Bayesian Model Averaging

David Madigan, Adrian E. Raftery,
& Chris T. Volinsky
Department of Statistics, Box 354322
University of Washington
Seattle, WA 98195-4322
{madigan,raftery,volinsky}@stat.washington.edu

Jennifer A. Hoeting
Department of Statistics
B259 Clark Building
Colorado State University
Fort Collins, CO 80523-1877
jah@stat.colostate.edu

Abstract

Standard statistical practice ignores model uncertainty. Data analysts typically select a model from some class of models and then proceed as if the model generated the data. This approach ignores a component of uncertainty, leading to over-confident inferences. Bayesian model averaging (BMA) provides a coherent mechanism for accounting for this model uncertainty. Several methods for implementing BMA have recently emerged. We discuss these methods and provide pointers to a number of applications. In these applications, BMA provides improved out-of-sample predictive performance. We provide a catalogue of currently available BMA software.

Introduction

Consider the following scenario: a researcher has gathered data concerning cancer of the esophagus. For each of a large number of patients, the researcher has recorded a variety of demographic and medical covariates, along with each patient's last known survival status. The researcher would like to assess the size of each covariate's effect on survival time with a view to designing future interventions, and additionally, would like to be able to predict the survival time of specific patients. The researcher decides to use proportional hazards regression models to analyze the data. Next the researcher conducts a data-driven search to select covariates for the specific proportional hazards regression model, M, that will provide the framework for subsequent inference. The researcher checks that M fits the data reasonably well and notes that the parameter estimates look sensible. The researcher proceeds to use M to estimate effect sizes and associated standard errors, and make predictions.

This may approximate standard statistical practice, but is it entirely satisfactory? Suppose there exists an alternative proportional hazards model, M^*, that also provides a good fit to the data but leads to substantively different estimated effect sizes or different standard errors or different predictions. What then? Models like M^* are commonplace – for particularly striking examples see (Regal & Hook 1991), (Draper 1995), and (Madigan & York 1995). Even if no single troublesome M^* exists, basing inference on M alone is risky; presumably, ambiguity over the model should dilute information about effect sizes and predictions, since "part of the evidence is spent to specify the model" (Leamer 1978, p.91). (Draper et al. 1987), (Hodges 1987), and (Raftery 1988) make essentially the same observation.

Bayesian model averaging (BMA) provides a solution to this problem. If Δ is the quantity of interest, such as an effect size, a future observable, 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, the posterior probability for model M_k is given by:

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

where

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

is the marginal likelihood of model M_k, θ_k is the vector of parameters of model M_k (e.g., for regression $\theta = (\beta, \sigma^2)$), $\mathrm{pr}(\theta_k \mid M_k)$ is the prior density of θ_k under model M_k, $\mathrm{pr}(D \mid \theta_k, M_k)$ is the likelihood, and $\mathrm{pr}(M_k)$ is the prior probability that M_k is the true model. All probabilities are implicitly conditional on $\mathcal{M}$, the set of all models being considered.

(Madigan & Raftery 1994) note that 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. Considerable empirical evidence supports this theoretical claim; we will review some of this evidence.

Implementation of Bayesian model averaging presents several difficulties:

- The integrals implicit in (1) can in general be hard to compute.
- The number of terms in (1) can be enormous, rendering exhaustive summation infeasible.
- Specification of $\text{pr}(M_k)$, the prior distribution over competing models, is challenging and has received scant attention.
- Choosing the class of models to average over becomes the fundamental modeling task and at least three competing schools of thought have emerged.

We will discuss these difficulties and present recent progress towards solutions. We will draw on the related reviews of (Draper 1995) and (Chatfield 1995), although we will place more emphasis on practical applications of the BMA methodology. Space does not permit us to review related work on "multiple models" from the machine learning, neural network, computational learning theory, and artificial intelligence communities.

Implementing Bayesian Model Averaging

Computing BMA's Integrals

For certain interesting classes of models such as discrete graphical models (Madigan & York 1995) and linear regression (Raftery, Madigan, & Hoeting 1993), closed form integrals *are* available. In other cases, Laplace approximations can often provide an excellent approximation to $\text{pr}(D \mid M_k)$. (Taplin 1993) suggested approximating $\text{pr}(\Delta \mid M_k, D)$ by $\text{pr}(\Delta \mid M_k, \hat{\theta}, D)$ where $\hat{\theta}$ is the maximum likelihood estimate of the parameter vector θ. (Draper 1995), (Raftery 1992), (Raftery, Madigan, & Volinsky 1996), and (Volinsky, Madigan, & Raftery 1996) demonstrate its usefulness in the BMA context. In later sections we discuss these approximations in more detail in the context of specific model classes.

Managing the Summation

The size of most interesting model classes renders the exhaustive summation of Equation 1 impractical. In our own work, we have adopted two distinct approaches to this problem. The Occam's Window method of (Madigan & Raftery 1994) does not attempt to approximate (1) but instead, appealing to standard norms of scientific investigation, averages over a set of parsimonious, data-supported models. To rapidly identify the models in Occam's Window, (Volinsky, Madigan, & Raftery 1996) use the "leaps and bounds" algorithm (see below).

Two basic principles underly the Occam's Window method. First, (Madigan & Raftery 1994) 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. Second, appealing to Occam's razor, they exclude complex models which receive less support from the data than their simpler counterparts.

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 Occam's Window. Two further principles underly the search strategy. First, when the algorithm compares two nested models and decisively rejects the simpler model, then all submodels of the simpler model are rejected. The second principle 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 is "smaller" than M_1. The essential idea is this: 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. (Madigan & Raftery 1994) adopted $\frac{1}{20}$ and 1 as the two extremes of the Window.

(Raftery, Madigan, & Volinsky 1996) adopted the strategy of averaging over *all* models with posterior probability within factor of 20 of the model with the highest posterior probability. In their example, this simpler strategy provided improved predictive performance.

Our second approach, Markov chain Monte Carlo model composition (MC^3), uses a Markov chain Monte Carlo method to directly approximate (1) (Madigan & York 1995). This generates a stochastic process which moves through model space. Specifically, let $\mathcal{M}$ denote the space of models under consideration. We can construct a Markov chain $\{M(t)\}, t = 1, 2, \ldots$ with state space $\mathcal{M}$ and equilibrium distribution $\text{pr}(M_i \mid D)$. Then for a function $g(M_i)$ defined on $\mathcal{M}$, if we simulate this Markov chain for $t = 1, \ldots, N$, the average:

$$\hat{G} = \frac{1}{N} \sum_{t=1}^{N} g(M(t)) \qquad (4)$$

is an estimate of $E(g(M))$. Applying the ergodic theorem for finite irreducible Markov chains,

$$\hat{G} \to \mathbf{E}(g(M)) \; a.s. \text{ as } N \to \infty.$$

To compute (1) in this fashion set $g(M) = \mathrm{pr}(\Delta \mid M, D)$.

To construct the Markov chain we define a neighbourhood $\mathrm{nbd}(M)$ for each $M \in \mathcal{M}$. For example, with graphical models the neighborhood might be the set of models with either one link more or one link fewer than M and the model M itself (Madigan et al. 1994). Define a transition matrix q by setting $q(M \rightarrow M') = 0$ for all $M' \notin \mathrm{nbd}(M)$ and $q(M \rightarrow M')$ non–zero for all $M' \in \mathrm{nbd}(M)$. If the chain is currently in state M, we proceed by drawing M' from $q(M \rightarrow M')$; if M' is a legitimate model (it contains no directed cycles in the directed case and is chordal in the undirected case) it is accepted with some positive probability chosen so that the process has the correct stationary distribution.

MC^3 offers considerable flexibility. For example, working with equivalence classes of acyclic directed graphical models, (Madigan et al. 1996) introduce a total ordering of the vertices into the stochastic process as an auxiliary variable, thereby providing a three-fold computational speed-up. (York et al. 1995) incorporate missing data and a latent variable into their MC^3 scheme. However, as with other Markov chain Monte Carlo methods, convergence issues can be problematic.

Earlier related work includes (Stewart 1987) who uses importance sampling to average across logistic regression models, and (Carlin & Polson 1991) who use Gibbs sampling to mix models with different error distributions. (Besag et al. 1995, Section 5.6) use a Markov chain Monte Carlo approach to average across families of t-distributions.

For linear models, (Raftery, Madigan, & Hoeting 1993) apply MC^3 to average across models with many predictors. (Clyde, DeSimone, & Parmigiani 1994) introduced an importance sampling strategy based on orthogonalizing the predictor space. Their approach is dramatically more efficient than MC^3 in that context. The Gibbs sampling method of (George & McCulloch 1993) also has application to BMA.

Stochastic methods that move simultaneously in model space and parameter space open up a limitless range of applications for BMA. Since the dimensionality of the parameter space generally changes with the model, standard methods do not apply. However, recent work by (Carlin & Chib 1994), (Phillips & Smith 1994), and (Green 1996) provides potential solutions.

Implementation Details for Specific Model Classes

In this Section we present various context-specific implementation details.

Linear Regression: Predictors, Outliers and Transformations

The selection of subsets of predictor variables is a basic part of building a linear regression model. The objective of variable selection is typically stated as follows: given a dependent variable Y and a set of a candidate predictors $X_1, X_2, \ldots, X_k$, find the "best" model of the form

$$Y = \beta_0 + \sum_{j=1}^{p} \beta_{i_j} X_{i_j} + \epsilon,$$

where $X_{i_1}, X_{i_2}, \ldots, X_{i_p}$ is a subset of $X_1, X_2, \ldots, X_k$. Here "best" may have any of several meanings, e.g., the model providing the most accurate predictions for new cases exchangeable with those used to fit the model.

Using the standard normal-gamma conjugate class of priors, (Raftery, Madigan, & Hoeting 1993) provide a closed form expression for the likelihood, an extensive discussion of hyperparameter choice in the situation where little prior information is available, and BMA implementation details for both Occam's Window and MC^3.

(Hoeting, Raftery, & Madigan 1995a,b; hereafter HRMa and HRMb) extend this framework to include transformations and outliers respectively.

HRMa used the Box-Cox class of power transformations for the response. The Box-Cox class of power transformations changes the problem of selecting a transformation into one of estimating a parameter. The model is $Y^{(\rho)} = X\beta + \epsilon$ where $\epsilon \sim N(0, \sigma^2 I)$ and

$$y^{(\rho)} = \begin{cases} \frac{y^\rho - 1}{\rho} & \rho \neq 0 \\ \log(y) & \rho = 0 \end{cases}.$$

While the class of power transformations is mathematically appealing, power transformations are typically not interpretable unless they are limited to a few possible values of ρ. RMHa averaged over the values (-1, 0, .5, 1), so that the transformed predictors can be interpreted as the reciprocal, the logarithm, the square root, and the untransformed response.

For transformation of the predictors, however, they proposed an approach consisting of an initial exploratory use of the Alternating Conditional Expectation algorithm (ACE), followed by change point transformations if needed. RMHa's BMA averages over all predictor transformations for which the evidence exceeds a user-specified level. This is accomplished simply by including the transformed predictors as extra covariates for consideration in potential models.

HRMb average over sets of predictors *and* possible outliers. They adopt a variance–inflation model for outliers as follows: Let $Y = X\beta + \epsilon$ where the observed data on the predictors are contained in the $n \times (p +$

1) matrix X and the observed data on the dependent variable are contained in the n-vector Y. They assume that the ϵ's in distinct cases are independent where

$$\epsilon \sim \begin{cases} \mathrm{N}\left(0,\sigma^2\right) & \text{w.p.} \quad (1-\pi) \\ \mathrm{N}\left(0,K^2\sigma^2\right) & \text{w.p.} \quad \pi. \end{cases} \tag{5}$$

Here π is the probability of an outlier and K^2 is the variance–inflation parameter.

Their simultaneous variable and outlier selection (SVO) method involves two steps. In a first exploratory step they use a robust technique to identify a set of potential outliers. The robust approach typically identifies a large number of potential outliers. In the second step, HRMb compute all possible posterior model probabilities or use MC^3, considering all possible subsets of the set of potential outliers. This two–step method is computationally feasible, and it allows for groups of observations to be considered simultaneously as potential outliers. RMHb provide evidence that SVO successfully identifies masked outliers.

Recently, Hoeting has developed SVOT which averages over variables, transformations, and outliers.

Generalized Linear Models

Model-building for generalized linear models involves choosing the independent variables, the link function, and the variance function (McCullagh & Nelder 1989). Each possible combination of choices defines a different model. (Raftery 1996) presents a series of methods for calculating approximate Bayes factors for generalized linear models. The Bayes factors, in turn, trivially yield posterior model probabilities.

The relative error of Raftery's approximation is $O(n^{-\frac{1}{2}})$. However, in the case where the canonical link function is used, the relative error improves to $O(n^{-1})$.

Survival Analysis and LEAPS

Methods for analyzing survival data often focus on modelling the hazard rate. The most popular way of doing this is to use the Cox proportional hazards model (Cox 1972), which allows different hazard rates for cases with different covariate vectors, and leaves the underlying common baseline hazard rate unspecified. The Cox model specifies the hazard rate for subject i with covariate vector $\mathbf{x_i}$ to be

$$\lambda(t \mid \mathbf{x_i}) = \lambda_0(t)\exp(\mathbf{x_i}^T\theta)$$

where $\lambda_0(t)$ is the baseline hazard function at time t, and θ is a vector of unknown parameters.

The estimation of θ is commonly based on the partial likelihood, namely

$$PL(\theta) = \prod_{i=1}^{n} \left(\frac{\exp(\mathbf{x}_i^T\theta)}{\sum_{\ell \in R_i}\exp(\mathbf{x}_\ell^T\theta)} \right)^{w_i},$$

where R_i is the risk set at time t_i (i.e. the set of subjects who have not yet experienced an event) and w_i is an indicator for whether or not patient i is censored.

Data analysts often use Cox regression models in conjunction with a variable selection method. Such methods usually proceed in a stepwise fashion and aim to choose the "best" subset of the full covariate list (Fleming & Harrington 1991). Subsequent inference conditions on this subset.

Since the integrals required for BMA do not have a closed form solution for Cox models, (Raftery, Madigan & Volinsky 1996) and (Volinsky, Madigan, Raftery, & Kronmal 1996, VMRK hereafter) adopted a number of approximations. In particular, VMRK used the MLE approximation:

$$\mathrm{pr}(\Delta \mid M_k, D) \approx \mathrm{pr}(\Delta \mid M_k, \hat{\theta}_k, D),$$

and the BIC approximation:

$$\log \mathrm{pr}(D \mid M_k) = \log \mathrm{pr}(D \mid \hat{\theta}_k, M_k) - d_k \log n + O(1).$$

VMRK used the Occam's Window approach to BMA for Cox models and adapted the "Leaps and Bounds" algorithm of (Furnival & Wilson 1974) to provide an efficient method for identifiying the models in the Window. The leaps and bounds algorithm provides the top q models of each model size, where q is designated by the user, plus the MLE $\hat{\theta}_k$, and var$(\hat{\theta}_k)$ for each model k returned. (Lawless & Singhal 1978) provided a modified algorithm for nonlinear regression models that provides an approximate likelihood ratio test statistic, and hence an approximate BIC value.

As long as q is large enough, this procedure returns the models in Occam's window ($\mathcal{A}$) plus many models not in $\mathcal{A}$. VMRK use the approximate LRT to reduce the remaining subset of models to those most likely to be in $\mathcal{A}$. This reduction step keeps only the models whose posterior probabilities fall within a factor C of the model with the best posterior model probabilty, where C is greater than 20. VMRK set $C = 20^2$ and almost no models in $\mathcal{A}$ were lost. A standard software can then analyze the remaining models, calculate the exact BIC value for each one, and eliminate those models not in $\mathcal{A}$.

For the models in $\mathcal{A}$, VMRK calculate a posterior model probability by normalizing over the model set. Parameter estimates and standard errors of those estimates derive from weighted averages of the estimates and errors from the individual models, using the posterior model probabilities as weights. The posterior probability that a regression coefficient for a variable is non-zero ("posterior effect probability") is simply the sum of posterior probabilities of the models which contain that variable. In the context of the Cardiovascular

Health Study, an on-going, multicenter investigation into risk factors for stroke, VMRK demonstrate that these posterior effect probabilities often lead to substantive interpretations that are at odds with usual the P-values, but admit more direct interpretation. VMRK also show that BMA provides substantively improved out-of-sample predictive performance over any single model that might reasonably have been selected.

Graphical Models: Missing Data and Auxilliary Variables

A *graphical model* is a statistical model embodying a set of conditional independence relationships which may be summarized by means of a graph. To date, most graphical models research has focused on acyclic digraphs, chordal undirected graphs, and chain graphs that allow both directed and undirected edges, but have no partially directed cycles. Some classes of graphical models coupled with particular parametric assumptions yield closed-form expressions for complete-data likelihoods and posterior model probabilities (Heckerman, Geiger, & Chickering 1994), and BMA proceeds in a straightforward manner.

(York et al. 1995) consider an example involving both missing data and latent variables. Let Y the observed data and Z the missing values (i.e., the latent variables and the missing data). The posterior distribution of the quantity of interest can then be re-expressed as follows :

$$\mathrm{pr}(\Delta \mid Y) = \mathbf{E}(\mathrm{pr}(\Delta \mid M_k, Y, Z) \mid Y).$$

This can be numerically approximated by simulating a process $\{Z(t), M(t)\}$ with stationary distribution $\mathrm{pr}(Z, M \mid Y)$.

(York et al. 1995) proceed by also including θ in the simulation scheme. At each iteration, they first draw $M(t)$ from the conditional distribution

$$\mathrm{pr}(M|Y,Z(t-1)) \propto \mathrm{pr}(M) \int \mathrm{pr}(Y,Z(t-1)|M,\theta)\mathrm{pr}(\theta|M)d\theta \tag{6}$$

A conjugate prior for θ makes this integration straightforward. Next, they draw $\theta(t)$ from $\mathrm{pr}(\theta|Y,Z(t-1),M)$, which amounts to simulating values from independent beta distributions. These two steps together provide values of M and θ that are drawn jointly from $\mathrm{pr}(M,\theta|Y,Z(t-1))$. Finally, they draw from the distribution $\mathrm{pr}(Z|Y,M(t),\theta(t))$ to obtain $Z(t)$.

The simulated values of Z and M can then provide an estimate of arbitrary posterior expectations of functions of Δ. Since the quantity Δ is a function of the model and its parameters, one can also use the simulated values of $\theta(t)$ and $M(t)$ to estimate the expectation of Δ :

$$\lim_{N\to\infty} \frac{1}{N} \sum_{t=1}^{N} \Delta(\theta(t), M(t)) \to \mathbf{E}(\Delta \mid Y) \tag{7}$$

by the same argument as above.

Specifying Prior Model Probabilities

When there is little prior information about the relative plausibility of the models considered, taking them all to be equally likely *a priori* may be a reasonable "neutral" choice. However, (Spiegelhalter et al. 1993) and (Lauritzen, Thiesson, & Spiegelhalter 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.

In the context of graphical models, (Madigan & Raftery 1994) and others have suggested eliciting a prior probability for the presence of each potential link and then multiplying these probabilities to provide the required prior distribution. Similar ideas apply in regression modeling.

(Madigan, Gavrin, & Raftery 1995) provide a simple method for discrete data applications and demonstrate that an informative $\mathrm{pr}(M_k)$ provides improved predictive performance in their particular application. The method elicits an informative prior distribution on model space via "imaginary data" (Good 1950). The basic idea is to start with a uniform prior distribution on model space, update it using imaginary data provided by the domain expert (the number of imaginary cases will depend on the application and the available resources), and then use the updated prior distribution as the actual prior distribution for the Bayesian analysis. (Ibrahim & Laud 1994) adopt a somewhat similar approach in the context of linear models.

Predictive Performance

A primary purpose of statistical analysis is to forecast the future. Measuring how well a model predicts future observations is one way to judge the efficacy of a BMA strategy. The *logarithmic scoring rule* is a *proper scoring rule* as defined by (Matheson & Winkler 1976) and others and provides an omnibus measure of predictive performance. The logarithmic scoring rule measures the predictive ability of an individual model, M, with:

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

and measures the predictive performance of BMA with:

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

Table 1 on the next page presents the out-of-sample log score for a variety of applications. We have recorded similar improvements in several other applications.

Software

Currently available software from StatLib includes:

- `glib` (BMA for Generalized Linear Models - user specified models: accurate Bayes Factors)
- `bicreg` (BMA for Linear Models)
- `bic.logit` (BMA for logistic regression - uses leaps)
- `bic.surv` (BMA for Cox models)
- `bic.glm` (BMA for Generalized Linear Models using leaps) - soon to appear.

Statlib is the premier Web repository for research statistical software. Statlib's address is: http://lib.stat.cmu.edu/.

Acknowledgments. Research supported in part by the U.S. National Science Foundation and the U.S. Office of Naval Research (N00014-91-J-1014). The authors are grateful to David Lewis and Robert Schapire for helpful advice.

References. Besag, J.E., Green, P., Higdon, D., and Mengerson, K. 1995. Bayesian computation and stochastic systems. *Statistical Science*, **10** 3-66.

Carlin, B.P. and Polson, N.G. 1991. Inference for nonconjugate Bayesian models using the Gibbs sampler. *The Canadian Journal of Statistics*, **19**, 399–405.

Carlin, B.P. and Chib, S. 1993. Bayesian model choice via Markov chain Monte Carlo. *Journal of the Royal Statistical Society (Series B)*, **55**, 473–484.

Chatfield, C. 1995. Model uncertainty, data mining, and statistica inference (with discussion). *Journal of the Royal Statistical Society (Series A)*, **158**, 419–466.

Clyde, M.A., DeSimone, H., and Parmigiani, G. 1994. Prediction via orthogonalized model mixing. *Technical Report DP-92-32*, ISDS, Duke University.

Cox, D.R. 1972. Regression models and life tables (with discussion). *Journal of the Royal Statistical Society (Series B)*, **34**, 187–220.

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. 1995. Assessment and propagation of model uncertainty (with discussion). *Journal of the Royal Statistical Society (Series B)*, **57**, 45–97.

Fleming T.R. and Harrington, D.H. 1991. *Counting Processes and Survival Analysis.* John Wiley and Sons.

Furnival, G. M. and Wilson, R.W. 1974. Regression by leaps and bounds. *Technometrics*, **16**, 499–511.

George, E.I. and McCulloch, R. 1993. Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, **88**, 881-889.

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

Green, P.J. 1996. Reversible jump MCMC computation and Bayesian model determination. *Biometrika*, to appear.

Heckerman, D., Geiger, D., and Chickering, D.M. 1994. Learning Bayesian networks: The combination of knowledge and statistical data. In *Uncertainty in Artificial Intelligence, Proceedings of the Tenth Conference* (R. Lopez de Mantaras and D. Poole, eds.), San Francisco: Morgan Kaufmann, p.293–301.

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

Hoeting, J.A., Raftery, A.E., and Madigan, D. 1995a) Simultaneous Variable and Transformation Selection in Linear Regression. *Computational Statistics and Data Analysis*, to appear.

Hoeting, J.A., Raftery, A.E., and Madigan, D. 1995b) A Method for Simultaneous Variable Selection and Outlier Identification in Linear Regression. Submitted for publication.

Ibrahim, J.G. and Laud, P.W. 1994. A predictive approach to the analysis of designed experiments. *Journal of the American Statistical Association*, **89**, 309–319.

Lauritzen, S.L., Thiesson, B., and Spiegelhalter, D.J. 1994. Diagnostic systems created by model selection methods - A case study. In: *Selecting Models from Data: Artificial Intelligence and Statistics IV*, P. Cheeseman and W. Oldford (Eds.), Springer Verlag, 143–152.

Lawless, J. and Singhal, K. 1978. Efficient screening of nonnormal regression models. *Biometrics*, **34**, 318–327.

Leamer, E. E. 1978. *Specification Searches. Ad Hoc Inference with Nonexperimental Data.* Wiley: New York.

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* **89**, 1335–1346.

Madigan, D., Raftery, A.E., York, J.C., Bradshaw, J.M., and Almond, R.G. 1994. Strategies for graphical model selection. In: *Selecting Models from Data: Artificial Intelligence and Statistics IV*, P. Cheeseman and W. Oldford (Eds.), Springer Verlag, 91–100.

Table 1: 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.
1. Coronary risk factors	Discrete graphical	29.8	1381	2.2
2. Women and mathematics	Discrete graphical	5.0	892	0.6
3. Scrotal swellings	Discrete graphical	11.1	224	5.1
4. Crime and punishment	Linear regression	11.0	23	61.3
5. Lung cancer trial	Exponential regression	1.1	62	1.8
6. PBC trial	Cox regression	2.7	155	1.8

Note: δ is the improvment in log predictive score; n_{test} is the number of individuals in the test data set; % increase in pred. prob. $= 100(\exp(\delta/n_{\text{test}}) - 1)$;
Sources: Data sets 1, 2, 3: (Madigan & Raftery 1994); Data set 4: (Raftery, Madigan, & Hoeting 1993); Data sets 5 and 6: (Raftery, Madigan, Volinsky, & Kronmal 1996).

Madigan, D., Gavrin, J., and Raftery, A.E. 1995. Eliciting prior information to enhance the predictive performance of Bayesian graphical models. *Communications in Statistics - Theory and Methods* **24**, 2271–2292.

Madigan, D. and York, J. 1995. Bayesian graphical models for discrete data. *International Statistical Review* **63**, 215–232.

Madigan, D., Andersson, S.A., Perlman, M.D., and Volinsky, C.T. 1996. Bayesian Model Averaging and Model Selection for Markov Equivalence Classes of Acyclic Digraphs. *Communications in Statistics - Theory and Methods*, to appear.

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

McCullagh, P. and Nelder, J.A. 1989. *Generalized Linear Models* (2nd ed.). London: Chapman and Hall.

Philips, D.B. and Smith, A.F.M. 1994) Bayesian model comparison via jump diffusions. *Technical Report 94-20*, Imperial College, London.

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

Raftery, A.E. 1992. Bayesian model selection in structural equation models. In *Testing Structural Equation Models* (eds. K.A. Bollen and J.S. Long), Beverly Hills: Sage.

Raftery, A.E. 1996. Approximate Bayes factors and accounting for model uncertainty in generalized linear models. *Biometrika*, to appear.

Raftery, A.E., Madigan, D., and Hoeting, J.1993. Model uncertainty and accounting for model uncertainty in linear regression models. *Technical Report 262*, Department of Statistics, University of Washington.

Raftery, A.E., Madigan, D., and Volinsky, C.T. 1996. Accounting for model uncertainty in survival analysis improves predictive performance. *Bayesian Statistics V*, to appear.

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.

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

Stewart, L. 1987. Hierarchical Bayesian analysis using Monte Carlo integration: Computing posterior distributions when there are many possible models. *The Statistician*, **36**, 211–219.

Taplin, R.H. 1993. Robust likelihood calculation for time series. *Journal of the Royal Statistical Society (Series B)*, **55**, 829–836.

Volinsky, C.T., Madigan, D., Raftery, A.E., and Kronmal, R.A. 1996. Bayesian model averaging in proportional hazards models: Predicting strokes. Technical Report, Department of Statistics, University of Washington.

York, J., Madigan, D., Heuch, I., and Lie, R.T. 1995. Estimation of the proportion of congenital malformations using double sampling: Incorporating covariates and accounting for model uncertainty. *Applied Statistics* **44**, 227–242.