Model Selection and Accounting for Model Uncertainty in Graphical Models Using Occam's Window

David MADIGAN and Adrian E. RAFTERY*

We consider the problem of model selection and accounting for model uncertainty in high-dimensional contingency tables, motivated by expert system applications. The approach most used currently is a stepwise strategy guided by tests based on approximate asymptotic P values leading to the selection of a single model; inference is then conditional on the selected model. The sampling properties of such a strategy are complex, and the failure to take account of model uncertainty leads to underestimation of uncertainty about quantities of interest. In principle, a panacea is provided by the standard Bayesian formalism that averages the posterior distributions of the quantity of interest under each of the models, weighted by their posterior model probabilities. Furthermore, this approach is optimal in the sense of maximizing predictive ability. But this has not been used in practice, because computing the posterior model probabilities is hard and the number of models is very large (often greater than 10^{11}). We argue that the standard Bayesian formalism is unsatisfactory and propose an alternative Bayesian approach that, we contend, takes full account of the true model uncertainty by averaging over a much smaller set of models. An efficient search algorithm is developed for finding these models. We consider two classes of graphical models that arise in expert systems: the recursive causal models and the decomposable log-linear models. For each of these, we develop efficient ways of computing exact Bayes factors and hence posterior model probabilities. For the decomposable log-linear models, this is based on properties of chordal graphs and hyper-Markov prior distributions and the resultant calculations can be carried out locally. The end product is an overall strategy for model selection and accounting for model uncertainty that searches efficiently through the very large classes of models involved.

Three examples are given. The first two concern data sets that have been analyzed by several authors in the context of model selection. The third addresses a urological diagnostic problem. In each example, our model averaging approach provides better out-of-sample predictive performance than any single model that might reasonably have been selected.

KEY WORDS: Chordal graph; Contingency table; Decomposable log-linear model; Expert system; Hyper-Markov distribution; Recursive causal model.

1. INTRODUCTION

Fruitful approaches to inference in high-dimensional contingency tables all involve choosing a broad class of models to be considered and then comparing them on the basis of how well they predict the data. Typically, the model classes are huge, and inference in the presence of the many competing models is not easy.

Here we consider two classes of graphical models: the recursive causal models of Kiiveri, Speed, and Carlin (1984) and the decomposable log-linear models introduced by Goodman (1970) and Haberman (1974). This work is motivated by applications in expert systems that use a belief network to represent knowledge and perform inference (Lauritzen and Spiegelhalter 1988). These are the two model classes that arise in such applications. Potentially the most important advantage of constructing expert systems in this fashion is the system's ability to modify itself as data become available. In a series of recent papers, Spiegelhalter and Lauritzen (1990a, 1990b), Dawid and Lauritzen (1993), and Spiegelhalter and Cowell (1991) have addressed the issue of updating the quantitative layer of such models. Building on this work, we address the issue of updating the qualitative layer: How can the graphical structure itself be updated as data becomes available?

Currently, the most commonly used approach to model selection in contingency tables is a stepwise one, adapted from stepwise regression by Goodman (1971); see also Bishop, Fienberg, and Holland (1975, sec. 4.5 and chap. 9). This consists of sequentially adding and deleting terms on the basis of approximate asymptotic likelihood ratio tests, leading to the selection of a single model. Inference about the quantities of interest is then made conditionally on the selected model.

There are several difficulties with this approach. The sampling properties of the overall strategy are complex because it involves multiple tests and, at least implicitly, the comparison of nonnested models (Fenech and Westfall 1988). The use of P values themselves is controversial, even when there are only two models to be compared, because of the so-called "conflict between P values and evidence" discussed by Berger and Sellke (1987) and Berger and Delampady (1987). One aspect of this is that tests based on P values tend to reject even apparently satisfactory models when the sample size is large; a dramatic example of this was discussed by Raftery (1986b). On the other hand, when the sample size is small and the table sparse, the asymptotic approximations on which the P values are based tend to break down.

Perhaps most fundamentally, conditioning on a single selected model ignores model uncertainty and so leads to un-

* David Madigan is Assistant Professor of Statistics and Adrian E. Raftery is Professor of Statistics and Sociology, Department of Statistics, GN-22, University of Washington, Seattle, WA 98195. Madigan's research was partially supported by the Graduate School Research Fund, University of Washington and by the National Science Foundation. Raftery's research was supported by Office of Naval Research Contract N-00014-91-J-1074. The authors are grateful to Gregory Cooper, Leo Goodman, Shelby Haberman, David Hinkley, Graham Upton, Jon Wellner, Nanny Wermuth, Jeremy York, Walter Zucchini, and two anonymous referees for helpful comments and discussions, and to Michael R. Butler for providing the data for the scrotal swellings example.

Journal of the American Statistical Association
December 1994, Vol. 89, No. 428, Theory and Methods

derestimation of the uncertainty about the quantities of interest. This underestimation can be large, as was shown by Regal and Hook (1991) in the contingency table context and by Miller (1984) in the regression context. One bad consequence is that it can lead to decisions that are too risky (Hodges 1987).

In principle, the standard Bayesian formalism provides a panacea for all these difficulties. If Δ is the quantity of interest, such as a parameter, a future observation, or the utility of a course of action, then its posterior distribution, given data D, is

$$\mathrm{pr}(\Delta \mid D) = \sum_{k=1}^{K} \mathrm{pr}(\Delta \mid M_k, D)\mathrm{pr}(M_k \mid D). \qquad (1)$$

This is an average of the posterior distributions under each of the models, weighted by their posterior model probabilities. In Equation (1), $M_1, \ldots, M_K$ are the models considered and

$$\mathrm{pr}(M_k \mid D) = \frac{\mathrm{pr}(D \mid M_k)\mathrm{pr}(M_k)}{\sum_{l=1}^{K} \mathrm{pr}(D \mid M_l)\mathrm{pr}(M_l)}, \qquad (2)$$

where

$$\mathrm{pr}(D \mid M_k) = \int \mathrm{pr}(D \mid \boldsymbol{\theta}_k, M_k)\mathrm{pr}(\boldsymbol{\theta}_k \mid M_k)\, d\boldsymbol{\theta}_k \qquad (3)$$

is the marginal likelihood of model M_k, $\boldsymbol{\theta}_k$ is the (vector) parameter of M_k, $\mathrm{pr}(\boldsymbol{\theta}_k \mid M_k)$ is the prior distribution of $\boldsymbol{\theta}_k$, $\mathrm{pr}(D \mid \boldsymbol{\theta}_k, M_k)$ is the likelihood, and $\mathrm{pr}(M_k)$ is the prior probability of M_k.

Furthermore, averaging over *all* the models in this fashion provides better predictive ability, as measured by a logarithmic scoring rule, than using any single model M_j:

$$-E\left[\log\left\{\sum_{k=1}^{K} \mathrm{pr}(\Delta \mid M_k, D)\mathrm{pr}(M_k \mid D)\right\}\right] \le -E[\log\{\mathrm{pr}(\Delta \mid M_j, D)\}] \qquad (j = 1, \ldots, K), \qquad (4)$$

where Δ is the observable to be predicted and the expectation is with respect to $\sum_{k=1}^{K} \mathrm{pr}(\Delta \mid M_k, D)\mathrm{pr}(M_k \mid D)$. This follows from the nonnegativity of the Kullback–Leibler information divergence. The logarithmic scoring rule suggested by Good (1952) assigns to each event A that occurs a score of $-\log\{\mathrm{pr}(A)\}$. (See Dawid 1986 for further discussion and Kass and Raftery 1995 for a review of the general approach.)

Cooper and Herskovits (1992) presented this approach in the context of recursive causal models. But the approach in general has not been adopted in practice. This appears to be because (a) the posterior model probabilities $\mathrm{pr}(M_k \mid D)$ are hard to compute, because they involve the very high-dimensional integrals in Equation (3), and (b) the number of models in the sum in Equation (1) can be huge. For example, with just 10 variables (small by expert system standards), there are approximately 4×10^{18} recursive causal models and 2×10^{11} decomposable models.

One might hope that most of the posterior probability would be accounted for by a small number of models so that the sum in Equation (1) would be well approximated by a small number of terms. Unfortunately, this is not typically the case, because, although a small number of models do have much higher posterior probabilities than all the others, the very many models with small posterior probabilities contribute substantially to the sum. For example, Moulton (1991) reported a regression example with $2^{12} = 4{,}096$ models where about 800 models were needed to account for 90% of the posterior probability.

We argue that the standard Bayesian formalism of Equation (1) is flawed. Adopting standard methods of scientific investigation, we contend that accounting for the true model uncertainty involves averaging over a much smaller set of models. We present simple and efficient ways of computing the exact posterior model probabilities for the two model classes considered. Our approach is to take advantage of the graphical structure to calculate the required probabilities very quickly, while representing prior opinion in an easily elicitable form. We also describe an efficient algorithm for searching the very large model space.

Putting all this together gives us a simple and computationally efficient way of selecting the best models and accounting for model uncertainty in recursive causal models and decomposable log-linear models. To demonstrate the generality of our approach, our discussion will be in the context of conventional statistical model selection rather than expert systems. In Section 2 we describe the principles underlying our approach to model selection. In Section 3 we apply those principles to the recursive causal models, and in Section 4 we consider the decomposable models.

2. MODEL SELECTION STRATEGY

2.1 General Principles and Occam's Razor

We argue that Equation (1) does not accurately represent model uncertainty. Science is an iterative process in which competing models of reality are compared on the basis of how well they predict what is observed; models that predict much less well than their competitors are discarded. Most of the models in Equation (1) have been discredited in the sense that they predict the data far less well than the best models and so they should be discarded. Hence they should not be included in Equation (1).

In our approach, if a model predicts the data far less well than the best model in the class, it will be discarded, so that initially we exclude from Equation (1) those models not belonging to the set

$$\mathcal{A}' = \left\{ M_k : \frac{\max_l \{\mathrm{pr}(M_l \mid D)\}}{\mathrm{pr}(M_k \mid D)} \le c \right\}, \qquad (5)$$

for some constant c. The value of c used will depend on the context. In our examples we used $c = 20$, by analogy with the popular .05 cutoff for P values; Jeffreys (1961, app. B) would suggest some number between 10 and 100, whereas Evett (1991) suggested a value of 1,000 for forensic evidence in criminal cases. Note that we use $\mathrm{pr}(M_k \mid D)$ rather than $\mathrm{pr}(D \mid M_k)$ as the measure of how well the model predicts the data. In this way the likelihood is weighted by the prior model probability $p(M_k)$, assumed to reflect past data. This results in a composite predictive probability for both past and present data.

Next we appeal to one of the most widely accepted norms of scientific investigation, Occam's razor. Let E represent the evidence and let $\mathrm{pr}(H|E)$ represent the probability of a specified hypothesis H given the evidence E. Occam's razor states that if

$$\mathrm{pr}(H_1|E) = \mathrm{pr}(H_2|E) = \cdots = \mathrm{pr}(H_k|E)$$

for hypotheses $H_1, \ldots, H_k$, then the simplest among $H_1, \ldots, H_k$ is to be preferred (Kotz and Johnson 1985). Thus we also exclude from Equation (1) models belonging to the set

$$\mathcal{B} = \left\{ M_k : \exists\, M_l \in \mathcal{A}', M_l \subset M_k, \frac{\mathrm{pr}(M_l|D)}{\mathrm{pr}(M_k|D)} > 1 \right\}, \quad (6)$$

and Equation (1) is replaced by

$$\mathrm{pr}(\Delta|D, \mathcal{A}) = \sum_{M_k \in \mathcal{A}} \mathrm{pr}(\Delta|M_k, D)\mathrm{pr}(M_k|D, \mathcal{A}), \quad (7)$$

where

$$\mathcal{A} = \mathcal{A}' \backslash \mathcal{B}.$$

This reduces considerably the number of models in the sum in Equation (1) and hence greatly simplifies the model uncertainty problem. Note that our argument is *not* an approximation adopted for computational convenience, but rather a solution based on accepted scientific methodology. Also, note also that our approach in Equation (7) will not necessarily give an answer close to that given by Equation (1) because, due to the very large number of models in the class, the models discarded may have a large total posterior probability $\sum_{M_k \notin \mathcal{A}} \mathrm{pr}(M_k|D)$, even though each individual model discarded has a very small posterior probability. Similarly, excluding the models not in the set $\mathcal{A}$ may result in violations of the inequality (4). But experience to date suggests that this will rarely happen and that averaging over the models in $\mathcal{A}$ provides better predictive performance than conditioning on a single model. See Section 5 for further discussion.

The problem thus reduces to finding the set $\mathcal{A}$. We now outline a computational strategy for doing this.

2.2 Model Selection Strategy

Our approach is heuristic in nature and is a variant of the greedy-search algorithm. The essentials of the approach are the same for the recursive causal models and the decomposable models and could be readily applied to more general graphical models. Posterior model probabilities are used as a metric to guide the search. The strategy proceeds out into model space away from the opening set of models, comparing models via ratios of posterior model probabilities in a series of nested comparisons. In what follows, M_0 will denote the smaller of the two models being compared and M_1 will denote the larger. In fact, M_0 and M_1 will differ by just one link throughout.

Our basic rule is that if M_0 is rejected, then so are all its submodels. Here we define M to be a submodel of M_0 if all the links in M are also in M_0. To see this, consider the (undirected) example in Figure 1. Suppose that we start with the saturated model $[ABC]$ of Figure 1a, and that when we compare it with the model of conditional independence $[AC][BC]$ of Figure 1b, we reject the smaller model decisively. Then we are precisely rejecting the conditional independence of A and B given C. This conditional independence also holds in all the submodels of $[AC][BC]$, and so we reject all of those as well, including the model $[A][BC]$ of Figure 1c. Thus if we reject a model, then we reject all its submodels. In the algorithm described in Section 2.3, even this rule is relaxed in the sense that $[A][BC]$ may be subsequently considered as a submodel of a different model.

Our basic rule is the first of two "coherence" rules proposed by Gabriel (1969) for sequential testing procedures based on monotone test statistics. His second rule was that if a model is not rejected, then no model that includes it is considered rejected; but this second rule does not apply here because posterior model probabilities are not monotone (i.e., unlike deviance, for example, the posterior probability for a particular model can be smaller than the posterior probability of its submodels). The model selection strategy of Edwards and Havránek (1985) is based on both these rules, whereas that of Havránek (1984) is based on the first rule alone.

2.3 Occam's Window

A crucial aspect of the strategy concerns the interpretation of the ratio of posterior model probabilities when comparing two models. Again we appeal to Occam's razor, which we implement as follows:

1. If the log posterior odds are positive (i.e., the data provide evidence for the smaller model), then we reject M_1 and consider M_0. We could generalize this by requiring the log posterior odds to be greater than some positive constant O_R before rejecting M_1.
2. If the log posterior odds is small and negative, providing evidence against the smaller model that is not very strong, then we consider both models.
3. If the log posterior odds is large and negative (i.e., smaller than $O_L = -\log(c)$ where c is defined by equation (5)), then we reject M_0 and consider M_1.

Thus there are three possible actions following each comparison; see Figure 2.

Now that the various elements of the strategy are in place, we outline the search technique. The search can proceed in two directions: "up" from each starting model by adding links, or "down" from each starting model by dropping links. When starting from a nonsaturated, nonempty model, we

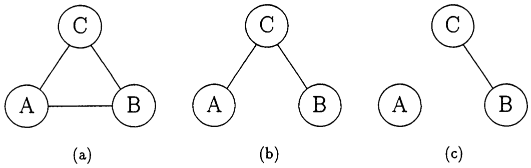

Figure 1. Model Selection Strategy—A Simple Example.

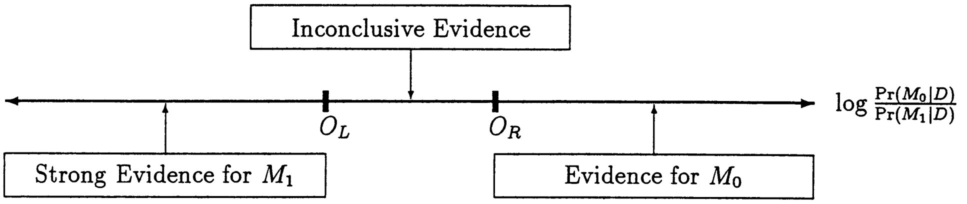
Inconclusive Evidence
$\log \frac{\Pr(M_0|D)}{\Pr(M_1|D)}$
O_L
O_R
Strong Evidence for M_1
Evidence for M_0

Figure 2. Occam's Window: Interpreting the Log Posterior Odds, $\log[Pr(M_0|D)/Pr(M_1|D)]$, Where M_0 is a Submodel of M_1.

first execute the "down" algorithm. Then we execute the "up" algorithm, using the models from the "down" algorithm as a starting point. Experience to date suggests that the ordering of these operations has little impact on the final set of models. Let $\mathcal{A}$ and $\mathcal{C}$ be subsets of model space $\mathcal{M}$, where $\mathcal{A}$ denotes the set of "acceptable" models and $\mathcal{C}$ denotes the models under consideration. For both algorithms, we begin with $\mathcal{A} = \emptyset$ and $\mathcal{C}$ = set of starting models.

Down Algorithm.

1. Select a model M from $\mathcal{C}$.
2. $\mathcal{C} \leftarrow \mathcal{C} \setminus \{M\}$ and $\mathcal{A} \leftarrow \mathcal{A} \cup \{M\}$.
3. Select a submodel M_0 of M by removing a link from M.
4. Compute $B = \log\{[\mathrm{pr}(M_0|D)]/[\mathrm{pr}(M|D)]\}$.
5. If $B > O_R$, then $\mathcal{A} \leftarrow \mathcal{A} \setminus \{M\}$ and if $M_0 \notin \mathcal{C}$, then $\mathcal{C} \leftarrow \mathcal{C} \cup \{M_0\}$.
6. If $O_L \le B \le O_R$, then, if $M_0 \notin \mathcal{C}$, $\mathcal{C} \leftarrow \mathcal{C} \cup \{M_0\}$.
7. If there are more submodels of M, then go to 3.
8. If $\mathcal{C} \neq \emptyset$, then go to 1.

Up Algorithm.

1. Select a model M from $\mathcal{C}$.
2. $\mathcal{C} \leftarrow \mathcal{C} \setminus \{M\}$ and $\mathcal{A} \leftarrow \mathcal{A} \cup \{M\}$.
3. Select a supermodel M_1 of M by adding a link to M.
4. Compute $B = \log\{[\mathrm{pr}(M|D)]/[\mathrm{pr}(M_1|D)]\}$.
5. If $B < O_L$, then $\mathcal{A} \leftarrow \mathcal{A} \setminus \{M\}$; if $M_1 \notin \mathcal{C}$, then $\mathcal{C} \leftarrow \mathcal{C} \cup \{M_1\}$.
6. If $O_L \le B \le O_R$, then if $M_1 \notin \mathcal{C}$, $\mathcal{C} \leftarrow \mathcal{C} \cup \{M_1\}$.
7. If there are more supermodels of M, then go to 3.
8. If $\mathcal{C} \neq \emptyset$, then go to 1.

On termination, $\mathcal{A}$ contains the set of potentially acceptable models. Finally, we remove all the models that satisfy Equation (6), where 1 is replaced by $\exp(O_R)$, and those models M_k for which

$$\frac{\max_l\{\mathrm{pr}(M_l|D)\}}{\mathrm{pr}(M_k|D)} > c. \tag{8}$$

The set $\mathcal{A}$ now contains the acceptable models.

3. THE DIRECTED CASE—RECURSIVE CAUSAL MODEL SELECTION

3.1 Implementation

Implementation for the recursive causal models proceeds in a straightforward fashion. Consider a recursive causal model for a set of random variables X_v, $v \in V$. The model is represented by a directed graph where each variable in V is represented by a node in the graph. For each variable $v \in V$, we define $\mathrm{pa}(v)$ to be the set of parent nodes of v; that is, nodes w for which there exists a directed link from w to v. The assumptions of the model imply that the joint distribution of X_v, $v \in V$, which we denote by $\mathrm{pr}(V)$, is given by

$$\mathrm{pr}(V) = \prod_{v \in V} \mathrm{pr}(v|\mathrm{pa}(v)).$$

Here, and in what follows, we are using node labels like v to represent the random variables which correspond to the nodes. In early implementations, $\mathrm{pr}(v|\mathrm{pa}(v))$ was assumed to be fully specified for all v by the expert/data analyst. Spiegelhalter and Lauritzen (1990a) introduced a parameterization for $\mathrm{pr}(v|\mathrm{pa}(v))$ whereby the relationship between a node v and its parents, $\mathrm{pa}(v)$, is fully specified by a vector parameter $\boldsymbol{\theta}_v \in \Theta_v$. This leads to a conditional distribution for V,

$$\mathrm{pr}(V|\boldsymbol{\theta}) = \prod_{v \in V} \mathrm{pr}(v|\mathrm{pa}(v), \boldsymbol{\theta}_v),$$

where $\boldsymbol{\theta}$ is a general parameter with components $\boldsymbol{\theta}_v$.

Spiegelhalter and Lauritzen (1990a) made two key assumptions that greatly simplify subsequent analysis. The first assumption was that of *global independence,* whereby the parameters $\boldsymbol{\theta}_v$ are assumed mutually independent a priori. This assumption alone allows us to calculate the likelihood for a single case:

$$\begin{aligned}\mathrm{pr}(V) &= \int \mathrm{pr}(V, \theta)\, d\boldsymbol{\theta} \\ &= \int \prod_{v} \mathrm{pr}(v|\mathrm{pa}(v), \boldsymbol{\theta}_v)\mathrm{pr}(\boldsymbol{\theta}_v)\, d\boldsymbol{\theta}_v \\ &= \prod_{v} \mathrm{pr}(v|\mathrm{pa}(v)),\end{aligned}$$

where

$$\mathrm{pr}(v|\mathrm{pa}(v)) = \int \mathrm{pr}(v|\mathrm{pa}(v), \boldsymbol{\theta}_v)\mathrm{pr}(\boldsymbol{\theta}_v)\, d\boldsymbol{\theta}_v.$$

The second assumption was that of *local independence,* whereby the parameter $\boldsymbol{\theta}_v$ breaks into components corresponding to the levels of the factors in $\mathrm{pa}(v)$. These components are assumed to be mutually independent a priori.

Now consider a conditional probability distribution $\mathrm{pr}(v|\mathrm{pa}(v)^+, \boldsymbol{\theta}_v^+) = \boldsymbol{\theta}_v^+$ for a specific set of levels, $\mathrm{pa}(v)^+$, of $\mathrm{pa}(v)$. We assume that $\boldsymbol{\theta}_v^+$ has a Dirichlet distribution $\mathcal{D}[\lambda_1^+, \ldots, \lambda_k^+]$, where k is the number of levels of v. Then we can show that

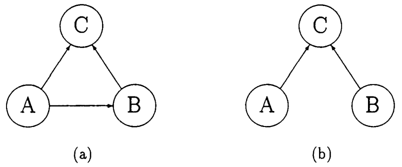

Figure 3. Simplifying the Likelihood Computations.

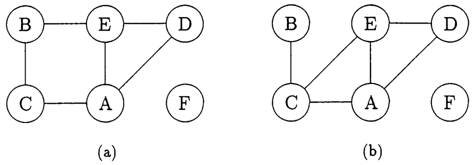

Figure 4. Models Selected by Edwards and Havránek.

$$\mathrm{pr}(v = j \mid \mathrm{pa}(v)^{+}) = \lambda_j^{+} \Big/ \sum_i \lambda_i^{+}, \qquad j = 1, 2, \ldots, k.$$

If we observe v to be at level x_{v_j} and the parent state to be $\mathrm{pa}(v)^{+}$, then we have

$$\theta_v^{+} \mid v \sim \mathscr{D}[\lambda_1^{+}, \ldots, \lambda_j^{+} + 1, \ldots, \lambda_k^{+}].$$

This provides a method for *sequentially* calculating the required ratios of posterior model probabilities and is simpler than the nonsequential approach. Furthermore, the sequential approach allows for efficient incorporation of new evidence. The elicitation of the required Dirichlet priors is feasible provided the cardinality of $\mathrm{pa}(v)$ is not too large. Computer-based methods for eliciting Dirichlet prior distributions have been described by Chaloner and Duncan (1987). If $\mathrm{pa}(v)$ is not observed, the updating becomes more complex (see Spiegelhalter and Lauritzen 1990a,b for details). A Jeffreys's prior density was used in the examples; that is, $\lambda_i^{+} = .5$, for $i = 1, 2, \ldots, k$. A uniform prior typically selects identical models.

A considerable computational saving is obtained by noting that the sequential updating of the distribution of θ_v depends on the levels of v and $\mathrm{pa}(v)$ *only*. Therefore, the likelihood for all qualitative layers (graphs) having the same set $\mathrm{pa}(v)$ of parent nodes of v will have identical contributions from v. For example, consider the two recursive causal models of Figure 3. When calculating the likelihood for the model of Figure 3a, we store the likelihood of each node/parent combination separately. Now when subsequently calculating the likelihood for the model of Figure 3b, only the likelihood for node B requires recalculation, as the sets of parent nodes of A and C have not changed.

To implement the model selection strategy described in Section 2 for the recursive causal models, an *ordering* of the nodes must be prespecified by the expert/data analyst. If v_i precedes v_j in the ordering, then a directed link from v_j to v_i is prohibited. In certain applications it may be possible to search over all possible orderings, but this typically will not be the case. Pearl's IC algorithm (Pearl and Verma 1991) induces directed "causal" structures from data. An ordering of the nodes is not required, but for each pair of nodes v_i and v_j, the algorithm does involve searching among all subsets of $V \backslash \{v_i, v_j\}$ for cutsets between v_i and v_j (sets that when conditioned on, render v_i and v_j independent.) Cooper and Herskovits (1992) provided a review of other approaches.

3.2 Examples

3.2.1 Coronary Heart Disease Risk Factors. First, we consider a data set that has been previously analyzed by Edwards and Havránek (1985). The data concern 1,841 men cross-classified according to six coronary heart disease risk factors: A, smoking; B, strenuous mental work; C, strenuous physical work; D, systolic blood pressure; E, ratio of β and α proteins; and F, family anamnesis of coronary heart disease. Their likelihood ratio–based model selection strategy selected two graphical log-linear models: $[AC][ADE][BC][BE][F]$, which is not decomposable and thus is not equivalent to any recursive causal model, and $[ACE][ADE][BC][F]$, which is decomposable. A striking feature of both models is the independence of F, family anamnesis. The models are shown in Figure 4.

To implement the Bayesian graphical model selection procedure, we started from the saturated model and used the "down" algorithm only. (Starting from the empty model and using the "up" algorithm produced the same set of models.) All qualitative structures were assumed equally likely a *priori.* A natural partial ordering of the variables suggests itself: F, (B, C), A, (E, D). 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. The four corresponding complete orderings produced strong evidence for the precedence of E over D and weak evidence for the precedence of C over B. Several further orderings were tried, but this "natural" ordering resulted in the models with highest posterior probabilities. The selected models are shown in Figure 5, and their posterior probabilities are listed in Table 1.

The two most likely models are shown in Figure 5, a and b. They are rather similar in that both contain the CB, CA, AE, ED, and AD links. The main difference between them lies in the way that they describe the effect of strenuous mental work (B) and strenuous physical work (C) on the ratio

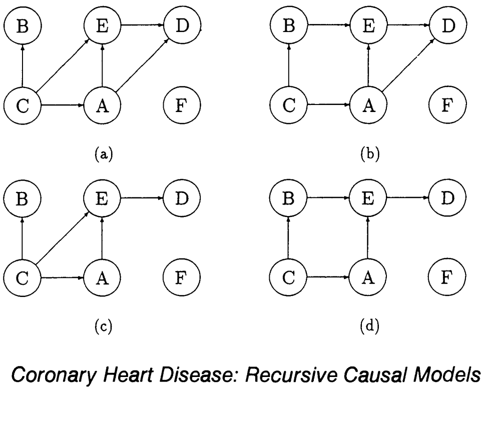

Figure 5. Coronary Heart Disease: Recursive Causal Models Selected.

Table 1. Coronary Heart Disease: Posterior Model Probabilities for Recursive Causal Models

Figure	Posterior probability %
5(a)	52
5(b)	40
5(c)	5
5(d)	4

of β and α proteins (E). Figure 5a says that C affects E both directly and indirectly via A, whereas Figure 5b says that the effect of C on E is solely indirect, being mediated by B and A. There is also some uncertainty about the presence of a link from smoking (A) to systolic blood pressure (D). The evidence favors the marginal independence of family anamnesis of coronary heart disease (F).

The four models selected are similar to the models selected by Edwards and Havránek (1985) and shown in Figure 4. We note that the AD link (smoking and systolic blood pressure) is present in both of the models of Figure 4 and also in the models of Figure 5, a and b, but it is absent from the models of Figure 5, c and d. In fact, the exact test for zero partial association of A and D reported by Edwards and Havránek (1985) had a significance level of .04, which was the largest of any of the links whose absence was rejected at the 5% level.

3.2.2 Women and Mathematics. Our second example concerns a survey reported by Fowlkes, Freeny, and Landwehr (1988) concerning the attitudes of New Jersey high school students toward mathematics. The data were further analyzed by Upton (1991). A total of 1,190 students in eight schools took part in the survey. Data on six dichotomous variables were collected:

A. Lecture attendance: attended or did not attend
B. Sex: female or male
C. School type: suburban or urban
D. "I'll need mathematics in my future work": agree or disagree
E. Subject preference: math/science or liberal arts
F. Future plans: college or job.

Upton (1991) reported that a model selection procedure based on the Akaike Information Criterion (AIC) (Akaike 1973) selects $[ABCE][CDF][BCD][DEF]$, whereas a procedure based on the BIC criterion (Raftery 1986a) selects the much simpler $[A][BE][CE][CF][BD][DE][DF]$. Clearly, an important difference between these two models is the treatment of A.

The Bayesian graphical model selection procedure started from the empty model and used the "up" algorithm. It is clear that B (sex) cannot be influenced by other variables and must be exogenous. Initially, it was also assumed that C (school type) was exogenous. An exhaustive search over all consequent orderings produced the single model shown in Figure 6.

The selected model is similar to the model selected by Upton's BIC procedure. The model selected by AIC clearly overfits the data (Upton 1991). It is of interest to note the direction of the link from D to F. Both Upton (1991) and Fowlkes et al. (1988) treated D as a response variable, and Upton's path diagram shows a directed link from F to D. But the data provide strong evidence that the direction of the influence is from D to F; that is, that students' attitudes toward mathematics influence their future plans, rather than the other way around. The ability of the selected model to predict is unaffected by the direction of the ED link.

Further analysis removed the restriction that C be exogenous. The data now provide some support for the presence of a link from E to C, although its interpretation is somewhat unclear.

4. THE UNDIRECTED CASE—DECOMPOSABLE MODEL SELECTION

4.1 Implementation

To implement the strategy for the decomposable models, we rely heavily on a recent fundamental paper by Dawid and Lauritzen (1993), hereafter denoted by DL. We consider three issues specific to model selection for the decomposable models.

First, how should we add and remove links while efficiently ensuring that all the models created are decomposable? Here we use a result that follows from Lemma 3 of Frydenberg and Lauritzen (1989): Let $\mathcal{G} = (V, E)$ be a chordal graph with vertices V and edges E and let $\mathcal{G}' = (V, E')$ be a chordal subgraph of $\mathcal{G}$ with exactly one edge, e, less. Then e is contained in exactly one clique of $\mathcal{G}$. Therefore, the model selection strategy must remove only links that are members of a single clique. When adding links, the strategy must not create any chordless four cycles.

Second, given any two decomposable models M and M^*, is it possible to generate M^* from M, adding or removing only one edge at a time but staying within the class of decomposable models? Lemma 5 of Frydenberg and Lauritzen (1989) shows that it is.

Finally, how do we calculate the required posterior model probabilities? Following DL, we consider a decomposable model M for a set of random variables X_v, $v \in V$, whose joint distribution is specified by a vector parameter, $\boldsymbol{\theta}$. $\boldsymbol{\theta}$, in turn, is determined by the clique marginal probability tables $\theta_C = (\theta_C)_{C\in\mathcal{C}}$, where $\mathcal{C}$ denotes the set of cliques of M:

$$\theta(i) = \frac{\prod_{C\in\mathcal{C}} \theta_C(i_C)}{\prod_{S\in\mathcal{S}} \theta_S(i_S)}, \qquad i \in \mathcal{I},$$

where $\mathcal{S}$ denotes the system of separators in an arbitrary perfect ordering of $\mathcal{C}$ and $\mathcal{I}$ denotes the set of possible configurations of X.

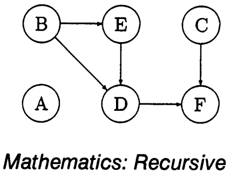

Figure 6. Women and Mathematics: Recursive Causal Model Selected.

For each clique $C \in \mathcal{C}$, let

$$\lambda_C = (\lambda_C(i_C))_{i_C \in \mathcal{I}_C}$$

be a given table of arbitrary positive numbers and let $\mathcal{D}(\lambda_C)$ denote the Dirichlet distribution for $\boldsymbol{\theta}_C$ with density

$$\pi(\boldsymbol{\theta}_C | \lambda_C) \propto \prod_{i_C \in \mathcal{I}_C} \theta_C(i_C)^{\lambda_C(i_C)-1},$$

where $\sum_{i_C} \theta_C(i_C) = 1$ and $\theta(i_C) > 0$.

Now let us suppose that the collection of specifications $\mathcal{D}(\lambda_C)$, $C \in \mathcal{C}$ is constructed in such a way that for any two cliques C and D in $\mathcal{C}$ we have

$$\lambda_C(i_{C\cap D}) = \lambda_D(i_{C\cap D}).$$

Then DL showed that there exists a unique *strong hyper-Markov* distribution for $\boldsymbol{\theta}$ over M that has density $\mathcal{D}(\lambda_C)$ for all $C \in \mathcal{C}$. DL called this the hyper-Dirichlet distribution for $\boldsymbol{\theta}$. A distribution for $\boldsymbol{\theta}$ is strong hyper-Markov if and only if $\boldsymbol{\theta}_{A|B}$, $\boldsymbol{\theta}_{B|A}$, and $\boldsymbol{\theta}_{A\cap B}$ are mutually independent whenever $A \cap B$ is complete and separates A from B. It follows that by letting $\lambda_0 = \sum_{i \in \mathcal{I}} \lambda_i$, the likelihood for a single case is given by

$$\text{pr}(v) = \frac{\prod_{C \in \mathcal{C}} \lambda_C}{\lambda_0(\prod_{S \in \mathcal{S}} \lambda_S)}.$$

From Proposition 1, we have that updating can be carried out one clique at a time.

Proposition 1. If the prior distribution $\mathcal{L}(\boldsymbol{\theta})$ is strong hyper-Markov, then the posterior distribution of $\boldsymbol{\theta}$ is the unique hyper-Markov distribution $\mathcal{L}^*$ specified by the clique-marginal distributions $\{\mathcal{L}^*_C : C \in \mathcal{C}\}$, where $\mathcal{L}^*_C$ is the posterior distribution of $\boldsymbol{\theta}_C$ based on its prior distribution $\mathcal{L}_C$ and the clique-specific data $X_C = x_C$.

Proof. This is corollary 9 of DL.

The posterior distribution for $\boldsymbol{\theta}_C$, given data n_C from the marginal table corresponding to clique C, is $\mathcal{D}(\lambda_C + n_C)$.

Consider the Bayes factor

$$B_{01} = \frac{\text{pr}(D|M_0)}{\text{pr}(D|M_1)},$$

where M_0 and M_1 are decomposable and M_0 is obtained from M_1 by deleting one edge e linking u with v. Because both models are decomposable, we have that e is contained in a single clique, C say, of M_1. Let $C_u = C \setminus \{v\}$, $C_v = C \setminus \{u\}$, and $C_0 = C \setminus \{u, v\}$. Then DL showed that the Bayes factor is given by

$$B_{01} = \frac{p_{C_u}(D_{C_u}) p_{C_v}(D_{C_v})}{p_{C_0}(D_{C_0}) p_C(D_C)}.$$

Thus the required decomposable model comparisons can be carried out very rapidly with calculations local to single cliques.

4.2 Examples

4.2.1 Coronary Heart Disease Risk Factors. First, we consider again the coronary heart disease risk factor data of Edwards and Havránek (1985). We note that the model of Figure 4a, which was selected by the Edwards and Havránek procedure, is not decomposable and hence will not be selected by our procedure.

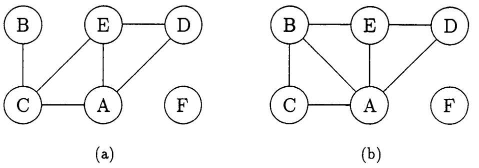

Figure 7. Coronary Heart Disease: Decomposable Models Selected.

The selection procedure started from the saturated model and used the "down" algorithm. All qualitative structures were assumed equally likely a priori. A standard Jeffreys prior was adopted for $\boldsymbol{\theta}_C$, $C \in \mathcal{C}$. Just two models were selected; they are shown in Figure 7. Starting from the empty model and using the "up" algorithm resulted in the same two models. The corresponding posterior probabilities are shown in Table 2.

The model of Figure 7a was also selected by the directed model selection procedure and by Edwards and Havránek (1985). The model of Figure 7b is essentially a decomposable version of the directed model of Figure 5b and Edwards and Havránek's model of Figure 4a.

Overall, the model selection exercise indicates that there is very strong evidence for the *BC*, *AC*, *AE*, *DE* and *AD* links. There is also some evidence for the *CE* and *BE* links, but it seems that one of these alone is enough to describe the data, and it is not fully clear which one is better. Of course the interpretation of these links is different in the two model classes. Again, as in the directed case, there is evidence for the marginal independence of F.

4.2.2 Women and Mathematics. We consider again the survey data previously analyzed by Fowlkes et al. (1988) and Upton (1991). We note that the models selected by Upton (1991) are not graphical and hence will not be selected by our procedure. The procedure adopted was identical to that adopted for the example of Section 4.2.1. The two models selected are shown in Figure 8, and the corresponding posterior probabilities are listed in Table 3.

As in the directed case, the selected models are close to the models selected by the BIC model selection procedure carried out by Upton (1991). But there is uncertainty about the *CD* link (school type and "I'll need mathematics in my future work"), which is not apparent in Upton's analysis. The odds in favor of the inclusion of the *CD* link are 3 to

Table 2. Coronary Heart Disease: Posterior Model Probabilities for Decomposable Models

Figure	*Model*	*Posterior probability %*
7(a)	[*BC*][*ACE*][*ADE*][*F*]	92
7(b)	[*ABC*][*ABE*][*ADE*][*F*]	8

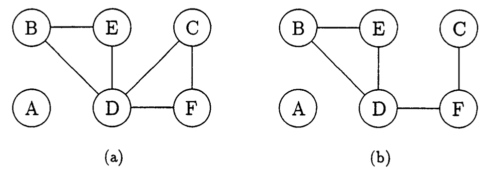

Figure 8. Women and Mathematics: Decomposable Models Selected.

1, which Jeffreys (1961, p. 432) would call evidence "not worth more than a bare mention." The data strongly supports the marginal independence of A.

4.2.3 Scrotal Swellings. Our final example concerns the diagnosis of scrotal swellings. Data on 299 patients were gathered at the Meath Hospital, Dublin, Ireland, under the supervision of Mr. Michael R. Butler. We consider a cross-classification of the patients according to one disease class, hernia (H), and seven binary indicants as follows: A, possible to get above the swelling; B, swelling transilluminates; C, swelling separate from testes; D, positive Valsalva/stand test; E, tenderness; F, pain; G, evidence of other urinary tract infections. The data are shown in Table 4. There are 28 possible links to be considered by the selection procedure in this example. In the absence of prior expert opinion, computation times can be prohibitive. Clearly, if the starting point for the selection procedure were close to the models for which the data provides evidence, then this problem could be overcome.

With this objective, we adopted the following heuristic procedure. First, Bayes factors for each of the 28 links are calculated by comparing the saturated model with the 28 submodels generated by removing single links. The model consisting of the links for which the data provides evidence in this manner is then used as a starting point for the selection procedure. If this model is not decomposable, some of the links may be removed or additional ones may be added. A similar approach was suggested by Goodman (1973). The starting model is shown in Figure 9.

Now the "up" algorithm is executed, followed by the "down" algorithm (or vice versa). Note that if the starting links are badly chosen, the complete procedure has the opportunity to remove them—although in this example, the final model contains all the links from the starting model. Two models were selected by this procedure; they are shown in Figure 10. The corresponding posterior probabilities are listed in Table 5.

The result of primary interest here is the importance of A (possible to get above swelling) and D (Valsalva/stand test) with respect to hernia diagnosis. Both indicants can be established through simple procedures at physical examination. The only real model uncertainty exhibited concerns the relationship between C (swelling separate from testes) and E (tenderness). The odds in favor of the inclusion of the CE link are 3 to 1 (evidence not worth more than a bare mention). Analysis of further cross-classifications extracted from this data base also yield similarly sparse models.

Table 3. Women and Mathematics: Posterior Model Probabilities for Decomposable Models

Figure	*Model*	*Posterior probability %*
8(a)	[*A*][*BDE*][*CDF*]	75
8(b)	[*A*][*BDE*][*DF*][*CF*]	25

Table 4. Scrotal Swelling Data

	Indicants							
Hernia	*A*	*B*	*C*	*D*	*E*	*F*	*G*	*Count*
N	N	N	N	N	Y	Y	N	1
N	Y	N	N	N	N	N	N	16
N	Y	N	N	N	N	Y	N	3
N	Y	N	N	N	Y	Y	N	51
N	Y	N	N	N	Y	Y	Y	17
N	Y	N	Y	N	N	N	N	30
N	Y	N	Y	N	N	N	Y	1
N	Y	N	Y	N	N	Y	N	3
N	Y	N	Y	N	Y	N	N	1
N	Y	N	Y	N	Y	Y	N	20
N	Y	N	Y	N	Y	Y	Y	4
N	Y	N	Y	Y	N	N	N	36
N	Y	N	Y	Y	N	Y	N	3
N	Y	Y	N	N	N	N	N	38
N	Y	Y	N	N	N	N	Y	1
N	Y	Y	N	N	N	Y	N	3
N	Y	Y	N	N	Y	Y	N	3
N	Y	Y	Y	N	N	N	N	21
N	Y	Y	Y	N	Y	Y	N	2
Y	N	N	Y	Y	N	N	N	39
Y	N	N	Y	Y	N	Y	N	5
Y	Y	N	Y	Y	N	N	N	1

5. PERFORMANCE

Following Dawid (1984), we contend that one of the primary purposes of statistical analysis is to make forecasts for the future. Therefore, one way we can judge the efficacy of a model selection strategy is to measure how well the resulting models predict future observations. In the case of Occam's window, our specific objective is to compare the quality of the predictions based on model averaging against those based on any single model that an analyst might reasonably have selected.

We examined the predictive performance for each of the examples considered previously as follows. We randomly split the complete data sets into two subsets. One subset, D^S, containing 25% of the data, was used to select models, while $D^T = D \backslash D^S$ was used as a set of test cases. We measured performance by the logarithmic scoring rule of Good (1952). Specifically, we measured the predictive ability of an individual model, M, with

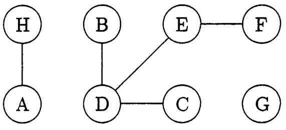

Figure 9. Starting Model for Scrotal Swelling Example.

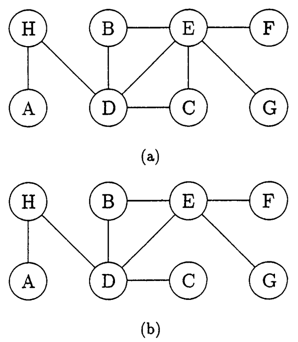

Figure 10. Scrotal Swellings: Decomposable Models Selected.

$$- \sum_{d \in D^T} \log \mathrm{pr}(d \mid M, D^S).$$

We measured the predictive performance of model averaging with

$$- \sum_{d \in D^T} \log \left\{ \sum_{M \in \mathcal{A}} \mathrm{pr}(d \mid M, D^S) \mathrm{pr}(M \mid D^S) \right\},$$

where $\mathcal{A}$ is the set of selected models.

We present results in Tables 6, 7, and 8 for each of the undirected examples of Section 4. In each case, we give the models selected and the performance measure (up to a normalizing constant) for each individual model and for model averaging. For the coronary heart disease example, we also include the score for the model selected by Whittaker (1990) on the basis of the full data set. The models selected by Upton (1991) and Fowlkes et al. (1988) are not included, because they are not decomposable.

In each case, the method that averages over the models selected provides predictive performance superior to the performance resulting from basing the inference on any single model that might reasonably have been selected. In the coronary heart disease data, for example, our model-averaging method outperforms the "best" model (i.e., that with the highest posterior probability) by 31 points of log-predictive probability, or 62 points on the scale of twice the log probability on which deviances are measured. Repeating the random split or varying the subset proportions produces very similar performance results.

We also carried out a ROC (receiver operating characteristic) analysis for each of the examples. In Figure 11 we show two ROC curves for the variable E, ratio of β and α proteins, in the coronary heart disease example. The solid ROC curve shows how well the single model with the highest posterior probability predicts variable E, whereas the dashed curve shows the performance achieved by averaging over the selected models. Again, we used 25% of the data to select models and the remainder of the data for testing.

Table 5. Scrotal Swellings: Posterior Model Probabilities for Decomposable Models

Figure	*Model*	*Posterior probability %*
8(a)	[*AH*][*DH*][*BDE*][*CDE*][*EF*][*EG*]	75
8(b)	[*AH*][*DH*][*BDE*][*CD*][*EF*][*EG*]	25

Table 6. Coronary Heart Disease: Predictive Performance

Model	*Posterior probability %*	*Logarithmic score*
[*AE*][*BC*][*BE*][*DE*][*F*]	26	4984.4
[*AC*][*BC*][*BE*][*DE*][*F*]	16	4990.2
[*AC*][*AE*][*BC*][*DE*][*F*]	13	4992.2
[*A*][*BC*][*BE*][*DE*][*F*]	9	4990.1
[*AE*][*BC*][*BE*][*D*][*F*]	8	4981.7
[*AE*][*BC*][*DE*][*F*]	7	4983.7
[*AC*][*BC*][*BE*][*D*][*F*]	5	4981.6
[*AC*][*BC*][*DE*][*F*]	4	4989.5
[*AC*][*AE*][*BC*][*D*][*F*]	4	4987.4
[*A*][*BC*][*BE*][*D*][*F*]	3	4989.4
[*A*][*BC*][*DE*][*F*]	2	4981.0
[*AE*][*BC*][*D*][*F*]	2	4980.9
[*AC*][*BC*][*D*][*E*][*F*]	1	4986.7
[*ABCE*][*ADE*][*BF*]	Whittaker	4984.8
Model averaging		4953.6

These ROC curves show the false-positive and true-positive proportions for different probability thresholds for variable E. The area above the curve in the unit square provides a measure of predictive ability. Model averaging provides substantially better predictive performance in this instance. The area above the curve is .08 for model averaging and .22 for the best single model. Thus model averaging reduces the average false-positive rate for a given true-positive rate by about two-thirds, where the false-positive rates are averaged over all true-positive rates. Model averaging is *not* guaranteed to provide superior predictive performance for each variable, although the situation in Figure 11 is typical.

6. DISCUSSION

6.1 General Comments and Other Approaches

We have outlined an overall strategy for model selection and accounting for model uncertainty in two important classes of models for high-dimensional contingency tables. This involves a redefinition of the Bayesian model uncertainty formalism, an efficient way of computing exact Bayes factors that exploits the graphical structure, and an algorithm for quickly searching through the very large model classes involved. The resulting procedure is quite efficient; for the example of Section 4.2.1, approximately 3,000 model comparisons per minute can be carried out on a Sun IPC.

Table 7. Women and Mathematics: Predictive Performance

Model	*Posterior probability %*	*Logarithmic score*
[*A*][*B*][*CDF*][*DE*]	75	3318.9
[*A*][*B*][*CF*][*DE*][*DF*]	21	3317.3
[*A*][*B*][*CF*][*DE*]	4	3320.4
Model averaging		3313.9

Table 8. Scrotal Swelling: Predictive Performance

Model	Posterior probability %	Logarithmic score
[AH][AD][BDE][CD][EF][FG]	3	605.3
[AH][DH][BDE][CD][EF][FG]	3	599.6
[AH][DH][BDE][CDE][EF][FG]	5	600.6
[AH][AD][BDE][CDE][EF][FG]	5	606.3
[AH][AD][BDE][CD][EF][EG]	15	603.4
[AH][DH][BDE][CD][EF][EG]	15	597.7
[AH][DH][BDE][CDE][EF][EG]	27	598.7
[AH][AD][BDE][CDE][EF][EG]	27	604.4
Model averaging		594.2

There is a considerable literature on model selection for multidimensional contingency tables; this is generally concerned with the selection of a single "best" model. Most of it is based on the asymptotic properties of goodness-of-fit statistics (Edwards and Havránek 1985; Fowlkes et al. 1988; Goodman 1973; Havránek 1984; Wermuth 1976; Whittaker 1984). There are also approaches based on information criteria and discrepancy measures (Gokhale and Kullback 1978; Linhart and Zucchini 1986; Sakamoto 1984). A recent review is provided by Upton (1991), who advocated the use of the BIC statistic. The calculation of Bayes factors for contingency table models has been considered by Spiegelhalter and Smith (1982), Raftery (1986a, 1988, 1993), Spiegelhalter and Lauritzen (1990a), and Spiegelhalter and Cowell (1992).

Pearl and Verma (1991) and Glymour, Scheines, Spirtes, and Kelly (1987) have proposed strategies for recovering causal structure from data. Although these authors' objectives differ from ours, their procedures for selecting directed graphical structures have much in common with our recursive causal model selection strategy.

Cooper and Herskovits (1992) and Anderson, Krebs, and Anderson (1991) have examined model selection in the context of probabilistic expert systems. In both cases the examples are based solely on data analysis, and the incorporation of prior expert opinion is not considered. Cooper and Herskovits (1992) described a general theory involving averaging over all models and suggested possible approximations. Their K2 strategy seeks out the "best" recursive causal model for the qualitative layer, where "best" is taken to mean the *single* model with maximum probability. The algorithm starts with a model with no links and at each stage adds the directed link that most increases the model probability. The user must prespecify an ordering of the nodes. Anderson et al. (1991) carried out their search in the undirected graphical model framework using a method introduced by Kreiner (1987). The difficulties with large sparse tables mentioned earlier are avoided by using exact tests when comparing models.

6.2 Model Priors

In the examples considered earlier, the prior model probabilities pr(M) were assumed equal; Cooper and Herskovits (1992) also assumed that models are equally likely a priori. In general this can be unrealistic and may also be expensive, and we will want to penalize the search strategy as it moves further away from the model(s) provided by the expert(s)/data analyst(s). Ideally, one would elicit prior probabilities for all possible qualitative structures from the expert, but this will be feasible only in trivial cases.

For models with fewer than 15 to 20 nodes, prior model probabilities may be approximated by eliciting prior probabilities for the presence of every possible link and assuming that the links are mutually independent, as follows. Let $\mathcal{E} = \mathcal{E}_P \cup \mathcal{E}_A$ denote the set of all possible links for the nodes of model M, where $\mathcal{E}_P$ denotes the set of links present in model M and $\mathcal{E}_A$ denotes the absent links. For every link $e \in \mathcal{E}$, we elicit pr(e), the prior probability that link e is included in M. The prior model probability is then approximated by

$$\mathrm{pr}(M) \propto \prod_{e\in\mathcal{E}_P} \mathrm{pr}(e) \prod_{e\in\mathcal{E}_A} (1-\mathrm{pr}(e)).$$

Prior link probabilities from multiple experts are treated as independent sources of information and are simply multiplied together to give pooled prior model probabilities. Clearly, the contribution from each expert/data analyst could be weighted.

For applications involving a larger number of nodes or where the elicitation of link probabilities is not possible, we could assume that the "evidence" in favor of each link included by the expert(s)/data analyst(s) in the elicited qualitative structure(s) is "substantial" or "strong" but not "very strong" or "decisive" (Jeffreys 1961). For example, we could assume that the evidence in favor of an included link lies at the center of Occam's window corresponding to a prior link probability for all $e \in \mathcal{E}_P$ of

$$\mathrm{pr}(e) = \frac{1}{1+\exp\left(\dfrac{O_L+O_R}{2}\right)}.$$

Similarly, the prior link probabilities for $e \in \mathcal{E}_A$ are given by

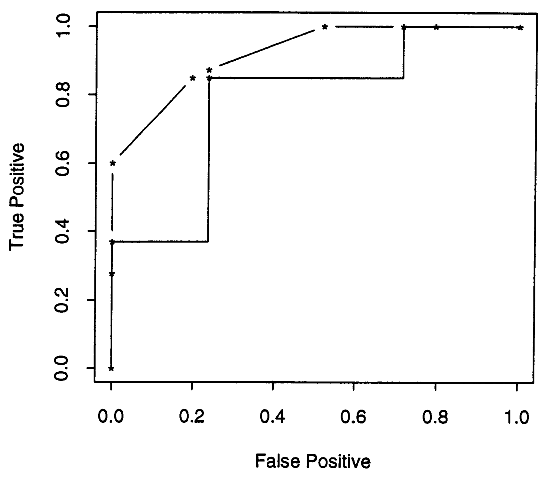

Figure 11. Coronary Heart Disease: ROC Curves for Node E.

$$\mathrm{pr}(e) = \frac{\exp\left(\dfrac{O_L + O_R}{2}\right)}{1 + \exp\left(\dfrac{O_L + O_R}{2}\right)}.$$

In the directed case it may be possible to construct a prior distribution on the space of orderings (see Critchlow 1985 for further discussion).

6.3 Remaining Issues

Although we believe that the methods we propose provide a workable approach to qualitative updating in expert systems, some issues remain. Spiegelhalter and Lauritzen (1990a) and other authors have expressed concerns about automatically updating the qualitative structure without reference to the domain expert. Such concerns need to be addressed in the context of real expert systems. Extension of the methods to include the more general graphical models of Wermuth and Lauritzen (1990) and Edwards (1990) will also be important. Missing data will frequently be a problem, and we are currently exploring a number of techniques for the incorporation of missing data in the model selection strategy.

In the examples we have used vague priors for the model parameters that do not incorporate specific prior information. But in expert system applications there will often be substantial prior information, and taking it into account would be expected to improve performance. How to elicit the required Dirichlet prior distributions thus is a major issue. Direct elicitation is typically intractable; this has been a barrier to the use of the DL approach.

Madigan and Raftery (1991) outlined a simple approach to the elicitation of the required priors. They regarded the parameters of the Dirichlet prior distribution as "equivalent prior samples," which are elicited subject to constraints that ensure consistency. The priors are elicited sequentially in a way that avoids the need to store the full "equivalent prior" table.

[Received July 1992. Revised October 1993.]

REFERENCES

Anderson, L. R., Krebs, J. H., and Anderson, J. D. (1991), "STENO: An Expert System for Medical Diagnosis Based on Graphical Models and Model Search," *Journal of Applied Statistics,* 18, 139–153.

Berger, J. O., and Delampady, M. (1987), "Testing Precise Hypotheses" (with discussion), *Statistical Science,* 2, 317–352.

Berger, J. O., and Sellke, T. (1987), "Testing a Point Null Hypothesis: The Irreconcilability of *P* Values and Evidence" (with discussion), *Journal of the American Statistical Association,* 82, 112–139.

Bishop, Y. M. M., Fienberg, S. E., and Holland, P. W. (1975), *Discrete Multivariate Analysis,* Cambridge, MA: MIT Press.

Chaloner, K. M., and Duncan, G. T. (1987), "Some Properties of the Dirichlet-Multinomial Distribution and Its Use in Prior Elicitation," *Communications in Statistics, Part A—Theory and Methods,* 16, 511–523.

Cooper, G. F., and Herskovits, E. (1992), "A Bayesian Method for the Induction of Probabilistic Networks From Data," *Machine Learning,* 9, 309–347.

Critchlow, D. E. (1985), *Metric Methods for Analyzing Partially Ranked Data,* New York: Springer-Verlag.

Dawid, A. P. (1984), "Statistical Theory—The Prequential Approach," *Journal of the Royal Statistical Society,* Ser. A, 147, 278–292.

——— (1986), "Probability Forecasting." in *Encyclopedia of Statistical Sciences* (Vol. 7), eds. Kotz, S., and Johnson, N. L., New York: John Wiley, pp. 210–218.

Dawid, A. P., and Lauritzen, S. L. (1993), "Hyper-Markov Laws in the Statistical Analysis of Decomposable Graphical Models," *The Annals of Statistics,* 21, 1272–1317.

Edwards, D. (1990), "Hierarchical Mixed Interaction Models," *Journal of the Royal Statistical Society,* Ser. B, 52, 3–20.

Edwards, D., and Havránek, T. (1985), "A Fast Procedure for Model Search in Multidimensional Contingency Tables," *Biometrika,* 72, 339–351.

Evett, I. W. (1991), "Implementing Bayesian Methods in Forensic Science," paper presented to the Fourth Valencia International Meeting on Bayesian Statistics.

Fenech, A., and Westfall, P. (1988), "The Power Function of Conditional Log-Linear Model Tests," *Journal of the American Statistical Association,* 83, 198–203.

Fowlkes, E. B., Freeny, A. E., and Landwehr, J. M. (1988), "Evaluating Logistic Models for Large Contingency Tables," *Journal of the American Statistical Association,* 83, 611–622.

Frydenberg, M., and Lauritzen, S. L. (1989), "Decomposition of Maximum Likelihood in Mixed Graphical Interaction Models," *Biometrika,* 76, 539–555.

Gabriel, K. R. (1969), "Simultaneous Test Procedures—Some Theory of Multiple Comparisons," *Annals of Mathematical Statistics,* 40, 224–250.

Glymour, C., Scheines, R., Spirtes, P., and Kelly, K. (1987), *Discovering Causal Structure,* New York: Academic Press.

Gokhale, D. V., and Kullback, S. (1978), *The Information in Contingency Tables,* New York: Marcel Dekker.

Good, I. J. (1952), "Rational Decisions," *Journal of the Royal Statistical Society,* Ser. B, 14, 107–114.

Goodman, L. A. (1970), "The Multivariate Analysis of Qualitative Data: Interaction Among Multiple Classifications," *Journal of the American Statistical Association,* 65, 226–256.

——— (1971), "The Analysis of Multidimensional Contingency Tables: Stepwise Procedures and Direct Estimation Methods for Building Models for Multiple Classifications," *Technometrics,* 13, 33–61.

——— (1973), "Guided and Unguided Methods for the Selection of Models for a Set of T Multidimensional Contingency Tables," *Journal of the American Statistical Association,* 68, 165–175.

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

Havránek, T. (1984), "A Procedure for Model Search in Multidimensional Contingency Tables," *Biometrics,* 40, 95–100.

Hodges, J. S. (1987), "Uncertainty, Policy Analysis and Statistics" (with discussion), *Statistical Science,* 2, 259–291.

Jeffreys, H. (1961), *Theory of Probability* (3rd ed.), Oxford, U.K.: Oxford University Press.

Kass, R. E., and Raftery, A. E. (1995), "Bayes Factors," submitted to *Journal of the American Statistical Association.*

Kiiveri, H., Speed, T. P., and Carlin, J. B. (1984), "Recursive Causal Models," *Journal of the Australian Mathematical Society,* Ser. A, 36, 30–52.

Kotz, S., and Johnson, N. L. (eds.) (1985), *Encyclopedia of Statistical Sciences* (Vol. 6), New York: John Wiley, pp. 578–579.

Kreiner, S. (1987), "Analysis of Multidimensional Contingency Tables by Exact Conditional Tests: Techniques and Strategies," *Scandinavian Journal of Statistics,* 14, 97–112.

Lauritzen, S. L., and Spiegelhalter, D. J. (1988), "Local Computations With Probabilities on Graphical Structures and Their Application to Expert Systems" (with discussion), *Journal of the Royal Statistical Society,* Ser. B, 50, 157–224.

Linhart, H., and Zucchini, W. (1986), *Model Selection,* New York: John Wiley.

Madigan, D., and Raftery, A. E. (1991), "Model Selection and Accounting for Model Uncertainty in Graphical Models Using Occam's Window," Technical Report 213, University of Washington, Dept. of Statistics.

Miller, A. J. (1984), "Selection of Subsets of Regression Variables" (with discussion), *Journal of the Royal Statistical Society,* Ser. A, 147, 389–425.

Moulton, B. R. (1991), "A Bayesian Approach to Regression Selection and Estimation With Application to a Price Index for Radio Services," *Journal of Econometrics,* 49, 169–193.

Pearl, J., and Vermà, T. S. (1991), "A Theory of Inferred Causation," in *Knowledge Representation and Reasoning: Proceedings of the Second International Conference,* eds. J. A. Allen, R. Fikes, and E. Sandewall, San Mateo, CA, Morgan Kaufmann, pp. 441–452.

Raftery, A. E. (1986a), "A Note on Bayes Factors for Log-Linear Contingency Table Models With Vague Prior Information," *Journal of the Royal Statistical Society,* Ser. B, 48, 249–250.

——— (1986b), "Choosing Models for Cross-Classifications," *American Sociological Review,* 51, 145–146.

——— (1988), "Approximate Bayes Factors for Generalized Linear Models," Technical Report 121, University of Washington, Dept. of Statistics.

——— (1993), "Approximate Bayes Factors and Accounting for Model Uncertainty in Generalized Linear Models," Technical Report 255, University of Washington, Dept. of Statistics.

Regal, R. R., and Hook, E. B. (1991), "The Effects of Model Selection on Confidence Intervals for the Size of a Closed Population," *Statistics in Medicine,* 10, 717–721.

Sakamoto (1984), "Efficient Use of Akaike's Information Criterion on High-Dimensional Contingency Table Analysis," *Metron,* 40, 257–275.

Spiegelhalter, D. J., and Cowell, R. G. (1991), "Learning in Probabilistic Expert Systems," in *Bayesian Statistics 4,* eds. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, Oxford: Clarendon Press, pp. 447–466.

Spiegelhalter, D. J., and Lauritzen, S. L. (1990a), "Sequential Updating of Conditional Probabilities on Directed Graphical Structures," *Networks,* 20, 579–605.

——— (1990b), "Techniques for Bayesian Analysis in Expert Systems," *Annals of Mathematics and Artificial Intelligence,* 2, 353–366.

Spiegelhalter, D. J., and Smith, A. F. M. (1982), "Bayes Factors for Linear and Log-Linear Models With Vague Prior Information," *Journal of the Royal Statistical Society,* Ser. B, 44, 377–387.

Upton, G. J. G. (1991), "The Exploratory Analysis of Survey Data Using Log-Linear Models," *The Statistician,* 40, 169–182.

Wermuth, N. (1976), "Model Search Among Multiplicative Models," *Biometrics,* 32, 253–263.

Wermuth, N., and Lauritzen, S. L. (1990), "On Substantive Research Hypotheses, Conditional Independence Graphs and Graphical Chain Models," *Journal of the Royal Statistical Society,* Ser. B, 52, 21–50.

Whittaker, J. (1984), "Fitting All Possible Decomposable and Graphical Models to Multiway Contingency Tables," in *Compstat 84,* eds. T. Havránek et al., Vienna: Physica-Verlag, pp. 401–406.

——— (1990), *Graphical Models in Applied Multivariate Statistics,* Chichester, U.K.: John Wiley.