Estimation of (causal?) structure

Steffen Lauritzen, University of Oxford

Graphical Models and Inference, Lecture 14, Michaelmas Term 2011

November 21, 2011

Causal interpretations are tied to the notion of *conditioning by intervention*

$$P(X = x \mid Y \leftarrow y) = P\{X = x \mid \mathrm{do}(Y = y)\} = p(x \,||\, y), \qquad (1)$$

which in general is quite different from conventional conditioning or *conditioning by observation* which is

$$P(X = x \mid Y = y) = P\{X = x \mid \mathrm{is}(Y = y)\} = p(x \mid y) = p(x, y)/p(y).$$

A causal interpretation of a Bayesian network involves giving (1) a special form.

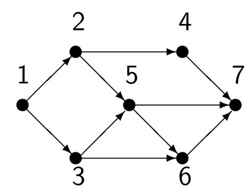
We say that a BN is *causal w.r.t. atomic interventions at* $B \subseteq V$ if it holds for any $A \subseteq B$ that

$$\begin{aligned} p(x \,||\, x_A^*) &= \prod_{v \in V \setminus A} p(x_v \,|\, x_{\mathrm{pa}(v)}) \Big|_{x_A = x_A^*} \\ &= \frac{\prod_{v \in V} p(x_v \,|\, x_{\mathrm{pa}(v)})}{\prod_{v \in A} p(x_v \,|\, x_{\mathrm{pa}(v)})} \Big|_{x_A = x_A^*}. \end{aligned}$$

For $A = \emptyset$ we obtain standard factorisation.

Note that *conditional distributions* $p(x_v \,|\, x_{\mathrm{pa}(v)})$ are *stable under interventions* which do not involve x_v. Such assumption must be justified in any given context.

An example



$$\begin{aligned} p(x \,||\, x_5^*) &= p(x_1)p(x_2 \mid x_1)p(x_3 \mid x_1)p(x_4 \mid x_2) \\ &\times p(x_6 \mid x_3, x_5^*)p(x_7 \mid x_4, x_5^*, x_6) \end{aligned}$$

whereas

$$\begin{aligned} p(x \mid x_5^*) &\propto p(x_1)p(x_2 \mid x_1)p(x_3 \mid x_1)p(x_4 \mid x_2) \\ &\times p(x_5^* \mid x_2, x_3)p(x_6 \mid x_3, x_5^*)p(x_7 \mid x_4, x_5^*, x_6) \end{aligned}$$

DAG $\mathcal{D}$ can also represent structural equation system:

$$X_v \leftarrow g_v(x_{\mathrm{pa}(v)}, U_v), v \in V, \tag{2}$$

where g_v are fixed functions and U_v are independent random disturbances.

Intervention in structural equation system can be made by *replacement*, i.e. so that $X_v \leftarrow x_v^*$ is replacing the corresponding line in 'program' (2).

Corresponds to *g_v and U_v being unaffected by the intervention* if intervention is not made on node v. Hence the equation is *structural*.

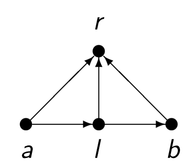
Intervention by replacement in structural equation system implies $\mathcal{D}$ causal for distribution of $X_v, v \in V$.

Occasionally used for *justification* of CBN.

Ambiguity in choice of g_v and U_v makes this problematic.

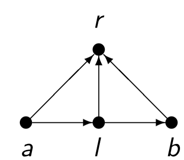
May take *stability of conditional distributions* as a primitive rather than structural equations.

Structural equations more expressive when choice of g_v and U_v can be externally justified.



a - treatment with AZT; l - intermediate response (possible lung disease); b - treatment with antibiotics; r - survival after a fixed period.
Predict survival if $X_a \leftarrow 1$ and $X_b \leftarrow 1$, assuming stable conditional distributions.

G-computation



$$\begin{aligned} p(1_r \,||\, 1_a, 1_b) &= \sum_{x_l} p(1_r, x_l \,||\, 1_a, 1_b) \\ &= \sum_{x_l} p(1_r \,|\, x_l, 1_a, 1_b) p(x_l \,|\, 1_a). \end{aligned}$$

V set of variables, assume DAG $\mathcal{D}$ unknown and P given.

Assume joint distribution P *faithful* to $\mathcal{D}$:

$$X_A \perp\!\!\!\perp X_B \mid X_S \iff A \perp_{\mathcal{D}} B \mid S$$

Most distributions are faithful

Find $\mathcal{D}$ which matches conditional independence relations of P.

$\mathcal{D}$ and $\mathcal{D}'$ are *Markov equivalent* if the separation relations $\perp_{\mathcal{D}}$ and $\perp_{\mathcal{D}'}$ are identical.

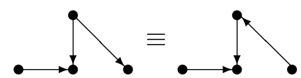
$\mathcal{D}$ can only be determined up to Markov equivalence.

Markov equivalence

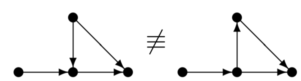
$\mathcal{D}$ and $\mathcal{D}'$ are equivalent if and only if:

1. $\mathcal{D}$ and $\mathcal{D}'$ have same *skeleton* (ignoring directions)
2. $\mathcal{D}$ and $\mathcal{D}'$ have same unmarried parents

so



but



Step 1: Identify skeleton, using that, for a faithful distribution

$$u \not\sim v \iff \exists S \subseteq V \setminus \{u, v\} : X_u \perp\!\!\!\perp X_v \mid X_S.$$

Begin with complete graph and check first for $S = \emptyset$ and remove edges when independence holds. Then continue for increasing cardinality of S.

PC-algorithm exploits that only S with $S \subseteq \text{ne}(u)$ or $S \subseteq \text{ne}(v)$ needs checking, where ne refers to current skeleton graph.

Step 2: Identify directions to be consistent with independence relations found in Step 1.

Exact properties of PC-algorithm

If P is faithful to DAG $\mathcal{D}$, PC-algorithm finds $\mathcal{D}'$ equivalent to $\mathcal{D}$.
It uses N independence checks where N is at most

$$N \leq 2\binom{|V|}{2}\sum_{i=0}^{d}\binom{|V|-1}{i} \leq \frac{|V|^{d+1}}{(d-1)!},$$

where d is the maximal degree of any vertex in $\mathcal{D}$.
So worst case complexity is exponential, but algorithm fast for sparse graphs.

Empirical independence checks

For finite samples, independence checks can be performed as

- significance tests for independence;
- asymptotic model selection criteria such as BIC, AIC, etc.

$$IC_{\kappa}(\mathcal{D}) = \log \hat{L}(\mathcal{D}) - \kappa \dim(\mathcal{D})$$

  with $\kappa = 1$ for AIC , or $\kappa = \frac{1}{2} \log N$ for BIC .
- Bayes factors in local Bayesian approach;

Data uncertainty and causal discovery

Situation less clear if P is not known, but estimated:

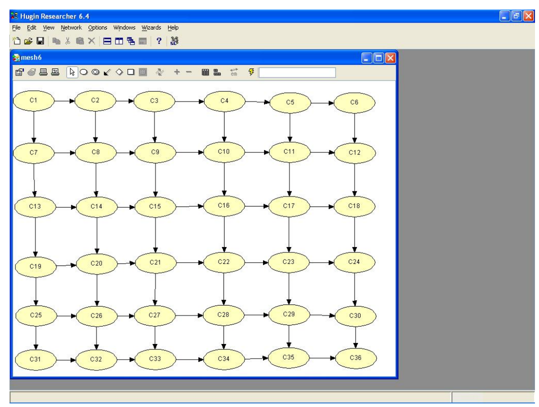
Constraint-based: Independence checks may randomly give errors.

Algorithms more robust than PC exist.

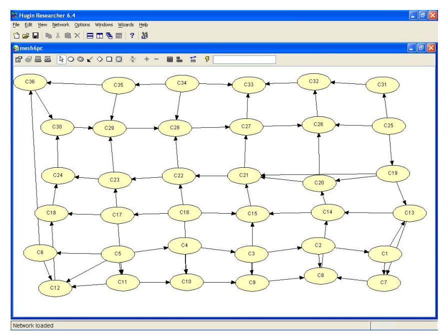
Most checks are made with separation set S small, so power high.

Asymptotically correct if e.g. marginal BIC or BF used in checks.

Markov mesh model

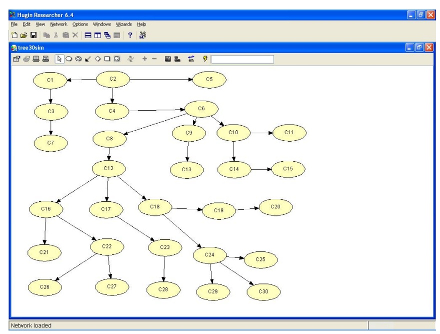


PC algorithm



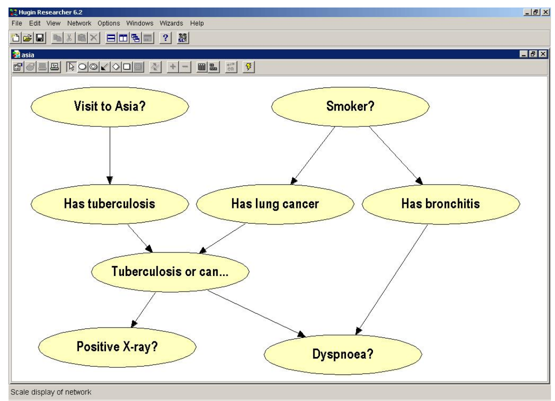
PC algorithm (HUGIN), 10000 simulated cases

Tree model

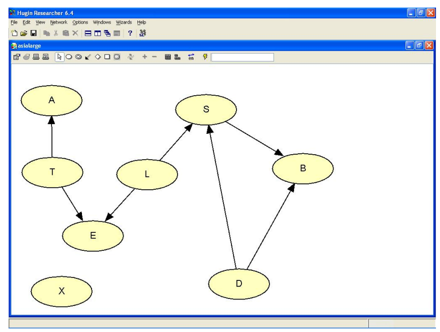


PC algorithm, 10000 cases, correct reconstruction

Chest clinic

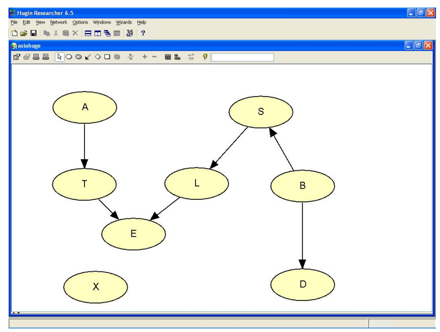


PC algorithm



10000 simulated cases

PC algorithm

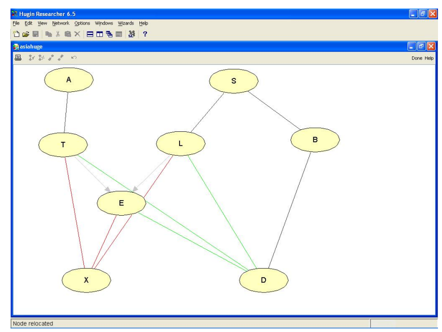


100000 simulated cases

This algorithm avoids early acceptance of conditional *in*dependences.

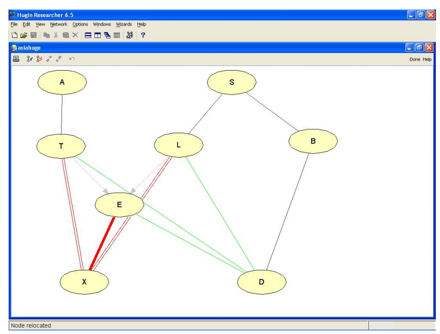
- if a dependence is established, believe it;
- if an independence is established, put it on hold for a while;
- proceed as in the PC algorithm, but insist on *necessary path condition* (NPC): if a conditional dependence is established at some point, there must be a connecting path explaining it.

Non-unique identification, involving *ambiguous regions*. User may resolve these.



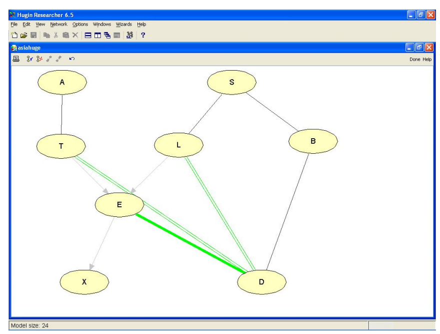
First stage

NPC algorithm



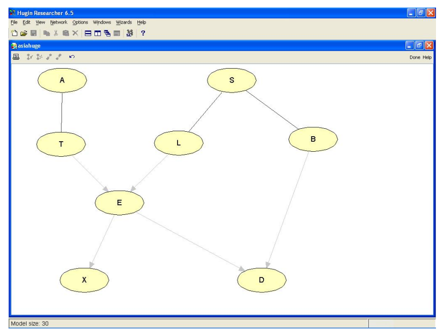
Resolving one ambiguity

NPC algorithm



Resolving another

NPC algorithm



Final model

Searches directly in equivalence classes of DAGS.

Define *score function* $\sigma(P, \mathcal{D})$, measuring the adequacy of $\mathcal{D}$ for P with the property that

$$\mathcal{D} \equiv \mathcal{D}' \Rightarrow \sigma(P, \mathcal{D}) = \sigma(P, \mathcal{D}').$$

Typically the score function will penalise $\mathcal{D}$ with unnecessary many links. BIC score satisfies condition. So does fully Bayesian score for certain classes of priors.

Equivalence class with maximal score is sought.

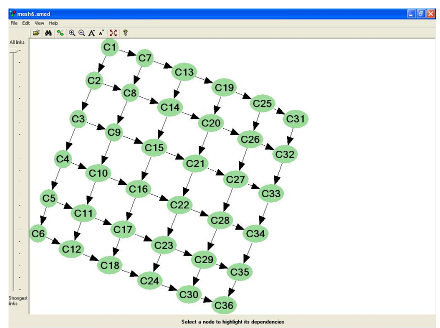
Greedy equivalence search

1. Initialize with empty DAG
2. Repeatedly search among equivalence classes with a single additional edge and go to class with highest score - until no improvement.
3. Repeatedly search among equivalence classes with a single edge less and move to one with highest score - until no improvement.

For suitable score functions, this algorithm identifies correct equivalence class for P.

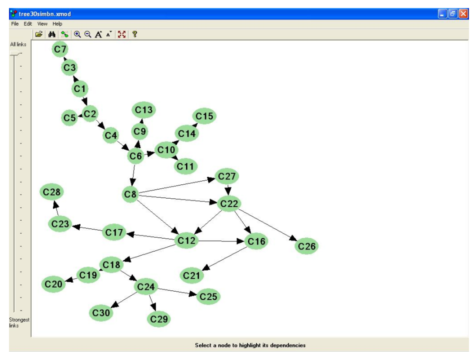
Asymptotically correct if using BIC or fully Bayesian approach.

Bayesian GES om Markov mesh

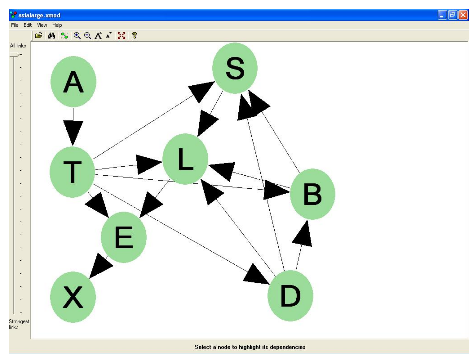


Crudest algorithm (WinMine), 10000 simulated cases

Bayesian GES on tree

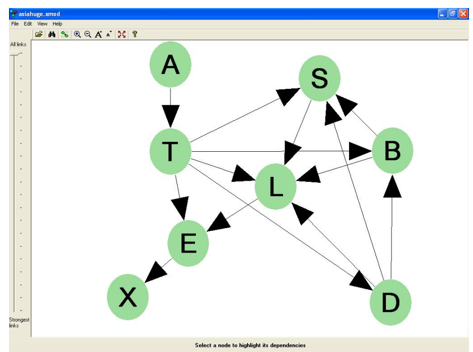


Bayesian GES on Chest Clinic



10000 cases

Bayesian GES on Chest Clinic



100000 cases

More serious that *one would rarely expect all causally relevant variables to be measured.* Selection effects are also an issue.

More relevant to assume data obtained from P by *marginalisation* to subset V and *conditioning* with subset C so $W = V \cup U \cup C$, data represents P_V^C, where P is faithful to some DAG $\mathcal{D}$.

Graphs that describe independence relations in such cases are *Maximal Ancestral Graphs. Constraint-based methods for identifying MAGs exist: FCI-algorithm.*

Bayesian approach for MAGs seems out of hand.