On searching for optimal classifiers among Bayesian networks

Robert G. Cowell
Department of Actuarial Science and Statistics
City University, London
Northampton Square
London EC1E 0HT

Abstract

There is much interest in constructing from datasets Bayesian networks which are efficient, or even optimal, for classification purposes. Most search strategies usually discriminate between networks by comparing their marginal likelihood score, but recently it has been suggested that search strategies for classifiers should instead select among models using alternative scores. This paper contributes to this discussion by presenting the results of simulations on the sets of all directed acyclic graphs on four and five nodes. Our results add evidence to earlier indications that the marginal likelihood is likely to be a poor criterion to use for classifier selection.

Keywords: Bayesian network, classifier, classification rate, model search, marginal likelihood, Kullback–Leibler divergence.

1 Introduction

Let C denote a discrete class random variable and A a set of discrete attribute random variables, taking values in the spaces $\mathcal{C}$ and $\mathcal{A}$ respectively. The classification problem is to predict the value $c \in \mathcal{C}$ of the class variable C given a value $a \in \mathcal{A}$ of the attributes A. If the joint probability distribution $P(C, A)$ were known, then the natural predictor for C given $A = a$ is that state $\hat{c}(a) \in \mathcal{C}$ which maximizes $p(c \mid a)$. (Indeed this is the Bayes act under a $0 - 1$ loss function.) If, furthermore, P is directed Markov with respect to a directed acyclic graph (DAG) g having the nodes $V = \{C, A\}$, then the Bayes act for selecting among all Bayesian networks having nodes V would be to select g endowed with the generating distribution $P(C, A)$. For a recent comprehensive treatment of Bayesian networks see Cowell *et al.* (1999).

Typically in classification problems neither g nor $P(C, A)$ will be known. Then one approach to the classification problem is to find a Bayesian network, consisting of the variables C and A, which provides a good fit to the data from which the conditional probabilities $p(c \mid a)$ may be estimated reliably. A commonly used score to judge the relative fits of alternative Bayesian network models is the *marginal likelihood*. For more details of these approaches see: Buntine(1994, 1996); Sanguesa and Cortes (1997); and Heckerman (1998). Unfortunately, the space of possible Bayesian network structures grows rapidly in size with the number of nodes in the graph (Robinson 1977), which makes an exhaustive comparison of all graphical models infeasible (except for graphs having a small number of nodes) and search heuristics become mandatory. One should expect a good heuristic search should find networks having a marginal likelihood close to the maximum, and one might hope that such networks would themselves be close to being optimal for use as classifiers.

Recently, however, Kontkanen *et al.* (1999) have questioned the logic of selecting among Bayesian network classifiers using the marginal likelihood, which is equivalent to using a log-scoring rule on the joint distribution, when a classifier will be judged by its predictive accuracy under a $0-1$ loss function. They point to examples for which such induced networks perform badly at classification. Instead they argue that better classifiers should be sought using scoring rules more appropriate to the classification task. Friedman *et al.* (1997) have voiced similar concerns.

In this paper we compare how the classification rates of alternative Bayesian networks vary with their predictive performance (or rather, the asymptotic increase in the marginal likelihood per case). We do this by taking a specific network, which we denote by m_0, on four binary nodes, and generating at random a set of probability distributions directed Markov with respect to m_0. For each simulated distribution, the asymptotic classification rate (r) and asymptotic expected increase per case in the marginal likelihood score (l) are computed for each of the 543 possible Bayesian networks on four nodes We perform a similar set of simulations on the set of graphs on five nodes. In essence, if there were a systematic improvement of classification rate r with predictive score l, then a search procedure based upon marginal likelihood could be justified.

The plan of the rest of the paper is as follows. In the next section we define the expected classification rate of a distribution, and give a simple proof that if $P(C, A)$ is the generating distribution, then $P(C \mid A)$ is also optimal in terms of expected classification. In Section 3 we relate the asymptotic increase in the marginal likelihood per case of a Bayesian network to its Kullback–Leibler divergence from the generating distribution. In Section 4 classification rates are illustrated for data generated from a particular Bayesian network on 20 nodes. Section 5 presents some essential discussion on Markov equivalent graphs. In Section 6 are presented the description and results of computer simulations on the sets of all four- and five-node labelled directed acyclic graphs (LDAGs). This is followed by the conclusion.

2 An optimal classification distribution

Suppose that a single sample is drawn from the probability distribution $P(C, A)$ having a given configuration $a \in \mathcal{A}$ of the set of variables A. Suppose that for each a we denote by $\hat{c}(a)$ the class for which $p(c \mid a)$ attains a maximum (if this is not unique, choose any such class arbitrarily); then for the sampled configuration the probability that $c = \hat{c}(a)$ will be $p(\hat{c}(a) \mid a)$; this will be the probability of a correct classification given a. We define the *expected classification rate* r_P of the distribution P, obtained on averaging over all attribute configurations a, to be

$$r_P := \sum_a p(\hat{c}(a) \mid a) p(a). \qquad (1)$$

Now suppose that $Q(C, A)$ is an alternative distribution, and let $\tilde{c}(a)$ denote the class for which $q(c \mid a)$ attains a maximum (again breaking ties arbitrarily). Then, given a, the distribution Q will correctly classify the case sampled from P with probability $p(\tilde{c}(a) \mid a) \leq p(\hat{c}(a) \mid a)$. Again, averaging over all a we may define expected classification rate $r_{Q \mid P}$ of Q with respect to P:

$$r_{Q \mid P} := \sum_a p(\tilde{c}(a) \mid a) p(a), \qquad (2)$$

Clearly $r_P \geq r_{Q \mid P}$, and $r_P = r_{P \mid P}$. Thus we have shown that a probability distribution which generates data is also the optimal choice (Bayes act) for classification the data: no other probability distribution can classify better in expectation (though some may do as well.) Put another way, the joint probability distribution *uniquely* chosen as the Bayes act under the log-scoring rule is *simultaneously* an optimal distribution to choose — though not necessarily uniquely so — as the basis for a Bayes act for classification under the $0 - 1$ scoring rule. Note that this result does not require that the distributions P or Q be directed Markov with respect to some pair of graphs; we now consider the case in which this occurs.

3 Optimal classification using Bayesian networks

Let $\mathcal{G}_P$ and $\mathcal{G}_Q$ be Bayesian networks in the discrete random variables $\{C, A\}$ representing directed Markov probability distributions P and Q respectively. Then, as shown in Section 2, if P is the distribution generating the data, $\mathcal{G}_P$ will be optimal for classifying the cases in the database. In Section 5 we give some sufficient conditions under which $\mathcal{G}_Q$ will also be optimal for classifying C given A, but first we introduce the marginal likelihood, and show how it is related to a Kullback–Leibler divergence for asymptotically large samples.

3.1 Asymptotic marginal likelihood and Kullback–Leibler divergence

Suppose that a complete database D_N of N independent cases is generated from $P(C, A)$, i.e. from the $\mathcal{G}_P$ whose structure is assumed unknown. Let $\mathcal{M}$ denote the set of all LDAGs having nodeset $V = \{C\} \cup A$. Each model $m \in \mathcal{M}$ is parameterized by a set of parameters, θ_m say. In the Bayesian approach to model selection, a prior $P(\theta_m \mid m)$ is specified for the parameters, and the marginal likelihood of the data D_N under the model is calculated; it is given by

$$p(D_N \mid m) = \int p(D_N \mid \theta_m, m) p(\theta_m \mid m) d\theta_m. \qquad (3)$$

If all models in $\mathcal{M}$ are considered a-priori equally likely, then $p(m \mid D_N) \propto p(D_N \mid m)$, and so choosing the model having maximum posterior probability becomes equivalent to choosing the model having the highest marginal likelihood.

Now suppose that a new complete case δ is observed. Then the change in the marginal likelihood, given the previous data, D_N, is given by the factor

$$\begin{aligned} p(\delta \mid D_N, m) &= p(\delta, D_N \mid m) / p(D_N \mid m) \\ &= \int p(\delta \mid \theta_m, m) p(\theta_m \mid D_N, m) d\theta_m. \end{aligned}$$

Now if the number of cases N in the dataset D_N is sufficiently large then, under mild conditions on the prior $P(\theta_m \mid m)$, the posterior density $p(\theta_m \mid D_N, m)$ will be very sharply peaked around the maximum likelihood estimate $\hat{\theta}_m(D_N)$, so that to a good approximation

$$p(\delta \mid D_N, m) = p(\delta \mid \hat{\theta}_m(D_N), m). \tag{4}$$

Writing $\hat{\theta}_m = \lim_{N\to\infty} \hat{\theta}_m(D_N)$, we thus obtain the asymptotic additive change in the log-marginal likelihood for the single observation δ to be

$$\log p(\delta \mid \hat{\theta}_m, m). \tag{5}$$

We can interpret (5) as the *logarithmic penalty* that the model m trained on a sufficiently large dataset would obtain for its probability prediction for the next case δ.

Now the data generating distribution $P(C, A)$ will generate δ with probability $p(\delta)$, and hence would itself accrue a logarithmic penalty $\log p(\delta)$. The difference between this penalty and that due to model m in (5), will be $\log p(\delta) - \log p(\delta \mid \hat{\theta}_m, m)$ having expectation

$$\sum_{\delta} p(\delta) \log \frac{p(\delta)}{p(\delta \mid \hat{\theta}_m, m)}, \tag{6}$$

where the sum is over all configurations δ: this expectation is simply the Kullback–Leibler divergence between the two distributions.

Thus we have related the expected change per case in the asymptotic log marginal likelihood of a model to the Kullback–Leibler divergence of its expected distribution from the generating distribution.

Now, it is shown by Cowell (1996) (see also Cowell *et al.* (1999), Theorem 11.1) that the distribution P_m directed Markov with respect the DAG m which minimizes the Kullback–Leibler divergence $K(P, P_m) = \mathrm{E}_P\{\log p(x)/p_m(x)\}$ is the one for which the conditional probabilities $p_m(x \mid \mathrm{pa}(x)^m)$ agree with each $p(x \mid \mathrm{pa}(x)^m)$ calculated from P for every node x in the graph m, (here $\mathrm{pa}(x)^m$ denotes the parents of x in m).

Thus, provided the parameterization of the model m exhibits local meta independence, (Dawid and Lauritzen 1993), we have the following result: *Given a generating distribution $P(X)$ and a DAG m in the discrete random variables X, the distribution $P_m(X)$ directed Markov with respect to m which minimizes the Kullback–Leibler divergence $K(P, P_m) = \mathrm{E}_P\{\log p(x)/p_m(x)\}$ is the same as would be obtained either as the maximum likelihood estimate, or as the posterior expectation of θ_m under mild conditions on the prior $P_m(\theta_m \mid m)$, using an asymptotically large number of independent observations drawn from $P(X)$.*

4 Example: Child

The Child network (Spiegelhalter and Cowell 1992; Spiegelhalter *et al.* 1993; Cowell *et al.* 1999), has twenty nodes: we shall take its Disease node as our classification node; it has six states.

Using the Child network we generated two independent datasets each of 10,000 cases. The first dataset was used to train Child together with the naïve Bayes network (in which all attributes are conditionally independent give the class variable). It was also used to search for and train the best scoring (with respect to log-score) Chow–Liu tree (Chow and Liu 1968) and TAN (Tree Augmented Naïve Bayes) network (Friedman *et al.* 1997). All four trained networks were then used to classify the data in the second data set, cumulating their successes; in addition their joint predictive log-scores were evaluated. The results are shown in Table 1.

We see that the Child network has the best classification rate of 88.56%; given the size of the data set this is probably an accurate estimate of the classification rate for Disease. The TAN network is second best, both in terms of log-score and classification accuracy. This small table suggests a direct link between predictive and classification accuracy. The simulations described in Section 6 indicates that things are not so simple, but before then we need to make some observations concerning Markov equivalence.

Table 1: Classification accuracy and logarithmic scores using data simulated from Child network.

Graphical Model	Log Score	Classification rate
Child	-121838	88.56%
TAN	-123147	87.90%
Chow–Liu Tree	-125112	86.26%
Naïve Bayes	-144072	84.69%

5 Markov equivalent optimal classifiers

Suppose that the generating model m_0 with distribution P were a subgraph of the DAG m. Then one could assign $p_m(x \mid \mathrm{pa}(x)^m) = p(x \mid \mathrm{pa}(x)^{m_0})$ and have $P_m = P$. One can indeed go further, by considering *Markov equivalent graphs*, that is, distinct graphs representing the same conditional independence relations. (Frydenberg (1990) gave conditions for two chain graphs — of which DAGs are a special subclass — to be Markov equivalent.) If m_0^e is any graph Markov equivalent to m_0 and m_0^e is a subgraph of m, then one can assign conditional probabilities $p_m(x \mid \mathrm{pa}(x)^m)$ to each node of m such that the distribution P_m exactly matches that of P. Hence, by the results of Section 3, any such distribution will also be an optimal classifier (for C given A).

We can go even further in identifying optimal classifiers, by considering the Markov blanket $B := \mathrm{bl}(C) \subseteq A$ of the node C in m_0. We have $P(C \mid A) = P(C \mid B)$; that is for classification purposes C is independent of $A \setminus B$, given B. For some sub-configuration $b \in \mathcal{B}$ of the configuration a, we will have there have $r_P = \sum_b p(\hat{c}(b) \mid b)p(b)$.

We have already seen that $r_P \geq r_{Q \mid P}$ for any distribution Q. Suppose that m_0 and m have the same *local structure* $L(C)$ at C defined as follows: in both models the node C has the same set of parents and children, and furthermore that each of node C's children has the same set of parents in both models. Then we can assign to m a distribution Q whose conditional probabilities of the nodes C and each of its children (given their respective parents) have the same values as arise in P. Under these conditions, then whatever assignment is made to the other conditional probabilities to ensure Q is directed Markov with respect to m, Q will be optimal as a classifier for C given A.

Combined with the two previous results, we have the following lemma.

Lemma 1.

Let P be directed Markov with respect to the DAG m_0 having nodes consisting of a class node C and set of attribute nodes A. Let $\mathcal{G}_P^{L(C),e}$ denote the set of all LDAGs which are Markov equivalent to any graph having the same local structure $L(C)$ at C as m_0. Then for any LDAG m which contains some graph of $\mathcal{G}_P^{L(C),e}$ as a subgraph, there exists a distribution Q directed Markov with respect to m, which has the same — and hence optimal — classification rate of C given A as the distribution P.

The characterization given in Lemma 1 is sufficient for optimal classifiers, but it cannot be necessary: indeed it is straightforward to construct some directed Markov distributions P for which there is some optimal classification distribution Q directed Markov over a DAG m where $m \notin \mathcal{G}_P^{L(C),e}$.

6 A simulation study

We ran two sets of similar simulation studies, one using all four node LDAGs, the other using all five node LDAGs. The experiments were similar in nature, so we first describe the experiments on four node networks first in detail, and then more briefly the experiments on five node Bayesian networks.

6.1 The four-node network experiments

For these, the DAG m_0 shown on the left Figure 1 was taken as the generating network in all four-node network simulations. The node C is the classification variable, the remaining nodes being the attributes. Note that all attributes are in the Markov blanket of the class node. All nodes were binary. In each simulation, a probability distribution directed Markov with respect to m_0 was generated at random, by the following method. Each of the conditional probabilities $P_0(C \mid A_1)$ and $P_0(A_3 \mid C, A_2)$ for each parent configuration, together with the unconditional probabilities, $P_0(A_1)$, $P_0(A_2)$, were sampled independently from Beta(1,1) (that is, flat) priors. The product of these simulated distributions defined the joint distribution $P_0(C, A_1, A_2, A_3)$. Next, for each graph $m \in \mathcal{G}_4$, where $\mathcal{G}_4$ is the set of 543 LDAGs on the four nodes C, A_1, A_2, A_3, the Kullback–Leibler projection of P_0 onto m, denoted by $P_m(C, A_1, A_2, A_3)$, was found. It was used to evaluate the Kullback–Leibler divergence $K(P_0, P_m)$ and the expected classification rate $r_{P_m \mid P_0}$ of P_m given P_0 (using (2)). Thus each run generated 543 pairs of numbers. Examples of scatterplots produced by such samples are shown in Figure 2. For each such set of number pairs, Spearman's rank correlation coefficient and Kendall's rank correlation coefficient were evaluated. In all, 3,000 such simulations were performed, leading to 3,000 pairs of values for the rank correlation coefficients. Their values are plotted as histogram density plots in Figure 3.

A similar set of 3,000 simulations was performed using a full subset $\mathcal{G}_4^e \subset \mathcal{G}_4$ of Markov inequivalent LDAGs (numbering 185 in all); histograms of the rank correlation coefficients are also plotted in Figure 3.

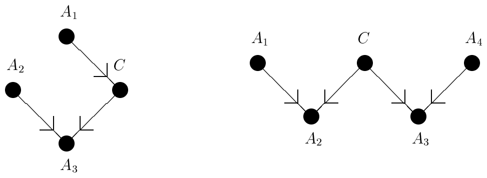

Figure 1: Generating models used in the four node (left) and five node (right) experiments.

6.2 Five node experiments

A similar set of simulations was performed using the set of all LDAGs on five nodes, $\mathcal{G}_5$, of which there are 29,281, and a complete subset $\mathcal{G}_5^e \subset \mathcal{G}_5$ of Markov inequivalent LDAGs on five nodes, there being 8,782 of these. 1,600 simulations were performed on each of these two sets of graphs, the data generating model, which we also call m_0, is on the right in Figure 1. Histogram density plots of the rank correlation coefficients are shown in Figure 4.

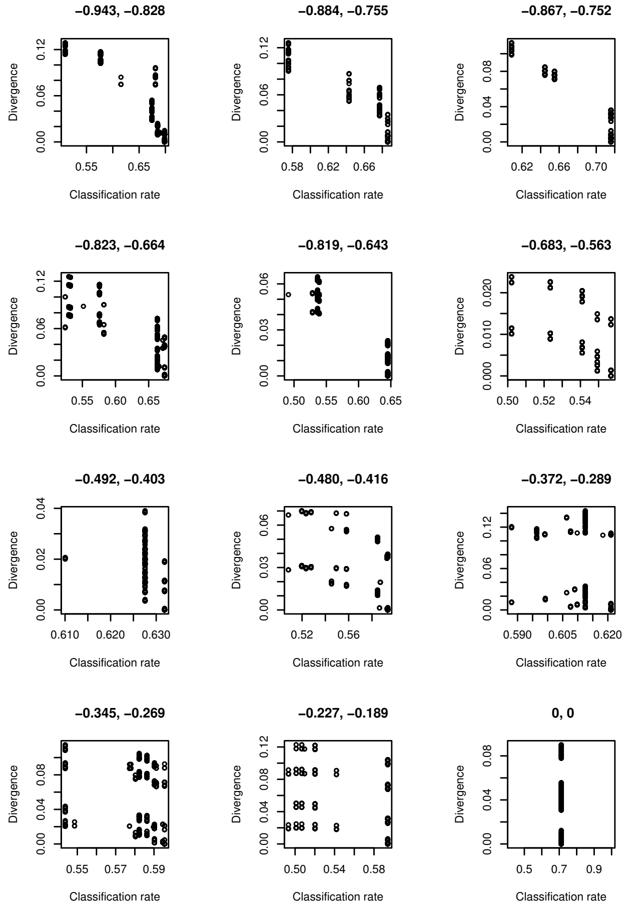

Figure 2: Scatterplots of Kullback–Leibler divergence vs. classification rate from a selection of twelve simulations on the set of all four node LDAGs. The figures above each plot are Spearman's and Kendall's rank correlation coefficients respectively.

6.3 Interpreting the results

We now comment on the plots obtained from the simulation runs. First note that the generating model is simple to locate in each scatterplot — it has zero Kullback–Leibler divergence and the highest possible classification rate, and hence is located at the bottom right in each scatterplot. Now for overall predictive performance to be indicative of good classification performance, low Kullback–Leibler divergence should be strongly associated with high classification. In order to quantify the dependence of classification rate against divergence for each simulation, we calculated Spearman's and Kendall's rank correlation coefficients; density plots in the forms of histograms of these values are shown in Figure 3 for the simulations based on four-node LDAGs, and in Figure 4 for the simulations based on five-node LDAGs, all bin-widths being 0.1.

Note the somewhat anomalous jump in the interval $(-0.1, 0]$. This is due to those simulations in which, with the sampled distribution, all graphs were optimal as classifiers (as in the final plot of Figure 2. (Incidentally, such distributions show that the conditions of Lemma 1 are sufficient but not necessary.) The fractions of such runs are shown in the plots by the horizontal line in the rectangle, with the area above the line representing these situations where all graphs are equally optimal as classifiers.

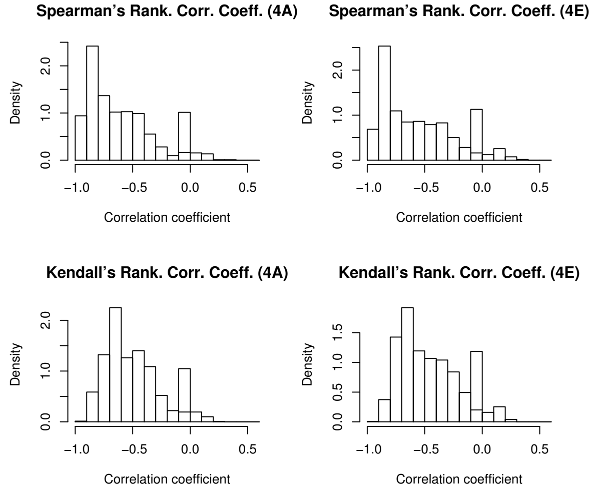

Figure 3: Density plots of Spearman's and Kendall's rank correlation coefficient obtained from pooling each of the 1600 simulations on the set $\mathcal{G}_4$ of all four node LDAGs (4A) and a set $\mathcal{G}_4^e$ all four node Markov inequivalent LDAGs (4E).

From the histograms we see that the mode (in the Spearman's rank plots) is in the range (-0.9,-0.8], with only around 20% of such runs in this range, and only around 30% of runs achieving a correlation of less than -0.8. The scatterplots shown in Figure 2 indicate that a correlation of -0.8 or less is generally required for finding good classifiers using a search based on marginal likelihood. These results indicate that a classifier search based on marginal likelihood may be effective on only around 30% or fewer of occasions. (Similar conclusions can be drawn from the Kendall's rank correlation coefficient histograms.) Even this may be an overestimate, because our analysis has been based on Kullback–Leibler projections of the generating distribution, corresponding to training networks with an infinite amount of data. However, real problems are characterized by finite data sets, sometimes

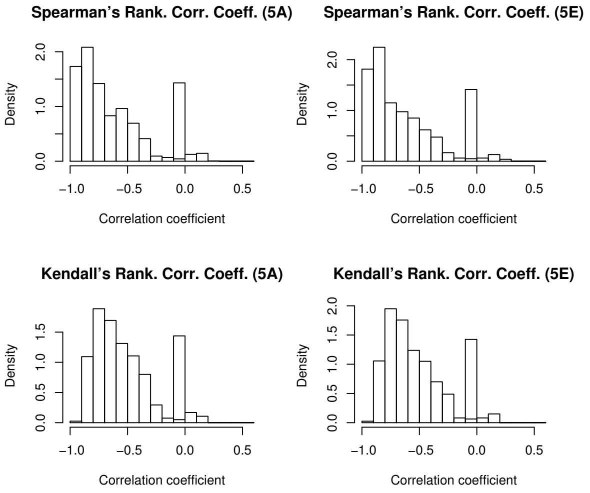

Figure 4: Density plots of Spearman's and Kendall's rank correlation coefficient obtained from pooling each of the 3000 simulations on the set $\mathcal{G}_5$ of all five node LDAGs (5A) and a set $\mathcal{G}_5^e$ all five node Markov inequivalent LDAGs (5E).

quite small, so that the vagaries of sample variation and sensitivity to priors, not to mention problems with missing data, will probably exacerbate the situation. Although we have used small networks, it seems to us likely that our conclusions will hold for networks with many more nodes, especially if in the generating network the Markov blanket of the class node is small.

7 Conclusions

Based on a simulation study on the set of all four node and five node LDAGs, and assuming an infinite training set, we have estimated that finding good or optimal classifiers using a model selection procedure based upon marginal likelihood might be effective in only 30% or fewer occasions, and in some cases could produce very bad classifiers. Thus, although the criterion of maximizing the marginal likelihood can in principle find optimal Bayesian network classifiers, given a sufficiently large amount of data, in practice it may not be the best score to use for guiding the search among Bayesian network classifiers. Our conclusions echo those of Kontkanen *et al.* (1999) and Friedman *et al.* (1997), who based their conclusions on analyses of real data sets.

References

Buntine, W. L. (1994). Operations for learning with graphical models. *Journal of Artificial Intelligence Research*, **2**, 159–225.

Buntine, W. L. (1996). A guide to the literature on learning probabilistic networks from data. *IEEE Transactions on Knowledge and Data Engineering*, **8**, 195–210.

Chow, C. K. and Liu, C. N. (1968). Approximating discrete probability distributions with dependence trees. *IEEE Transactions On Information Theory*, **14**, 462–7.

Cowell, R. G. (1996). On compatible priors for Bayesian networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **18**, 901–11.

Cowell, R. G., Dawid, A. P., Lauritzen, S. L., and Spiegelhalter, D. J. (1999). *Probabilistic Networks and Expert Systems.* Springer.

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

Friedman, N., Geiger, D., and Goldszmidt, M. (1997). Bayesian network classifiers. *Machine Learning*, **29**, 131–63.

Frydenberg, M. (1990). The chain graph Markov property. *Scandinavian Journal of Statistics*, **17**, 333–53.

Heckerman, D. (1998). A tutorial on learning with Bayesian networks. In *Learning in Graphical Models*, (ed. M. I. Jordan), pp. 301–54. Kluwer Academic Publishers, Dordrecht, The Netherlands.

Kontkanen, P., Myllymaki, P., Silander, T., and Tirri, H. (1999). On supervised selection of Bayesian networks. In *Proceedings of the Fifteenth Annual Conference on Uncertainty in Artificial Intelligence (UAI–99)*, pp. 334–42. Morgan Kaufmann Publishers, San Francisco, CA.

Robinson, R. W. (1977). Counting unlabelled acyclic digraphs. In *Lecture Notes in Mathematics: Combinatorial Mathematics V*, (ed. C. H. C. Little). Springer–Verlag, New York.

Sanguesa, R. and Cortes, U. (1997). Learning causal networks from data: a survey and a new algorithm for recovering possibilistic causal networks. *AI Communications*, **10**, 31–61.

Spiegelhalter, D. J. and Cowell, R. G. (1992). Learning in probabilistic expert systems. In *Bayesian Statistics 4*, (ed. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith), pp. 447–65. Clarendon Press, Oxford, United Kingdom.

Spiegelhalter, D. J., Dawid, A. P., Lauritzen, S. L., and Cowell, R. G. (1993). Bayesian analysis in expert systems (with discussion). *Statistical Science*, **8**, 219–83.