Decision Networks* or *Influence Diagrams*.
- BNs and decision networks can be extended to allow explicit reasoning about changes over time.

Causal Discovery

- Parameterization
  - Linear models: see path modeling
- Structure Learning
  - Constraint-based learning = CI learning
  - Metric learning: Bayesian (or non-Bayesian) scoring function plus search

Parameterization

Spiegelhalter & Lauritzen (1990) learning CPTs:

- assume parameter independence
- each CPT cell *i* = a parameter in a Dirichlet distribution

$$D[\alpha_1, \ldots, \alpha_i, \ldots, \alpha_K]$$

  for *K* parents

- prob of outcome *i* is $\alpha_i / \Sigma_{k=1}^{K} \alpha_k$
- observing outcome *i* update *D* to

$$D[\alpha_1, \ldots, \alpha_i + 1, \ldots, \alpha_K]$$

Learning without parameter independence:

- Decision trees to learn structure within CPTs (Boutillier et al. 1996).
- Hybrid model learning (CPTs, d-trees) (O'Donnell et al. 2006a)

Main research problems: dealing with noise & missing data.

Learning Causal Structure

This is the *harder* problem.

Size of the dag space is superexponential:

- Number of possible orderings: $n!$
- Times number of ways of pairing up (for arcs): $2^{C_2^n}$
- Minus number of possible cyclic graphs

Without the subtraction (which is a small proportion):

n	$n!2^{C_2^n}$
1	1
2	4
3	48
4	1536
5	12280
10	127677049435953561600
100	[too many digits to show]

Constraint-Based Learning

Verma-Pearl Algorithm

IC algorithm (Verma and Pearl, 1991)

Suppose you have an Oracle who can answer yes or no to any question of the type:

$$X \perp\!\!\!\perp Y | \mathbf{S}?$$

(Is X conditional independent Y given $\mathbf{S}$?)

Then you can learn the correct causal model, to within its statistical equivalence class (pattern).

Verma-Pearl Algorithm

Their IC algorithm allows the discovery of the set of causal models consistent with all such answers ("patterns"):

Step 1 Put an undirected link between any two variables X and Y iff
for every **S** s.t. $X, Y \notin \mathbf{S}$

$$X \not\perp\!\!\!\perp Y | \mathbf{S}$$

Step 2 For every undirected, uncovered collision $X - Z - Y$ orient the arcs $X \rightarrow Z \leftarrow Y$ iff

$$X \not\perp\!\!\!\perp Y | \mathbf{S}$$

for **every S** s.t. $X, Y \notin \mathbf{S}$ and $Z \in \mathbf{S}$.

Verma-Pearl Algorithm

Step 3 For each connected pair X–Y, both:

1. if $X \rightarrow Y$ would introduce a cycle, then put $X \leftarrow Y$,
2. if $X \rightarrow Y$ would introduce $X \rightarrow Y \leftarrow Z$ where X and Z are disconnected, then put $X \leftarrow Y$.

Repeat this Step until no changes can be made to any connected pair.

PC: TETRAD

Spirtes, Glymour and Scheines (1993) made this approach practical.
Replace the Oracle with statistical tests:

- for linear models a significance test on partial correlation

$$X \perp\!\!\!\perp Y | \mathbf{S} \text{ iff } \rho_{XY \cdot \mathbf{S}} = 0$$

- for discrete models a χ^2 test on the difference between CPT counts expected with independence (E_i) and observed (O_i)

$$X \perp\!\!\!\perp Y | \mathbf{S} \text{ iff } \sum_i O_i \ln \left(\frac{O_i}{E_i} \right)^2 \approx 0$$

Implemented in their **PC Algorithm**

PC

- Heuristics used to speed up search.
- Result: discovered pattern.
- Current version is in TETRAD IV
- PC is also (being) implemented by numerous BN tools, including Weka and Hugin
- Advantages: simple, quick and free

Metric Causal Discovery

A very different approach is *metric* learning of causality:

- Develop a score function which evaluates any Bayesian network *as a whole* relative to the evidence.
- Originally this was done in a brute force Bayesian computation of

$$P(dag|data)$$

  by counting methods (Cooper & Herskovits, 1991)
- CD then means: search the space of dags looking for that dag which maximizes the score.

Metric Discovery Programs

K2 (Cooper & Herskovits)
Greedy search. Mediocre performance.

MDL (Lam & Bacchus, 1993; Friedman, 1997)
An information-theoretic scoring function with various kinds of search, such as beam search. Friedman allows for hybrid local structure.

BDe/BGe (Heckerman & Geiger, 1995)
A Bayesian score; edit-distance priors supported; returns a pattern. Good performance.

CaMML (Korb & Nicholson, 2004; ch 8)
A Bayesian information-theoretic scoring function with MCMC (sampling search); returns dags and patterns. Performance similar to BDe/BGe. Supports priors and hybrid local structure.

GES

Greedy Equivalence Search (GES)

- Product of the CMU-Microsoft group (Meek, 1996; Chickering, 2002)
- Two-stage greedy search: Begin with unconnected pattern
  1. Greedily add single arcs until reaching a local maximum
  2. Prune back edges which don't contribute to the score
- Uses a Bayesian score over patterns only
- Implemented in TETRAD and Murphy's BNT

Recent Extensions to CaMML

Two significant enhancements have been added in the last few years.

Expert priors (O'Donnell et al., 2006b)

- Being Bayesian, it is relatively easy to incorporate non-default priors into CaMML. We've done this in various ways, specifying strengths for:
  - A prior dag, computing a prior distribution via edit distance
  - Arc densities
  - Topological orders, total or partial

Hybrid model learning (O'Donnell et al., 2006a)

- Allowing varying representations of local structure (CPTs, d-trees, logit model) throughout the network

Causal Discovery: Summary

- Constraint-based learning is simple and intuitive
- Metric learning is neither, but generally more effective
- CaMML uses an efficient coding for BNs and stochastic search
  - though the TOM space, not dag space
  - with a default prior rewarding richer dag models
  - with extensions allowing incorporation of expert prior information

FIN

References I

C. Boutillier, N. Friedman, M. Goldszmidt, D. Koller (1996) "Context-specific independence in Bayesian networks," in Horvitz & Jensen (eds.) *UAI 1996*, 115-123.

N. Cartwright (1979) Causal laws and effective strategies. *Noûs, 13,* 419-437.

D.M. Chickering (1995). A tranformational characterization of equivalent Bayesian network structures. *UAI* (pp. 87-98). Morgan Kaufmann.

D.M. Chickering (2002) . Optimal structure idenification with greedy search. *Journal of Machine Leaning Research, 3,* 507-559.

G.F. Cooper (1990). The computational complexity of probabilistic inference using belief networks. *Artificial Intelligence, 42,* 393-405.

G.F. Cooper and E. Herskovits (1991) A Bayesian method for constructing Bayesian belief networks from databases. *UAI,* 86-94.

References II

H. Dai, K.B. Korb, C.S. Wallace and X. Wu (1997) A study of casual discovery with weak links and small samples. *15th International Joint Conference on Artificial Intelligence (IJCAI),* pp. 1304-1309. Morgan Kaufmann.

C. Fell (2006) *Causal discovery: The incorporation of latent variables in causal discovery using experimental data.* Honours Thesis, Clayton School of IT, Monash University.

N. Friedman (1997) The Bayesian structural EM algorithm. *UAI* (pp. 129-138). Morgan Kaufmann.

N. Friedman, K. Murphy and S. Russell (1998) Learning the structure of dynamic probabilistic networks. *UAI* (pp. 139-148).

T. Handfield, C.R. Twardy, K.B. Korb and G. Oppy (2008) The metaphysics of causasl models: Where's the biff? *Erkenntnis, 68,* 149-168.

D. Heckerman and D. Geiger (1995) Learning Bayesian networks: A unification for discrete and Gaussian domains.*UAI,* 274-284.

References III

K.B. Korb, C. Kopp and L. Allison (1997). *A statement on higher education policy in Australia*. Technical Report 97/318, Dept Computer Science, Monash University.

K.B. Korb and A.E. Nicholson (2004) *Bayesian Artificial Intelligence.* CRC/Chapman Hall.

K.B. Korb and E. Nyberg (2006) The power of intervention. *Minds and Machines, 16,* 289-302.

W. Lam and F. Bacchus (1993) Learning Bayesian belief networks: An approach based on the MDL principle. *Computational Intelligence, 10,* 269-293.

D. Madigan, S.A. Andersson, M.D. Perlman & C.T. Volinsky (1996) Bayesian model averaging and model selection for Markov equivalence classes of acyclic digraphs. *Comm in Statistics: Theory and Methods, 25,* 2493-2519.

C. Meek (1996) *Graphical models: Selectiong causal and statistical models.* PhD disseration, Philosophy, Carnegie Mellon University.

References IV

N. Metropolis, Rosenbluth, Rosenbluth, Teller and Teller (1953) Equations of state calculations by fast computing machines. *Jrn Chemical Physics, 21,* 1087-1091.

J.R. Neil and K.B. Korb (1999) The Evolution of Causal Models: A Comparison of Bayesian Metrics and Structure Priors. *Knowledge Discovery and Data Mining: Third Pacific-Asia Conference* (pp. 432-437). Springer Verlag.

R. Neapolitan (1990) *Probabilistic Reasoning in Expert Systems*. Wiley.

R. O'Donnell, L. Allison and K.B. Korb (2006a) Learning hybrid Bayesian networks by MML. *Australian Joint Conference on AI* pp. 192-203. Springer.

R. O'Donnell, A.E. Nicholson, L. Allison and K.B. Korb (2006b) Causal discovery with prior information. *Australian Joint Conference on AI* pp. 1162-1167. Springer.

R. O'Donnell, K.B. Korb and L. Allison (2007) *Causal KL.* Technical Report 2007/207, Clayton School of IT, Monash University.

References V

J. Pearl (1988) *Probabilistic Reasoning in Intelligent Systems*, Morgan Kaufmann.

H. Reichenbach (1956). *The Direction of Time*. Univ of Calif.

W. C. Salmon (1984) *Scientific Explanation and the Causal Structure of the World.* Princeton Univ.

Elliott Sober (1988) The principle of the common cause. In J. Fetzer (ed.) *Probability and Causality* (pp. 211-28). Kluwer.

D. Spiegelhalter & S. Lauritzen (1990) "Sequential Updating of Conditional Probabilities on Directed Graphical Structures," *Networks, 20,* 579-605.

P. Spirtes, C. Glymour and R. Scheines (1993) *Causation, Prediction and Search.* Springer.

P. Spirtes, C. Glymour and R. Scheines (2000) *Causation, Prediction and Search,* 2nd ed. MIT.

D. Steel (2006) Homogeneity, selection and the faithfulness condition. *Minds and Machines, 16*, 303-317.

P. Suppes (1970) *A Probabilistic Theory of Causality*. North Holland.

References VI

A. Tucker and X. Liu (2003) Learning dynamic Bayesian networks from multivariate timer series with changing dependencies. *Advances in Intelligent Data Analysis* (pp. 100-110). Springer.

T.S. Verma and J. Pearl (1991) Equivalence and synthesis of causal models. *Uncertainty in Artificial Intelligence 6* (pp. 255-268). Elsevier.

C.S. Wallace and D. Boulton (1968) An information measure for classification. *Computer Journal, 11,* 185-194.

C. S. Wallace and K. B. Korb (1999) Learning linear causal models by MML sampling. In A. Gammerman (ed.) *Causal Models and Intelligent Data Management.* Springer.

C. S. Wallace, K. B. Korb, and H. Dai (1996) Causal discovery via MML. *13th International Conference on Machine Learning* (pp. 516-524). Morgan Kaufmann.

J. Williamson (2005) *Bayesian Nets and Causality*. Oxford.

J. H. Wigmore (1913). The problem of proof. *Illinois Law Journal 8,* 77-103.

References VII

S. Wright (1921). Correlation and causation. *Journal of Agricultural Research, 20*, 557-585.

S. Wright (1934). The method of path coefficients. *Annals of Mathematical Statistics, 5*, 161-215.