Structure estimation and Bayes Factors

Lecture 8
Saint Flour Summerschool, July 13, 2006

Steffen L. Lauritzen, University of Oxford

Overview of lectures

1. Conditional independence and Markov properties
2. More on Markov properties
3. Graph decompositions and junction trees
4. Probability propagation and related algorithms
5. Log-linear and Gaussian graphical models
6. Hyper Markov laws
7. More on Hyper Markov Laws
8. *Structure estimation and Bayes factors*
9. More on structure estimation.

Hyper Markov Laws

Identify $\theta \in \Theta$ and $P_\theta \in \mathcal{P}$, so e.g. θ_A denotes the marginal distribution of X_A under P_θ and $\theta_{A \mid B}$ the family of conditional distributions of X_A given X_B, etc.

For a law $\mathcal{L}$ on Θ we write

$$A \perp\!\!\!\perp_{\mathcal{L}} B \,|\, S \iff \theta_{A \mid S} \perp\!\!\!\perp_{\mathcal{L}} \theta_{B \mid S} \,|\, \theta_S.$$

A law $\mathcal{L}$ on Θ is *hyper Markov* w.r.t. $\mathcal{G}$ if

(i) All $\theta \in \Theta$ are globally Markov w.r.t. $\mathcal{G}$;

(ii) $A \perp\!\!\!\perp_{\mathcal{L}} B \,|\, S$ whenever S is *complete* and $A \perp_{\mathcal{G}} B \,|\, S$.

Hyper Markov property

The hyper Markov property has a simple formulation in terms of junction trees:

Arrange the prime components $\mathcal{Q}$ of $\mathcal{G}$ in a junction tree $\mathcal{T}$ with complete separators $\mathcal{S}$ and consider the *extended junction tree* $\overline{\mathcal{T}}$ which is the (bipartite) tree with $\mathcal{Q} \cup \mathcal{S}$ as vertices and edges from separators to prime components so that $C \sim S \sim D$ in $\overline{\mathcal{T}}$ if and only if $C \sim D$ in $\mathcal{T}$.

Next, associate θ_A to A for each $A \in \mathcal{Q} \cup \mathcal{S}$. It then holds that

$\mathcal{L}$ is hyper Markov on $\mathcal{G}$ if and only if $\{\theta_A, A \in \mathcal{Q} \cup \mathcal{S}\}$ is globally Markov w.r.t. the extended junction tree $\overline{\mathcal{T}}$.

Directed hyper Markov property

$\mathcal{L} = \mathcal{L}(\theta)$ is *directed hyper Markov* w.r.t. a DAG $\mathcal{D}$ if θ is directed Markov on $\mathcal{D}$ for all $\theta \in \Theta$ and

$$\theta_{v \mid \mathrm{pa}(v)} \perp\!\!\!\perp_{\mathcal{L}} \theta_{\mathrm{nd}(v)} \mid \theta_{\mathrm{pa}(v)}.$$

If $\mathcal{D}$ is perfect, $\mathcal{L}$ is directed hyper Markov w.r.t. $\mathcal{D}$ if and only if $\mathcal{L}$ is hyper Markov w.r.t. $\mathcal{G} = \sigma(\mathcal{D}) = \mathcal{D}^m$.

Meta Markov models

For $A, B \subseteq V$ identify

$$\theta_{A\cup B} = (\theta_{B\,|\,A}, \theta_A) = (\theta_{A\,|\,B}, \theta_B).$$

A and B are *meta independent* w.r.t. $\mathcal{P}$ given S, denoted $A \ddagger_{\mathcal{P}} B \,|\, S$, if the pair of conditional distributions $(\theta_{A\,|\,S}, \theta_{B\,|\,S})$ vary in a product space when θ_S is fixed.

The family $\mathcal{P}$, or Θ, is *meta Markov* w.r.t. $\mathcal{G}$ if

(i) All $\theta \in \Theta$ are globally Markov w.r.t. $\mathcal{G}$;

(ii) $A \perp_{\mathcal{G}} B \,|\, S \implies A \ddagger_{\mathcal{P}} B \,|\, S$ whenever S is complete.

Hyper Markov laws and meta Markov models

Hyper Markov laws live on meta Markov models.

A Gaussian graphical model with graph $\mathcal{G}$ is meta Markov on $\mathcal{G}$.

A log-linear model $\mathcal{P}_{\mathcal{A}}$ is meta Markov on its dependence graph $\mathcal{G}(\mathcal{A})$ if and only if $S \in \mathcal{A}$ for any minimal complete separator S of $\mathcal{G}(\mathcal{A})$.

In particular, *if $\mathcal{A}$ is conformal, $\mathcal{P}_{\mathcal{A}}$ is meta Markov.*

Maximum likelihood in meta Markov models

If the following conditions are satisfied:

(i) Θ is meta Markov w.r.t. $\mathcal{G}$;

(ii) For any prime component Q of $\mathcal{G}$, Θ_Q is a full and regular exponential family,

the MLE $\hat{\theta}$ of the unknown distribution θ will follow a hyper Markov law over Θ under P_θ.

In particular, *this holds for any Gaussian graphical model and any meta Markov log-linear model.*

Strong hyper and meta Markov properties

A meta Markov model is *strongly meta Markov* if $\theta_{A \mid S} \ddagger_{\mathcal{P}} \theta_S$ for all complete separators S.

Similarly, a hyper Markov model is *strongly hyper Markov* if $\theta_{A \mid S} \perp\!\!\!\perp_{\mathcal{L}} \theta_S$ for all complete separators S.

A directed hyper Markov model is *strongly directed hyper Markov* if $\theta_{v \mid \mathrm{pa}(v)} \perp\!\!\!\perp_{\mathcal{L}} \theta_{\mathrm{pa}(v)}$ for all $v \in V$.

Gaussian graphical models and log-linear meta Markov models are strong meta Markov models.

Conjugacy of hyper Markov properties

If $\mathcal{L}$ is a prior law over Θ and $X = x$ is an observation from θ, $\mathcal{L}^* = \mathcal{L}(\theta \,|\, X = x)$ denotes the *posterior law* over Θ.

If $\mathcal{L}$ is hyper Markov w.r.t. $\mathcal{G}$ so is $\mathcal{L}^*$.

If $\mathcal{L}$ is strongly hyper Markov w.r.t. $\mathcal{G}$ so is $\mathcal{L}^*$.

In the latter case, *the update of $\mathcal{L}$ is local to prime components*, i.e.

$$\mathcal{L}^*(\theta_Q) = \mathcal{L}^*_Q(\theta_Q) = \mathcal{L}_Q(\theta_Q \,|\, X_Q = x_Q)$$

and *the marginal distribution p of X is globally Markov w.r.t. $\overline{\mathcal{G}}$,* where

$$p(x) = \int_\Theta P(X = x \,|\, \theta) \mathcal{L}(d\theta).$$

Conjugate exponential families

For a k-dimensional exponential family

$$p(x \mid \theta) = b(x)e^{\theta^\top t(x) - \psi(\theta)}$$

the *standard conjugate family* is given as

$$\pi(\theta \mid a, \kappa) \propto e^{\theta^\top a - \kappa\psi(\theta)}$$

for $(a, \kappa) \in \mathcal{A} \subseteq \mathcal{R}^k \times \mathcal{R}_+$, where $\mathcal{A}$ is determined so that the normalisation constant is finite.

Posterior updating from $(x_1, \ldots, x_n)$ with $t = \sum_i t(x_i)$ is then made as $(a^*, \kappa^*) = (a + t, \kappa + n)$.

Hyper inverse Wishart and Dirichlet laws

Gaussian graphical models are canonical exponential families. The standard family of conjugate priors have densities

$$\pi(K \mid \Phi, \delta) \propto (\det K)^{\delta/2} e^{-\operatorname{tr}(K\Phi)}, K \in \mathcal{S}^+(\mathcal{G}).$$

These laws are termed *hyper inverse Wishart laws* as Σ follows an inverse Wishart law for complete graphs. *For chordal graphs, each marginal law $\mathcal{L}_C, \mathcal{C}$ of Σ_C is inverse Wishart.*

The standard conjugate prior law for log-linear meta Markov models are termed *hyper Dirichlet laws*. If $\mathcal{G}$ is *chordal, each induced marginal law $\mathcal{L}_C, C \in \mathcal{C}$ is a standard Dirichlet law*.

Conjugate prior laws are strong hyper Markov

If Θ is meta Markov and Θ_Q are full and regular exponential families for all prime components Q, the standard conjugate prior law is strongly hyper Markov w.r.t. $\mathcal{G}$.

This is in particular true for the hyper inverse Wishart laws and the hyper Dirichlet laws.

Thus, for the hyper inverse and hyper Dirichlet laws we have simple *local updating* based on *conjugate priors* for Bayesian inference.

Estimation of structure

Previous lectures have considered the *graph $\mathcal{G}$ defining the model as known* and inference was concerning an unknown P_θ with $\theta \in \Theta$.

The last two lectures are concerned with *inference concerning the graph $\mathcal{G}$,* specifying only a family Γ of possible graphs.

Methods must scale well with data size, as *many* structures and *huge* collections of data are to be considered.

Structure estimation is also known as *model selection* (mainstream statistics) *system identification* (engineering), *structural learning* (AI or machine learning.)

Examples of structural assumptions

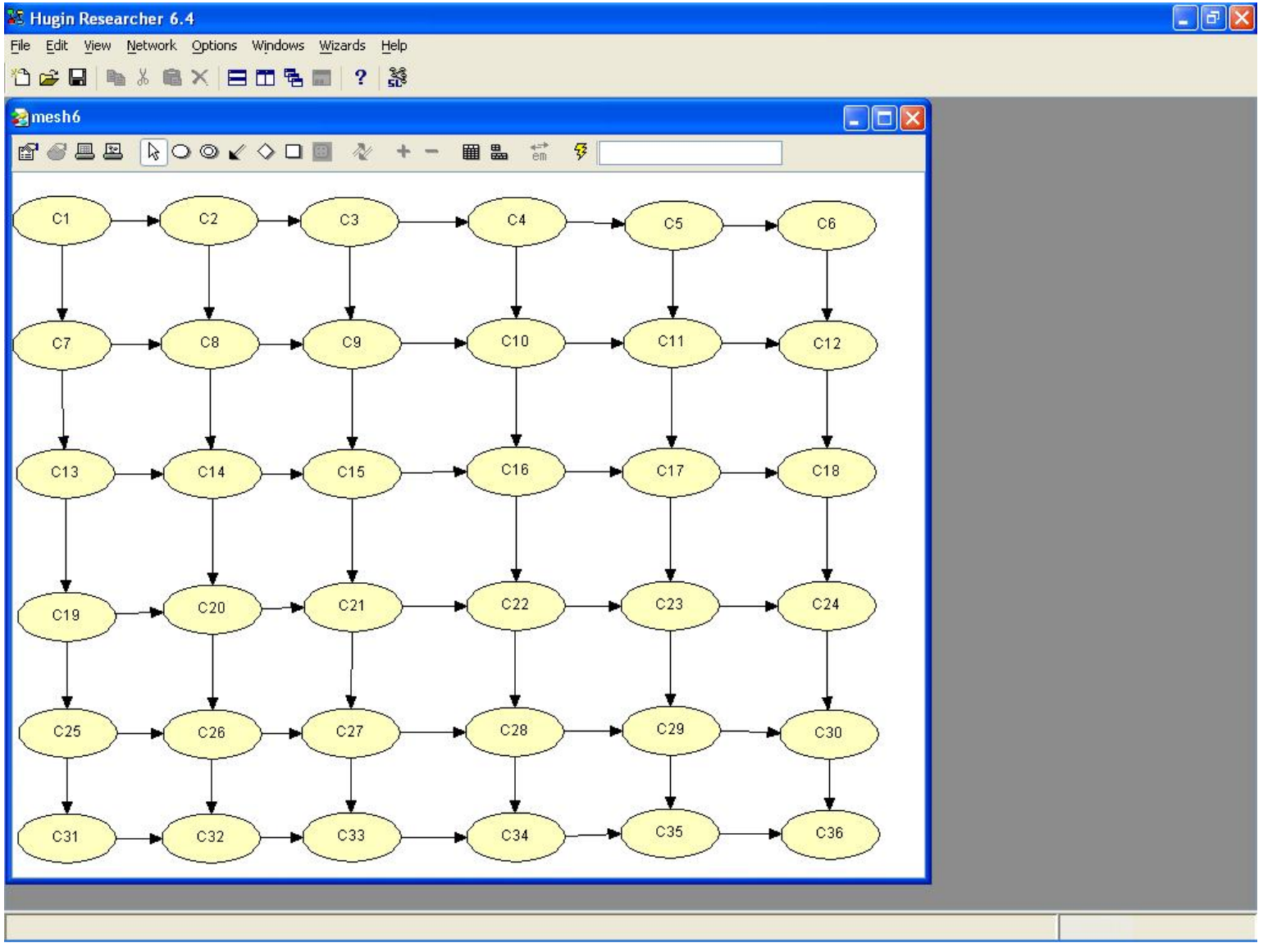
Different situations occur depending on the type of assumptions concerning Γ.

1. Γ is the set of *undirected graphs* over V;
2. Γ is the set of *chordal graphs* over V;
3. Γ is the set of *forests* over V;
4. Γ is the set of *trees* over V;
5. Γ is the set of *directed acyclic graphs* over V;
6. Other conditional independence structures

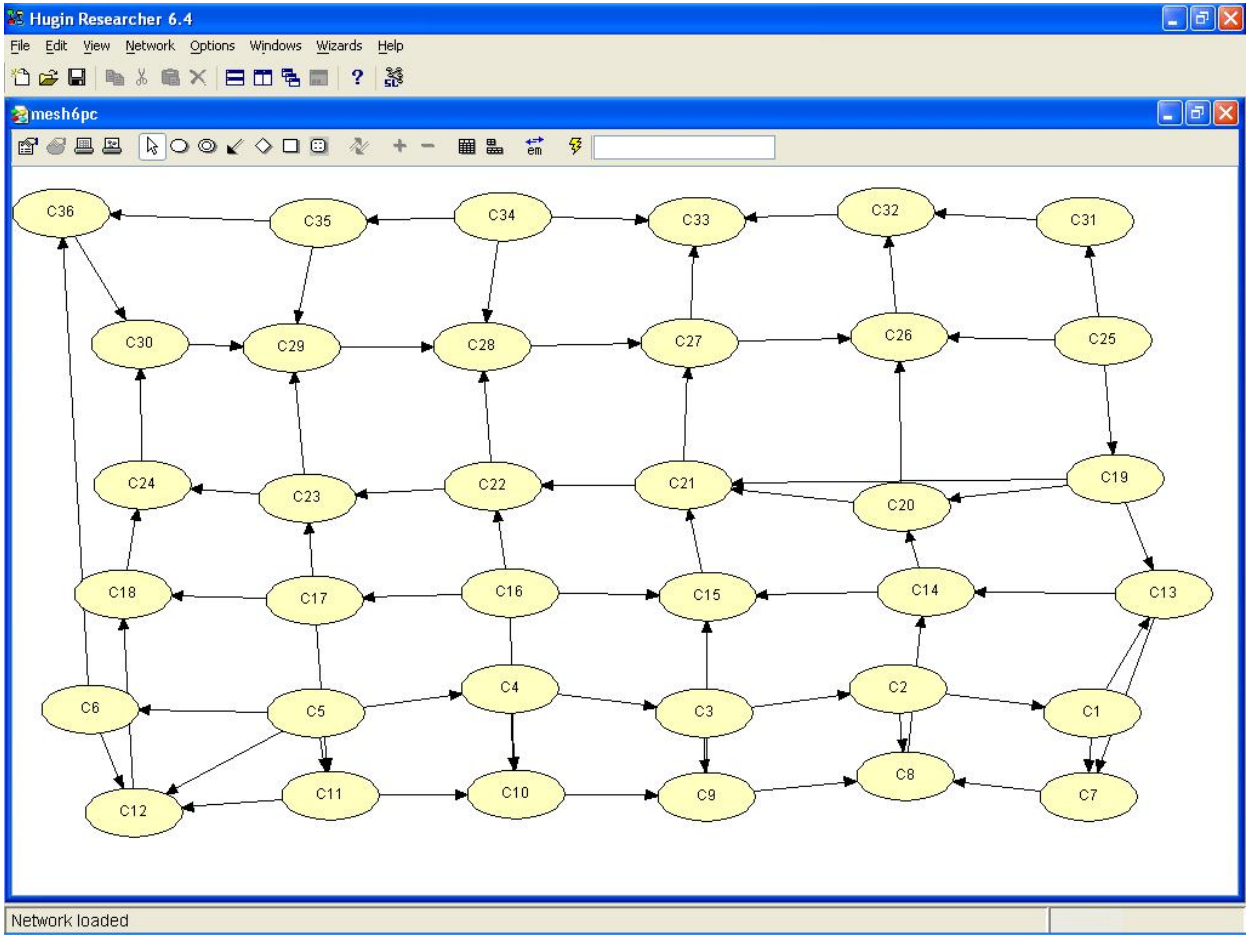
Why estimation of structure?

- Parallel to e.g. density estimation
- Obtain quick overview of relations between variables in complex systems
- Data mining
- Gene regulatory networks
- Reconstructing family trees from DNA information
- Methods exist, but need better understanding of their *statistical properties.*

Markov mesh model

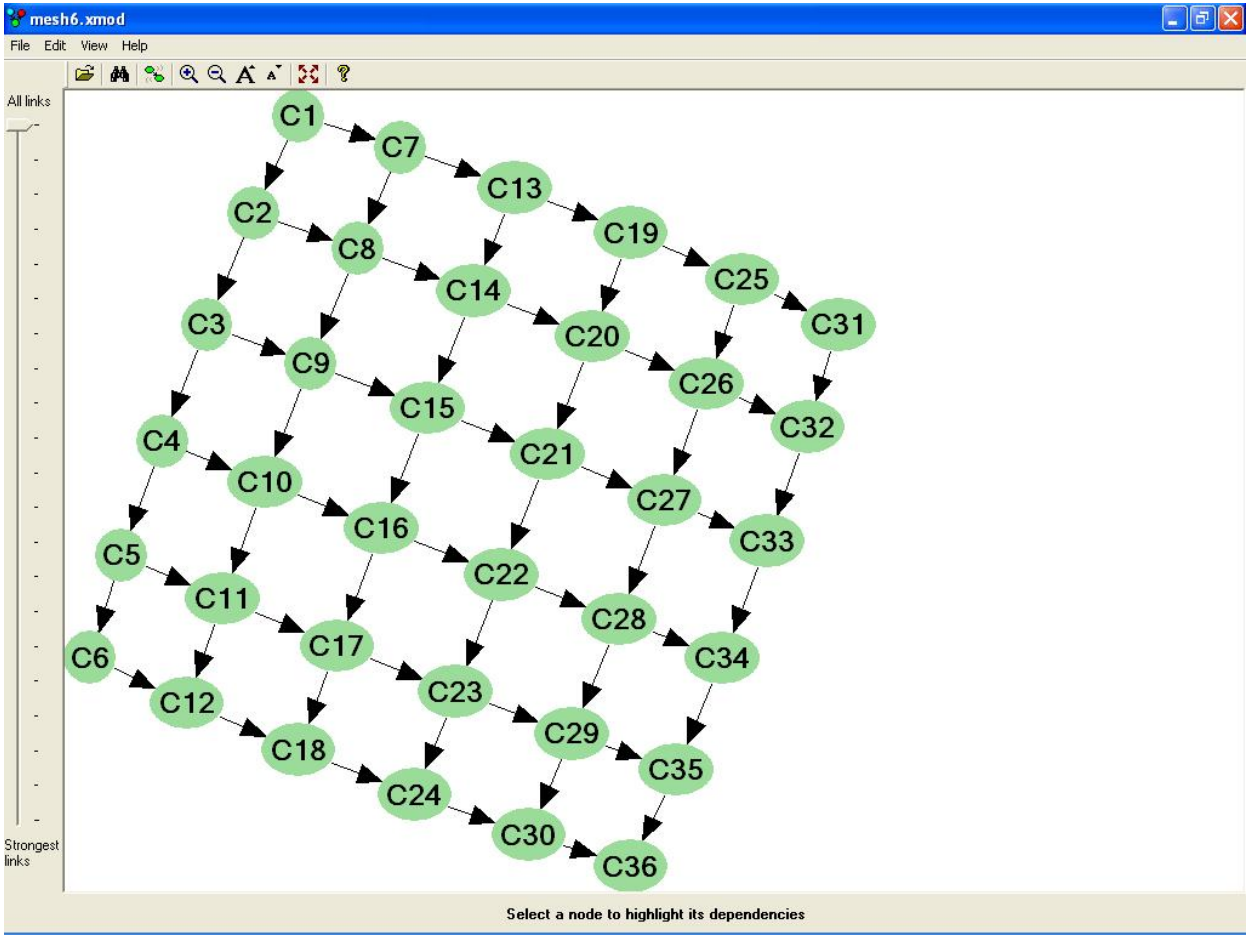


PC algorithm



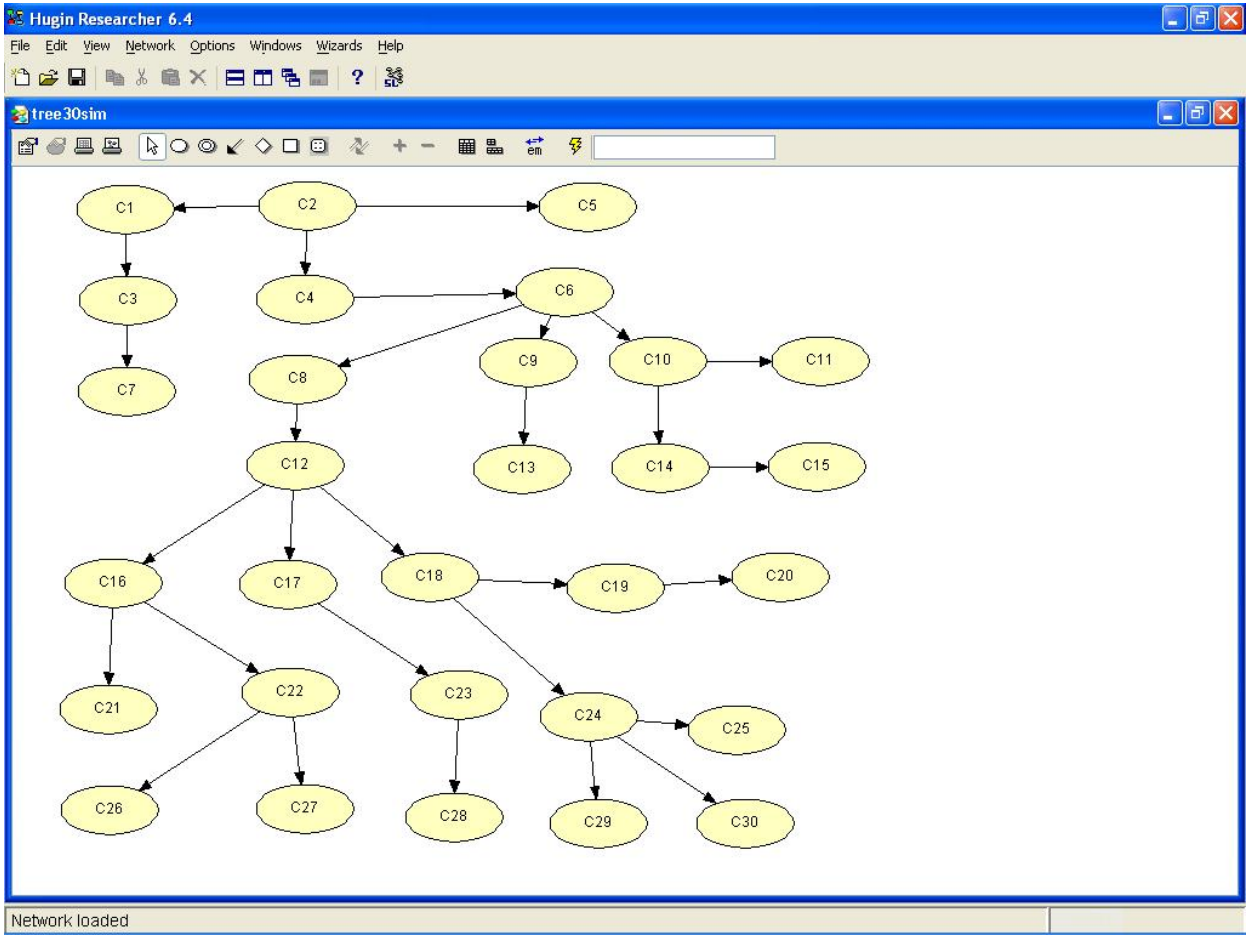
Crudest algorithm (HUGIN), 10000 simulated cases

Bayesian GES



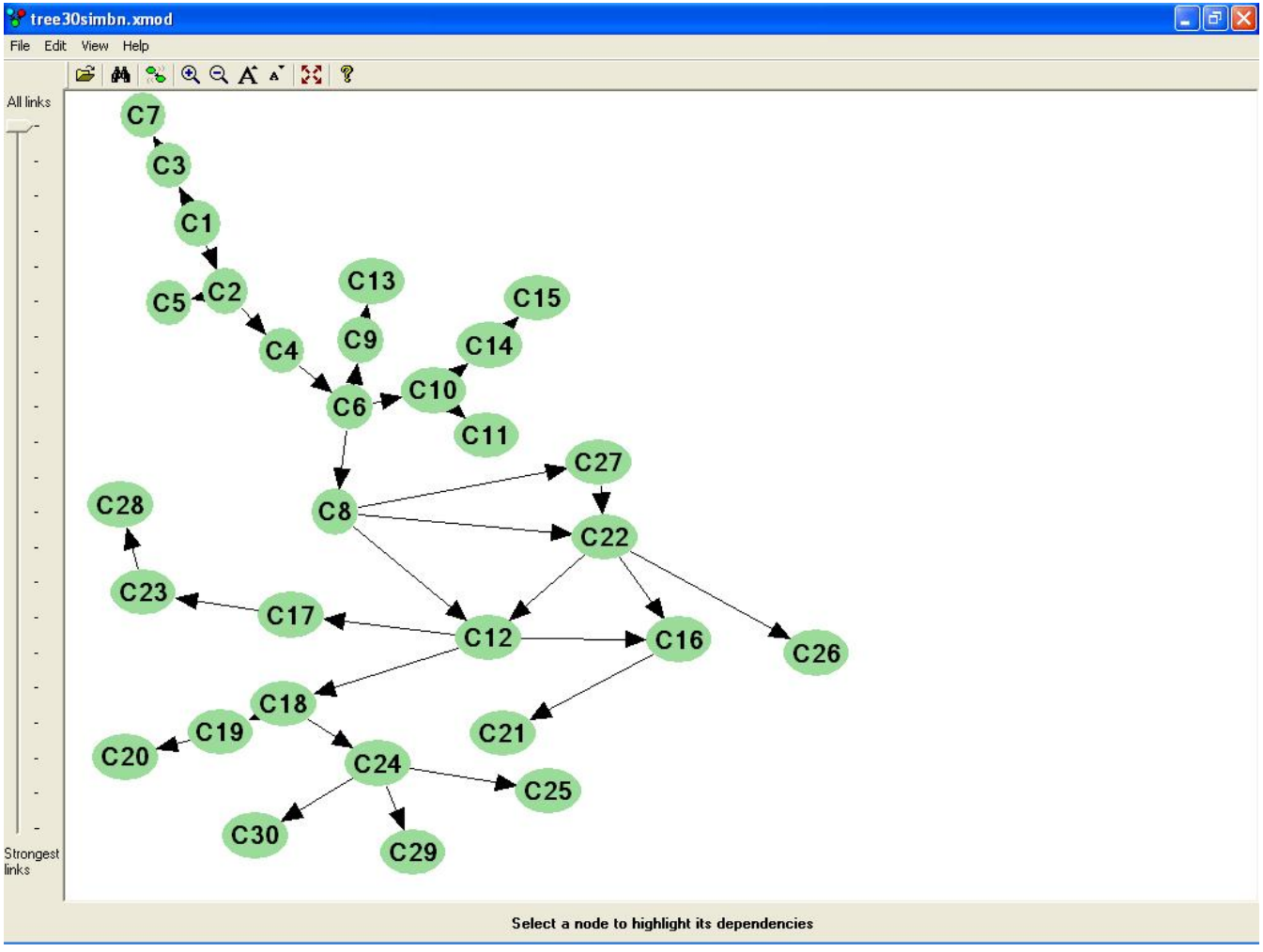
Crudest algorithm (WinMine), 10000 simulated cases

Tree model

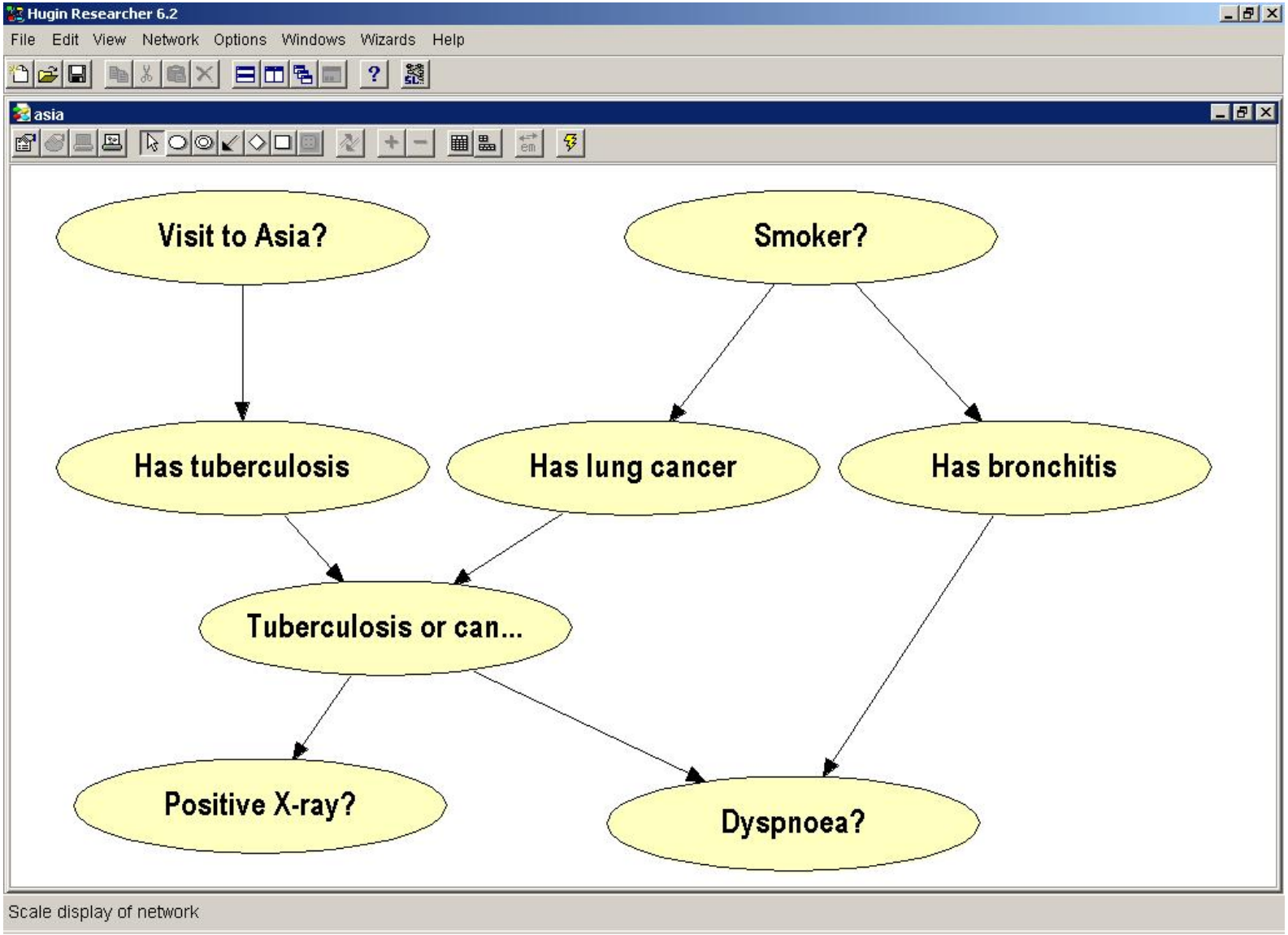


PC algorithm, 10000 cases, correct reconstruction

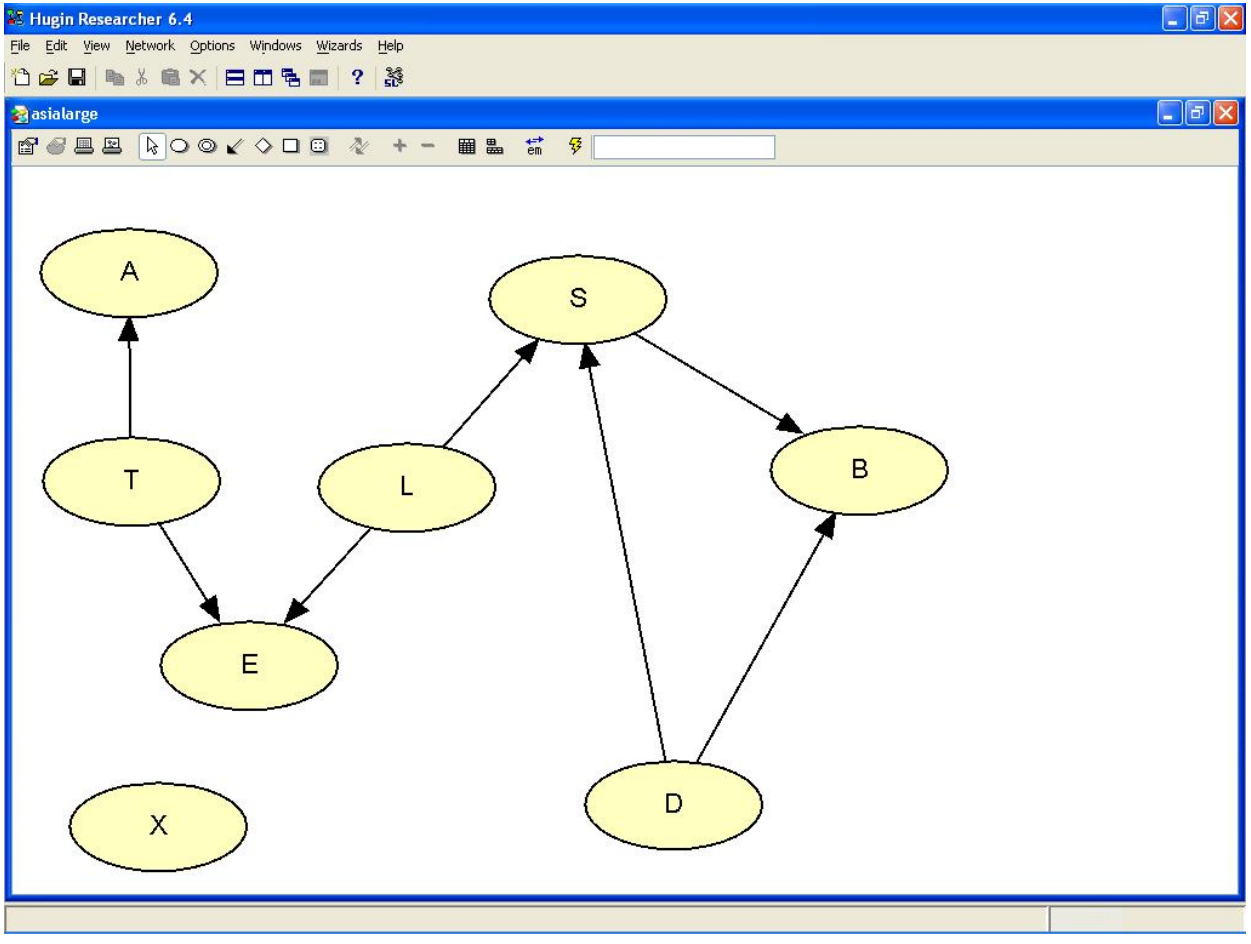
Bayesian GES on tree



Chest clinic

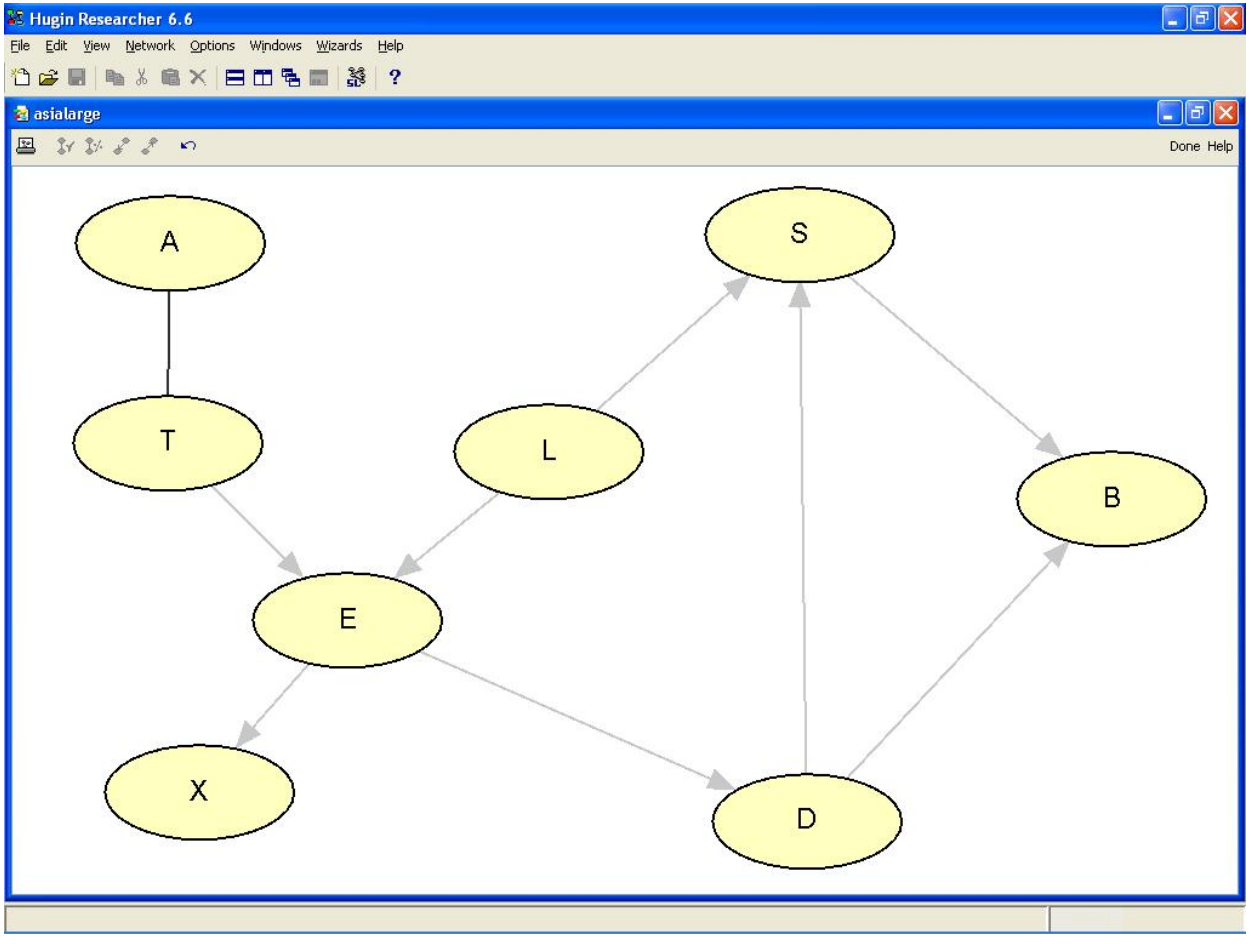


PC algorithm



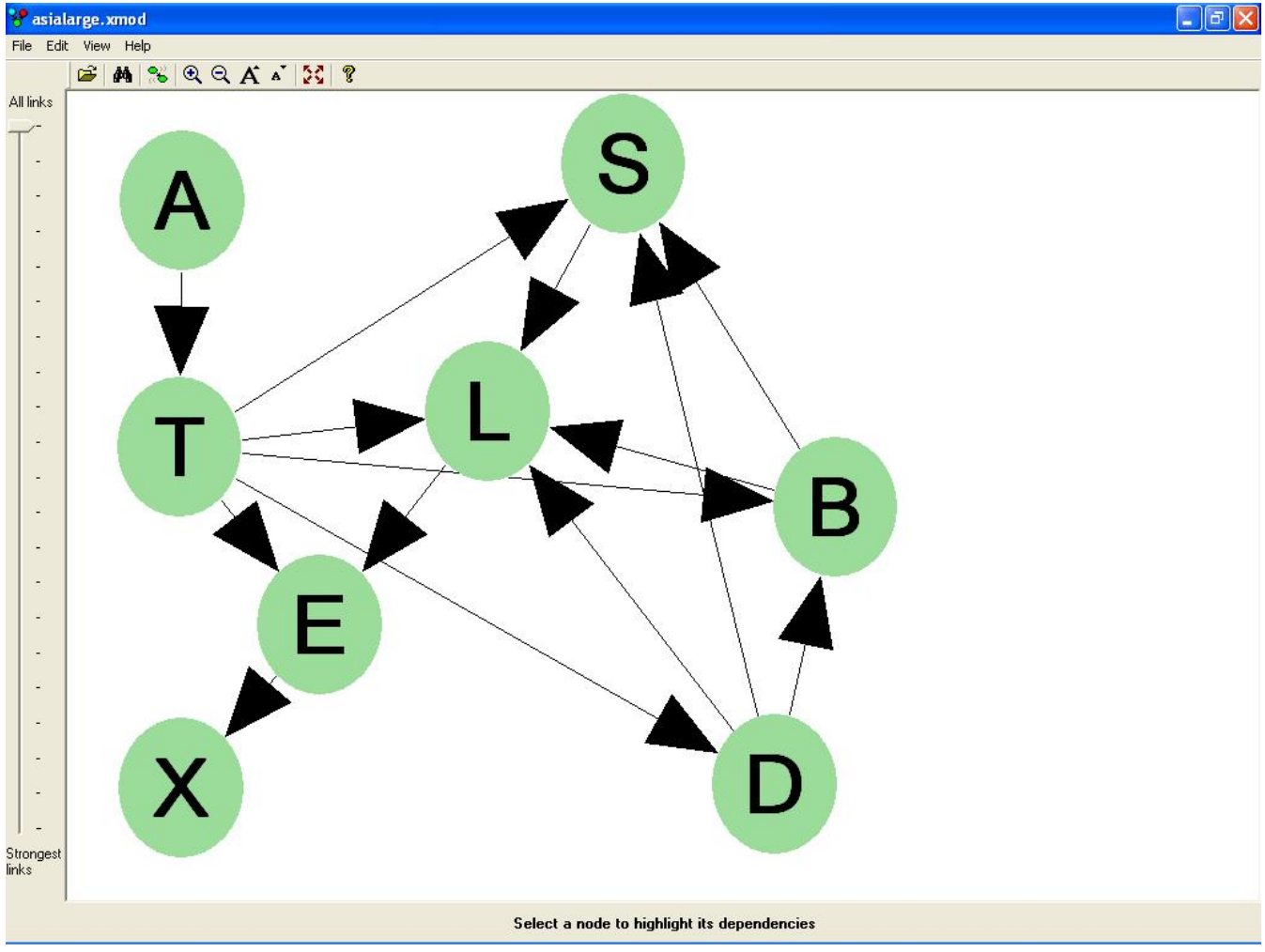
10000 simulated cases

NPC algorithm



10000 simulated cases

Bayesian GES



Types of approach

- Methods for *judging adequacy of structure* such as
  - Tests of significance
  - Penalised likelihood scores

    $$I_\kappa(\mathcal{M}) = \log \hat{L} - \kappa \dim(\mathcal{M})$$

    with $\kappa = 1$ for AIC Akaike (1974), or
    $\kappa = \frac{1}{2} \log N$ for BIC (Schwarz 1978).
  - *Bayesian posterior probabilities.*
- *Search strategies* through space of possible structures, more or less based on *heuristics.*

Estimating trees

Assume P factorizes w.r.t. an unknown *tree* τ.

Chow and Liu (1968) showed *MLE* $\hat{\tau}$ *of* $\mathcal{T}$ *has maximal weight*, where the *weight* of τ is

$$w(\tau) = \sum_{e \in E(\tau)} w_n(e) = \sum_{e \in E(\tau)} H_n(e)$$

and $H_n(e)$ is the empirical *cross-entropy* or *mutual information* between endpoint variables of the edge $e = \{u, v\}$:

$$H_n(e) = \sum \frac{n(x_u, x_v)}{n} \log \frac{n(x_u, x_v)/n}{n(x_u)n(x_v)/n^2}.$$

Extensions

Results are easily *extended to Gaussian graphical models,* with the weight of a tree determined as

$$w_n(e) = -\frac{1}{2}\log(1 - r_e^2),$$

where r_e^2 is *correlation coeffient* along edge $e = \{u, v\}$.

Highest AIC or BIC scoring forest *also available as MWSF,* with modified weights

$$w_n^{\text{pen}}(e) = nw(e) - \kappa_n \text{df}_e,$$

with $\kappa_n = 2$ for AIC, $\kappa_n = \log n$ for BIC and df_e the *degrees of freedom for independence* along e.

More on trees

Fast algorithms (Kruskal Jr. 1956) compute maximal weight spanning tree (or forest) from weights $W = (w_{uv}, u, v \in V)$.

Chow and Wagner (1978) show *a.s. consistency in total variation of* $\hat{P}$*:* If P factorises w.r.t. τ, then

$$\sup_x |p(x) - \hat{p}(x)| \to 0 \text{ for } n \to \infty,$$

so *if* τ *is unique for* P, $\hat{\tau} = \tau$ *for all* $n > N$ *for some* N.

If P does not factorize w.r.t. a tree, $\hat{P}$ *converges to closest tree-approximation* $\tilde{P}$ *to* P (Kullback-Leibler distance).

Bayes factors

For $\mathcal{G} \in \Gamma$, $\Theta_{\mathcal{G}}$ is associated parameter space so that P factorizes w.r.t. $\mathcal{G}$ if and only if $P = P_\theta$ for some $\theta \in \Theta_{\mathcal{G}}$. $\mathcal{L}_{\mathcal{G}}$ is prior law on $\Theta_{\mathcal{G}}$.

The *Bayes factor* (likelihood ratio) for discriminating between $\mathcal{G}_1$ and $\mathcal{G}_2$ based on observations $X^{(n)} = x^{(n)}$ is

$$\mathrm{BF}(\mathcal{G}_1 : \mathcal{G}_2) = \frac{f(x^{(n)} \,|\, \mathcal{G}_1)}{f(x^{(n)} \,|\, \mathcal{G}_2)},$$

where

$$f(x^{(n)} \,|\, \mathcal{G}) = \int_{\Theta_{\mathcal{G}}} f(x^{(n)} \,|\, \mathcal{G}, \theta)\, \mathcal{L}_{\mathcal{G}}(d\theta)$$

is known as the *marginal likelihood* of $\mathcal{G}$.

Posterior distribution over graphs

If $\pi(\mathcal{G})$ is a prior probability distribution over a given set of graphs Γ, the posterior distribution is determined as

$$\pi^*(\mathcal{G}) = \pi(\mathcal{G} \,|\, x^{(n)}) \propto f(x^{(n)} \,|\, \mathcal{G})\pi(\mathcal{G})$$

or equivalently

$$\frac{\pi^*(\mathcal{G}_1)}{\pi^*(\mathcal{G}_2)} = \mathrm{BF}(\mathcal{G}_1 : \mathcal{G}_2)\frac{\pi(\mathcal{G}_1)}{\pi(\mathcal{G}_2)}.$$

Bayesian analysis looks for the *MAP estimate* $\mathcal{G}^*$ maximizing $\pi^*(\mathcal{G})$ over Γ, or attempts to *sample from the posterior* using e.g. Monte-Carlo methods.

Strong hyper Markov prior laws

For strong hyper Markov prior laws, $X^{(n)}$ is itself marginally Markov so

$$f(x^{(n)} \mid \mathcal{G}) = \frac{\prod_{Q \in \mathcal{Q}} f(x_Q^{(n)} \mid \mathcal{G})}{\prod_{S \in \mathcal{S}} p(x_S^{(n)} \mid \mathcal{G})^{\nu_{\mathcal{G}}(S)}}, \tag{1}$$

where $\mathcal{Q}$ are the prime components and $\mathcal{S}$ the minimal complete separators of $\mathcal{G}$.

Hyper inverse Wishart laws

Denote the normalisation constant of the hyper inverse Wishart density as

$$h(\delta, \Phi; \mathcal{G}) = \int_{\mathcal{S}^+(\mathcal{G})} (\det K)^{\delta/2} e^{-\operatorname{tr}(K\Phi)}\, dK,$$

i.e. the usual Wishart constant if $Q = C$ is a clique.

Combining with the Gaussian likelihood, it is easily seen that for Gaussian graphical models we have

$$f(x^{(n)} \,|\, \mathcal{G}) = \frac{h(\delta + n, \Phi + W^n; \mathcal{G})}{h(\delta, \Phi; \mathcal{G})}.$$

Comparing with (1) leads to a similar factorization of the

normalising constant

$$h(\delta, \Phi; \mathcal{G}) = \frac{\prod_{Q \in \mathcal{Q}} h(\delta, \Phi_Q; \mathcal{G}_Q)}{\prod_{S \in \mathcal{S}} h(\delta, \Phi_S; S)^{\nu_{\mathcal{G}}(S)}}.$$

For *chordal graphs* all terms in this expression reduce to known Wishart constants, and we can thus calculate the normalization constant explicitly.

In general, Monte-Carlo simulation or similar methods must be used (Atay-Kayis and Massam 2002).

The marginal distribution of $W^{(n)}$ is (weak) *hyper Markov* w.r.t. $\mathcal{G}$. It was termed the *hyper matrix F law* by Dawid and Lauritzen (1993).

Bayes factors for forests

Trees and forests are decomposable graphs, so for a forest ϕ we get

$$f(\phi \,|\, x^{(n)}) \propto \frac{\prod_{e \in E(\phi)} f(x_e^{(n)})}{\prod_{v \in V} f(x_v^{(n)})^{d_\phi(v)-1}},$$

since all minimal complete separators are singletons and $\nu_\phi(\{v\}) = d_\phi(v) - 1$.

Multiplying the right-hand side with $\prod_{v \in V} f(x_v^{(n)})$ yields

$$\frac{\prod_{e \in E(\phi)} p(x_e^{(n)})}{\prod_{v \in V} f(x_v^{(n)})^{d_\phi(v)-1}} = \prod_{v \in V} f(x_v^{(n)}) \prod_{e \in \phi} \mathrm{BF}(e),$$

where $\mathrm{BF}(e)$ is the *Bayes factor* for independence along the edge e:

$$\mathrm{BF}(e) = \frac{f(x_u^{(n)}, x_v^{(n)})}{p(x_u^{(n)})p(x_v^{(n)})}.$$

Bayesian analysis

MAP estimates of forests can thus be computed using an MWSF algorithm.

Algorithms exist for generating random spanning trees (Aldous 1990), so *full posterior analysis is in principle possible for trees*.

These work less well for weights occurring with typical Bayes factors, as most of these are essentially zero, so methods based on the *Matrix Tree Theorem* seem currently more useful.

Only heuristics available for MAP estimators or maximizing penalized likelihoods such as AIC or BIC, for other than trees.

Some challenges for undirected graphs

- Find *feasible algorithm for (perfect) simulation* from a distribution over chordal graphs as

$$p(\mathcal{G}) \propto \frac{\prod_{C \in \mathcal{C}} w(C)}{\prod_{S \in \mathcal{S}} w(S)^{\nu_{\mathcal{G}}(S)}},$$

  where $w(A), A \subseteq V$ are a prescribed set of positive weights.

- Find *feasible algorithm for obtaining MAP* in decomposable case. This may not be universally possible as problem most likely is NP-complete.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, **19**, 716–23.

Aldous, D. (1990). A random walk construction of uniform spanning trees and uniform labelled trees. *SIAM Journal on Discrete Mathematics*, **3**, (4), 450–65.

Atay-Kayis, A. and Massam, H. (2002). The marginal likelihood for decomposable and non-decomposable graphical Gaussian models. Technical report, Department of Mathematics, York University.

Chow, C. K. and Liu, C. N. (1968). Approximating discrete probability distributions with dependence trees. *IEEE Transactions on Information Theory*, **14**, 462–

7.

Chow, C. K. and Wagner, T. J. (1978). Consistency of an estimate of tree-dependent probability distributions. *IEEE Transactions on Information Theory*, **19**, 369–71.

Dawid, A. P. and Lauritzen, S. L. (1993). Hyper Markov laws in the statistical analysis of decomposable graphical models. *Annals of Statistics*, **21**, 1272–317.

Kruskal Jr., J. B. (1956). On the shortest spanning subtree of a graph and the travelling salesman problem. *Proceedings of the American Mathematical Society*, **7**, 48–50.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, **6**, 461–4.