On learning of structures for Bayesian networks
Timo Koski
Mathematical statistics, KTH

ERNSI Workshop on System Identification

Sigtuna 02.10.2008
Bayesian networks are a special case of multivariate (discrete) statistical distributions embodying a collection of marginal and conditional independencies which may be represented by means of a directed acyclic graph.
A Bayesian Network

$W$ is a parent of $X$ and $Y$, $W$ and $X$ are parents of $Y$, and $X$ and $Y$ are parents of $Z$.

$$P(X, Y, Z, W) = P(Z \mid X, Y) P(Y \mid X, W) P(X \mid W) P(W)$$

We specify the simultaneous distribution by a set of simpler conditional distributions (modularity).
A Bayesian Network: learning structure

ERNSI Workshop on System Identification
Learning of Graph Structures
Graph Structure (DAG), random walk

omgivning

omgivning
The main challenges are:

- the increase in the number of graphical model structures as a function of the number of nodes.
The main challenges are:

- the increase in the number of graphical model structures as a function of the number of nodes.
- the equivalence of the statistical models determined by different graphical models.
Robinson (1973) has given the following recursive function for computing the number $f(d)$ of directed acyclic graphs with $d$ nodes:

$$f(d) = \sum_{i=1}^{d} (-1)^{i+1} \binom{d}{i} 2^{i(d-1)} f(d - i).$$  \hspace{1cm} (1)

This number grows exponentially, for $d = 5$ it is 29000, and for $d = 10$ it is approximately $4.2 \times 10^{18}$.
There are two approaches of model selection for Bayesian networks:

- **maximize a score function**, e.g., the posterior probability of the graph $P(G|X)$, where $G$ is a graph and $X$ is a set of data (a sample of a set of random variables). The algorithms search the space of all possible structures (in fact, one should search over representors of equivalence classes, as discussed in the sequel).
There are two approaches of model selection for Bayesian networks:

- **maximize a score function**, e.g., the posterior probability of the graph $P(G|X)$, where $G$ is a graph and $X$ is a set of data (a sample of a set of random variables). The algorithms search the space of all possible structures (in fact, one should search over representors of equivalence classes, as discussed in the sequel).

- **constraint based**: algorithms estimate from data whether certain conditional independencies between the variables hold.

- the correct dimension of a model when applying Bayesian Information Criterion (BIC) for model selection with hidden variables. $\text{BIC} = \text{an asymptotic expansion of } P(G|X)$. 
Consider the random variables $X_1, \ldots, X_d$. Each variable $X_i$ assumes values in the discrete and finite alphabet $\mathcal{X}_i$. The set of all possible configurations of $X_1, \ldots, X_d$ is denoted by $\mathcal{X} = \times_{i=1}^d \mathcal{X}_i$. $|\mathcal{X}_i| = n_i$. Let the generic element of $\mathcal{X}$ be denoted by $x$. $\mathbb{R}^\mathcal{X}$ denotes the real vector space of $d$–dimensional tables of the format $n_1 \times n_2 \ldots \times n_d$. 
We assign a joint probability to \( x \in X \),

\[
P(x) = P(X_1 = x_1, \ldots, X_d = x_d) = p_{x_1 \ldots x_d}
\]

We can regard \( p_{x_1 \ldots x_d} \) as indeterminates (real numbers) that generate the ring \( R[X] \) of polynomial functions on the space of tables \( R^X \). Actually we need not restrict them to be probabilities.
Let $X$, $Y$ and $Z$ be pairwise disjoint subsets of the random variables $X_1, \ldots, X_d$. We say that $X$ and $Y$ are conditionally independent given $Z$ if

$$P(X = x, Y = y \mid Z = z) = P(X = x \mid Z = z) \cdot P(Y = y \mid Z = z)$$

for all $x, y, z$. We write this with the symbol

$$X \perp Y \mid Z$$
B. Sturmfels\textsuperscript{1} observed that

\[ X \perp Y \mid Z \iff p_{xyz}p_{x'y'z} - p_{x'yz}p_{xy'z} = 0 \]

where \( p_{xyz} \) e.t.c. are obtained by marginalization from the joint probability \( p_{x_1...x_d} \). Hence \( p_{xyz} \) e.t.c. are polynomials of degree 1 in \( R(\mathcal{X}) \).

\textsuperscript{1}Solving Systems of Polynomial Equations, AMS, 2002
Therefore the cross product differences (cpd)

\[ cpd := p_{xyz}p_{x'y'z} - p_{x'yz}p_{xy'z} \]

correspond to homogeneous quadratic polynomials. Hence all the statements of conditional independence for a distribution \( P \) translate into a system of homogeneous quadratic polynomials (Sturmfels, loc.cit. 2002, Theorem 8.1.). We write \( I_{X \perp Y \| Z} \) for the ideal generated by these polynomials.
Bayesian networks, as many other statistical models, can be described by a finite list of conditional independence statements,

\[ \mathcal{M} = \{ X^{(1)} \perp Y^{(1)} \parallel Z^{(1)}, X^{(2)} \perp Y^{(2)} \parallel Z^{(2)}, \ldots, X^{(m)} \perp Y^{(m)} \parallel Z^{(m)} \} \]

and the ideal of the model \( \mathcal{M} \) is

\[ I_{\mathcal{M}} = I_{X^{(1)} \perp Y^{(1)} \parallel Z^{(1)}} + I_{X^{(2)} \perp Y^{(2)} \parallel Z^{(2)}} + \cdots + I_{X^{(m)} \perp Y^{(m)} \parallel Z^{(m)}} \]
The independence variety is $V(I_M)$ is the set of common (complex) zeros of the polynomials in $I_M$, or, equivalently, the set of all $n_1 \times n_2 \ldots \times n_d$ tables which satisfy the conditional independence statements in $\mathcal{M}$. 

$$I_M = I_{X(1) \perp Y(1)} \parallel Z(1) + I_{X(2) \perp Y(2)} \parallel Z(2) + \cdots + I_{X(m) \perp Y(m)} \parallel Z(m)$$
Let now $V(I_M + < \sum p_{x_1 \ldots x_d} - 1 >)$ denote the subset of the independence variety $V(I_M)$ which consists of non-negative tables whose entries sum to one. This is the subset of the probability simplex specified by the model $M$. 
A graph $G$ is a pair $(V, E)$, where $E \subseteq V \times V \setminus \{(i, i) \in V \times V\}$ is the set of ordered pairs of nodes denoted as edges.

- If both $(i, j) \in E$ and $(j, i) \in E$ then there is an undirected edge between $i$ and $j$ written as $i - j$.
- If $(i, j) \in E$ but $(j, i) \not\in E$ there is a directed edge between $i$ and $j$ written as $i \rightarrow j$.
- A path is a sequence $v_0, ..., v_n$ of different nodes such that $v_i \neq v_j$ and $(v_{i-1}, v_i) \in E$ for all $i, j = 1, ..., n$.
- A cycle is a path with the only exception that $v_0 = v_n$. A directed cycle is a cycle which has at least one directed edge.
Let $G = (V, E)$ be a directed acyclic graph (DAG) and the set of nodes $V = \{1, \ldots, d\}$ index the random variables $X_1, \ldots, X_d$. Let $P$ be a joint probability distribution of $X_1, \ldots, X_d$. We say that the pair $(G, P)$ is a (discrete) Bayesian Network if $(G, P)$ satisfies the following independence model a.k.a. the (directed) local Markov property

$$local\ (G) = \{X_i \perp nd\ (X_i) \setminus pa\ (X_i) \mid \| \ pa\ (X_i), i = 1, 2, \ldots, d\}$$

where $nd\ (X_i)$ is the set of nondescendants of $X_i$ and $pa\ (X_i)$ is the set of parents of $X_i$.

---

2Convention: We are going to identify nodes of a graph with random variables.
Bayesian Network

\[ local (G) = \{ X_i \perp nd (X_i) \setminus pa (X_i) \parallel pa (X_i), i = 1, 2, \ldots, d \} \]

- \( X_j \) is a nondescendant of \( X_i \) if there is no directed path from \( X_i \) to \( X_j \).
Bayesian Network

\[ \text{local}(G) = \{ X_i \perp \text{nd}(X_i) \setminus \text{pa}(X_i) \parallel \text{pa}(X_i), i = 1, 2, \ldots, d \} \]

- \( X_j \) is a nondescendant of \( X_i \) if there is no directed path from \( X_i \) to \( X_j \).
- \( \text{pa}(X_i) \) is the set of parents of \( X_i \).
Global Markov and d-separation

There may be more conditional independencies in \((G, P)\) than those listed in \(local(G)\). The (directed) global Markov property of \((G, P)\) is the list of independence statements

\[
global(G) = \{X \perp Y \| Z \mid X \perp Y \|_G Z\},
\]

where

\[X \perp Y \|_G Z\]

says that \(X\) and \(Y\) are \(d\)-separated by \(Z\). \(d\)-separation is a graphical algorithm to find conditional independence statements in a Bayesian network.
$X$ and $Y$ are $d$-separated by $Z$, if and only if all trails from $X$ to $Y$ are blocked by $Z$. A trail $\pi$ from $X_i$ to $X_j$ is blocked in $G$ by (the nodes in) $Z$ if there is a node $X_b$ such that either

- $X_b \in Z$ and arrows of the trail $\pi$ do not meet head-to-head at $X_b$
- or $X_b \notin Z$ and $X_b$ has no descendants in $Z$, and arrows of $\pi$ do meet head-to-head at $X_b$. 

d-separation
d-separation
We should assume that \((G, P)\) is such that

\[ X \perp Y \parallel_G Z \iff X \perp Y \parallel Z \]

This is the starting point of the constraint based methods of model choice.
Assume that

\[ \text{pa}(X_j) = \{X_{i_1}, X_{i_2}, \ldots, X_{i_r}\} \]

Then we set

\[ q_{x_0x_1\ldots x_r}^j = P(X_j = x_0 \mid X_{i_1} = x_1, X_{i_2} = x_2, \ldots, X_{i_r} = x_r) \]

Let \( E \) denote the set of these unknowns \( q_{x_0x_1\ldots x_r}^j, j = 1, 2, \ldots, d \) and let \( R[E] \) denote the polynomial ring they generate.
\[ \phi : R^E \mapsto R^X \text{ is defined by} \]
\[ \prod_{j=1}^{d} q_{x_1 \cdots x_r}^j \mapsto p_{x_1 \cdots x_d} \]

The image of \( \phi \) is contained in the independence variety \( V_{\text{global}}(G) \).

The ring homomorphism \( \Phi : R[X] \mapsto R[E]/J \), where
\[ J = \langle q_{\sum_{x_1 \cdots x_r}^j} - 1 \rangle \]
\[ p_{x_1 \cdots x_d} \mapsto \prod_{j=1}^{d} q_{x_1 \cdots x_r}^j \]
The following four subsets of the probability simplex coincide:

\[ V \left( I_{\text{local}}(G) + \langle \sum p_{x_1...x_d} - 1 \rangle \right) = \]

\[ V \left( I_{\text{global}}(G) + \langle \sum p_{x_1...x_d} - 1 \rangle \right) = \]

\[ = V(\ker(\Phi)) = \text{image}(\phi). \]

There may be several graphs $G$ such that

$$(G, P)$$

is a Bayesian network for a given $P$. In other words, $G_1$ and $G_2$ may have

$$\text{global}(G_1) = \text{global}(G_2)$$

for a given $P$. This cannot be ignored in the construction of algorithms. We shall discuss this within the context of scoring algorithms by introduction of partially directed acyclic graphs.
A *partially directed acyclic graph* (PDAG)\(^3\) is a graph \(G\) that does not contain any directed cycles. A *directed acyclic graph* (DAG) is a PDAG with only directed edges and an *undirected graph* (UG) is a PDAG with only undirected edges.


\(^{3}\text{a.k.a. CHAIN GRAPH}\)
Lauritzen, Wermuth and Frydenberg (LWF) Markov properties. \( P, L, G \) relative to a (DAG, UG, PDAG) are defined as

1. the pairwise Markov property \( P \), if for any pair \((i, j)\) of nodes \((i, j) \notin E\) with \( j \in \text{nd}(i) \), \( i \perp j | \text{nd}(i) \setminus \{i, j\} \).

2. the local Markov property \( L \), for any \( i \in V \), \( i \perp \text{nd}(i) \setminus \text{bd}(i) | \text{bd}(i) \).

3. the global Markov property \( G \), if for any triple of disjoint subsets \((A, B, S \subseteq V)\) such that \( S \) separates \( A \) from \( B \) in \((G_{\text{An}(A \cup B \cup S)})^m, A \perp B | S\).
A complex in a PDAG is a sequence of nodes $v_1, \ldots, v_k$, $k \geq 3$ such that $v_1 \rightarrow v_2$, $v_i \leftarrow v_{i+1}$ for $i = 2, \ldots, k - 2$, $v_{k-1} \leftarrow v_k$ and there are no other edges $(v_i, v_j) \in E$ for all $0 \leq i, j \leq k$. 

A Complex
The following is proved in (Frydenberg 1990), see also (Andersson et al 1997).

**Theorem**

*Two PDAGs have the same (LWF) Markov properties iff they have the same undirected graph and the same complexes.*
Markov equivalence

\[ \{ D_1, D_2, D_3 \} \]

\[ D_4 \]
A graph $G = (V, E)$ is larger than the graph $\hat{G} = (V, \hat{E})$ if $G$ might have edges where $\hat{G}$ has arrows.

**Theorem**

*(Frydenberg 1990)* For any PDAG $G$ there exists a unique PDAG $\tilde{G} = (\tilde{V}, \tilde{E})$ with the same Markov properties as $G$ such that $\tilde{G}$ is larger than any other graph $\hat{G} = (\hat{V}, \hat{E})$ that has the same Markov properties as $G$. 
Each Markov-equivalence class is uniquely determined by a single PDAG (a chain graph), simultaneously equivalent to each graph in the class, hence the learning of structure should be based on these representatives. The question is, how to construct them.

Volf and Studený define

- a *descending path*, if either $i - i + 1$ or $i \rightarrow i + 1$ for all $i = 1, \ldots, n - 1$.
- an *ancestor* of a node $j$ as a node $i$ such that there exists a descending path from $i$ to $j$ in $G$. 
A directed edge $X_i \to X_j$ covers $X_y \to X_z$ if $X_i$ is an ancestor of $X_y$ and $X_z$ is an ancestor of $X_j$ in a PDAG.

An edge is protected iff it covers a directed edge which belongs to a complex.
Theorem

(Volf & Studený) A LPDAG is the largest PDAG of a Markov equivalent set of PDAGs iff every directed edge is protected. The largest graph is constructed by undirecting every non-protected arrow.
Markov equivalence
Markov equivalence

Note however that it is possible to represent a set of Markov properties in a PDAG that cannot be represented in a DAG or UG, such as $1 \rightarrow 2 \leftarrow 3 \leftarrow 4$. Hence we need Gröbner bases for $I_M$ to check this.

For small $d$ the known number of LPDAGs that cannot be represented by UGs is 0 for $d = 2, 3$ (Volf and Studený 1999). For $d = 4, 5$ the percentage of LPDAGs that cannot be represented as UGs is 6 and 22.
The presence of equivalence suggests the use of search space $S$, the states of which are equivalence classes represented by its LPDAG, that is, graphical models that all have the same Markov properties according to Definition LWF. Then the size of a search space, where the states are equivalence classes, is smaller than, e.g., the space of all DAGs for $d \leq 10$ (Gillispie and Perlman 2001).
Milan Studený and his co-workers are currently developing (have already developed (?)) and algebraic representation, using objects called imsets, of the equivalence classes that so that an equivalence class corresponds uniquely to (a long) vector (with lots of zeros). Then the polytope generated by the imsets may be the basis of learning Bayesian networks by linear programming.

J. Vomlel & M. Studený: Graphical and Algebraic Representatives of Conditional Independence Models.

http://staff.utia.cas.cz/studeny/studeny_home.html
In the current approach the next issue is to compute the marginal data distribution

\[ P(X \mid G) \]

taking into account the equivalence classes. We follow/present the solution due to The symbol \( X \) designates a set of data, or, \( r \) samples of each of the variables \( X_1, \ldots, X_d \), i.e., \( X = \{ x^{(l)} \}_{l=1}^r, x^{(l)} \in \mathcal{X} \), with neither missing values nor hidden variables.

For a subset \( A \subset V \), the corresponding random variable is \( X_A = (X_i)_{i \in A} \).
(Frydenberg 1990) A probability distribution on a discrete and finite sample space with strictly positive density \( P \), w.r.t. a product measure, satisfies \( P \) (see Definition LWF) over a PDAG \( G \) if and only if it factorizes as

\[
P(x) = \prod_{\tau \in \mathcal{T}(G)} P(x_\tau | x_{bd(\tau)}) ,
\]

where each factor \( P(x_\tau | x_{bd(\tau)}) \) further factorizes along the undirected closure graph \( G_{cl(\tau)}^m \).
A *clique* is a complete subgraph \( G'_A \), such that there exists no (other) complete subgraph \( G'_{A'} \), \( A \subset A' \subset V \) such that \( G_A \neq G'_{A'} \). A hyper-Dirichlet density is defined for given real positive numbers \( \lambda_c = \{\lambda_c, x_c\}_{x_c \in \mathcal{X}_c} \) let \( \mathcal{D}(\lambda_c) \) for \( \theta_c \) by the density

\[
\pi (\theta_c | \lambda_c) = \frac{\Gamma (\lambda_0)}{\gamma_p} \prod_{x_c \in \mathcal{X}_c} \theta_{c, x_c}^{\lambda_c, x_c - 1} \tag{3}
\]

on the set \( \{\theta_c | \sum_{x_c \in \mathcal{X}_c} \theta_{c, x_c} = 1, \theta_{c, x_c} > 0\} \), where \( \Gamma (x) \) is the Euler gamma function, \( \lambda_0 = \sum_{x_c \in \mathcal{X}_c} \lambda_{c, x_c} \), \( \gamma_p = \prod_{x_c \in \mathcal{X}_c} \Gamma (\lambda_{c, x_c}) \).
Let $G$ be a PDAG, and let us set

$$P(X | \theta, G) = \prod_{\tau \in T(G)} \prod_{l=1}^{r} P \left( x^l_{\tau} | \theta_{\tau}, x^l_{bd(\tau)} \right),$$

by (2), which is formally rewritten to include the parameters.
Next note that in (2) the probabilities $P \left( x_\tau | \theta_\tau, x_{bd(\tau)} \right)$ satisfy the LWF Markov property $\mathbf{P}$ on the undirected closure graph $G_{cl(\tau)}^m$. Recall that a probability $P(x|\theta) > 0$ satisfying the property $\mathbf{P}$ on an undirected graph factorizes as

$$P(x|\theta) = \frac{\prod_{c \in C} \theta_{c,x_c}}{\prod_{s \in S} \theta_{s,x_s}},$$

where $C \subset \{ A|A \subset V \}$ is the set of cliques and $S \subset \{ A|A \subset V \}$ is the multiset (including repetitions of elements) of separators (for two cliques $C_i$ and $C_j$ the separator $S = C_i \cap C_j$), respectively. Thus

$$\prod_{l=1}^r P \left( x^l_\tau | \theta_\tau, x^l_{bd(\tau)} \right)$$

is a function of a product of expressions of the following form

$$\prod_{l=1}^r \theta(x^l_c) = \prod_{x_c \in X_c} \theta_{c,x_c}^{n_{c,x_c}},$$

where $n_{c,x_c}$ is the number of times the configuration $x_c$ occurs in

$$X_c = \left\{ x^{(l)}_c \right\}_{l=1}^r.$$
On any clique $c$ in $G_{cl}^m(\tau)$ we introduce the integration with respect to $D(\lambda_c,i)$ as follows,

$$\int \left\{ \theta_c \mid \sum_{x_c \in X_c} \theta_{c,x_c} = 1, \theta_{c,x_c} > 0 \right\} \prod_{x_c \in X_c} \theta_{c,x_c}^{n_c,x_c} \pi (\theta_c | \lambda_c) \, d\theta_c.$$
By standard properties of the Dirichlet integral one gets from (3) that

\[ P_c (X_c) = \frac{\Gamma (\lambda_0)}{\Gamma (r + \lambda)} \prod_{x_c \in X_c} \frac{\Gamma (n_{c,x_c} + \lambda_{c,x_c})}{\Gamma (\lambda_{c,x_c})}, \]  

(4)

where

\[ r + \lambda = \sum_{x_c \in X_c} (n_{c,x_c} + \lambda_{c,x_c}). \]
Hence, if $C(\tau)$ and $S(\tau)$ are the set of cliques and separators in the undirected graph $G_\tau$

$$
\prod_{l=1}^{r} P \left( \mathbf{x}_\tau^l | \mathbf{x}_{bd(\tau)}^l \right) = \frac{\prod_{c \in C(\tau)} P_c (\mathbf{X}_c)}{\prod_{s \in S(\tau)} P_s (\mathbf{X}_c)}.
$$

(5)

The probability $P_s (\mathbf{X}_c)$ can be constructed for a separator by marginalizing over a clique that includes $S$. For the validity of this argument and the properties of the marginal data distribution, see (Dawid and Lauritzen 1993). The expression (5) must be multiplied over all the chain components $\tau$ in order to yield the overall marginal data distribution.
\[ P(\mathbf{X} | \mathbf{G}) = \prod_{\tau \in T(\mathbf{G})} \frac{\prod_{c \in C(\tau)} P_c(\mathbf{X}_c)}{\prod_{s \in S(\tau)} P_s(\mathbf{X}_c)}. \]  

(6)

From (6) we define the final scoring function as the posterior probability

\[ P(\mathbf{G} | \mathbf{X}) = \frac{P(\mathbf{X} | \mathbf{G}) P(\mathbf{G})}{\sum_{\mathbf{G} \in S} P(\mathbf{X} | \mathbf{G}) P(\mathbf{G})}. \]  

(7)

For a given set of nodes \( V \) let \( S = \{ \text{LPDAG} \} \) be the search space. One particular goal for the topology learning can be stated as the identification of a structure \( \mathbf{G}^{\text{opt}} \in S \) having highest posterior probability (7), i.e.

\[ \mathbf{G}^{\text{opt}} \in \arg \max_{\mathbf{G} \in S} P(\mathbf{G} | \mathbf{X}). \]
As noted by (Andersson et al 1997), the factors in (2) 
\[ P \left( x_\tau | x_{bd(\tau)} \right) \] satisfy also \( P \) on \( G_\tau \), i.e., the graph induced by the chain component\(^4\) \( \tau \). When the so-called hyper-Dirichlet distributions are used as priors the marginal data likelihood has the LWF Markov properties over graphs (Dawid and Lauritzen 1993).

\(^4\)the connected components of a graph with arrows removed
We need to construct an irreducible Markov chain searching the space. For PDAGs we observe the next result, which is Proposition 4.5 of (Andersson et al 1997).

**Theorem**

Consider two graphs $G$ and $H$ with same set of nodes in $S = \{\text{LPDAG}\}$. Then there exists a finite sequence of LPDAGs $G \equiv G_1, \ldots, G_k \equiv H$ such that each consecutive pair $G_i, G_{i+1}$ differs by either (i) exactly one undirected edge, (ii) exactly by one directed edge, or (iii) by two edges that form an immorality.
Non-reversible MCMC

An algorithm, based on parallell Markov chains $G_j = \{G_{tj}\}$ with the transition kernels, with the probability of a transition from a current state $S$ to a proposed (by the mechanism above) new state $S^*$, as

$$\min\left(1, \frac{p(X|G^*)}{p(X|G)}\right).$$

The chains draw at independent random times a value from $P(G_j \parallel X)$. This forces the chains to move to regions of maximal posterior probability.
Define the sequence of probabilities \( \{\alpha_t, t = 2, 3, \ldots\} \) according to

\[
\alpha_t = \frac{1}{q \log t},
\]

where \( q \geq 1 \) can be chosen suitably, for instance \( q \in [5, 10] \).

\( Z_0 = 0 \), and \( P(Z_t = 1) = \alpha_t \), \( P(Z_t = 0) = 1 - \alpha_t \), independently for \( t = 1, 2, \ldots \).
For each $t = 0, 1, \ldots$, define the distribution

$$P_t(G_{tj}) = \frac{p(X|G_{tj})}{\sum_{j=1}^{m} p(X|G_t)},$$

over the space of the current states $\{G_{t1}, G_{t2}, \ldots, G_{tm}\}$. For each $t = 0, 1, \ldots$ such that $Z_t = 1$, the transition to the next state is determined according to this distribution, such that the next state for each chain is sampled to the non-reversible proposal-acceptance formulae, independently for $j = 1, \ldots, m$. 
For each $t$, such that $Z_t = 0$, transition to the next state $S_{(t+1)j}$ is determined according to the non-reversible proposal-acceptance formulae above independently for $j = 1, \ldots, m$. 
Table: Prognostic factors in coronary heart disease.

<table>
<thead>
<tr>
<th>F</th>
<th>E</th>
<th>D</th>
<th>C</th>
<th>B</th>
<th>A</th>
<th>yes</th>
<th>no</th>
<th>yes</th>
<th>no</th>
</tr>
</thead>
<tbody>
<tr>
<td>neg</td>
<td>&lt; 3</td>
<td>&lt; 140</td>
<td>no</td>
<td>44</td>
<td>40</td>
<td>112</td>
<td>67</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>yes</td>
<td>129</td>
<td>145</td>
<td>12</td>
<td>23</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>&gt; 140</td>
<td>no</td>
<td></td>
<td>35</td>
<td>12</td>
<td>80</td>
<td>33</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>yes</td>
<td></td>
<td>109</td>
<td>67</td>
<td>7</td>
<td>9</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>&gt; 3</td>
<td>&lt; 140</td>
<td>no</td>
<td>23</td>
<td>32</td>
<td>70</td>
<td>66</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>yes</td>
<td></td>
<td>50</td>
<td>80</td>
<td>7</td>
<td>13</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>&gt; 140</td>
<td>no</td>
<td></td>
<td>24</td>
<td>25</td>
<td>73</td>
<td>57</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>yes</td>
<td></td>
<td>51</td>
<td>63</td>
<td>7</td>
<td>16</td>
<td></td>
<td></td>
</tr>
<tr>
<td>pos</td>
<td>&lt; 3</td>
<td>&lt; 140</td>
<td>no</td>
<td>5</td>
<td>7</td>
<td>21</td>
<td>9</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>yes</td>
<td></td>
<td>9</td>
<td>17</td>
<td>1</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>&gt; 140</td>
<td>no</td>
<td></td>
<td>4</td>
<td>3</td>
<td>11</td>
<td>8</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>yes</td>
<td></td>
<td>14</td>
<td>17</td>
<td>5</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>&gt; 3</td>
<td>&lt; 140</td>
<td>no</td>
<td>7</td>
<td>3</td>
<td>14</td>
<td>14</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>yes</td>
<td></td>
<td>9</td>
<td>16</td>
<td>2</td>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>&gt; 140</td>
<td>no</td>
<td></td>
<td>4</td>
<td>0</td>
<td>13</td>
<td>11</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>yes</td>
<td></td>
<td>5</td>
<td>14</td>
<td>4</td>
<td>4</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table: Explanations of the Labels in Table 1

<table>
<thead>
<tr>
<th>Label</th>
<th>Meaning</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>smoking</td>
<td>no, yes</td>
</tr>
<tr>
<td>B</td>
<td>strenuous mental work</td>
<td>no, yes</td>
</tr>
<tr>
<td>C</td>
<td>strenuous physical work</td>
<td>no, yes</td>
</tr>
<tr>
<td>D</td>
<td>systolic blood pressure</td>
<td>$&lt; 140, &gt; 140$</td>
</tr>
<tr>
<td>E</td>
<td>ratio of $\beta$ and $\alpha$ lipoproteins</td>
<td>$&lt; 3, &gt; 3$</td>
</tr>
<tr>
<td>F</td>
<td>family anamnesis$^5$ of coronary heart disease</td>
<td>no, yes</td>
</tr>
</tbody>
</table>

$^5$Information concerning a medical patient and his/her background for use in analysis of her/his condition
The optimal equivalence class, having the estimated posterior probability .32
The optimal equivalence class, having the estimated posterior probability .32 and the cliques

\[
\{F\}, \{B, C\}, \{A, C, E\}, \{A, D, E\}.
\]
The method enables also consistent estimation of the marginal posterior probabilities of any edges being present in the graphs, that is we estimate the marginal probabilities as

\[
\hat{P}(e|X) = \frac{\sum\{G=(V,E)\in\{UG\}|G\in S_t, e\in E\} P(G|X)}{\sum\{G\in\{UG\}|G\in S_t\} P(G|X)}.
\]

(c.f. M Koivisto: *Advances in Exact Bayesian Structure Discovery in Bayesian Networks* 2006.) These are illustrated in Table 3 where the probability of adjacency is given for all pairs of the considered variables.
**Table:** The estimated marginal posterior probabilities of edges for the coronary heart disease data.

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>-</td>
<td>0.1630</td>
<td>0.9999</td>
<td>0.9655</td>
<td>0.9997</td>
<td>0.0487</td>
</tr>
<tr>
<td>B</td>
<td>0.1630</td>
<td>-</td>
<td>1.0000</td>
<td>0.0002</td>
<td>0.1864</td>
<td>0.3003</td>
</tr>
<tr>
<td>C</td>
<td>0.9999</td>
<td>1.0000</td>
<td>-</td>
<td>0.0010</td>
<td>0.9313</td>
<td>0.0164</td>
</tr>
<tr>
<td>D</td>
<td>0.9655</td>
<td>0.0002</td>
<td>0.0010</td>
<td>-</td>
<td>0.9839</td>
<td>0.0553</td>
</tr>
<tr>
<td>E</td>
<td>0.9997</td>
<td>0.1864</td>
<td>0.9313</td>
<td>0.9839</td>
<td>-</td>
<td>0.1273</td>
</tr>
<tr>
<td>F</td>
<td>0.0487</td>
<td>0.3003</td>
<td>0.0164</td>
<td>0.0553</td>
<td>0.1273</td>
<td>-</td>
</tr>
</tbody>
</table>
For the special case of decomposable UGs, (Frydenberg and Lauritzen 1989) shows that search operators adding and deleting of single edges in decomposable UGs is sufficient to guarantee irreducibility.

**Corollary**

The Markov chain $G^{(i)}$ with transition mechanism defined by the kernel, Algorithm and the search operator in the Algorithm is irreducible with respect to $S = \{UG\}$. 
The example above

\[ P \propto \psi_{BC} \psi_{CAE} \psi_{EAD} \psi_{F} \]
parts of the talk are based on a paper to appear in Data Mining and Knowledge Discovery, 2008, joint work with Jukka Corander and Magnus Ekdahl.