

BAYESIAN MODEL AVERAGING AND MODEL SELECTION FOR MARKOV EQUIVALENCE CLASSES OF ACYCLIC DIGRAPHS

David Madigan¹, Steen A. Andersson², Michael D. Perlman¹,
and Chris T. Volinsky¹

¹Department of Statistics, Box 354322,
University of Washington,
Seattle, WA 98195, USA.

²Department of Mathematics,
Indiana University,
Bloomington, IN 47405, USA.

KEYWORDS: *Bayesian graphical model; Essential graph; model uncertainty; model averaging; Markov equivalence; Markov chain Monte Carlo.*

ABSTRACT

Acyclic digraphs (ADGs) are widely used to describe dependences among variables in multivariate distributions. In particular, the likelihood functions of ADG models admit convenient recursive factorizations that often allow explicit maximum likelihood estimates and that are well suited to building Bayesian networks for expert systems. There may, however, be many ADGs that determine the same dependence (= Markov) model. Thus, the family of all ADGs with a given set of vertices is naturally partitioned into Markov-equivalence classes, each class being associated with a unique statistical model. Statistical procedures, such as model selection or model averaging, that fail to take into account these equivalence classes, may incur substantial computational or other inefficiencies. Recent results have shown that each Markov-equivalence class is uniquely determined by a single chain graph, the *essential graph*, that is itself Markov-equivalent simultaneously to all ADGs in the equivalence class. Here we propose two stochastic Bayesian model averaging and selection algorithms for essential graphs and apply them to the analysis of three discrete-variable data sets.

1. Introduction.

The use of directed graphs to represent possible dependencies among random variates dates back to Wright (1921) and has generated considerable research activity in the social and natural sciences. Since 1980, particular attention has been directed at graphical Markov models specified by conditional independence relations among the variables, i.e.,

by the Markov properties determined by the graph. The recent books by Whittaker (1990) and Lauritzen (1996) conveniently summarize the statistical perspective on these developments.

Graphical Markov models determined by acyclic directed graphs (ADGs) admit especially simple statistical analyses. In particular, ADG models admit convenient recursive factorizations of their joint probability density functions (Lauritzen *et al.* (1990)), provide an elegant framework for Bayesian analysis (Spiegelhalter and Lauritzen (1990)), and, in expert system applications, allow simple causal interpretations (Lauritzen and Spiegelhalter (1988)). In the multinomial and multivariate normal cases, the likelihood function (i.e., both the joint probability density function and the parameter space) factorizes and admits explicit maximum likelihood estimates. Furthermore, the only undirected graphical (UDG) models that provide these conveniences are the decomposable models, i.e., the UDG models which have the same Markov properties as ADG models (Dawid and Lauritzen (1993), Andersson *et al.* (1995a)).

For these reasons, ADG models have become popular across an extraordinary range of applications; see, for example, Heckerman *et al.* (1992), Lauritzen and Spiegelhalter (1988), Pearl (1988), Neapolitan (1990), Spiegelhalter and Lauritzen (1990), Spiegelhalter *et al.* (1993), Madigan and Raftery (1994), and York *et al.* (1995). Indeed, the vigorous “Uncertainty in Artificial Intelligence” community focuses much of its effort on ADG models.

Much of this applied work has adopted a Bayesian perspective: “experts” specify a prior distribution on competing ADG models. These prior distributions are combined with likelihoods (typically integrated over parameters) to give posterior model probabilities. Model selection algorithms seek out the ADG models with the highest posterior probability, and subsequent inference proceeds conditionally on these selected models (Cooper and Herskovits (1990), Buntine (1994), Spiegelhalter *et al.* (1993), Heckerman *et al.* (1994), Madigan and Raftery (1994)). Non-Bayesian model selection methods proceed in a similar manner, replacing posterior model probabilities by, for example, penalized maximum likelihoods (Chickering (1995)).

Heckerman *et al.* (1994) highlight a fundamental problem with these approaches. Because several different ADGs may determine the *same* statistical model, i.e., may determine the same set of conditional independence restrictions among a given set of random variates, the collection of all possible ADGs for these variates naturally coalesces into one or more classes of *Markov-equivalent* ADGs, where all ADGs within a

Markov-equivalence class determine the *same* statistical model. Model selection algorithms that ignore these equivalence classes face three main difficulties:

1. Repeating analyses for equivalent ADGs leads to significant computational inefficiencies.
2. Ensuring that equivalent ADGs have equal posterior probabilities imposes severe constraints on prior distributions.
3. Bayesian model averaging procedures that average across ADGs assign weights to statistical models that are proportional to equivalence class sizes.

Treating each Markov-equivalence class as a single model would overcome these difficulties. Andersson *et al.* (1995b) show that for every ADG D , the equivalence class $[D]$ can be uniquely represented by a certain Markov-equivalent chain graph D^* , the *essential graph* associated with the equivalence class. (Chain graphs may have both directed and undirected edges but may contain no partially directed cycles; they include both ADGs and UDGs as special cases.) They provided an explicit characterization of those graphs G such that $G = D^*$ for some ADG D , and provide a polynomial-time algorithm for constructing D^* from D . Meek (1995) and Chickering (1995) have independently provided alternative constructions.

This characterization and construction enable more efficient model selection and model averaging procedures for ADG models, based on essential graphs. Such procedures are not immediate, however, and Section 3 describes some of the difficulties that arise, as well as two possible solutions.

Section 2 provides basic results concerning graphical models, and their Markov equivalence. Section 3 discusses Bayesian model averaging for essential graphs and Section 4 illustrates and evaluates the proposed methods in the context of three applications. We refer the reader to Andersson *et al.* (1995a or 1995b) for definitions and basic graph theoretic results.

2. Markov Equivalence of Acyclic Digraphs; the Essential Graph D^* .

We begin with the well-known graph-theoretic criterion for the Markov equivalence of ADGs. This was first discovered by Verma and Pearl (1992, Corollary 3.2) and, independently, by Frydenberg (1990, Theorem 5.6) for the more general class of chain graphs (also see Andersson *et al.* (1995a, Theorem 3.1)).

Theorem 2.1. Two ADGs are Markov equivalent if and only if they have the same skeleton and the same immoralities (see Figure 2.1).

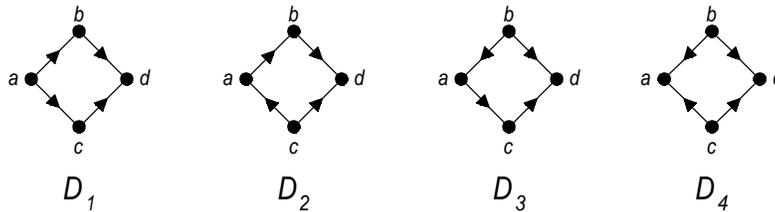


Figure 2.1: The four ADGs with the same skeleton as D_1 and the immortality (b, d, c) . The ADGs $D_1, D_2,$ and D_3 have no other immoralities, hence are Markov equivalent by Theorem 2.1. The ADG D_4 has the additional immortality (b, a, c) , hence is not Markov equivalent to the others. Thus, $[D_1] = \{D_1, D_2, D_3\}$.

The equivalence class containing D is denoted by $[D]$.

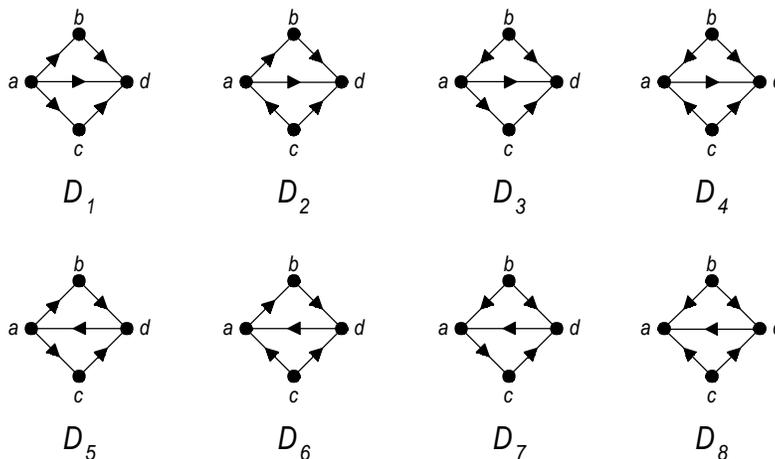


Figure 2.2: The $2^3 = 8$ possible digraphs with the same skeleton as D_1 and the immortality (b, d, c) . Of these 8, $D_5, D_6,$ and D_7 are not acyclic, while D_4 and D_8 are acyclic but possess the additional immortality (b, a, c) , so $[D_1] = \{D_1, D_2, D_3\}$.

While Theorem 2.1 provides a practical criterion for deciding whether two given ADGs are Markov equivalent, it does not directly yield a characterization of the entire equivalence class $[D]$ for a given ADG D . Consider, for example, the ADG D_1 in Figure 2.2. Theorem 2.1 implies that each digraph in $[D_1]$ must have the same skeleton as D_1 and Figure 2.2 shows all such digraphs. Since (b, d, c) is an immorality in D_1 , the arrows $b \rightarrow d$ and $c \rightarrow d$ are *essential* in D_1 , i.e., these arrows must occur in every member of $[D_1]$. The remaining three edges of D_1 might be oriented in $2^3 = 8$ possible ways, as shown in Figure 2.2; of these 8 digraphs, only 5 are acyclic, and of these 5, only three (D_1, D_2, D_3) possess the same immoralities as D_1 . Thus, $[D_1] = \{D_1, D_2, D_3\}$.

Since the number of possible orientations of all arrows that do not participate in any immorality of an ADG D grows exponentially with the number of such arrows, hence super-exponentially with the number of vertices, determination of the equivalence class $[D]$ by exhaustive enumeration of possibilities rapidly becomes computationally infeasible as the size of D increases. A closer examination of this example reveals, however, that the arrow $a \rightarrow d$ occurs in every member of $[D_1]$, hence is an essential arrow of D_1 even though it is not involved in any immorality of D_1 . Had we been able to identify all 3 essential arrows of D_1 directly from D_1 itself, it would not have been necessary to consider $D_5 - D_8$ in order to determine $[D_1]$. On the other hand, it appears necessary to determine $[D_1]$ before we can identify the essential arrows of D_1 .

Fortunately, this is not the case. Andersson *et al.* (1995b) present a polynomial-time algorithm for determining all essential arrows of an ADG D . This is done by introducing and characterizing the *essential graph* D^* associated with D :

Definition 2.1. The *essential graph* D^* associated with D is the graph with the same skeleton as D , but where an edge is directed in D^* if and only if it occurs as a directed edge (\equiv arrow) with the same orientation in *every* $D' \in [D]$; all other edges of D^* are undirected (See Figure 2.3 for examples). The directed edges in D^* are called the *essential arrows* of D .

Clearly, every arrow that participates in an immorality in D is essential, but D may contain other essential arrows as well, e.g., the arrow $a \rightarrow d$ in the second graph in Figure 2.3 and the arrows $a \rightarrow d$ and $b \rightarrow d$ in the third graph in Figure 2.3. Andersson *et al.* (1995b) show that D^* is a *chain graph* that is itself Markov equivalent to D , so that D^* contains the same

statistical information as D . Definition 2.2 and Theorem 2.2 provide a complete characterization of essential graphs.

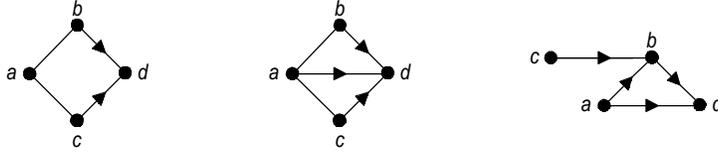
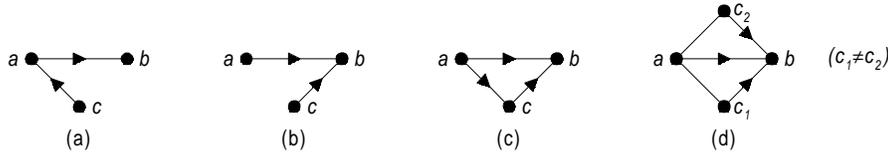


Figure 2.3: Three examples of essential graphs D^* . In the first example, D is the ADG D_1 of Figure 2.1. In the second example, D is the ADG D_1 of Figure 2.2. In the third example, $D = D^*$.

Definition 2.2. Let G be a graph. An arrow $a \rightarrow b \in G$ is *strongly protected* in G if $a \rightarrow b$ occurs in at least one of the following four configurations as an induced subgraph of G :



Theorem 2.2 (Characterization of D^* ; Andersson *et al.* (1995b)). A graph $G \equiv (V, E)$ is equal to D^* for some ADG D if and only if G satisfies the following four conditions:

- (i) G is a chain graph;
- (ii) for every chain component τ of G , G_τ is chordal;
- (iii) the configuration $a \rightarrow b - c$ does not occur as an induced subgraph of G ;
- (iv) every arrow $a \rightarrow b \in G$ is strongly protected in G .

3. Bayesian Model Averaging and Model Selection for Essential Graphs

Madigan and York (1995) introduced Markov chain Monte Carlo model composition (MC^3) for approximate Bayesian model averaging (BMA). MC^3 generates a stochastic process that moves through the class of models under consideration. Specifically, let \mathcal{m} denote the model class. MC^3 constructs an aperiodic and irreducible Markov chain, $\{M(t), t=1,2,\dots\}$, with state space \mathcal{m} and equilibrium distribution $\Pr(M | \delta)$, where δ denotes the data. If we simulate this Markov chain for $t=1,2,\dots,N$, then under mild regularity conditions, for any function $g(M)$ defined on \mathcal{m} , the average:

$$\frac{1}{N} \sum_{t=1}^N g(M(t))$$

is a simulation-consistent estimate of the expectation of $g(M)$ with respect to $\Pr(M | \delta)$ (Smith and Roberts, 1993). To estimate the posterior distribution of some quantity of interest, Δ , in this manner, set $g(M) = \Pr(\Delta | M, \delta)$. Madigan and York (1995) provide details about parametrizations, prior distributions, and likelihood calculations for graphical models.

For graphical models, selection and averaging algorithms typically move through model space by changing one edge at a time (see, for example, Edwards and Havránek (1985) or Madigan and Raftery (1994)). However, when \mathcal{M} is the class of essential graphs with a specified vertex set V , Markov chains that change one edge at a time will not be irreducible. Consider, for example, the essential graph M of Figure 2.4. Changing a single edge of M leads to one of M_1 through M_9 , none of which is an essential graph. Therefore, it is not possible to get from M to *any* other three-vertex essential graph by changing just one edge.

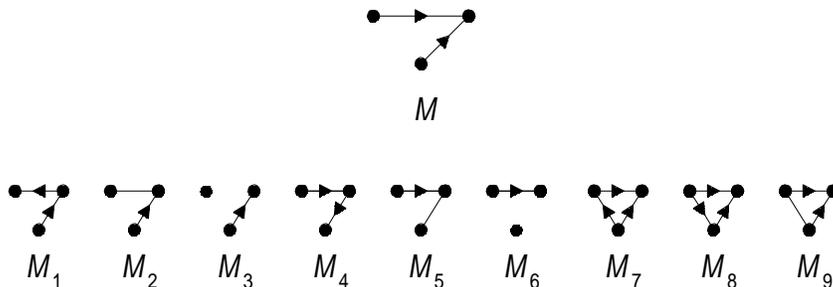


Figure 2.4: M is an essential graph. M_1 through M_9 are the graphs reachable from M by changing a single edge.

A chain that moves *one or two* edges at a time, while computationally more complex than a single edge algorithm, does overcome this irreducibility problem. Section 3.1 describes one such approach. Section 3.2 adopts an auxiliary variables approach to provide an alternative scheme that may be more efficient.

3.1 Gibbs MC³

Here we denote essential graphs by $M \equiv (V, E)$ and the data by δ . Andersson *et al.* (1995b) note that it is possible to traverse the space of essential graphs by changing one or two edges at a time. Here we propose

a corresponding Gibbs sampling scheme that is related to the SSVS scheme of George and McCulloch (1994). For every pair of vertices $(v_i, v_j) \in (V \times V)$, define E_{ij} as follows:

$$E_{ij} = \begin{cases} 0, & \text{if } (v_i, v_j) \notin E \text{ and } (v_j, v_i) \notin E \\ 1, & \text{if } (v_i, v_j) \in E \text{ and } (v_j, v_i) \notin E \\ 2, & \text{if } (v_i, v_j) \notin E \text{ and } (v_j, v_i) \in E \\ 3, & \text{if } (v_i, v_j) \in E \text{ and } (v_j, v_i) \in E. \end{cases}$$

Note that a graph $M \equiv (V, E)$ is fully specified by the collection $\Omega \equiv \{E_{ij}\}$. Our Gibbs sampler proceeds by choosing three vertices v_i , v_j , and v_k from V either according to some systematic irreducible scheme or at random, and then drawing from

$\Pr(E_{ij}, E_{jk} \mid \Omega \setminus \{E_{ij}, E_{jk}\}, \delta)$. Calculation of this conditional distribution is computationally demanding. It requires that we construct the 16 graphs corresponding to the possible states of (E_{ij}, E_{jk}) and, using Theorem 2.2 above, check whether each graph is an essential graph, assign zero probability to non-essential graphs, and compute posterior model probabilities for the remaining essential graphs.

Andersson *et al.* (1995b) suggest that it may be possible to develop more efficient Gibbs samplers. It follows from the proof of their Proposition 4.5 that the Markov chain on \mathcal{m} will be irreducible whenever the chain has positive probability of moving from the current essential graph to any essential graph:

- (a) that differs by *exactly one edge* from the current graph; or
- (b) that is obtained from the current graph by *deleting both arrows* in an immorality $a \rightarrow b \leftarrow c$, where b is a terminal vertex of the current graph and where a and c are the only parents of b in the current graph; or
- (c) that is obtained from the current graph by *adding two arrows* to form an immorality $a \rightarrow b \leftarrow c$, where b is an isolated vertex of the current graph and where a and c are not adjacent in the current graph.

We will pursue this approach in a future paper.

3.2 Augmented MC³ (AMC³)

To circumvent the computational overhead associated with the Gibbs sampler algorithm defined above, we introduce an auxiliary variable Markov chain Monte Carlo model composition scheme. Rather than construct a Markov chain with equilibrium distribution $\Pr(M \mid \delta)$, we define an auxiliary variable T taking values in \mathcal{I} , the set of total orderings

of V , and construct a Markov chain with equilibrium distribution $\Pr(M, T | \delta) \equiv \Pr(M | \delta) \Pr(T | M, \delta)$, where $\Pr(M | \delta)$ is the usual posterior model distribution and $\Pr(T | M, \delta)$ can be chosen arbitrarily. A Gibbs sampler that draws in turn from $\Pr(M | T, \delta)$ and $\Pr(T | M, \delta)$ defines the required Markov chain for $\Pr(M, T | \delta)$ and thence $\Pr(M | \delta)$ (Besag and Green (1993)). We use Markov chain methods to draw from both conditional distributions (Tierney, 1995, Smith and Roberts, 1993); we devise Hastings-Metropolis algorithms for each one.

A total ordering T is said to be *compatible* with an essential graph M , if $v_i <_T v_j$ whenever (v_i, v_j) is a directed edge in M and T_τ is a perfect ordering for each chain component τ of M (so that orienting the lines in τ according to T_τ generates an ADG that is Markov equivalent to M). Furthermore, for technical reasons we require that T is restricted to those total orderings that the maximum cardinality search (MCS) algorithm can generate. MCS provides a convenient method of generating perfect orderings for chordal graphs but does not generate *all* possible perfect orderings. Now define $\Pr(T | M, \delta)$ as follows:

$$\Pr(T | M, \delta) = \begin{cases} \text{constant, if } T \text{ is compatible with } M \\ 0, \text{ otherwise} \end{cases}$$

Note that by construction, $T \perp \delta \mid M$, so that $\Pr(M | T, \delta) \propto \Pr(\delta | M) \times \Pr(M | T)$.

The Gibbs sampler begins at an arbitrary $M \in \mathcal{M}$ and a total ordering T compatible with M , generated by the following algorithm:

1. For each chain component τ of M :
 - 1.1 Choose a vertex $v_1^\tau \in \tau$ at random.
 - 1.2 Using MCS, order the remaining vertices $v_2^\tau, \dots, v_{m_\tau}^\tau \in \tau$ (if any) breaking ties by random selection.
 - 1.3 Store p , the probability of having chosen this ordering.
2. Now generate T as follows:
 - 2.1 Let $T = \emptyset$.
 - 2.2 Choose a chain component $\tau \in M$ at random such that τ is initial in M . Append $v_1^\tau, \dots, v_{m_\tau}^\tau$ to T .
 - 2.3 Remove τ from M and if $M \neq \emptyset$, go to 2.2.
 - 2.4 Store p , the probability of choosing this sequence of chain components.

Note that given M , the probability of generating this T is $p_T \equiv p \prod_{\tau} p_\tau$.

To draw from $\Pr(T|M,\delta)$, use the above algorithm to generate T' , a candidate total ordering compatible with M . T' is then accepted with probability

$$\min\left\{1, \frac{p_{T'}}{p_T}\right\};$$

otherwise the chain remains at T . This is an example of an “independence chain” (Tierney, 1995).

To draw from $\Pr(M|T,\delta)$, first generate an ADG, $D \equiv (V,F)$, by orienting all lines in M in accordance with T . Next, generate an ADG $D' \equiv (V,F')$ as follows. Randomly choose a pair of vertices $(v_i, v_j) \in V \times V$, with $v_i <_T v_j$. If $(v_i, v_j) \in F$, then let $F' = F \setminus (v_i, v_j)$. If $(v_i, v_j) \notin F$, let $F' = F \cup (v_i, v_j)$. Next generate the essential graph $M' \equiv (V, E')$ corresponding to D' and accept it with probability

$$\min\left\{1, \frac{\Pr(M|T)\Pr(\delta|M')}{\Pr(M|T)\Pr(\delta|M)}\right\}.$$

Otherwise, the chain stays at M . Note that since D and D' have different skeletons, they are not Markov equivalent, so that $M' \neq M$. Furthermore, $\Pr(M \rightarrow M') = \Pr(M' \rightarrow M) = 2/n(n-1)$ so that the Metropolis algorithm applies.

The calculation of $\Pr(M'|T)/\Pr(M|T)$ presents a potential difficulty with this algorithm. First note that

$$(1) \quad \frac{\Pr(M|T)}{\Pr(M|T)} = \frac{\Pr(T|M')\Pr(M')}{\Pr(T|M)\Pr(M)}.$$

$\Pr(T|M)$ is the reciprocal of the number of compatible orderings associated with M and can be formidable to compute. However, since M and M' correspond to ADGs that differ by just one edge, $\Pr(T|M')/\Pr(T|M)$ is typically close to one. A further refinement is provided by noting that p_T is a consistent estimator of $\Pr(T|M)$. Since p_T will be calculated for every ordering, this can provide an estimate of $\Pr(T|M')/\Pr(T|M)$ at no extra computational cost. In what follows, we refer to the algorithm with $\Pr(T|M')/\Pr(T|M)$ in the right hand side of (1) replaced by the approximation $p_{T'}/p_T$, as the *adjusted* AMC³ algorithm.

For both MC³ algorithms, aperiodicity is guaranteed since the chain always has positive probability of remaining in its current state. Both algorithms can also be used to find the essential graph with the maximum

posterior probability, although the introduction of an annealing parameter would then hasten convergence. We note that it may be possible to develop more efficient Gibbs MC³ algorithms by combining the algorithms of Andersson *et al.* (1995b, Section 5) and Meek (1995) to transform proposed non-essential graphs to essential graphs.

Non-stochastic model selection and model averaging schemes based on essential graphs also can be developed, analogous to those proposed by Heckerman *et al.* (1994), Højsgaard and Thiesson (1995), and Madigan and Raftery (1994) for ADGs.

4. Applications

We have applied both MC³ algorithms to two much-studied datasets, both involving six binary variables, as well as a third example involving 14 binary variables. Even for the six vertex examples there are approximately 10^6 essential graphs. Thus an exhaustive search over the space of essential graphs would be laborious.

In each case we started the Markov chain at the empty model and ran the chain for 100,000 iterations, discarding the first 10,000. The prior distributions on the parameters of each of the models used an equivalent prior sample size of one. We assumed that all models were equally likely *a priori*.

4.1 Coronary Heart Disease

Our first example concerns data on 1,841 men cross-classified according to risk factors for Coronary Heart Disease. This data set was previously analyzed by Edwards and Havránek (1985) and others. The risk factors are as follows: *A*, smoking; *B*, strenuous mental work; *C*, strenuous physical work; *D*, systolic blood pressure; *E*, ratio of beta and alpha proteins; *F*, family anamnesis of coronary heart disease. Figure 2.5 shows the leading three models from the adjusted AMC³ algorithm:

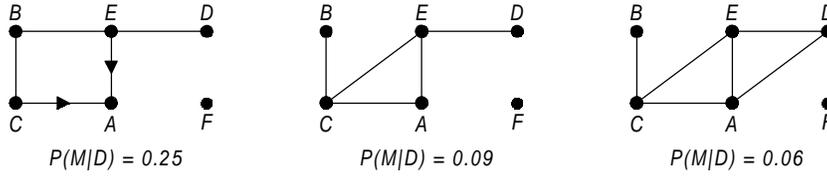


Figure 2.5: The top three models for the Coronary Heart Disease example visited by the adjusted AMC³ algorithm, and their respective estimated posterior probabilities.

For comparison purposes, Figure 2.6 shows the four ADG models selected by the Occam’s Window procedure of Madigan and Raftery (1994) and Figure 2.7 shows the corresponding essential graphs:

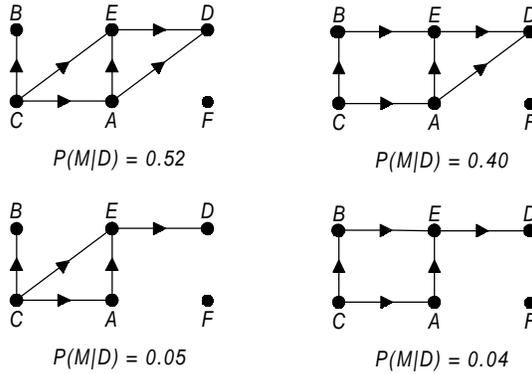


Figure 2.6: ADG models selected by Madigan and Raftery (1994) for the Coronary Heart Disease example. The model probabilities have been normalized.

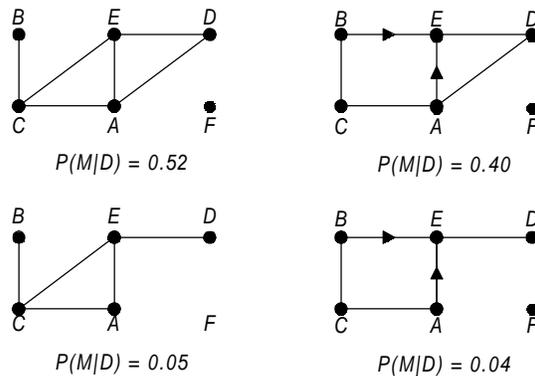


Figure 2.7: The essential graphs corresponding to the ADG models selected by Madigan and Raftery (1994) for the Coronary Heart Disease example.

By summing the posterior model probabilities for models in which a particular arrow or line occurs, we obtain corresponding posterior probabilities. Table 1 shows these probabilities for the Coronary Heart

Disease example. We believe that this output may be of interest from a causal modeling perspective.

Similarly by summing the posterior model probabilities for models in which a particular conditional independence occurs, we obtain corresponding posterior probabilities. Table 2 shows these probabilities for three medically interesting conditional independences for the Coronary Heart Disease example. There is a 0.77 probability that F (family anamnesis) is marginally independent of the remaining variables, a 0.60 probability that D (systolic blood pressure) is conditionally independent of A (smoking), B (strenuous mental work), and C (strenuous physical work) given E (proteins), and a 0.35 probability that B (strenuous mental work) and E (proteins) are conditionally independent given C (strenuous physical work). These posterior probabilities are not conditional on a

Table 1: Edge probabilities for the Coronary Heart Disease example provided by the adjusted AMC³ algorithm.

Vertices		Posterior Probabilities			
		• •	•→•	•←•	•—•
A	B	1.00	0.00	0.00	0.00
A	C	0.00	0.03	0.41	0.56
A	D	0.65	0.09	0.01	0.25
A	E	0.11	0.17	0.41	0.31
A	F	0.98	0.01	0.00	0.01
B	C	0.00	0.00	0.04	0.96
B	D	1.00	0.00	0.00	0.00
B	E	0.37	0.15	0.00	0.47
B	F	0.89	0.00	0.00	0.10
C	D	1.00	0.00	0.00	0.00
C	E	0.68	0.00	0.03	0.29
C	F	0.99	0.00	0.00	0.01
D	E	0.27	0.02	0.11	0.59
D	F	0.97	0.00	0.00	0.03
E	F	0.95	0.01	0.00	0.04

model and may be more useful to a data analyst than the list of independences in a single selected model. Our approach may also provide a better “oracle” for inferring causation from data (Spirtes, *et al.*, 1993).

Table 2: Marginal probabilities associated with specific conditional independences for the Coronary Heart Disease example provided by the adjusted AMC³ algorithm.

Conditional Independence	Posterior Probability
$F \perp (A,B,C,D,E)$	0.77
$D \perp (A,B,C) E$	0.60
$B \perp E C$	0.35

Madigan and Raftery (1994) in their analysis of the Coronary Heart Data impose a partial ordering on the variables: $F, (B, C), A, (E, D)$ noting that “The variables B, F , or C could not be influenced by the other factors and must be exogenous, although the ordering of B and C is unclear. Similarly D or E could hardly influence A , although the ordering of E and D is unclear.” While Table 1 does suggest that the data supports the precedence of C over A , it also suggests that the data does not support the precedence of A (smoking) over E (protein ratio).

Madigan and Raftery (1994) further report that their data analysis provided “strong evidence for the precedence of E over D and weak evidence for the precedence of C over B ”. From Table 1, the odds in favor of $E \rightarrow D$ as against $D \rightarrow E$ are 5.5 and there is no support for a $B \rightarrow C$ edge, so that our data-driven results are in agreement with those of Madigan and Raftery (1994).

Both the non-adjusted AMC³ algorithm and the Gibbs MC³ algorithm produce essentially identical results to those in Table 1 and Figure 2.5, although the Gibbs MC³ algorithm takes approximately fifteen times more CPU time. This discrepancy may be somewhat illusory however, since the Gibbs MC³ algorithm accepts all proposed moves, whereas the AMC³ algorithm typically accepts 10-20% of the proposed moves.

4.2 Women and Mathematics

Our second example concerns a survey which was reported in Fowlkes *et al.* (1988) concerning the attitudes of New Jersey high-school students towards mathematics. A total of 1,190 students in eight schools took part in the survey. The variables collected were: *A*, lecture attendance; *B*, Sex; *C*, School Type (suburban or urban); *D*, “I’ll need mathematics in my future work” (agree or disagree); *E*, Subject Preference (maths/science or liberal arts); *F*, Future Plans (college or job). In what follows we refer to this as the “Women and Mathematics” example.

Figure 2.8 shows the leading three models from the adjusted AMC³ algorithm:

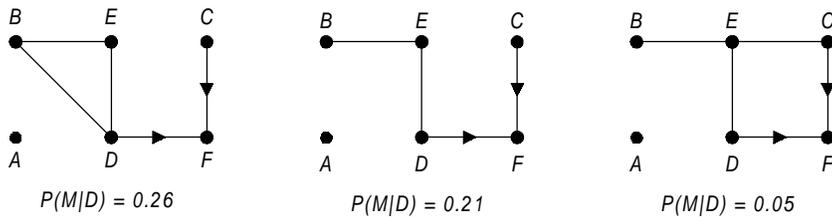


Figure 2.8: The top three models for the Women and Mathematics example visited by the adjusted AMC³ algorithm, and their respective estimated posterior probabilities.

Again, for comparison purposes, Figure 2.9 shows the sole ADG model selected by the Occam’s Window procedure of Madigan and Raftery (1994):

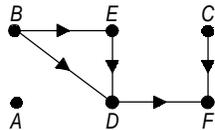


Figure 2.9: ADG model selected by Madigan and Raftery (1994) for the Women and Mathematics example. The essential graph corresponding to this ADG is the leftmost (and MAP) graph in Figure 2.8.

Table 3 shows the edge probabilities for the Women and Mathematics example.

Table 3: Edge probabilities for the Women and Mathematics example provided by the adjusted AMC³ algorithm.

Vertices		Posterior Probabilities			
		• •	•→•	•←•	•—•
A	B	0.98	0.00	0.00	0.02
A	C	0.98	0.00	0.00	0.02
A	D	0.98	0.00	0.00	0.01
A	E	0.98	0.00	0.00	0.02
A	F	0.99	0.00	0.00	0.00
B	C	0.99	0.00	0.00	0.01
B	D	0.42	0.00	0.00	0.58
B	E	0.00	0.00	0.00	0.99
B	F	1.00	0.00	0.00	0.00
C	D	0.98	0.00	0.00	0.02
C	E	0.82	0.00	0.00	0.17
C	F	0.00	0.70	0.00	0.30
D	E	0.00	0.00	0.03	0.97
D	F	0.00	0.70	0.03	0.27
E	F	1.00	0.00	0.00	0.00

Madigan and Raftery (1994) in their analysis of the Women and Mathematics assume that B (Sex) and C (School Type) were exogenous. Our analysis provides some support for this assumption - the posterior probability that B and C are marginally independent is 0.64. They later removed the restriction that C be exogenous and found some support for a link from E to C , “although its interpretation is somewhat unclear”. Table 3 suggests that if there is an edge connecting E and C , it is an undirected edge. Our analysis does provide strong support for directed links from C (School Type) to F (Future Plans) and D (I’ll need math) to F .

4.3 Coronary Artery Disease

Our final example concerns a study of risk factors for coronary artery disease reported by Hansen (1980) and reanalyzed by Anderson et al. (1991) using decomposable graphical models and Højsgaard and Thiesson (1995) using ADG-equivalent chain graphs. Coronary artery disease is a disease caused by a reduction in the ability of the coronary arteries to

supply the heart muscle. Physicians refer patients suspected of having coronary artery disease to coronary arteriography if their presenting features, disease manifestations, and non-invasive tests are indicative of coronary artery disease. Due to the morbidity and cost associated with arteriography, Hansen's (1980) study sought to provide improved screening. Hansen (1980) reports data for 236 patients on 14 binary variables ("the learning cases") and data on 67 patients on subsets of the 14 variables ("the test cases"). Hansen (1980) actually reports Angina Pectoris with three levels: none, typical, and atypical. In our analysis we combined the typical and atypical categories. Table 4 presents the fourteen variables:

Table 4: The fourteen variables in the coronary artery disease example.

Symbol	Variable Name
<i>s</i>	Sex
<i>S</i>	Smoking
<i>H</i>	Hypercholesterolaemia
<i>I</i>	Hereditary Predispositions
<i>w</i>	Workload (adequate ECG)
<i>A</i>	Previous Myocardial Infarction
<i>a</i>	Angina Pectoris
<i>h</i>	Left Ventricular Hypertrophy
<i>K</i>	Congenital Heart or Valve Disease
<i>Q</i>	Q-wave in ECG
<i>T</i>	ST Segment Shift in ECG
<i>q</i>	Q-wave informative
<i>t</i>	ST-shift Informative
<i>c</i>	Coronary Artery Disease

Figure 2.10 shows the leading model from the adjusted AMC³ algorithm. Note that this graph contains no essential arrows. One possible interpretation of this model is that the data provide little evidence about possible causal relationships between these variables.

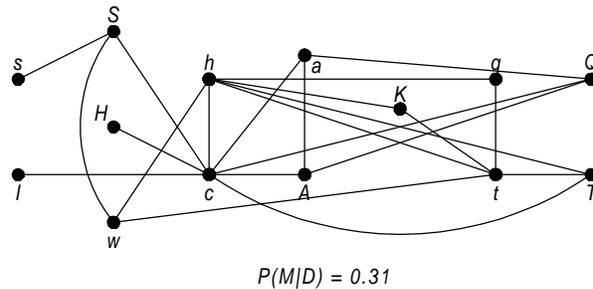


Figure 2.10: The top model for the Coronary Artery Disease example visited by the adjusted AMC³ algorithm, and its estimated posterior probability.

In consultation with a physician, Højsgaard and Thiesson (1995) *a priori* impose a block recursive structure on the variables, and also constrained their model selection algorithm to include certain links and exclude others. Figure 2.11 displays these constraints.

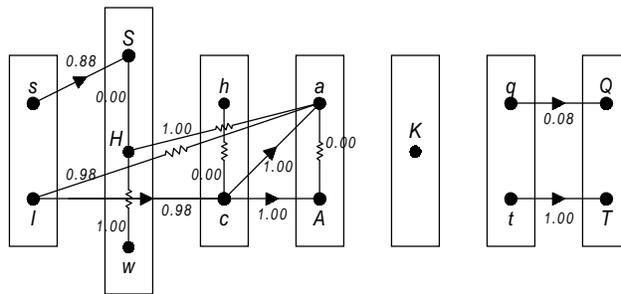


Figure 2.11: A priori constraints imposed by Højsgaard and Thiesson (1995). The block recursive structure shown constrained their model selection algorithm. Edges drawn as solid lines were forced to be present; edges drawn as zigzag lines were forced to be absent. The probabilities attached to each edge are the estimated posterior probability that the purported event (i.e., edge present regardless of orientation or edge absent) occurred. The data provides widely varying support for their prior assumptions.

Using different selection criteria, Højsgaard and Thiesson (1995) select several ADG-equivalent chain graph models for this example. Among these, a selection criterion similar to BIC produced the single model with the best predictive performance on the test cases and Figure 2.12 shows this chain graph model. This model is similar to the leading model from the AMC³ algorithm shown in Figure 2.10. However, the model of Figure 2.10 has edges connecting q and Q , and S and H , and no edges connecting a and A , and c and h , contrary to Højsgaard and Thiesson's prior constraints. Furthermore, the model of Figure 2.10 contains edges

connecting s and w , K and t , a and Q , and c and S , and does not have an edge connecting t and c .

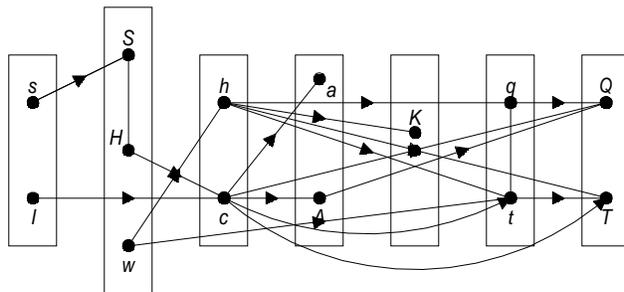


Figure 2.12: Of the models selected by Højsgaard and Thiesson (1995), this model provided the best predictive performance on the test cases.

We note that via $\Pr(M)$, the MC^3 algorithms can incorporate *a priori* constraints such as those imposed by Højsgaard and Thiesson (1995).

4.4 Predictive Performance

An effective method to judge a modeling strategy is to see how well the resulting models predict future observations. We have assessed the predictive performance of the adjusted AMC^3 method for the three examples. For the Coronary Heart Disease and Women and Mathematics examples, predictive performance, measured by the average predictive probability, is assessed by randomly splitting the complete data sets into two subsets. One subset, δ_S , containing 50% of the data, is used to select models with the other subset, $\delta_T \equiv \delta \setminus \delta_S$, being used as set of test cases. For the Coronary Artery Disease example, δ_S consists of the 236 learning cases and δ_T consists of the 67 test cases referred to in Section 4.3.

Specifically, we measure the predictive ability of an individual model, M , by

$$\frac{1}{\#(\delta_T)} \sum_{d \in \delta_T} \Pr(d|M, \delta_S),$$

while we measure the predictive performance of BMA by

$$\frac{1}{\#(\delta_T)} \sum_{d \in \delta_T} \left\{ \sum_M \Pr(d|M, \delta_S) \Pr(M|\delta_S) \right\}.$$

The intuition here is that effective models should assign high probability to the observations in δ_T and hence have a higher predictive score.

Table 5 shows the predictive performance for the Coronary Heart Disease example. The numbers can be compared with the probability of 0.0156 that a uniform distribution would assign to each observation. Adjusted AMC³ assigns probabilities to the observations in the test dataset that are, on average, 1.5% higher than those assigned by the single model with the (estimated) largest posterior probability.

Table 5: Predictive performance for the Coronary Heart Disease example showing the average predictive probability for the three models with the highest posterior probability, the maximum *a posteriori* (MAP) model from Madigan and Raftery (1994), and the Adjusted AMC³ method. Note that this analysis uses half the data to select models so that the leading models are different from those in Figure 2.5.

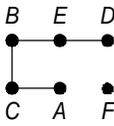
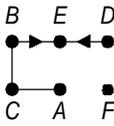
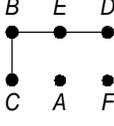
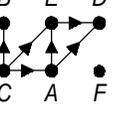
Model	Posterior Probability %	Average Predictive Probability
	0.35	0.0263
	0.02	0.0264
	0.02	0.0264
	Madigan and Raftery MAP ADG model.	0.0265
Adjusted AMC ³	BMA	0.0267

Table 6 shows the predictive performance for the Women and Mathematics example. Adjusted AMC³ assigns probabilities to the observations in the test dataset that are, on average, 3% higher than those assigned by the single model with the largest posterior probability.

Table 6: Predictive performance for the Women and Mathematics example showing the average predictive probability for the three models with the highest posterior probability, MAP model from Madigan and Raftery (1994), and the Adjusted AMC³ method.

Model	Posterior Probability %	Average Predictive Probability
	0.29	0.0235
	0.07	0.0231
	0.07	0.0233
	Madigan and Raftery MAP ADG model.	0.0241
Adjusted AMC ³	BMA	0.0242

Table 7: Predictive performance for the Coronary Artery Disease example showing the average predictive probability for the model with the highest posterior probability, the model selected by Højsgaard and Thiesson (1995), and the Adjusted AMC³ method.

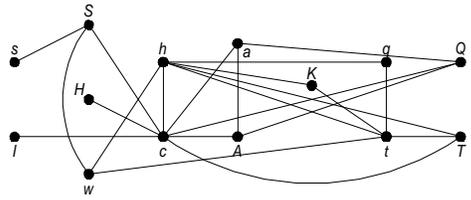
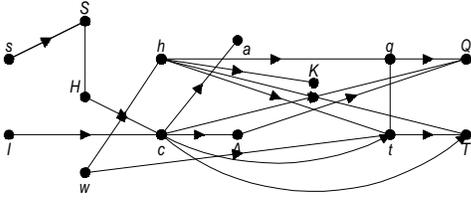
Model	Posterior Probability %	Average Predictive Probability
	0.31	0.00063
	Model from Højsgaard and Thiesson (1995) with best predictive performance	0.00059
Adjusted AMC ³	BMA	0.00066

Table 7 shows the predictive performance for the Coronary Artery Disease example. Adjusted AMC³ assigns probabilities to the observations in the test dataset that are, on average, 4% higher than those assigned by the single model with the largest posterior probability.

Repeating the random split, varying the subset proportions, or starting the Markov chain from a different location produces very similar results, so here we have reported the results from a single split for each application.

5. Discussion

By focusing on Markov-equivalence classes of ADGs rather than on the individual ADGs themselves, data analysts and expert system builders can overcome several difficulties associated with ADG models. Three

such difficulties were listed in Section 1 - here we examine these in more detail and indicate how the introduction of essential graphs can help to overcome them.

Heckerman *et al.* (1994) and Chickering (1995) argue that statistical inference for ADG models should be “score equivalent”: in the absence of *a priori* causal knowledge, Markov-equivalent ADGs should have identical posterior model probabilities (Bayesian) or identical penalized likelihoods (non-Bayesian). Under this criterion, therefore, model selection and model averaging algorithms need to visit each Markov-equivalence class only once. However, standard algorithms (e.g., Madigan and Raftery (1994), Madigan and York (1995), Heckerman *et al.* (1994)) fail to treat each Markov-equivalence class of ADGs as a single statistical model and search in the space of ADGs, introducing considerable computational inefficiency. For example, an exhaustive search amongst all ADGs on four variables would require the calculation of posterior probabilities for all 453 such ADGs, whereas a search over the space of essential graphs (in 1-1 correspondence with the equivalence classes) would require only 185 such calculations. For five variables the numbers become 29,281 and 8,782, respectively (see Andersson *et al.* (1995b, Section 6)).

For a Bayesian analysis over the space of all *individual* ADG models with a fixed vertex set V , score equivalence imposes severe restrictions on the prior distributions that may be used to represent prior knowledge about the parameters in these models. For any individual ADG D , the joint pdf (if it exists) of a global D -Markovian distribution admits the factorization (cf. Lauritzen *et al.* (1990, Theorem 1)):

$$f(V) = \prod (f(a|\text{pa}_D(a)) | a \in V).$$

For categorical data, where each conditional pdf $f(a|\text{pa}_D(a))$ is multinomial, Spiegelhalter and Lauritzen (1990) proposed the now-widely accepted conjugate family of Dirichlet prior distributions for the parameters occurring in these conditional multinomial distributions. However, Heckerman *et al.* (1994) show that score equivalence *requires that the sum of the parameters of all the Dirichlet distributions associated with each $a \in V$ (ie, the Dirichlet distributions for each of the levels of $\text{pa}_D(a)$) be identical for all $a \in V$* . Since these sums behave as “equivalent sample sizes” in subsequent Bayesian updating, this constraint severely restricts an “expert” with more prior knowledge about some variables than others - he must use a single equivalent sample size for *each* of the Dirichlet

distributions occurring in the conjugate prior, and is therefore unable to fully utilize his prior knowledge.

We overcome this difficulty by constructing prior distributions over Markov-equivalence classes of ADG models, rather than over the individual ADG models themselves, and since score equivalence is no longer an issue, *no constraints are required* on the parameters of these hyper-Dirichlet priors.

Furthermore, although the Dirichlet and hyper-Dirichlet families provide considerable flexibility for modeling prior knowledge in the Bayesian analysis of categorical data, more general priors, such as mixtures of Dirichlet distributions, sometimes may be needed to adequately reflect prior knowledge (Bernardo and Smith (1994), p.279). When working in the space of individual ADG models, however, Geiger and Heckerman (1995) show that the Dirichlet family is the *only family of prior distributions* that can be used to achieve score equivalence. Working in the space of Markov-equivalence classes, conveniently represented by essential graphs, eliminates the issue of score equivalence and therefore allows the adoption of *arbitrary* prior distributions on the associated parameters, at least in principle.

Madigan and Raftery (1994) and others argue that basing inference on a single model ignores model uncertainty and leads to poorly calibrated predictions. Bayesian model averaging provides a remedy: current BMA procedures average inferences or predictions over all models in the class under consideration, or at least over a subset of the models that receive substantial posterior weight (see Madigan and York (1995) for a review.) When applied naively to ADG models, however, BMA assigns a weight to each Markov-equivalence class that is proportional to its size. Instead, averaging directly over equivalence classes overcomes this problem.

Acknowledgments

The U. S. National Science Foundation supported the research of Andersson, Madigan and Perlman. The U.S. Office of Naval Research supported Volinsky's research (N00014-91-J-1014). The authors are grateful to Julian Besag and Adrian Raftery for helpful discussions

References

- Andersen, L.R., Krebs, J.H., and Damgaard, J. (1991) STENO: an expert system for medical diagnosis based on graphical models and model search. *Journal of Applied Statistics*, **18**, 139-153.
- Andersson, S. A., D. Madigan, and M. D. Perlman (1995a). On the Markov equivalence of chain graphs, undirected graphs, and acyclic digraphs. *Scandinavian Journal of Statistics*, to appear.
- Andersson, S. A., D. Madigan, and M. D. Perlman (1995b). A characterization of Markov equivalence classes for acyclic digraphs. Submitted for publication.
- Bernardo, J. M. and A. F. M. Smith (1994). *Bayesian Theory*. Wiley, Chichester.
- Besag, J. and P.J. Green, (1993). Spatial statistics and Bayesian computation (with discussion). *Journal of the Royal Statistical Society (Series B)* **55**, 25-37.
- Buntine, W. (1994). Operations for learning with graphical models. *Journal of Artificial Intelligence Research* **2**, 159-225.
- Chickering, D. M. (1995). A transformational characterization of equivalent Bayesian network structures. In *Uncertainty in Artificial Intelligence, Proceedings of the Eleventh Conference* (P.Besnard and S. Hanks, eds.), San Francisco: Morgan Kaufman, pp. 87-98.
- Cooper, G. F. and E. Herskovits (1992). A Bayesian method for the induction of probabilistic networks from data. *Machine Learning* **9**, 309-347.

- Dawid, A. P. and S. L. Lauritzen (1993). Hyper Markov laws in the statistical analysis of decomposable graphical models. *Annals of Statistics* **21**, 1272-1317.
- Edwards, D. and T. Havránek (1985). A fast procedure for model search in multidimensional contingency tables. *Biometrika* **72**, 339-351.
- Fowlkes, E.B., A.E. Freeny, and J.M. Landwehr (1988). Evaluating logistic models for large contingency tables. *Journal of the American Statistical Association* **83** 611-622.
- Frydenberg, M. (1990). The chain graph Markov property. *Scandinavian Journal of Statistics* **17**, 333-353.
- Geiger, D. and D. Heckerman (1995). A characterization of the Dirichlet distribution with applications to learning Bayesian networks. In *Uncertainty in Artificial Intelligence, Proceedings of the Eleventh Conference* (P. Besnard and S. Hanks, eds.), San Francisco: Morgan Kaufman, pp. 196-207.
- George, E. and R. McCulloch (1994). Variable selection via Gibbs sampling. *Journal of the American Statistical Association* **88**, 881-889.
- Hansen, J.F. (1980) The clinical diagnosis of ischaemic heart disease due to coronary artery disease. *Danish Medical Bulletin*, **27**, 280-186.
- Heckerman, D., E. Horvitz, and B. Nathwani (1992). Toward normative expert systems I: The PATHFINDER project. *Methods of Information in Medicine* **31**, 90-105.
- Heckerman, D., D. Geiger, and D. M. Chickering (1994). Learning Bayesian networks: the combination of knowledge and statistical data. In *Uncertainty in Artificial Intelligence, Proceedings of the Tenth Conference* (B. Lopez de Mantaras and D. Poole, eds.), San Francisco: Morgan Kaufman, pp. 293-301.

- Højsgaard, S. and B. Thiesson (1995). BIFROST - Block recursive models induced from relevant knowledge, observations, and statistical techniques. *Computational Statistics and Data Analysis* **19**, 155-175.
- Lauritzen, S. L. (1996). *Graphical Association Models*. Oxford University 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.
- Madigan, D., J. Gavrin, and A. E. Raftery (1995). Eliciting prior information to enhance the predictive performance of Bayesian graphical models. *Communications in Statistics: Theory and Methods*, **24**, 2271-2292.
- Madigan, D., A. E. Raftery, J.C. York, J.M Bradshaw, and R.G. Almond (1994). Strategies for graphical model selection. In *Selecting Models from Data: Artificial Intelligence and Statistics IV*, (P. Cheeseman and W. Oldford, eds.), Springer Verlag, pp. 91-100.
- Madigan, D. and A. E. Raftery (1994). Model selection and accounting for model uncertainty in graphical models using Occam's window. *Journal of the American Statistical Association* **89** 1535-1546.
- Madigan, D. and J. York. (1995). Bayesian graphical models for discrete data. *International Statistical Review*, **63**, 215-232.

- Meek, C. (1995). Causal inference and causal explanation with background knowledge. In *Uncertainty in Artificial Intelligence, Proceedings of the Eleventh Conference* (P. Besnard and S. Hanks, eds.), San Francisco: Morgan Kaufman, pp. 403-410.
- Neapolitan, R. E. (1990). *Probabilistic Reasoning in Expert Systems*. Wiley, New York.
- Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, San Mateo, CA.
- Smith, A.F.M. and G.O. Roberts, (1993). Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods (with discussion). *Journal of the Royal Statistical Society (Series B)* **55**, 3-23.
- Spiegelhalter, D. J., A. P. Dawid, S. L. Lauritzen, and R. G. Cowell (1993). Bayesian analysis in expert systems (with discussion). *Statistical Science* **8**, 219-283.
- Spiegelhalter, D. J. and S. L. Lauritzen (1990). Sequential updating of conditional probabilities on directed graphical structures. *Networks* **20**, 579-605.
- Spirtes, P., Glymour, C., and Scheines, R. (1993) *Causation, Prediction, and Search*. Springer-Verlag.
- Tierney, L. (1995). Markov chains for exploring posterior distributions (with discussion). *Annals of Statistics*, **22**, 1701-1762.
- Verma, T. and J. Pearl (1992). An algorithm for deciding if a set of observed independencies has a causal explanation. In *Uncertainty in Artificial Intelligence, Proceedings of the Eighth Conference* (D. Dubois, M. P. Wellman, B. D'Ambrosio, and P. Smets, eds.), San Francisco: Morgan Kaufman, pp. 323-330.

Whittaker, J. L. (1990). *Graphical Models in Applied Multivariate Statistics*. Wiley, New York.

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

York, J., D. Madigan, I. Heuch, and R. T. Lie (1995). Estimating a proportion of birth defects by double sampling: a Bayesian approach incorporating covariates and model uncertainty. *Applied Statistics* **44**, 227-242.