Graphical Models

Steffen Lauritzen, University of Oxford

Graduate Lectures Hilary Term 2011

January 27, 2011

- Precursors originate mostly from Physics (Gibbs, 1902), Genetics (Wright, 1921, 1934), and Economics (Wold, 1954);

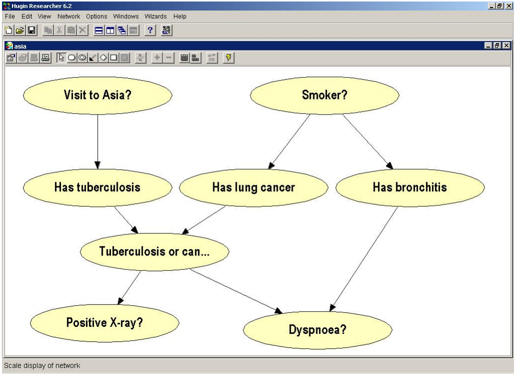
- Precursors originate mostly from Physics (Gibbs, 1902), Genetics (Wright, 1921, 1934), and Economics (Wold, 1954);
- Early graphical models in statistics include covariance selection models (Dempster, 1972) and log-linear models (Haberman, 1974);

- Precursors originate mostly from Physics (Gibbs, 1902), Genetics (Wright, 1921, 1934), and Economics (Wold, 1954);
- Early graphical models in statistics include covariance selection models (Dempster, 1972) and log-linear models (Haberman, 1974);
- Papers setting the scene include Darroch et al. (1980), Wermuth and Lauritzen (1983), and Lauritzen and Wermuth (1989).

- Precursors originate mostly from Physics (Gibbs, 1902), Genetics (Wright, 1921, 1934), and Economics (Wold, 1954);
- Early graphical models in statistics include covariance selection models (Dempster, 1972) and log-linear models (Haberman, 1974);
- Papers setting the scene include Darroch et al. (1980), Wermuth and Lauritzen (1983), and Lauritzen and Wermuth (1989).
- Subject took off after Pearl (1988) and Lauritzen and Spiegelhalter (1988), and in particular after Whittaker (1990) and Lauritzen (1996).

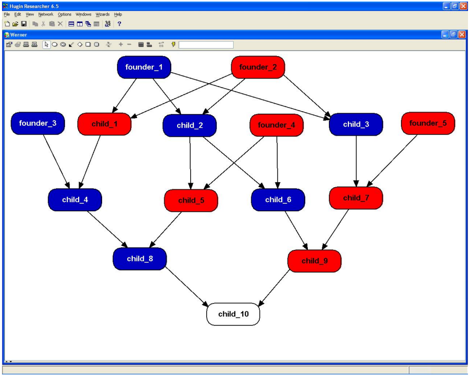
- Precursors originate mostly from Physics (Gibbs, 1902), Genetics (Wright, 1921, 1934), and Economics (Wold, 1954);
- Early graphical models in statistics include covariance selection models (Dempster, 1972) and log-linear models (Haberman, 1974);
- Papers setting the scene include Darroch et al. (1980), Wermuth and Lauritzen (1983), and Lauritzen and Wermuth (1989).
- Subject took off after Pearl (1988) and Lauritzen and Spiegelhalter (1988), and in particular after Whittaker (1990) and Lauritzen (1996).
- Developments now prolific and it is largely impossible to keep track. Google gives *7 420 000* hits.

A directed graphical model



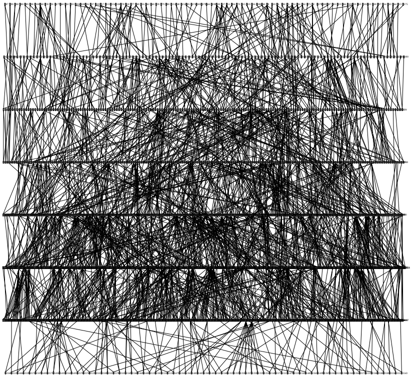
Directed graphical model (Bayesian network) showing relations between risk factors, diseases, and symptoms.

A pedigree



Graphical model for a pedigree from study of Werner's syndrome. Each node is itself a graphical model.

A large pedigree



Family relationship of 1641 members of Greenland Eskimo population.

Random variables X and Y are *conditionally independent* given the random variable Z if

$$\mathcal{L}(X \mid Y, Z) = \mathcal{L}(X \mid Z).$$

We then write $X \perp\!\!\!\perp Y \mid Z$ (or $X \perp\!\!\!\perp_P Y \mid Z$)
Intuitively:
Knowing Z renders Y *irrelevant* for predicting X.
Factorisation of densities:

$$\begin{aligned} X \perp\!\!\!\perp Y \mid Z \quad &\iff \quad f(x,y,z)f(z) = f(x,z)f(y,z) \\ &\iff \quad \exists a, b : f(x,y,z) = a(x,z)b(y,z). \end{aligned}$$

Fundamental properties

For random variables X, Y, Z, and W it holds

(C1) If $X \perp\!\!\!\perp Y \,|\, Z$ then $Y \perp\!\!\!\perp X \,|\, Z$;

(C2) If $X \perp\!\!\!\perp Y \,|\, Z$ and $U = g(Y)$, then $X \perp\!\!\!\perp U \,|\, Z$;

(C3) If $X \perp\!\!\!\perp Y \,|\, Z$ and $U = g(Y)$, then $X \perp\!\!\!\perp Y \,|\, (Z, U)$;

(C4) If $X \perp\!\!\!\perp Y \,|\, Z$ and $X \perp\!\!\!\perp W \,|\, (Y, Z)$, then $X \perp\!\!\!\perp (Y, W) \,|\, Z$;

If density w.r.t. product measure $f(x, y, z, w) > 0$ also

(C5) If $X \perp\!\!\!\perp Y \,|\, (Z, W)$ and $X \perp\!\!\!\perp Z \,|\, (Y, W)$ then $X \perp\!\!\!\perp (Y, Z) \,|\, W$.

A distribution P is said to *factorize* w.r.t. and undirected graph if its joint density f can be written as

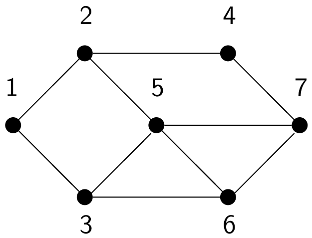
$$f(x) = Z^{-1} \prod_{A \in \mathcal{A}} \phi_A(x_A), \tag{1}$$

where $\mathcal{A}$ are complete subsets of the graph.

Here $x = (x_v, v \in V)$, $x_A = (x_v, v \in A)$ so ϕ_A only depends the A-coordinates of x.

The factorization is matched by a *global Markov property,* ie that *$A \perp\!\!\!\perp B \mid S$ if S separates A from B* in $\mathcal{G}$, written as $A \perp_{\mathcal{G}} B \mid S$ (Hammersley and Clifford, 1971).

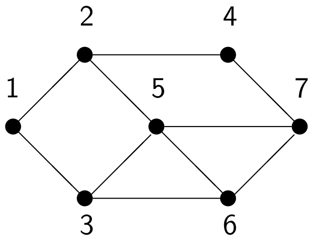
Factorization example



The graph above corresponds to a factorization as

$$\begin{aligned} f(x) &= \psi_{12}(x_1,x_2)\psi_{13}(x_1,x_3)\psi_{24}(x_2,x_4)\psi_{25}(x_2,x_5) \\ &\times \psi_{356}(x_3,x_5,x_6)\psi_{47}(x_4,x_7)\psi_{567}(x_5,x_6,x_7). \end{aligned}$$

Global Markov property



To find conditional independence relations, one should look for separating sets, such as $\{2,3\}$, $\{4,5,6\}$, or $\{2,5,6\}$
For example, it follows that $1 \perp\!\!\!\perp 7 \mid \{2,5,6\}$ and $2 \perp\!\!\!\perp 6 \mid \{3,4,5\}$.

Pairwise and local Markov properties

$\mathcal{G} = (V, E)$ simple undirected graph; A distribution P satisfies

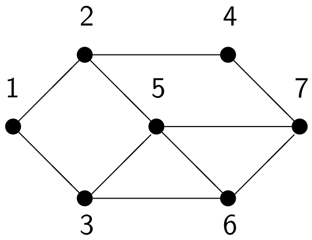
(P) *the pairwise Markov property* if

$$\alpha \not\sim \beta \Rightarrow \alpha \perp\!\!\!\perp_P \beta \mid V \setminus \{\alpha, \beta\};$$

(L) *the local Markov property* if

$$\forall \alpha \in V : \alpha \perp\!\!\!\perp_P V \setminus \mathrm{cl}(\alpha) \mid \mathrm{bd}(\alpha);$$

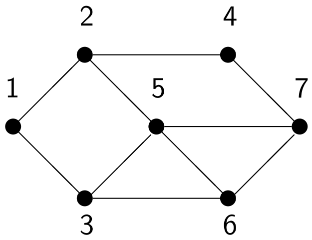
Pairwise Markov property



Any non-adjacent pair of random variables are conditionally independent given the remaning.

For example, $1 \perp\!\!\!\perp 5 \mid \{2,3,4,6,7\}$ and $4 \perp\!\!\!\perp 6 \mid \{1,2,3,5,7\}$.

Local Markov property



Every variable is conditionally independent of the remaining, given its neighbours.

For example, $5 \perp\!\!\!\perp \{1, 4\} \mid \{2, 3, 6, 7\}$ and $7 \perp\!\!\!\perp \{1, 2, 3\} \mid \{4, 5, 6\}$.

Let (F) denote the property that f factorizes w.r.t. $\mathcal{G}$ and let (G), (L) and (P) denote the Markov properties as defined. *Then it holds that*

$$(\mathrm{F}) \Rightarrow (\mathrm{G}) \Rightarrow (\mathrm{L}) \Rightarrow (\mathrm{P}).$$

All *reverse implications are false in general*.

If $f(x) > 0$ for all x it further holds that

$$(\mathrm{P}) \Rightarrow (\mathrm{F})$$

so then

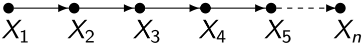
$$(\mathrm{F}) \iff (\mathrm{G}) \iff (\mathrm{L}) \iff (\mathrm{P})$$

(Lauritzen, 1996, Chap. 3).

A probability distribution P over $\mathcal{X} = \mathcal{X}_V$ *factorizes* over a DAG $\mathcal{D}$ if its density or probability mass function f has the form

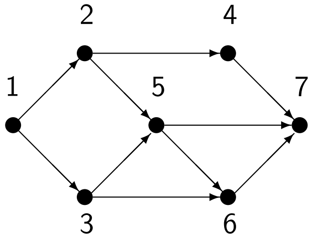
$$f(x) = \prod_{v \in V} f_v(x_v \mid x_{\mathrm{pa}(v)}).$$

A well-known example is a Markov chain:



with $X_{i+1} \perp\!\!\!\perp (X_1, \ldots, X_{i-1}) \mid X_i$ for $i = 3, \ldots, n$.

Example of DAG factorization



The above graph corresponds to the factorization

$$\begin{aligned} f(x) &= f(x_1)f(x_2 \mid x_1)f(x_3 \mid x_1)f(x_4 \mid x_2) \\ &\times f(x_5 \mid x_2, x_3)f(x_6 \mid x_3, x_5)f(x_7 \mid x_4, x_5, x_6). \end{aligned}$$

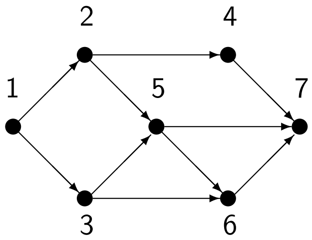
Local directed Markov property

A distribution P satisfies *the local Markov property* (L) w.r.t. a directed acyclic graph $\mathcal{D}$ if

$$\forall \alpha \in V : \alpha \perp\!\!\!\perp_P \{\mathrm{nd}(\alpha) \setminus \mathrm{pa}(\alpha)\} \mid \mathrm{pa}(\alpha).$$

Here $\mathrm{nd}(\alpha)$ are the *non-descendants* of α.

Local directed Markov property



For example, the local Markov property says

$4 \perp\!\!\!\perp \{1, 3, 5, 6\} \mid 2,$

$5 \perp\!\!\!\perp \{1, 4\} \mid \{2, 3\}$

$3 \perp\!\!\!\perp \{2, 4\} \mid 1.$

A distribution satisfies the *global Markov property* w.r.t. $\mathcal{D}$ if

$$A \perp_{\mathcal{D}} B \mid S \Rightarrow A \perp\!\!\!\perp B \mid S.$$

Here $\perp_{\mathcal{D}}$ is *d-separation,* which is somewhat subtle.

It is *always* true for a DAG that

$$\text{(F)} \iff \text{(G)} \iff \text{(L)}$$

(Pearl, 1986; Geiger and Pearl, 1990; Lauritzen et al., 1990).

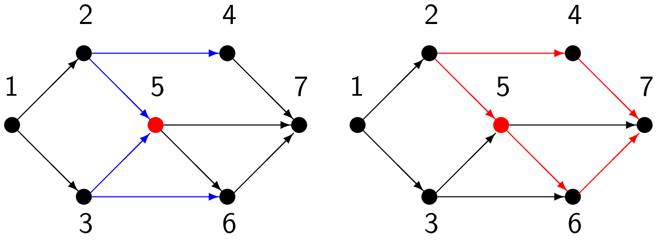
Separation in DAGs

A *trail* τ from vertex α to vertex β in a DAG $\mathcal{D}$ is *blocked* by S if it contains a vertex $\gamma \in \tau$ such that

- either $\gamma \in S$ and edges of τ do not meet head-to-head at γ, or
- γ and all its descendants are not in S, and edges of τ meet head-to-head at γ.

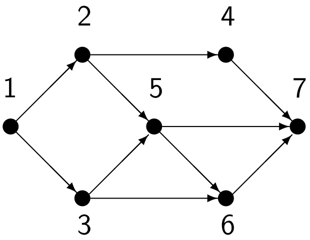
A trail that is not blocked is *active.* Two subsets A and B of vertices are *d-separated* by S if all trails from A to B are blocked by S. We write $A \perp_{\mathcal{D}} B \mid S$.

Separation by example



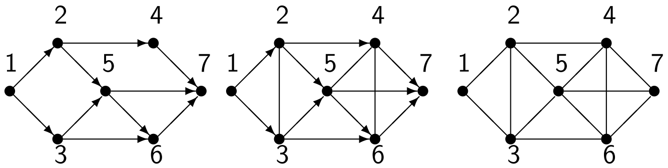
For $S = \{5\}$, the trail $(4, 2, 5, 3, 6)$ is *active*, whereas the trails $(4, 2, 5, 6)$ and $(4, 7, 6)$ are *blocked*.
For $S = \{3, 5\}$, they are all blocked.

Returning to example



Hence $4 \perp_{\mathcal{D}} 6 \mid 3, 5$, but it is *not* true that $4 \perp_{\mathcal{D}} 6 \mid 5$ nor that $4 \perp_{\mathcal{D}} 6$.

The *moral graph* $\mathcal{D}^m$ of a DAG $\mathcal{D}$ is obtained by adding undirected edges between unmarried parents and subsequently dropping directions, as in the example below:



Undirected factorizations

If P factorizes w.r.t. $\mathcal{D}$, it factorizes w.r.t. the moralised graph $\mathcal{D}^m$.

This is seen directly from the factorization:

$$f(x) = \prod_{v \in V} f(x_v \mid x_{\mathrm{pa}(v)}) = \prod_{v \in V} \psi_{\{v\} \cup \mathrm{pa}(v)}(x),$$

since $\{v\} \cup \mathrm{pa}(v)$ are all complete in $\mathcal{D}^m$.

Hence if *P satisfies any of the directed Markov properties w.r.t. $\mathcal{D}$, it satisfies all Markov properties for $\mathcal{D}^m$.*

Alternative equivalent separation

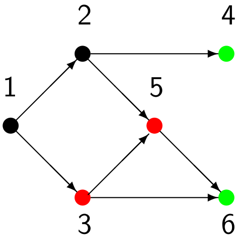
To resolve query involving three sets A, B, S:

1. Reduce to subgraph induced by ancestral set $\mathcal{D}_{\mathrm{An}(A \cup B \cup S)}$ of $A \cup B \cup S$;
2. Moralize to form $(\mathcal{D}_{\mathrm{An}(A \cup B \cup S)})^m$;

It then holds that *$A \perp_{\mathcal{D}} B \mid S$ if and only if S separates A from B in this undirected graph.*

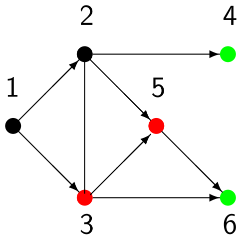
Proof in Lauritzen (1996) needs to allow self-intersecting paths to be correct.

Forming ancestral set



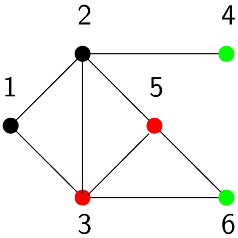
The subgraph induced by all ancestors of nodes involved in the query $4 \perp_{\mathcal{D}} 6 \mid 3, 5$?

Adding links between unmarried parents



Adding an undirected edge between 2 and 3 with common child 5 in the subgraph induced by all ancestors of nodes involved in the query $4 \perp_{\mathcal{D}} 6 \,|\, 3, 5$?

Dropping directions



Since $\{3, 5\}$ separates 4 from 6 in this graph, we can conclude that $4 \perp_{\mathcal{D}} 6 \mid 3, 5$

A particular successful development is associated with BUGS, (Gilks et al., 1994) (WinBUGS, OpenBUGS).

- enables a Bayesian analyst to focus on substantive modelling whereas the technical model specification and computational side is taken care of automatically,

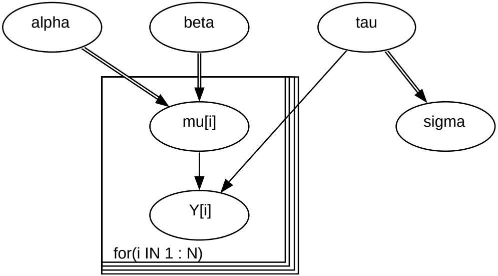
A particular successful development is associated with BUGS, (Gilks et al., 1994) (WinBUGS, OpenBUGS).

- enables a Bayesian analyst to focus on substantive modelling whereas the technical model specification and computational side is taken care of automatically,
- exploiting modularity, factorization, and MCMC methodology, including the Gibbs and Metropolis–Hastings sampler.

A particular successful development is associated with BUGS, (Gilks et al., 1994) (WinBUGS, OpenBUGS).

- enables a Bayesian analyst to focus on substantive modelling whereas the technical model specification and computational side is taken care of automatically,
- exploiting modularity, factorization, and MCMC methodology, including the Gibbs and Metropolis–Hastings sampler.
- Conforming with Bayesian paradigm, parameters and observations are explicitly represented in model as nodes in graph, all being observables;

Linear regression



Linear regression as a full Bayesian graphical model.

Linear regression

```
model
    {
        for( i in 1 : N ) {
            Y[i] ~ dnorm(mu[i],tau)
            mu[i] <- alpha + beta * (x[i] - xbar)
        }
        tau ~ dgamma(0.001,0.001) sigma <- 1 / sqrt(tau)
        alpha ~ dnorm(0.0,1.0E-6)
        beta ~ dnorm(0.0,1.0E-6)
    }
```

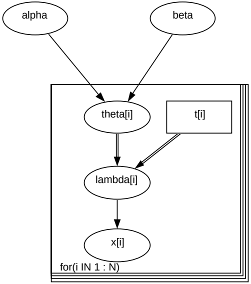
Data and BUGS model for pumps

The number of failures X_i is assumed to follow a Poisson distribution with parameter $\theta_i t_i, i = 1, \ldots, 10$
where θ_i is the failure rate for pump i and t_i is the length of operation time of the pump (in 1000s of hours). The data are shown below.

Pump	1	2	3	4	5	6	7	8	9	10
t_i	94.5	15.7	62.9	126	5.24	31.4	1.05	1.05	2.01	10.5
x_i	5	1	5	14	3	19	1	1	4	22

A gamma prior distribution is adopted for the failure rates:
$\theta_i \sim \Gamma(\alpha, \beta), i = 1, \ldots, 10$

Gamma model for pumpdata



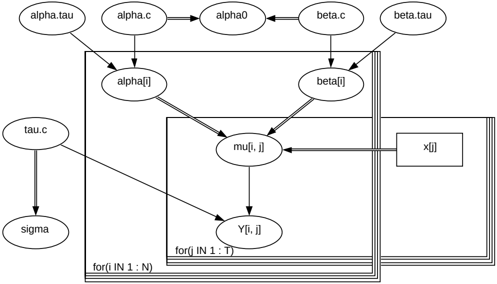
Failure of 10 power plant pumps.

BUGS program for pumps

With suitable priors the program becomes

```
model
    {
        for (i in 1 : N) {
            theta[i] ~ dgamma(alpha, beta)
            lambda[i] <- theta[i] * t[i]
            x[i] ~ dpois(lambda[i])
        }
        alpha ~ dexp(1)
        beta ~ dgamma(0.1, 1.0)
    }
```

Growth of rats



Growth of 30 young rats.

Finding full conditionals for Gibbs sampler

Inference in Bayesian complex graphical models as above uses the Gibbs sampler.

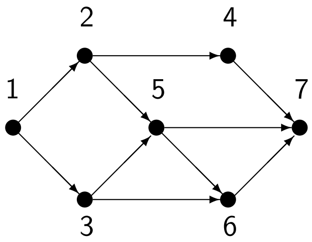
For a *DAG the densities of full conditional distributions are*:

$$\begin{aligned} f(x_i \mid x_{V\setminus i}) &\propto \prod_{v\in V} f(x_v \mid x_{\mathrm{pa}(v)}) \\ &\propto f(x_i \mid x_{\mathrm{pa}(i)}) \prod_{v\in \mathrm{ch}(i)} f(x_v \mid x_{\mathrm{pa}(v)}) \\ &= f(x_i \mid x_{\mathrm{bl}(i)}), \end{aligned}$$

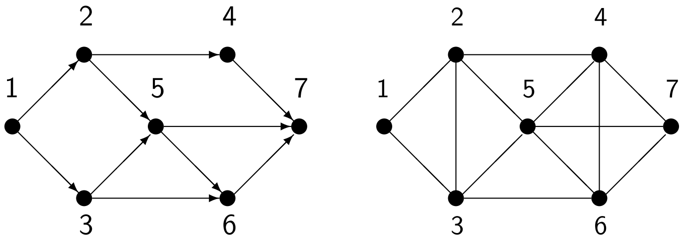
x where $\mathrm{bl}(i)$ is the *Markov blanket* of node i:

$$\mathrm{bl}(i) = \mathrm{pa}(i) \cup \mathrm{ch}(i) \cup \left\{\cup_{v\in \mathrm{ch}(i)} \mathrm{pa}(v) \setminus \{i\}\right\}.$$

Markov blanket



Markov blanket of 6 is $\mathrm{bl}(6) = \{3, 5, 7, 4\}$.



The Markov blanket is just the neighbours of in the moral graph:
$\mathrm{bl}(v) = \mathrm{ne}^m(v)$ so $\mathrm{bl}(6) = \{3, 5, 7, 4\}$ and $\mathrm{bl}(3) = \{1, 5, 6, 2\}$.

The *DAG is used for modular specification* of the model, and *the moral graph for local computation.*

- Is a huge conceptual extension of so-called Bayesian hierarchical models;

- Is a huge conceptual extension of so-called Bayesian hierarchical models;
- distinction prior/likelihood and parameter/random variable less well defined;

- Is a huge conceptual extension of so-called Bayesian hierarchical models;
- distinction prior/likelihood and parameter/random variable less well defined;
- If founder nodes in network are considered fixed and unknown, *no reason not to consider models in Fisherian paradigm.*

Darroch, J. N., S. L. Lauritzen, and T. P. Speed (1980). Markov fields and log-linear interaction models for contingency tables. *The Annals of Statistics 8*, 522–539.

Dempster, A. P. (1972). Covariance selection. *Biometrics 28*, 157–175.

Geiger, D. and J. Pearl (1990). On the logic of causal models. In R. D. Shachter, T. S. Levitt, L. N. Kanal, and J. F. Lemmer (Eds.), *Uncertainty in Artificial Intelligence 4*, pp. 3–14. Amsterdam, The Netherlands: North-Holland.

Gibbs, W. (1902). *Elementary Principles of Statistical Mechanics*. NewHaven, Connecticut: Yale University Press.

Gilks, W. R., A. Thomas, and D. J. Spiegelhalter (1994). BUGS: a language and program for complex Bayesian modelling. *The Statistician 43*, 169–178.

Haberman, S. J. (1974). *The Analysis of Frequency Data*. Chicago, Illinois: University of Chicago Press.

Hammersley, J. M. and P. E. Clifford (1971). Markov fields on finite graphs and lattices. Unpublished manuscript.

Lauritzen, S. L. (1996). *Graphical Models*. Oxford, United Kingdom: Clarendon Press.

Lauritzen, S. L., A. P. Dawid, B. N. Larsen, and H.-G. Leimer (1990). Independence properties of directed Markov fields. *Networks 20*, 491–505.

Lauritzen, S. L. and D. J. Spiegelhalter (1988). Local computations with probabilities on graphical structures and their application to expert systems (with discussion). *Journal of the Royal Statistical Society, Series B 50*, 157–224.

Lauritzen, S. L. and N. Wermuth (1989). Graphical models for associations between variables, some of which are qualitative and some quantitative. *The Annals of Statistics 17*, 31–57.

Pearl, J. (1986). A constraint–propagation approach to probabilistic reasoning. In L. N. Kanal and J. F. Lemmer (Eds.), *Uncertainty in Artificial Intelligence*, Amsterdam, The Netherlands, pp. 357–370. North-Holland.

Pearl, J. (1988). *Probabilistic Inference in Intelligent Systems*. San Mateo, CA: Morgan Kaufmann Publishers.

Wermuth, N. and S. L. Lauritzen (1983). Graphical and recursive models for contingency tables. *Biometrika 70*, 537–552.

Whittaker, J. (1990). *Graphical Models in Applied Multivariate Statistics*. Chichester, United Kingdom: John Wiley and Sons.

Wold, H. O. A. (1954). Causality and econometrics. *Econometrica 22*, 162–177.

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

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