Enhancing the Predictive Performance of Bayesian Graphical Models

David Madigan, Department of Statistics, University of Washington*
and Fred Hutchinson Cancer Research Center

Jonathan Gavrin, Department of Anesthesiology, University of Washington
and Fred Hutchinson Cancer Research Center

Adrian E. Raftery, Departments of Statistics and Sociology,
University of Washington

Abstract

Both knowledge-based systems and statistical models are typically concerned with making predictions about future observables. Here we focus on assessment of predictive performance and provide two techniques for improving the predictive performance of Bayesian graphical models. First, we present Bayesian model averaging, a technique for accounting for model uncertainty. Second, we describe a technique for eliciting a prior distribution for competing models from domain experts. We explore the predictive performance of both techniques in the context of a urological diagnostic problem.

KEYWORDS: Prediction; Bayesian graphical model; Bayesian network; Decomposable model; Model uncertainty; Elicitation.

1 Introduction

Both statistical methods and knowledge-based systems are typically concerned with combining information from various sources to make inferences about prospective measurements. Inevitably, to combine information, we must make modeling assumptions. It follows that we should carefully assess and optimize the predictive performance of the models that we build (Draper *et al.*, 1993, Bradshaw, *et al.*, 1993). In practice, assessment of predictive performance is difficult, which may explain the dearth of careful predictive analyses in the

*(madigan@stat.washington.edu) Department of Statistics, GN-22, University of Washington, Seattle WA 98195. This work is supported in part by a NSF grant to the University of Washington. The authors are grateful to Russell Almond, John Boose, Jeff Bradshaw, C. Richard Chapman, David Draper, Denise Draper, Peter Dunbar, and Jennifer Hoeting for stimulating discussions.

literature (see Spiegelhalter *et al.*, 1993 and Murphy and Winkler, 1977, for notable exceptions). However, when the primary purpose of a model is to predict the future, predictive performance should act as the primary guide for the model building process (Dawid, 1984).

In what follows, our predictions take the form of probability distributions. To assess predictive performance we use the logarithmic scoring rule suggested by Good (1952). This assigns to each event A which occurs a score of $-\log\{\mathrm{pr}(A)\}$. The higher the probability assigned by the model to the events that actually occur, the better the log score. See Dawid (1986) for further discussion. In Section 2 below we provide the implementation details.

In Section 3 we examine two techniques for improving predictive performance in the context of Bayesian graphical conditional independence models. First we describe Bayesian Model Averaging (BMA). Here the idea is to account for model uncertainty, specifically uncertainty about the graphical structure, by averaging over different models. Second we describe a technique for eliciting informative prior distributions for abstract entities such as models. We demonstrate the efficacy of the techniques in the context of an application to urology. Our results are especially relevant to the increasing number of knowledge-based system builders and statistical analysts using Bayesian graphical models (Spiegelhalter *et al.*, 1993, 1994).

2 Measuring predictive performance

For probabilistic predictions, there exist two types of discrepancies between observed and predicted values (Gaver *at al.*, 1992): *predictive bias* (a systematic tendency to mispredict on the low or high side) and *lack of calibration* (a systematic tendency to over- or understate predictive accuracy). The logarithmic score combines these discrepancies into a single measure. We note that the logarithmic scoring rule is a *proper scoring rule* as defined by Matheson and Winkler (1976) and others. Considering predictive bias and calibration separately can also be useful–see for example Madigan and Raftery (1994), Raftery, Madigan, and Hoeting, 1993, Madigan *et al.* (1994), and Spiegelhalter (1986). In particular, a predictive model which merely assigns the prior probability to each future observable may be well calibrated but of no practical use.

We examine the predictive performance for the urological application considered below as follows: we randomly split the complete data set into two subsets. We use one subset, D^S, containing 50% of the data, to select models, while we use $D^T = D \setminus D^S$, as a set of test cases, where D denotes the complete data set. We measure performance by the logarithmic

scoring rule. Specifically, we measure the predictive ability of an individual model, M, with:

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

We measure the predictive performance of BMA with:

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

where BMA is operating over the models in $\mathcal{A}$. We examine different splits to assess sensitivity.

Gaver *et al.* (1992) refer to this type of approach as *retrospective predictive validation*. An alternative approach is *prospective predictive validation* in which the modeler makes testable predictions about future observations, and these predictions are compared with what actually happens; we are currently conducting such a study.

3 Model Uncertainty

A typical approach to data analysis is to initially carry out a model selection exercise leading to a single "best" model and then to make inference as if the selected model were the true model. However, as a number of authors have pointed out, this paradigm ignores a major component of uncertainty, namely uncertainty about the model itself (Raftery, 1988, Chatfield, 1994, Cooper and Herskovits, 1992, Draper *et al.*, 1987, Hodges, 1987, Self and Cheeseman, 1987). This will lead to poorly calibrated predictions: time will reveal that one's uncertainty bands were not wide enough. For striking examples of this see York and Madigan (1992), York *et al.* (1994), Regal and Hook (1991), and Draper (1994).

In principle, BMA provides a straightforward way around this problem. If Δ is the quantity of interest, such as a future observation, a structural characteristic of the system being studied, or the utility of a course of action, then its posterior distribution given data D is

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

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, the posterior probability for model M_k is given by

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

where

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

θ is a vector of parameters, $\mathrm{pr}(\theta \mid M_k)$ is the prior for θ under model M_k, $\mathrm{pr}(D \mid \theta, M_k)$ is the likelihood, and $\mathrm{pr}(M_k)$ is the prior probability that M_k is the true model. In the graphical model context, θ is the vector of probabilities associated with the model.

Hodges (1987) argues that "when the time comes for betting on what the future holds, one's uncertainty about that future should be fully represented and model mixing is the only tool around". Furthermore, averaging over *all* the models in this fashion provides better predictive ability, as measured by the logarithmic scoring rule, than using any single model M_j (Madigan and Raftery, 1994, hereafter referred to as MR).

However, implementation of the above strategy proves difficult for two reasons: first, the integrals in (3) can be hard to compute, and second, the number of terms in (1) can be enormous.

For graphical models for discrete data, Spiegelhalter and Lauritzen (1990) and Dawid and Lauritzen (1993) have developed efficient solutions to the former problem (see Kass and Raftery, 1994, for solutions in more general contexts.) Several approaches to the latter problem, i.e. the enormous number of terms in (1), have emerged recently, and here we discuss two of them (see also, Almond, 1994). MR do not attempt to approximate (1) but instead, appealing to standard norms of scientific investigation, adopt a model selection procedure. This involves averaging over a much smaller set of models than in (1) and delivers a parsimonious set of models to the data analyst, thereby facilitating effective communication of model uncertainty. Madigan and York (1993) on the other hand suggest directly approximating (1) with a Markov chain Monte Carlo method. Both MR and Madigan and York (1993) provide implementation details for directed (i.e., Bayesian networks) and undirected decomposable models, although in this paper we focus on the latter.

3.1 Model Selection and Occam's Window

Two basic principles underly the approach presented in MR. First, if a model receives much less support than the model with maximum posterior probability, then one should drop it. Thus models not belonging to:

$$\mathcal{A}' = \left\{ M_k : \frac{\max_l\{\mathrm{pr}(M_l \mid D)\}}{\mathrm{pr}(M_k \mid D)} \leq C \right\}, \tag{4}$$

should be excluded from equation (1) where C is chosen by the data analyst. Second, appealing to Occam's razor, one should exclude complex models which receive less support

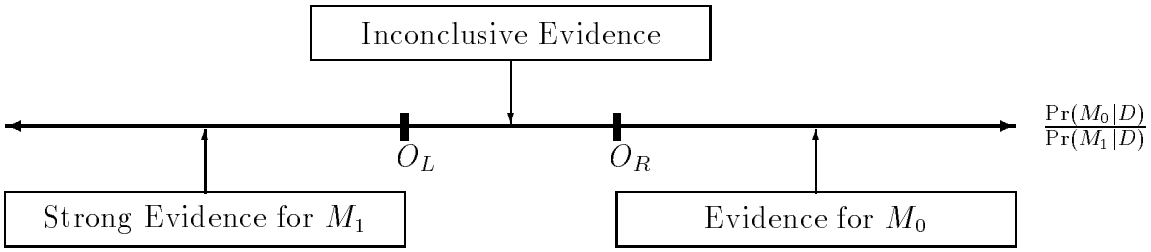


Figure 1: Occam's Window: Interpreting the posterior odds, $\frac{\Pr(M_0|D)}{\Pr(M_1|D)}$, where M_0 is a sub-model of M_1

from the data than their simpler counterparts. More formally exclude also from (1) models belonging to:

$$\mathcal{B} = \left\{ M_k : \exists M_l \in \mathcal{A}, M_l \subset M_k, \frac{\text{pr}(M_l \mid D)}{\text{pr}(M_k \mid D)} > 1 \right\} \tag{5}$$

and equation (1) is replaced by

$$\text{pr}(\Delta \mid D) = \frac{\sum_{M_k \in \mathcal{A}} \text{pr}(\Delta \mid M_k, D)\text{pr}(D \mid M_k)\text{pr}(M_k)}{\sum_{M_k \in \mathcal{A}} \text{pr}(D \mid M_k)\text{pr}(M_k)} \tag{6}$$

where

$$\mathcal{A} = \mathcal{A}' \backslash \mathcal{B}. \tag{7}$$

This greatly reduces the number of models in the sum in equation (1) so that now one need only search to identify the models in $\mathcal{A}$. Two further principles underly the search strategy. First, if a model is rejected then all its submodels obtained by deleting links are rejected. The independence properties of the models justify this. The second principle — "Occam's Window" — concerns the interpretation of the ratio of posterior model probabilities $\text{pr}(M_0 \mid D)/\text{pr}(M_1 \mid D)$. Here M_0 and M_1 are models such that M_0 is one link "smaller" than M_1. If there is evidence for M_0 then M_1 is rejected but to reject M_0 we require strong evidence *for* the larger model, M_1. If the evidence is inconclusive (falling in Occam's Window–see Figure 1) neither model is rejected. MR set the edges of the window at $\frac{1}{20}$ and 1, although our recent experience suggests that a window from $\frac{1}{20}$ to 2 provides better performance.

These principles fully define the strategy. Typically the number of terms in (1) is reduced to fewer than 20 models and often to as few as two. MR provide a detailed description of the algorithm.

3.2 Markov Chain Monte Carlo Model Composition

Our second approach is to approximate (1) using Markov chain Monte Carlo methods, such as in Hastings (1970) and Neal (1993), generating a process which moves through model space. Specifically, let $\mathcal{M}$ denote the space of models under consideration. We can construct an irreducible Markov chain $\{M(t)\}, t = 1, 2, \ldots$ with state space $\mathcal{M}$ and equilibrium distribution $\mathrm{pr}(M_i \mid D)$. Then for any well-behaved function $g(M(i))$ defined on $\mathcal{M}$, if we simulate this Markov chain for $t = 1, \ldots, N$, the average:

$$\hat{G} = \frac{1}{N} \sum_{t=1}^{N} g(M(t)) \tag{8}$$

converges with probability one to $E(g(M))$ as N goes to infinity. To compute (1) in this fashion we set $g(M) = \mathrm{pr}(\Delta \mid M, D)$.

To construct the Markov chain we define a neighborhood $\mathrm{nbd}(M)$ for each $M \in \mathcal{M}$ which is the set of models with either one link more or one link fewer than M and the model M itself. Define a transition matrix q by setting $q(M \to M') = 0$ for all $M' \notin \mathrm{nbd}(M)$ and $q(M \to M')$ constant for all $M' \in \mathrm{nbd}(M)$. If the chain is currently in state M, we proceed by drawing M' from $q(M \to M')$. If the model is legal then we accept it with probability:

$$\min\left\{1, \frac{\mathrm{pr}(M' \mid D)}{\mathrm{pr}(M \mid D)}\right\}.$$

Otherwise the chain stays in state M. We find this process highly mobile and runs of 10,000 or less typically prove adequate.

3.3 Analysis

As we argued above, we can judge the efficacy of these BMA strategies by examining their predictive performance. We have assessed the predictive performance of Occam's Window and Markov chain Monte Carlo model composition (MC^3) in the context of a challenging medical application.

The application concerns the diagnosis of scrotal swellings, a common urological disorder. MR presented data on 299 patients cross-classified according to one disease class, Hernia (H), and 7 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, tender; F, pain; G, evidence of other urinary tract infections. MR provide the complete data set. The first author gathered the data at the Meath Hospital, Dublin, Ireland under the supervision of urologist Michael R. Butler.

For simplicity, we confine our analyses in this paper to undirected decomposable graphical models. MR's analyses in the context of directed acyclic graphical models achieve similar performance results.

Specifically, we consider a decomposable model M for a set of random variables $X_v, v \in V$. Let $\mathcal{I} = \prod_{v \in V} \mathcal{X}_v$ denote the set of possible configurations of X. Denote by $\theta(i)$ the probability of a state $i \in \mathcal{I}$. Then the clique marginal probability tables $\theta_C, C \in \mathcal{C}$ determine $\theta(i)$, where $\mathcal{C}$ denotes the set of cliques of M:

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

$\mathcal{S}$ denotes the system of clique separators in an arbitrary perfect ordering of $\mathcal{C}$.

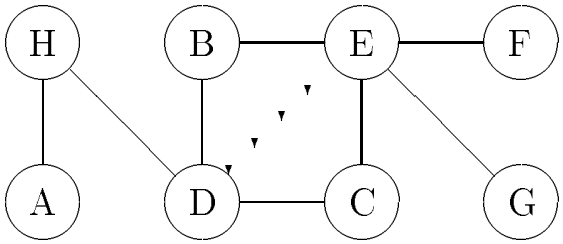
We construct a "hyper–Dirichlet" distribution on θ as follows: first identify a perfect ordering of the cliques $\{C_1, \ldots, C_n\}$; second, place a Dirichlet distribution $\mathcal{D}(\lambda_{C_1})$ on θ_{C_1}; third, place a Dirichlet distribution $\mathcal{D}(\lambda_{C_2})$ on θ_{C_2}, with parameters constrained by $\lambda_{C_1}(i_{C_1 \cap C_2}) = \lambda_{C_2}(i_{C_1 \cap C_2})$ and realizations constrained so that $\theta_{C_1 \cap C_2}$ is identical for θ_{C_1} and θ_{C_2}. For each subsequent clique C_i, place a Dirichlet on θ_{C_i} such that the parameters and the realizations of that distribution are consistent with those specified for the previous cliques.

Dawid and Lauritzen (1993) have shown that there exists a unique hyper-Dirichlet distribution for θ over M such that θ_C has the marginal density $\mathcal{D}(\lambda_C)$ for all $C \in \mathcal{C}$. This prior distribution is conjugate with multinomial sampling. Dawid and Lauritzen (1993) provide simple expressions for posterior distributions and likelihoods.

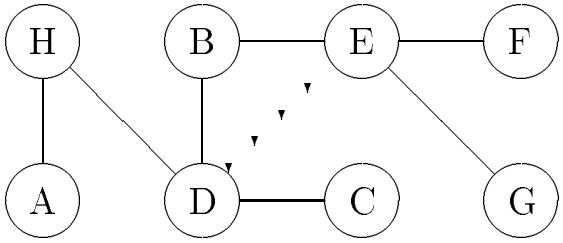
The model selection procedure can consider 28 possible links in the scrotal swelling application. The Occam's Window procedure selects two models. We show these in Figure 2 and the corresponding posterior probabilities in Table 1.

In large model spaces, if the starting model for the Occam's Window procedure is very different from the selected models (for instance, starting from the empty model in this example), computation times can be prohibitive. MC^3 on the other hand typically moves very rapidly to models with high posterior probability. For the scrotal swelling application we combined the two techniques: we ran the MC^3 algorithm for 1,000 cycles to find a reasonable starting model for Occam's Window.

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 which is exhibited concerns the relationship between C (swelling separate from testes) and E (tender). The odds in favor of the inclusion of the CE link are 3 to 1 ("evidence not worth more than a bare mention" in Jeffreys' terminology). Analysis of further cross-classifications extracted from this database also yields similarly sparse models.



(a)



(b)

Figure 2: Scrotal Swellings: Decomposable Models Selected

Table 1: Scrotal Swellings: Posterior Model Probabilities for Decomposable Models

Figure	*Model*	*Posterior probability*
2(a)	$[AH][DH][BDE][CDE][EF][EG]$	0.75
2(b)	$[AH][DH][BDE][CD][EF][EG]$	0.25

Table 2: Scrotal Swellings: Predictive Performance

Model	*Posterior probability*	*Logarithmic Score*
$[AGH][AFG][ACDF][BCD][BDF][DEF]$	0.02	421.5
$[AFH][FGH][ACDF][BCD][BDF][DEF]$	0.10	420.9
$[AH][GH][ACD][BCD][BDE][EF]$	0.11	413.8
$[ADH][GH][ACD][BCD][BDE][EF]$	0.15	414.5
$[ACH][GH][CDH][BCD][BDE][EF]$	0.15	414.5
$[ACH][GH][ACD][BCD][BDE][EF]$	0.15	415.7
$[ACH][FGH][CDFH][BCDF][DEF]$	0.32	419.6
BMA: Occam's Window		405.8
BMA: Markov Chain Monte Carlo Model Composition		390.6

In Table 2 we provide an assessment of predictive performance as described in Section 2 above. Note that the models selected by Occam's Window are different from those in Figure 2 above since we used only 50% of the data to select them.

In this application, as in others reported by Madigan and Raftery (1994), Raftery, Madigan, and Hoeting (1993), Raftery, Madigan, and Volinsky (1994), and Madigan *et al* (1994), BMA provides superior predictive performance to any single model which the analyst might reasonably have selected. For example, MC^3 outperforms the "best" model (i.e. the model with the highest posterior probability that the standard model selection procedures would select) by 29 points of log predictive probability, or 58 points on the scale of twice the log probability on which deviances are measured. Repeating the random split or varying the subset proportions typically produces similar results.

MC^3 generally provides superior predictive performance to Occam's Window (Raftery, Madigan, and Volinsky, 1994). However, the insight into model uncertainty provided by the Occam's Window method will be important in many applications. In particular, presentation of the models in Occam's Window provides a useful method for communicating model uncertainty to end users.

3.4 Interpreting the Log Score

The improvement in predictive score for Occam's Window over the model with highest posterior probability is $\delta = 13.8$ points. The test data has $n_{\text{test}} = 149$ cases, and so this means that, on average, the predictive probability of what was actually observed was bigger for Occam's Window than for the best single model by a factor of $\exp(\delta/n_{\text{test}}) = 1.097$, or by about 10%. Similarly, the predictive probability of what was actually observed was bigger

on average for MC^3 than for the best single model by a factor of $\exp(\delta/n_{\text{test}}) = 1.215$, or by about 22%.

The user can obtain this improvement at little cost, and may find it clinically useful, as the following "biased coin" analogy shows (Raftery, Madigan, and Volinsky, 1994). Suppose that we have to estimate the probability of success in a Bernouilli trial when the true probability is π. In the absence of any other information, we would guess $\hat{\pi} = 1/2$, for which our expected or large sample predictive score per observation would be $\pi \log \hat{\pi} + (1-\pi)\log(1-\hat{\pi}) = -\log 2$. Getting π exactly right (i.e., guessing $\hat{\pi} = \pi$), yields an average improvement in predictive score per observation of $\pi \log \pi + (1-\pi)\log(1-\pi) - (-\log 2)$. The value of π for which this is equal to δ/n_{test} gives an intuitive interpretation of the gain in predictive ability due to model averaging. For Occam's Window this is $\pi = 0.72$, an improvement of 0.22 on the biased coin prediction scale. For MC^3 this is $\pi = 0.82$, an improvement of 0.32 on the biased coin prediction scale.

In other words, *not* using Occam's Window is like predicting the outcome of a coin toss as if it were a fair coin when in fact it is a "72/28" coin; not using MC^3 is like predicting the outcome of a coin toss as if it were a fair coin when in fact it is an "82/18" coin. These improvements are all the more striking since model selection seems clearcut in these data and one would not necessarily expect model averaging to improve things by much.

We have noted improvements for BMA in several applications that range from modest to substantial in regression analysis and survival analysis as Table 3 (adapted from Raftery, Madigan, and Volinsky, 1994) shows.

4 Eliciting Informative Prior Distributions

When there is little prior information about the relative plausibility of the models considered, taking them all to be equally likely *a priori* is a reasonable "neutral" choice. However, Spiegelhalter *et al.* (1993) and Lauritzen *at al.* (1994) provide a detailed analysis of the benefits of incorporating informative prior distributions in Bayesian knowledge-based systems, and demonstrate improved predictive performance with informative priors. Here we consider informative prior distributions for BMA.

4.1 Prior Distributions on Model Space

BMA requires the specification of a prior distribution on the space of possible models, i.e., $\text{pr}(M_k), k = 1, \ldots, K$. All of the reported applications of BMA use a uniform prior distribution, thereby implying that all models are equally likely *a priori*. This prior distribution is intended to be "non-informative" or "vague" or "objective" in some sense, and indeed,

Table 3: Summary of improvements in predictive performance from model averaging (via Occam's Window), relative to the model with highest posterior probability.

Data	Model	δ	n_{test}	% increase in pred. prob.	Biased coin probability
1. Coronary risk factors	Discrete graphical	29.8	1381	2.2	0.60
2. Women and mathematics	Discrete graphical	5.0	892	0.6	0.56
3. Scrotal swellings	Discrete graphical	13.8	149	9.7	0.72
4. Crime and punishment	Linear regression	11.0	23	61.3	0.95
5. Lung cancer trial	Exponential regression	1.1	62	1.8	0.60
6. PBC trial	Cox regression	2.7	155	1.8	0.60

NOTE: δ is the improvement in (partial) predictive score;
n_{test} is the number of individuals in the test data set;
% increase in pred. prob. $= 100(\exp(\delta/n_{\text{test}}) - 1)$; and
% Biased coin probability is its equivalent on the biased coin scale of Section 3.
SOURCES: Data sets 1, 2: Madigan and Raftery (1994); Data set 3: this article; Data set 4: Raftery, Madigan and Hoeting (1993); Data sets 5 and 6: Raftery, Madigan, and Volinsky (1994).

much research on Bayesian methods is taken up with devising such priors. We argue however, that in the context of knowledge-based systems, or indeed in any context where the primary aim of the modeling effort is to predict the future, such prior distributions are often inappropriate; one of the primary advantages of the Bayesian approach is that it provides a practical framework for harnessing *all* available resources including prior expert knowledge.

Notwithstanding the preceding remarks, eliciting an informative prior distribution on model space from a domain expert is challenging. Madigan and Raftery (1991) suggested eliciting a prior probability for the presence of each potential link in the model and then multiplying these probabilities to provide the required prior distribution. This however makes the possibly unreasonable assumption that the presence or absence of each link is independent of the presence or absence of other links. Furthermore, it requires the domain expert to express knowledge through unobservable quantities, i.e., links in graphical models. Lichtenstein and Fischhoff (1980) suggest that it is possible to train domain experts to provide direct estimates of such quantities. However, our experience is that observable quantities are easier for domain experts to think about than abstract entities such as parameters and models. See Kadane *et al.* (1980) for a persuasive presentation of this viewpoint.

Researchers have developed methods for eliciting informative distributions via observable quantities for several different modeling situations, for example, the Bernoulli process (Winkler, 1967, Chaloner and Duncan, 1983), the normal linear model (Kadane *et al.*, 1980, Garthwaite and Dickey, 1990 and 1991), the analysis of variance (Laskey and Black, 1989),

generalized linear models (Laud *et al.*, 1992), and survival analysis (Chaloner *et al.*, 1993).

4.2 Elicitation Method

Here we propose a method for eliciting an informative prior distribution on model space via "imaginary data" (Good, 1950). Our approach is simple to implement and provides a further improvement in predictive performance for the scrotal swelling application discussed above. The basic idea is to start with a uniform prior distribution on model space, update it using imaginary data provided by the domain expert, and then use the updated prior distribution as the prior distribution for the Bayesian analysis. Ibrahim and Laud (1994) and Laud *et al.* (1992) adopt a somewhat similar approach, but in the context of linear models. Gavasakar (1988) uses imaginary data to elicit a prior distribution for a binomial parameter.

For the scrotal swelling application, one of us (DM) elicited imaginary data from the domain expert (JG) using a simple but effective computer program. It is worth noting that the domain expert has 16 years of experience practicing medicine and has considerable expertise in the diagnosis of scrotal swellings.

The program proceeds as follows:

1. Select a variable at random and chose its state (true or false) at random.

2. Chose one of the remaining variables *at random* and prompt the domain expert for a value for that variable.

3. If there are variables remaining, go to 2.

With this program, the domain expert created 95 imaginary cases in two hours (all data are available on request).

Note that the approach effectively precludes the possibility of unrealistic cases (although it could pick a rare value for the first variable). The program randomly selects from 128 (= 8 X 2 X 8) possible combinations before requiring a response from the user. This, together with the random order of presentation, is crucial. It stimulates the domain expert and prevents him from slipping into an answering pattern. In the event that the domain expert was unsure of a value, we instructed him to choose whichever state he thought was most likely for that variable. In general, this was the case for the first couple of questions on each case, but not for the later questions.

We show the predictive performance of the technique in Table 3. For ease of comparison, we reproduce the "non-informative" BMA scores from Table 2 above. Incorporation of the expert's opinion provides an improvement in predictive performance.

Table 4: Scrotal Swellings: Predictive Performance with Imaginary Data

BMA Method	*Non-informative Log Score*	*Expert Log Score*
Occam's Window	405.8	403.8
Markov Chain Monte Carlo Model Composition	390.6	387.0

The improvement in predictive score for Occam's Window with an informative prior on models over Occam's Window with a uniform prior on models is $\delta = 2.0$ points. Thus, on average, the predictive probability of what was actually observed was bigger for the informative prior by a factor of 1.014, or by about 1.4%. For MC^3 the log score improvement was $\delta = 3.6$ points and the predictive probability of what was actually observed was bigger for the informative prior by a factor of 1.024, or by about 2.4%.

Appealing to the biased coin analogy introduced in Section 3.4 above, not using the informative prior with Occam's Window is like predicting the outcome of a biased coin with true $\pi = 0.58$ as if it were a fair coin. Not using the informative prior with MC^3 is like predicting the outcome of a biased coin with true $\pi = 0.61$ as if it were a fair coin.

In our first implementation of the program, the domain expert also provided a confidence rating for each answer. However, this was very cumbersome to elicit and did not improve predictive performance, so we did not pursue it further in the present study.

Concerning how many imaginary cases to elicit, we have experimented with randomly chosen subsets of the 95 and almost always find inferior predictive performance. We are currently eliciting a larger number of imaginary cases for the prospective study mentioned in Section 2.

5 Discussion

We have considerable experience with BMA using Occam's Window and MC^3 and have mounting evidence that either method provides superior predictive performance to a single-model approach. Incorporation of prior knowledge provides a further improvement. For the scrotal swelling application with MC^3 *and* an informative prior distribution on models, we achieved a total out-of-sample predictive performance improvement of 32.6 points over the single "best" model. Thus, on average, the predictive probability of what was actually observed was bigger for the MC^3–informative prior combination by a factor of 1.245, or by about 25% over the single best model. Using the standard single model approach as against the MC^3–informative prior combination is like predicting a biased coin with true $\pi = 0.83$

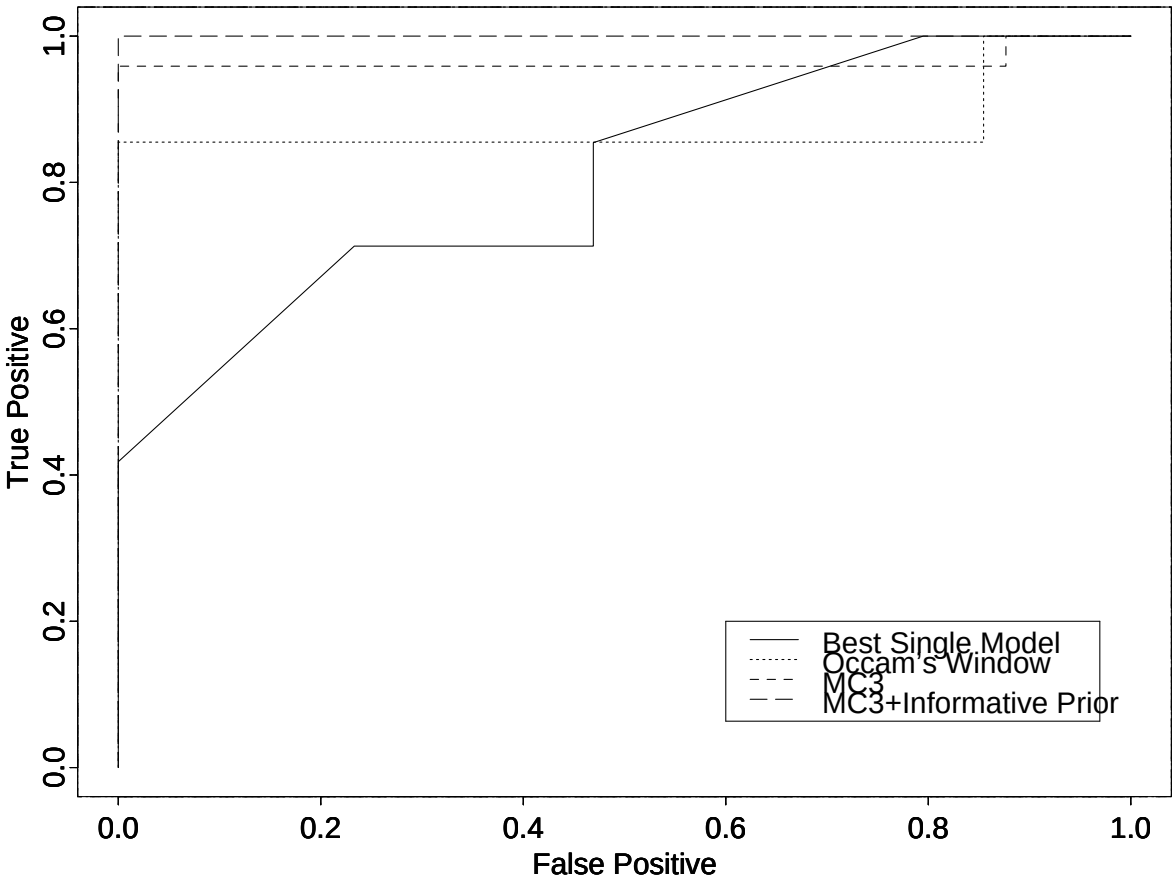


Figure 3: ROC Curves for the Scrotal Swelling exmaple.

as if it were a fair coin.

We also carried out a ROC (receiver operating characteristic) analysis for the scrotal swelling application. In Figure 3 we show ROC curves for the diagnosis variable H (Hernia). The solid ROC curve shows how well the single model with the highest posterior probability predicts variable H while the dashed curves shows the performance achieved by Occam's Window, MC^3, and the MC^3–informative prior combination.

These ROC curves show the false-positive and true-positive proportions for different probability thresholds ranging from zero to one for variable H. The MC^3–informative prior combination again provides the best performance.

We find these results especially encouraging since 8,000 miles and 8 years separated the domain expert and the data gathering.

Many issues require further study. Could graphical elicitation tools such as ElToY (Almond, 1992) be used in this context? What kind of sensitivity analysis (Laskey, 1993) is

required? Could and should the technique be used to elicit prior distributions for parameters such as probabilities? There is a large literature indicating that human judgement is frequently malleable, inconsistent and subject to bias (Kahneman *et al.* 1982, Hink and Woods, 1987); to what extent will the imaginary data technique suffer from some of these difficulties? How well would the method work with a large number of variables? However, the improved predictive performance demonstrated above suggests the technique is useful.

References

Almond, R.G. (1992) ElToY. Software available from StatLib Electronic Archive (statlib@stat.cmu.edu)

Almond, R.G. (1994) Hypergraph grammars for knowledge based model construction. *Technical Report 23*, Statistical Sciences, Inc., Seattle.

Bradshaw, J.M., Chapman, C.R., Sullivan, K.M., Almond, R.G., Madigan, D., Zarley, D., Gavrin, J., Nims, J. and Bush, N. (1992) KS-3000: An application of DDUCKS to bone-marrow transplant patient support. In *Proceedings of the Sixth Annual Florida AI Research Symposium (FLAIRS '93)*, Ft. Lauderdale, FL, 78–83.

Chaloner, K.M. and Duncan, G.T. (1983) Assessment of a beta prior distribution. *The Statistician*, **32**,174–180.

Chaloner, K.M., Church, T., Louis, T.A., and Matts, J.P. (1993) Graphical elicitation of a prior distribution for a clinical trial. *The Statistician*, **42**,341–353.

Chatfield, C. (1994) Model uncertainty, data mining, and statistical inference. *Statistics Research Report 94-01*, School of Mathematical Statistics, University of Bath.

Cooper, G.F. and Herskovits, E. (1992). A Bayesian method for the induction of probabilistic networks from data. *Machine Learning*, **9**:309–347.

Dawid, A.P. (1984) Statistical theory–The prequential approach *Journal of the Royal Statistical Society (Series A)*, **147**,278–292.

Dawid, A.P. (1986) Probability Forecasting. In *Encyclopedia of Statistical Sciences, Volume 7*, (ed. Kotz, S. and Johnson, N.L.). Wiley: New York. 210–218.

Dawid, A.P. and Lauritzen, S.L. (1993) Hyper Markov laws in the statistical analysis of decomposable graphical models. *Annals of Statistics*, **21**,1272–1317.

Draper, D., Hodges, J.S., Leamer, E.E., Morris, C.N. and Rubin, D.B. (1987) A research agenda for assessment and propagation of model uncertainty. Rand Note N-2683-RC, The RAND Corporation, Santa Monica, California.

Draper, D., Hodges, J.S., Mallows, C.L., and Pregibon, D. (1993) Exchangeability and data analysis (with discussion). *Journal of the Royal Statistical Society (Series A)*, **155**, 9-38.

Draper, D. (1994) Assessment and propagation of model uncertainty (with discussion). *Journal of the Royal Statistical Society (Series B)*, to appear.

Edwards, D. and Havránek, T. (1985) A fast procedure for model search in multidimensional contingency tables. *Biometrika*, **72** ,339–351.

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.

Gavasakar, U. (1988) A comparison of two elicitation methods for a prior distribution for a binomial parameter. *Management Science*, **34**,784–790.

Garthwaite, P.H. and Dickey, J.M. (1990) Quantifying expert opinion in linear regression models. *Journal of the Royal Statistical Society (Series B)*, **50**,462–474.

Garthwaite, P.H. and Dickey, J.M. (1991) An elicitation method for multiple linear regression models. *Journal of Behavioral Decision Making*, **4**,17–31.

Gaver, D.P., Jr., Draper, D., Goel, P.K., Greenhouse, J.B., Hedges, L.V., Morris, C.N., and Waternaux, C. (1992) *National Research Council Panel on Statistical Issues and Opportunities for Research in the Combination of Information.* National Academy Press, Washington.

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

Good, I.J. (1950) *Probability and the weighing of evidence.* Charles Griffin, London.

Hastings, W.K. (1970) Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, **57**,97–109.

Hink, R.F. and Woods, D.L. (1987) How humans process uncertain information. *AI Magazine*, **8**,41–53.

Hodges, J.S. (1987) Uncertainty, policy analysis and statistics. *Statistical Science*, **2**,259–291.

Ibrahim, J.G. and Laud, P.W. (1994) A predictive approach to the analysis of designed experiments. *Journal of the American Statistical Association*, to appear.

Kadane, J.B., Dickey, J.M., Winkler, R.L., Smith, W.S., and Peters, S.C. (1980) Interactive elicitation of opinion for a normal linear model. *Journal of the American Statistical Association*, **75**,845–854.

Kahneman, D., Slovic, P., and Tversky, A. (eds.) (1982) *Judgement under Uncertainty: Heuristics and Biases.* Cambridge: Cambridge University Press.

Kass, R.E. and Raftery, A.E. (1994). Bayes factors and model uncertainty. *Journal of the*

American Statistical Association, to appear.

Laskey, K.B. and Black, P.K. (1989) Models for elicitation in Bayesian analysis of variance. In *Computer Science and Statistics: Proceedings of the Eight Conference on the Interface*, 242–247.

Laskey, K.B. (1993) Sensitivity analysis for probability assessments for Bayesian networks. In *Proceedings of the Ninth Conference on Uncertainty in Artificial Intelligence*, D. Heckerman and A. Mamdani, (Eds.), Morgan Kaufman, San Mateo, 136–137.

Laud, P.W., Ibrahim, J.G., Gopalan, R., and Ramgopal, P. (1992) Predictive variable selection in generalized linear models. *Technical Report*, Division of Statistics, Northern Illinois University.

Lauritzen, S.L., Thiesson, B., and Spiegelhalter, D.J. (1994) Diagnostic systems created by model selection methods - A case study. In *Proceedings of the 4th International Workshop on Artificial Intelligence and Statistics*, to appear.

Lichtenstein, S. and Fischhoff, B. (1980) Training for calibration. *Organisational Behaviour and Human Performance*, **26**,149–171.

Madigan, D. and Raftery, A.E. (1991) Model selection and accounting for model uncertainty in graphical models using Occam's window. *Technical Report 213*, Department of Statistics, University of Washington.

Madigan, D. and York, J. (1993) Bayesian graphical models for discrete data. *Technical Report 259*, Department of Statistics, University of Washington. Submitted for publication.

Madigan, D., Raftery, A.E., York, J.C., Bradshaw, J.M., and Almond, R.G. (1994) Strategies for graphical model selection. In *Proceedings of the 4th International Workshop on Artificial Intelligence and Statistics*, to appear.

Madigan, D. and Raftery, A.E. (1994) Model selection and accounting for model uncertainty in graphical models using Occam's window. *Journal of the American Statistical Association*, to appear.

Matheson, J.E. and Winkler, R.L. (1976) Scoring rules for continuous probability distributions. *Management Science* **22**, 1087–1096.

Murphy, A.H. and Winkler, R.L. (1977) Reliability of subjective probability forecasts of precipitation and temperature. *Applied Statistics* **26**, 41–47.

Neal, R.M. (1993) Probabilistic inference using Markov chain Monte Carlo methods. *Technical Report CRG-TR-93-1*, Department of Computer Science, University of Toronto.

Raftery, A.E. (1988) Approximate Bayes factors for generalised linear models. *Technical Report 121*, Department of Statistics, University of Washington.

Raftery, A.E., Madigan, D., and Hoeting, J. (1993) Accounting for model uncertainty in linear regression. *Technical Report 262*, Department of Statistics, University of Washington. Submitted for publication.

Raftery, A.E., Madigan, D., and Volinsky, C.T. (1994) Accounting for model uncertainty in survival analysis improves predictive performance. In: J.O. Berger, J.M. Bernardo, A.P. Dawid, and Smith, A.F.M. (Eds.), *Bayesian Statistics V*, Oxford University Press, to appear.

Regal, R. and Hook, E. (1991) The effects of model selection on confidence intervals for the size of a closed population. *Statistics in Medicine* **10**, 717–721.

Self, M. and Cheeseman, P. (1987) Bayesian prediction for artificial intelligence. In *Proceedings of the Third Workshop on Uncertainty in Artificial Intelligence*, Seattle, 61–69.

Spiegelhalter, D.J. (1986) Probabilistic prediction in patient management and clinical trials. *Statistics in Medicine*, **5**,421–433.

Spiegelhalter, D.J. and Lauritzen, S.L. (1990) Sequential updating of conditional probabilities on directed graphical structures. *Networks*, **20**,579–605.

Spiegelhalter, D.J., Dawid, A.P., Lauritzen, S.L., and Cowell, R.G. (1993) Bayesian analysis in expert systems. *Statistical Science*, **8**,219–283.

Spiegelhalter, D.J., Thomas, A., and Best, N.G. (1994) Computation on Bayesian graphical models. In: J.O. Berger, J.M. Bernardo, A.P. Dawid, and Smith, A.F.M. (Eds.), *Bayesian Statistics V*, Oxford University Press, to appear.

Whittaker, J. (1990) *Graphical models in Applied Mathematical Multivariate Statistics*, Wiley.

Winkler, R.L. (1967) The assessment of prior distributions in Bayesian analysis. *Journal of the American Statistical Association*, **62**,1105–1120.

York, J. and Madigan, D. (1992) Bayesian methods for estimating the size of a closed population. *Technical Report 234*, Department of Statistics, University of Washington. Submitted for publication.

York, J., Madigan, D., Heuch, I. and Lie, R.T. (1994) Estimation of the proportion of congenital malformations using double sampling: Incorporating covariates and accounting for model uncertainty. *Applied Statistics*, to appear.