Strategies for Graphical Model Selection

David Madigan, University of Washington*
Adrian E. Raftery, University of Washington
Jeremy C. York, Carnegie-Mellon University
Jeffrey M. Bradshaw, Fred Hutchinson Cancer Research Center
Russell G. Almond, Statistical Sciences Inc.

1 Introduction

A typical approach to data analysis is to initially carry out a model selection exercise leading to a single "best" model and to then 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, Breslow, 1990, Draper *et al.*, 1987, Hodges, 1987, Self and Cheeseman, 1987). As a consequence uncertainty about quantities of interest can be underestimated. For striking examples of this see York and Madigan (1992), Regal and Hook (1991) and Draper *et al.* (1987).

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

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

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.

Hodges (1987) argues that "what is clear is 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 a logarithmic scoring rule, than using any single model M_j (Madigan and Raftery, 1991, hereafter referred to as MR).

However, implementation of the above strategy is difficult. There are two primary reasons for this: firstly, the integrals in (3) can be hard to compute, and secondly, the number of terms in (1) can be enormous.

*Department of Statistics, GN-22, University of Washington, Seattle WA 98195 (madigan@stat.washington.edu). This work is supported in part by a NSF grant to the University of Washington and by a NIH Phase I SBIR Award "Computing environments for graphical belief modeling" to Statistical Sciences.

For graphical models for discrete data, efficient solutions to the former problem have been developed. Two approaches to the latter problem, i.e. the enormous number of terms in (1), have recently been proposed. 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 (1992) on the other hand suggest directly approximating (1) with a Markov chain Monte Carlo method.

MR examined the predictive performance of their method. Our purpose in this paper is to examine the predictive performance of the Markov chain Monte Carlo method and compare the predictive performance of both approaches. This work is of direct relevance to probabilistic knowledge-based systems systems where model uncertainty abounds (Bradshaw *et al.*, 1992).

2 Model Selection and Occam's Window

Two basic principles underly the approach presented in MR. Firstly, they argue that if a model predicts the data far less well than the model which provides the best predictions, then it has effectively been discredited and should no longer be considered. 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. Secondly, appealing to Occam's razor, they exclude complex models which receive less support from the data than their simpler counterparts. More formally they also exclude from (1) models belonging to:

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

and equation (1) is replaced by

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

where

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

This greatly reduces the number of models in the sum in equation (1) and now all that is required is a search strategy to identify the models in $\mathcal{A}$. Two further principles underly the search strategy. Firstly, if a model is rejected then all its submodels are rejected. This is justified by appealing to the independence properties of the models. The second principle — "Occam's Window" — concerns the interpretation of the ratio of posterior model probabilities $\mathrm{pr}(M_0 \mid D)/\mathrm{pr}(M_1 \mid D)$. Here M_0 is one link "smaller" than M_1. The essential idea is that 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) neither model is rejected. MR set the edges of the window at $\frac{1}{20}$ and 1.

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 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 Tierney (1991), 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 $\text{pr}(M_i \mid D)$. Then for any 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 $g(M) = \text{pr}(\Delta \mid M, D)$.

To construct the Markov chain we define a neighborhood $\text{nbd}(M)$ 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 $q(M \rightarrow M') = 0$ for all $M' \notin \text{nbd}(M)$ and $q(M \rightarrow M')$ constant for all $M' \in \text{nbd}(M)$. If the chain is currently in state M, we proceed by drawing M' from $q(M \rightarrow M')$. If the model is decomposable it is then accepted with probability:

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

Otherwise the chain stays in state M. It has been our experience that this process is highly mobile and runs of 10,000 or less are typically adequate.

4 Analysis

The efficacy of a modeling strategy can be judged by how well the resulting "models" predict future observations (Self and Cheeseman, 1987). We have assessed the predictive performance of Markov chain Monte Carlo model composition (MC^3) method for the three examples considered by MR. Their results are reproduced for comparison purposes. In each case we have started the Markov chain at a randomly chosen model, ran the chain for 100,000 iterations discarding the first 10,000. The data sets each have between six and eight binary variables. Performance, measured by the logarithmic scoring rule, is assessed by randomly splitting the complete data sets into two subsets. One subset, containing 25% of the data, is used to select models with the other subset being used as set of test cases. Repeating the random split, varying the subset proportions, or starting the Markov chain from a different location produces very similar results.

The first example concerns data on 1,841 men cross-classified according to risk factors for Coronary Heart Disease. This data set was previously analysed by Edwards and Havránek (1985) and others. The risk factors are as follows: A, smoking; B, strenuous mental work; C, strenuous physical work; D, systolic blood pressure; E, ratio of β and α proteins; F, family anamnesis of coronary heart disease.

The second example concerns a survey which was reported in Fowlkes *et al.* (1988) concerning the attitudes of New Jersey high-school students towards mathematics. A total of 1190 students in eight schools took part in the survey. The variables collected were: A, lecture attendance; B, Sex; C, School Type (suburban or urban); D, "I'll need mathematics in my future work" (agree or disagree); E, Subject Preference (maths/science or liberal arts); F, Future Plans (college or job);

The final example concerns the diagnosis of scrotal swellings. Data on 299 patients were presented in MR, 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.

Results are presented in Tables 1, 2 and 3 for each of the examples. Given in each case are the models selected by MR and the logarithmic scoring rule summed over the test cases for each individual model. Next the score resulting from averaging over these models is given. For the Coronary Heart Disease example, the score is also included for the model selected by Whittaker (1990) on the basis of the full data set. This represents the score that would result from using a typical model selection procedure. Finally the score for MC^3 is given.

In each case, methods that average over models, provide predictive performance which is superior to the performance resulting from basing the inference on any single model which might reasonably have been

Table 1: Coronary Heart Disease: Predictive Performance

Model	*Posterior probability %*	Logarithmic Score
$[AE][BC][BE][DE][F]$	26	4986.7
$[AC][BC][BE][DE][F]$	16	4980.9
$[AC][AE][BC][DE][F]$	13	4981.0
$[A][BC][BE][DE][F]$	9	4989.4
$[AE][BC][BE][D][F]$	8	4987.4
$[AE][BC][DE][F]$	7	4989.5
$[AC][BC][BE][D][F]$	5	4981.6
$[AC][BC][DE][F]$	4	4983.7
$[AC][AE][BC][D][F]$	4	4981.7
$[A][BC][BE][D][F]$	3	4990.1
$[A][BC][DE][F]$	2	4992.2
$[AE][BC][D][F]$	2	4990.2
$[AC][BC][D][E][F]$	1	4984.4
$[ABCE][ADE][BF]$	Whittaker	4990.2
Model Averaging		4953.6
Markov Chain Monte Carlo Model Composition		4933.7

Table 2: 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
Markov Chain Monte Carlo Model Composition		3271.5

Table 3: Scrotal Swellings: 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
Markov Chain Monte Carlo Model Composition		590.1

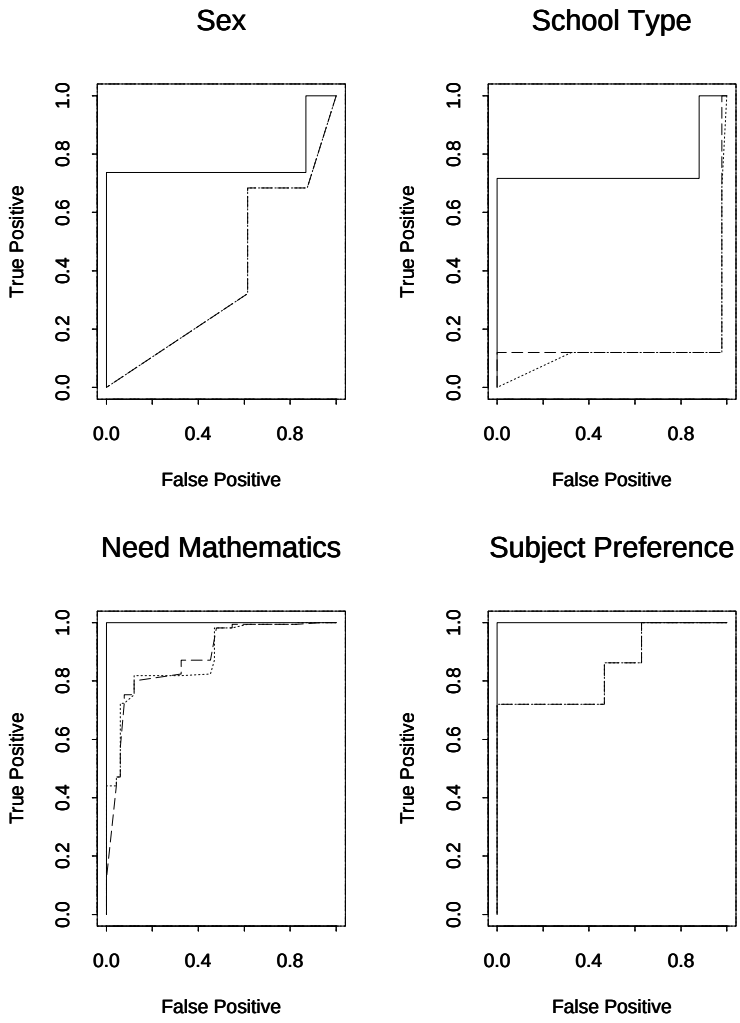

Figure 1: Women and Mathematics: ROC Curves

selected. In the coronary heart disease data for example, the Occam's window models outperform the "best" model (i.e. that with the highest posterior probability) by 33 points of log predictive probability, or 66 points on the scale of twice the log probability on which deviances are measured. MC^3 provides a further performance improvement of 20 points (or 40 points on the deviance scale).

A ROC analysis was also carried out for each of the examples and in Figure 1 we show the ROC curves for four of the variables in the women and mathematics data set. Here 25% of the data was used for testing. The dashed ROC curves show how well the model with the highest posterior probability performs, the dotted curves show the performance averaging over the models in Occam's window, while the solid curves are for MC^3. For each of the variables, MC^3 provides substantially improved performance. Such clear differences do not occur in each of the examples, although typically, methods which average over models provide superior ROC curves.

It is clear that model averaging improves predictive performance. MC^3 generally provides superior performance. However, the insight into model uncertainty provided by the Occam's window method will be important in many applications.

References

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. Submitted to the *Sixth Annual Florida AI Research Symposium (FLAIRS '93)*, Ft. Lauderdale, FL.

Breslow, N. (1991) Biostatistics and Bayes. *Statistical Science*, **5**,269–298.

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.

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.

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

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

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. (1992) Bayesian graphical models for discrete data. Submitted for publication.

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

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.

Tierney, L. (1991) Markov chains for exploring posterior distributions *Technical Report 560*, School of Statistics, University of Minnesota.

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

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