Bayesian Model Averaging: A Tutorial*†

Jennifer A. Hoeting
Colorado State University

David Madigan, Adrian E. Raftery
University of Washington

Chris T. Volinsky
AT&T Labs

Appeared in: *Statistical Science* (1999) **14**:4, with discussion
Note: this version does not have the typesetting errors that occurred in the printing of the version in *Statistical Science*.

Abstract

Standard statistical practice ignores model uncertainty. Data analysts typically select a model from some class of models and then proceed as if the selected model had generated the data. This approach ignores the uncertainty in model selection, leading to over-confident inferences and decisions that are more risky than one thinks they are. 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 present a number of examples. In these examples, BMA provides improved out-of-sample predictive performance. We also provide a catalogue of currently available BMA software.

KEYWORDS: *Bayesian model averaging; Bayesian graphical models; Learning; Model uncertainty; Markov chain Monte Carlo*

*Research supported in part by the U.S. National Science Foundation (DMS-9806243) and the U.S. Office of Naval Research (N00014-96-1-1092). We thank Leon Gleser (Executive Editor) and two anonymous referees for valuable comments.

† *Address for correspondence:* Department of Statistics, Colorado State University, Fort Collins, CO 80526, USA (jah@stat.colostate.edu).

Contents

List of Figures

List of Tables

1 Introduction

Consider the following scenario: a researcher has gathered data concerning cancer of the esophagus. For each of a large number of patients, she has recorded a variety of demographic and medical covariates, along with each patient's last known survival status. She 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 for future patients. She decides to use proportional hazards regression models to analyze the data. Next she conducts a data-driven search to select covariates for the specific proportional hazards regression model, M^*, that will provide the framework for subsequent inference. She checks that M^* fits the data reasonably well and notes that the parameter estimates are sensible. Finally, she 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, different standard errors, or different predictions. In this situation, how should the researcher proceed? Models like M^{**} are commonplace – for striking examples see Regal and Hook (1991), Draper (1995), Madigan and York (1995), Kass and Raftery (1995), and Raftery (1996). Basing inferences on M^* alone is risky; presumably, ambiguity about model selection 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) and Hodges (1987) make essentially the same observation.

Bayesian model averaging provides a way around 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:

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

This is an average of the posterior distributions under each of the models considered, weighted by their posterior model probability. 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

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

is the integrated 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 (given that one of the models considered is true). All probabilities are implicitly conditional on $\mathcal{M}$, the set of all models being considered.

The posterior mean and variance of Δ are as follows:

$$\begin{aligned} E[\Delta|D] &= \sum_{k=0}^{K} \hat{\Delta}_k \mathrm{pr}(M_k|D), \\ \mathrm{Var}[\Delta|D] &= \sum_{k=0}^{K} \left(\mathrm{Var}[\Delta|D, M_k] + \hat{\Delta}_k^2\right) \mathrm{pr}(M_k|D) - E[\Delta|D]^2, \end{aligned}$$

where $\hat{\Delta}_k = E[\Delta|D, M_k]$ (Raftery, 1993; Draper, 1995).

Madigan and Raftery (1994) note that averaging over *all* the models in this fashion provides better average predictive ability, as measured by a logarithmic scoring rule, than using any single model M_j, conditional on $\mathcal{M}$. Considerable empirical evidence now exists to support this theoretical claim; in Section 7 we will present some of this evidence.

While BMA is an intuitively attractive solution to the problem of accounting for model uncertainty, it is not yet part of the standard data analysis tool kit. This is, in part, due to the fact that implementation of BMA presents several difficulties:

- The number of terms in (1) can be enormous, rendering exhaustive summation infeasible (Section 3.1).

- The integrals implicit in (1) can in general be hard to compute. Markov chain Monte Carlo methods have partly overcome the problem, but challenging technical issues remain (Section 3.2).

- Specification of $\mathrm{pr}(M_k)$, the prior distribution over competing models, is challenging and has received little attention (Section 5).

- After these difficulties are overcome, choosing the class of models over which to average becomes the fundamental modeling task. At least three competing schools of thought have emerged (Section 8).

This paper will provide a tutorial introduction to BMA and discuss several solutions to these implementation difficulties. We will also briefly discuss related work.

2 Combining Models: A Historical Perspective

Barnard (1963) provided the first mention of model combination in the statistical literature in a paper studying airline passenger data. However, most of the early work in model combination was not in statistical journals. The seminal forecasting paper by Bates and Granger (1969) stimulated a flurry of articles in the economics literature of the 1970's about combining predictions from different forecasting models. See Clemen (1989) for a detailed review.

In the statistical literature, early work related to model averaging includes Roberts (1965) who suggested a distribution which combines the opinions of two experts (or models). This distribution, essentially a weighted averaged of posterior distributions of two models, is similar to BMA. Leamer (1978) expanded on this idea and presented the basic paradigm for BMA. He also pointed out the fundamental idea that BMA accounts for the uncertainty involved in selecting the model. After Leamer's book was published little attention was given to BMA for some time. The drawbacks of ignoring model uncertainty were recognized by many authors (e.g., the collection of papers edited by Dijkstra, 1988), but little progress was made until new theoretical developments and computational power enabled researchers to overcome the difficulties related to implementing BMA (Section 1). George (1999) reviews Bayesian model selection and discusses BMA in the context of decision theory. Draper (1995), Chatfield (1995), and Kass and Raftery (1995) all review BMA and the costs of ignoring model uncertainty. These papers focus more on Bayesian interpretation, whereas in this paper we will emphasize implementation and other practical matters.

3 Implementing Bayesian Model Averaging

In this section, we discuss general implementation issues for BMA. In Section 4, we will discuss specific model classes.

3.1 Managing the Summation

The size of interesting model classes often renders the exhaustive summation of Equation (1) impractical. We describe two distinct approaches to this problem.

The first approach is to average over a subset of models that are supported by the data. The Occam's Window method of Madigan and Raftery (1994) averages over a set of parsimonious, data-supported models, selected by applying standard norms of scientific investigation.

Two basic principles underly the Occam's Window method. First, Madigan and Raftery (1994) argued 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. Their second, optional, principle, appealing to Occam's razor, led them 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

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

where $\mathcal{A} = \mathcal{A}' \backslash \mathcal{B}$ and all probabilities are implicitly conditional on the set of models in $\mathcal{A}$.

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 $\mathcal{A}$. Madigan and Raftery (1994) proposed one possible search strategy, based on two main ideas. 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 idea — "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 "smaller" than M_1. The essential idea is shown in Figure 1: If there is evidence for M_0 then M_1 is rejected, but rejecting M_0 requires strong evidence *for* the larger model, M_1. If the evidence is inconclusive (falling in Occam's Window), neither model is rejected. Madigan and Raftery (1994) adopted $\frac{1}{20}$ and 1 for O_L and O_R, respectively (see Figure 1). Raftery *et al.* (1996) show that adopting $\frac{1}{20}$ and 20 for O_L and O_R, respectively, may provide improved predictive performance; this specifies $O_L = O_R^{-1}$ which amounts to using only the first Occam's Window principle and not the second one.

These principles fully define the strategy. In most model classes the number of terms in (1) is typically reduced to fewer than 100 models and often to fewer than 10; a reduction to one or two models is not unusual. Madigan and Raftery (1994) provide a detailed description of the algorithm.

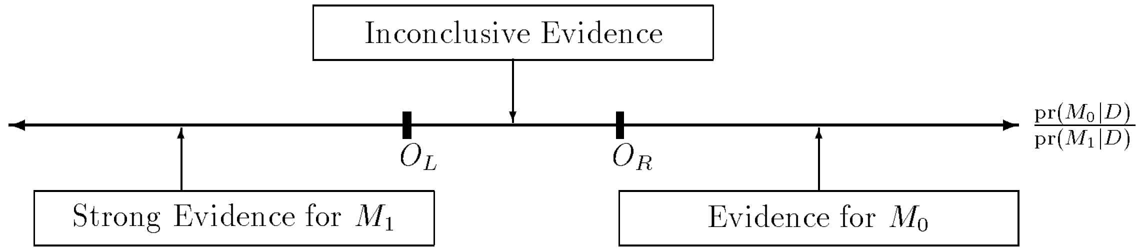

Figure 1: Occam's Window: Interpreting the Posterior Odds

Another way to search for the models in $\mathcal{A}$ is suggested by Volinsky *et al.* (1997). They use the "leaps and bounds" algorithm (Furnival and Wilson, 1974) to rapidly identify models to be used in the summation of Equation (1).

The second approach, Markov chain Monte Carlo model composition (MC^3), uses a Markov chain Monte Carlo method to directly approximate (1) (Madigan and York, 1995). Specifically, let $\mathcal{M}$ denote the space of models under consideration. One can construct a Markov chain $\{M(t)\}, t = 1, 2, \ldots$ with state space $\mathcal{M}$ and equilibrium distribution $\mathrm{pr}(M_i \mid D)$ and simulate this Markov chain to obtain observations $M(1), \ldots, M(N)$. Then for any function $g(M_i)$ defined on $\mathcal{M}$, the average:

$$\hat{G} = \frac{1}{N} \sum_{t=1}^{N} g(M(t)) \tag{7}$$

is an estimate of $E(g(M))$. Applying standard Markov chain Monte Carlo results,

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

(e.g., Smith and Roberts, 1993). To compute (1) in this fashion set $g(M) = \mathrm{pr}(\Delta \mid M, D)$.

To construct the Markov chain, define a neighborhood $\mathrm{nbd}(M)$ 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, plus the model M itself (Madigan *et al.*, 1994). Define a transition matrix q by setting $q(M \to M') = 0$ for all $M' \notin \mathrm{nbd}(M)$ and $q(M \to M')$ non–zero for all $M' \in \mathrm{nbd}(M)$. If the chain is currently in state M, proceed by drawing M' from $q(M \to M')$. M' is accepted with probability

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

Otherwise the chain remains in state M. For a basic introduction to the Metropolis-Hastings algorithm, see Chib and Greenberg (1995).

MC^3 offers considerable flexibility. For example, working with equivalence classes of graphical models, Madigan *et al.* (1996a) introduced a total ordering of the vertices into the stochastic process as an auxiliary variable, thereby providing a three-fold computational speed-up (see Section 4.4). York *et al.* (1995) incorporated missing data and a latent variable into their MC^3 scheme. For linear models, Raftery, *et al.* (1997) applied MC^3 to average across models with many predictors. However, as with other Markov chain Monte Carlo methods, convergence issues can be problematic.

The stochastic search variable selection (SSVS) method of George and McCulloch (1993) is similar in spirit to MC^3. In SSVS, a predictor is not actually removed from the full model; instead these predictors are set close to zero with high probability. A Markov chain Monte Carlo procedure is then used to move through model space and parameter space at the same time.

Clyde *et al.* (1996) introduced an importance sampling strategy based on a re-expression of the space of models in terms of an orthogonalization of the design matrix. Their goal is to implement model mixing for problems with many correlated predictors. One advantage to this approach is that orthogonalizing can reduce the number of competing plausible models. When orthgonalized model mixing is appropriate, it can be much more efficient than MC^3.

Earlier related work includes Stewart (1987) who used importance sampling to average across logistic regression models, and Carlin and Polson (1991) who used 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. Buntine (1992) applied BMA to classification trees (CART). Rather than average over all possible trees, his algorithm seeks out trees with high posterior probability and averages over those. Earlier related work includes Kwok and Carter (1990).

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 and Chib (1993), Philips and Smith (1994), and Green (1995) provides potential solutions.

3.2 Computing Integrals for BMA

Another difficulty in implementing BMA is that the integrals of the form (3) implicit in (1) can be hard to compute. For certain interesting classes of models such as discrete graphical models (e.g., Madigan and York, 1995) and linear regression (e.g., Raftery *et al.*, 1997), closed form integrals for the marginal likelihood, Equation (3), are available. The Laplace method (Tierney and Kadane, 1986) can provide an excellent approximation to $\mathrm{pr}(D|M_k)$; in certain circumstances this yields the very simple BIC approximation (Schwarz, 1978; Kass and Wasserman, 1995; Raftery, 1995). Taplin (1993) suggested approximating $\mathrm{pr}(\Delta \mid M_k, D)$ by $\mathrm{pr}(\Delta \mid M_k, \hat{\theta}, D)$ where $\hat{\theta}$ is the maximum likelihood estimate of the parameter vector θ; we refer to this as the "MLE approximation". Draper (1995), Raftery *et al.* (1996), and Volinsky *et al.* (1997) show its usefulness in the BMA context. Section 4 discusses these approximations in more detail in the context of specific model classes.

4 Implementation Details for Specific Model Classes

In this section we describe the implementation of the general strategy of the last section for specific model classes.

4.1 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 $X_1, \ldots, X_k$, find the "best" model of the form

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

where $X_{i_1}, \ldots, X_{i_p}$ is a subset of $X_1, \ldots, X_k$. 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.

BMA, on the other hand, seeks to average over all possible sets of predictors. Raftery *et al.* (1997) 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. Fernández *et al.* (1997; 1998) offer an alternative prior structure aiming at a more automatic choice of hyperparameters.

Hoeting *et al.* (1996, 1999; hereafter HRM96 and HRM99) extend this framework to include transformations and outliers respectively. Largely for reasons of convenience, HRM99 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 typically do not have a biological or physical interpretation unless they are limited to a few possible values of ρ. HRM99 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, HRM99 proposed a novel approach consisting of an initial exploratory use of the Alternating Conditional Expectation algorithm (ACE), followed by change point transformations if needed. The ACE algorithm (Breiman and Friedman, 1985) provides nonlinear, nonparametric transformations of the variables in a regression model. ACE chooses the transformations to maximize the correlation between the transformed response and the sum of the transformed predictors. HRM99 used ACE to *suggest* parametric transformations of the predictors. The transformations suggested by ACE often have roughly the form of a change point, a threshold or a saturation effect, with no change in the expected value of the response above (or below) a certain value. This type of transformation often better describes the assumed physical or biological context of the experiment than the commonly used power transformations discussed above. To choose the change point and to determine the evidence for the change point, HRM99 provided an approximate Bayes factor. HRM99'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.

HRM96 averaged over sets of predictors *and* possible outliers. They adopted 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 assumed that the ϵ's in distinct cases are independent where

$$\epsilon \sim \begin{cases} \mathrm{N}\left(0, \sigma^2\right) & \text{w.p.} \quad (1-\pi) \\ \mathrm{N}\left(0, K^2\sigma^2\right) & \text{w.p.} \quad \pi. \end{cases} \tag{8}$$

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 used a highly robust technique to identify a set of potential outliers. The robust approach typically identifies a large number of potential outliers. In the second step, HRM96 computed all possible posterior model probabilities or used 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. HRM96 provided evidence that SVO successfully identifies masked outliers. A simultaneous variable, transformation, and outlier selection approach (SVOT) which combines SVO and SVT has also been proposed (Hoeting, 1994). A faster but less exact implementation of BMA for variable selection in linear regression via the leaps-and-bound algorithm is available in the BICREG software (Section 4.5).

4.2 Generalized Linear Models

Model-building for generalized linear models involves choosing the independent variables, the link function, and the variance function (McCullagh and Nelder, 1989). Each possible combination of choices defines a different model. Raftery (1996) presents methods for calculating approximate Bayes factors for generalized linear models. The Bayes factor, B_{10} for a model M_1 against another model M_0 given data D is the ratio of posterior to prior odds, namely

$$B_{10} = \mathrm{pr}(D|M_1)/\mathrm{pr}(D|M_0),$$

the ratio of the marginal likelihoods. The Bayes factors, in turn, yield posterior model probabilities for all the models, and enable BMA, as follows. Suppose that $(K+1)$ models, $M_0, M_1, \ldots, M_K$, are being considered. Each of $M_1, \ldots, M_K$ is compared in turn with M_0, yielding Bayes factors $B_{10}, \ldots, B_{K0}$. Then the posterior probability of M_k is

$$\mathrm{pr}(M_k|D) = \alpha_k B_{k0} / \sum_{r=0}^{K} \alpha_r B_{r0}, \tag{9}$$

where $\alpha_k = \mathrm{pr}(M_k)/\mathrm{pr}(M_0)$ is the prior odds for M_k against M_0 $(k = 0, \ldots, K)$.

Raftery's derivation proceeds as follows. Suppose that Y_i is a dependent variable, and that $X_i = (x_{i1}, ..., x_{ip})$ is a corresponding vector of independent variables, for $i = 1, ..., n$. A generalized linear model M_1 is defined by specifying $\mathrm{pr}(Y_i|X_i, \beta)$ in such a way that $E[Y_i|X_i] = \mu_i$, $\mathrm{Var}\,[Y_i|X_i] = \sigma^2 v(\mu_i)$, and $g(\mu_i) = X_i\beta$, where $\beta = (\beta_1, ..., \beta_p)^T$; here g is called the link function. The $n \times p$ matrix with elements x_{ij} is denoted by X, and it is assumed that $x_{i1} = 1$ $(i = 1, ..., n)$. Here we assume that σ^2 is known; Raftery (1996) deals with the unknown σ^2 case.

Consider the Bayes factor for the null model M_0, defined by setting $\beta_j = 0$ $(j = 2, ..., p)$, against M_1. The likelihoods for M_0 and M_1 can be written down explicitly, and so, once the prior has been fully specified, the following (Laplace) approximation can be computed:

$$p(D|M_k) \approx (2\pi)^{p_k/2}|\Psi|^{\frac{1}{2}}\mathrm{pr}(D|\tilde{\beta}_k, M_k)\mathrm{pr}(\tilde{\beta}_k|M_k), \tag{10}$$

where p_k is the dimension of β_k, $\tilde{\beta}_k$ is the posterior mode of β_k, and Ψ_k is minus the inverse Hessian of $h(\beta_k) = \log\{\mathrm{pr}(D|\beta_k, M_k)\mathrm{pr}(\beta_k|M_k)\}$, evaluated at $\beta_k = \tilde{\beta}_k$. Arguments similar to those in the Appendix of Tierney and Kadane (1986) show that in regular statistical models the relative error in Equation (10), and hence in the resulting approximation to B_{10}, is $O(n^{-1})$.

However, this approximation is not easy to compute for generalized linear models using readily available software and Raftery (1996) presents three convenient but less accurate approximations. We reproduce here the most accurate of these approximations.

Suppose that the prior distribution of β_k is such that $E[\beta_k|M_k] = \omega_k$ and $\mathrm{Var}\,[\beta_k|M_k] = W_k$. Then approximating the posterior mode, $\tilde{\beta}_k$, by a single step of the Newton-Raphson algorithm (e.g., Kincaid and Cheney, 1991, p. 26) starting from the MLE, $\hat{\beta}_k$, and substituting the result into Equation (10) yields the approximation

$$2\log B_{10} \approx \chi^2 + (E_1 - E_0). \tag{11}$$

In Equation (11), $\chi^2 = 2\{\ell_1(\hat{\beta}_1) - \ell_0(\hat{\beta}_0)\}$, where $\ell_k(\hat{\beta}_k) = \log(\mathrm{pr}(D|\beta_k, M_k))$ is the log-likelihood when M_0 is nested within M_1 and χ^2 is the standard likelihood-ratio test statistic. Also,

$$E_k = 2\lambda_k(\hat{\beta}_k) + \lambda_k'(\hat{\beta}_k)^T(F_k + G_k)^{-1}\{2 - F_k(F_k + G_k)^{-1}\}\lambda_k'(\hat{\beta}_k) - \log|F_k + G_k| + p_k\log(2\pi),$$

where F_k is the expected Fisher information matrix, $G_k = W_k^{-1}$, $\lambda_k(\beta_k) = \log\mathrm{pr}(\beta_k|M_k)$ is the log-prior density, and $\lambda_k'(\hat{\beta}_k)$ is the p_k-vector of derivatives of $\lambda_k(\beta_k)$ with respect to the elements of β_k $(k = 0, 1)$. In general, the relative error in this approximation is $O(n^{-\frac{1}{2}})$. However, if the canonical link function is used, the observed Fisher information is equal to the expected Fisher information, and the relative error improves to $O\left(n^{-1}\right)$.

Raftery (1996) describes a useful parametric form for the prior parameters ω_k and W_k that involves only one user-specified input, and derives a way of choosing this when little prior information is available. The prior distribution for β has three user specified parameters and Raftery (1996) discusses possible choices in the situation where little prior information is available.

4.3 Survival Analysis

Methods for analyzing survival data often focus on modeling 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 X_i to be

$$\lambda(t \mid X_i) = \lambda_0(t) \exp(X_i \beta), \tag{12}$$

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

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

$$PL(\beta) = \prod_{i=1}^{n} \left(\frac{\exp(X_i \beta)}{\sum_{\ell \in R_i} \exp(X_\ell^T \beta)} \right)^{w_i},$$

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.

Since the integrals required for BMA do not have a closed form solution for Cox models, Raftery *et al.* (1996) and

Volinsky *et al.* (1997, VMRK hereafter) adopted a number of approximations. In particular, VMRK used the MLE approximation:

$$\text{pr}(\Delta \mid M_k, D) \approx \text{pr}(\Delta \mid M_k, \hat{\beta}_k, D),$$

and the Laplace approximation:

$$\log \text{pr}(D \mid M_k) \approx \log \text{pr}(D \mid \hat{\beta}_k, M_k) - d_k \log n, \tag{13}$$

where d_k is the dimension of β_k. This is the Bayesian Information Criterion (BIC) approximation. In Equation (13), n is usually taken to be the total number of cases. Volinsky (1997) provides evidence that n should be the total number of *uncensored* cases (i.e., deaths or events).

To implement BMA for Cox models, VMRK used an approach similar to the Occam's Window method described in Section 3.1. To efficiently identify good models, VMRK adapted the "Leaps and Bounds" algorithm of Furnival and Wilson (1974) which was originally created for linear regression model selection. The leaps and bounds algorithm provides the top q models of each model size, where q is designated by the user, plus the MLE $\hat{\beta}_k$, $\text{var}(\hat{\beta}_k)$, and R_k^2 for each model M_k returned. Lawless and Singhal (1978) and Kuk (1984) provided a modified algorithm for non-normal regression models that gives 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 used the approximate likelihood ratio test to reduce the remaining subset of models to those most likely to be in $\mathcal{A}$. This reduction step keeps only the models whose approximate posterior model probabilities fall within a factor C' of the model with the highest posterior model probability, where C' is greater than C, the cut-off in Equation (4). (VMRK set $C' = C^2$ and almost no models in $\mathcal{A}$ were lost in the examples they considered). A standard survival analysis program 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 calculated posterior model probabilities by normalizing over the model set, as in Equation (9). Model-averaged parameter estimates and standard errors of those estimates derive from weighted averages of the estimates and standard 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 a real example based on the Cardiovascular Health Study (Fried *et al.*, 1991), VMRK showed that these posterior effect probabilities can lead to substantive interpretations that are at odds with the usual p-values.

Prior probabilities on both model space and parameter space are implicitly defined by this procedure. All models are considered equally likely *a priori* by the leaps and bounds algorithm. Using the BIC approximation to the integrated likelihood defines an inherent prior on all of the regression parameters, as outlined in Kass and Wasserman (1995). This prior is a sensible one to take in the absence of substantial prior information; it is a normal distribution centered at the null hypothesized value (usually 0) with the amount of information in the prior equal to the average amount of information in one observation.

4.4 Graphical Models: Missing Data and Auxiliary Variables

A *graphical model* is a statistical model embodying a set of conditional independence relationships that can 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 (Lauritzen, 1996).

Here we focus on acyclic directed graphs (ADGs) and discrete random variables. In an

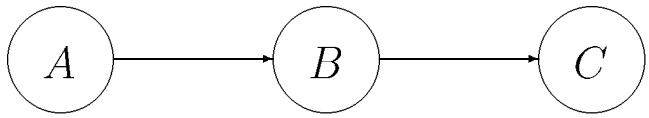

Figure 2: A Simple Discrete Graphical Model

ADG, *all* the edges are directed and are shown as arrows (see, e.g., Figure 2). A directed graph is acyclic if it contains no directed cycles. Each vertex in the graph will correspond to a random variable $X_v, v \in V$ taking values in a sample space $\mathcal{X}_v$. To simplify notation, we use v in place of X_v in what follows. In an ADG, the parents of a vertex v, $\mathrm{pa}(v)$, are those vertices from which edges point into v. The *descendants* of a vertex v are the vertices which are reachable from v along a directed path. The parents are taken to be the only direct influences on v, so that v is independent of its non-descendants given its parents. This property implies a factorization of the joint distribution of $X_v, v \in V$, which we denote by $\mathrm{pr}(V)$, given by:

$$\mathrm{pr}(V) = \prod_{v \in V} \mathrm{pr}(v \mid \mathrm{pa}(v)). \tag{14}$$

Figure 2 shows a simple example. This directed graph represents the assumption that C and A are conditionally independent given B. The joint density of the three variables factors accordingly:

$$\mathrm{pr}(A, B, C) = \mathrm{pr}(A)\mathrm{pr}(B \mid A)\mathrm{pr}(C \mid B). \tag{15}$$

Spiegelhalter and Lauritzen (1990) showed how independent Dirichlet prior distributions placed on these probabilities can be updated locally to form posterior distributions as data become available. Heckerman *et al.* (1994) provided corresponding closed-form expressions for complete-data likelihoods and posterior model probabilities.

The application of BMA and Bayesian graphical models to problems involving missing data and/or latent variables generally requires the use of either analytical or numerical approximations. Madigan and York (1995) and York *et al.* (1995) provide extensive implementation details. An especially useful approach derives from the following re-expression of the usual Bayes factor comparing two models, M_0 and M_1:

$$\frac{\mathrm{pr}(D \mid M_0)}{\mathrm{pr}(D \mid M_1)} = \mathbf{E}\left(\left. \frac{\mathrm{pr}(D, Z \mid M_0)}{\mathrm{pr}(D, Z \mid M_1)} \right| D, M_1 \right).$$

Here the expectation is over Z, which denotes the missing data and/or latent variables. This expectation can be numerically approximated by simulating the missing data from its predictive distribution under *only one* of the two models being compared. A similar formula appears in Thompson and Wijsman (1990) and its use in the present context was suggested by Augustine Kong.

4.5 Software for BMA

Software to implement several of the approaches described above is available on the internet. These programs, all written in S-Plus©, can be obtained free of charge via the Web address `www.research.att.com/~volinsky/bma.html`.

bic.glm performs BMA for generalized linear models using the Leaps and Bounds algorithm and the BIC approximation. [Volinsky].

bic.logit performs Bayesian model selection and accounting for model uncertainty using the BIC approximation for logistic regression models [Raftery].

bicreg does Bayesian model selection and accounting for model uncertainty in linear regression models using the BIC approximation [Raftery].

bic.surv does BMA for proportional hazard models using the BIC approximation [Volinsky].

BMA implements the MC^3 algorithm for linear regression models [Hoeting].

glib carries out Bayesian estimation, model comparison and accounting for model uncertainty in generalized linear models, allowing user-specified prior distributions [Raftery].

5 Specifying Prior Model Probabilities

Before implementing any of the BMA strategies described above, prior model probabilities must be assigned for Equation (2). When there is little prior information about the relative plausibility of the models considered, the assumption that all models are 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.

When prior information about the importance of a variable is available for model structures with a coefficient associated with each predictor (e.g., linear regression models and Cox proportional hazards models), a prior probability on model M_i can be specified as

$$\mathrm{pr}(M_i) = \prod_{j=1}^{p} {\pi_j}^{\delta_{ij}} \, (1 - \pi_j)^{1-\delta_{ij}}, \tag{16}$$

where $\pi_j \in [0, 1]$ is the prior probability that $\beta_j \neq 0$ in a regression model, and δ_{ij} is an indicator of whether or not variable j is included in model M_i. Assigning $\pi_j = 0.5$ for all j corresponds to a uniform prior across model space, while $\pi_j < 0.5$ for all j imposes a penalty for large models. Using $\pi_j = 1$ ensures that variable j is included in all models. This approach is used to specify model priors for variable selection in linear regression in George and McCulloch (1993) and suggested for model priors for BMA in Cox models in VMRK.

In the context of graphical models, Madigan and Raftery (1991) 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. This approach is similar to Equation (16). However, both approaches make the possibly unreasonable assumption that the presence or absence of each component (variable or link) is independent *a priori* of the presence or absence of other components.

Madigan, Gavrin and Raftery (1995) provide a simple method for informative prior elicitation in discrete data applications and show that their approach provides improved predictive performance for their 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 and Laud (1994) adopt a somewhat similar approach in the context of linear models.

6 Predictive Performance

Before presenting two examples, we briefly discuss methods for assessing the success of various modeling strategies. A primary purpose of statistical analysis is to make forecasts (Dawid, 1984). Similarly, Bernardo and Smith (1994, p. 238) argue that when comparing rival modeling strategies, all other things being equal, we are more impressed with a modeling strategy that consistently assigns higher probabilities to the events that actually occur. Thus, measuring how well a model predicts future observations is one way to judge the efficacy of a BMA strategy.

In the examples below we assess predictive performance as follows. First, we randomly split the data into two halves, and then we apply each model selection method to the first half of the data, called the *build data* (D^B). Performance is then measured on the second half of the data (*test data*, or D^T).

One measure of predictive ability is the logarithmic scoring rule of Good (1952) which is based on the conditional predictive ordinate (Geisser, 1980). Specifically, the predictive log score measures the predictive ability of an individual model, M, using the sum of the logarithms of the observed ordinates of the predictive density for each observation in the test set:

$$-\sum_{d \in D^T} \log \mathrm{pr}(d \mid M, D^B), \tag{17}$$

and measures the predictive performance of BMA with:

$$-\sum_{d \in D^T} \log\{\sum_{M \in \mathcal{A}} \mathrm{pr}(d \mid M, D^B)\mathrm{pr}(M \mid D^B)\}. \tag{18}$$

The smaller the predictive log score for a given model or model average, the better the predictive performance. We note that the logarithmic scoring rule is a *proper scoring rule* as defined by Matheson and Winkler (1976) and others. Several other measures of predictive performance are described in the examples below.

For probabilistic predictions, there exist two types of discrepancies between observed and predicted values (Draper *et al.*, 1993): *predictive bias* (a systematic tendency to predict on the low side or the high side), and *lack of calibration* (a systematic tendency to over- or understate predictive accuracy). The predictive log score is a combined measure of bias and calibration. Considering predictive bias and calibration separately can also be useful—see for example Madigan and Raftery (1994) and Madigan *et al.* (1994), Hoeting (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.

7 Examples

In this section we provide two examples where BMA provides additional insight into the problem of interest and improves predictive performance. For other applications of BMA see (Chatfield, 1995; Draper, 1995; Fernández *et al.*, 1997; Hoeting *et al.*, 1999; Hoeting *et al.*, 1996; Madigan *et al.*, 1996b; Madigan and Raftery, 1994; Raftery *et al.*, 1997; Raftery, 1996; Volinsky *et al.*, 1997).

7.1 Example 1: Primary Biliary Cirrhosis

7.1.1 Overview

From 1974 to 1984 the Mayo Clinic conducted a double-blind randomized clinical trial involving 312 patients to compare the drug DPCA with a placebo in the treatment of primary biliary cirrhosis (PBC) of the liver (Dickinson, 1973; Grambsch *et al.*, 1989; Markus *et al.*, 1989; Fleming and Harrington, 1991) The goals of this study were twofold: (a) to assess DPCA as a possible treatment through randomization, and (b) to use other variables to develop a natural history model of the disease. Such a model is useful for prediction (counseling patients and predicting the course of PBC in untreated patients) and inference (historical control information to assess new therapies). Fleming and Harrington (1991) — hereafter FH — developed such a model. Starting with DPCA plus 14 covariates, they selected a Cox regression model with five of the covariates. The analysis of FH represents the current best practice in survival analysis. However, we argue here that the model uncertainty is substantial and that procedures such as theirs can underestimate uncertainty about quantities of interest, leading to decisions that are riskier than one thinks they are.

Raftery *et al.* (1996) analyzed a subset of these data by averaging over all possible models in a much smaller model space. Here, we apply the leaps-and-bounds approach described in Section 4.3 to quickly approximate averaging over a much larger model space. Of the 312 patients, we omit eight due to incomplete data. Of the remaining 304 patients, 123 were followed until death and the other 181 observations were censored. There are 14 prognostic variables of interest in the natural history model, plus the treatment variable DPCA. Table 1 shows the independent and dependent variables. Subjects were observed for up to 12.5 years with a mean observation time of 5.5 years.

Following FH, we used logarithmic transformations of bilirubin, albumen, prothrombin time and urine copper. FH used a multistage variable selection method and concluded that the best model was the one with the five independent variables age, edema, bilirubin, albumin and prothrombin time.

Table 1: PBC Example: Summary Statistics and BMA Estimates.

Variable	Range	Mean	Mean $\beta\|D$	SD $\beta\|D$	$\mathrm{P}(\beta \neq 0\|D)$
Bilirubin (log)	-1.20–3.33	0.60	0.784	0.129	100
Albumen (log)	0.67–1.54	1.25	-2.799	0.796	100
Age (years)	26–78	49.80	0.032	0.010	100
Edema	0 = no edema	$n = 263$	0.736	0.432	84
	.5 = edema but no diuretics	$n = 29$			
	1 = edema despite diuretics	$n = 20$			
Prothrombin Time	2.20–2.84	2.37	2.456	1.644	78
Urine Copper (log)	1.39–6.38	4.27	0.249	0.195	72
Histologic Stage	1–4	3.05	0.096	0.158	34
SGOT	3.27–6.13	4.71	0.103	0.231	22
Platelets	62–563	262.30	-0.000	0.000	5
Sex	0=male	0.88	-0.014	0.088	4
Hepatomegaly	1=present	0.51	0.006	0.051	3
Alkaline Phosphates	5.67–9.54	7.27	-0.003	0.028	3
Ascites	1=present	0.08	0.003	0.047	2
Treatment (DPCA)	1=DPCA	0.49	0.002	0.028	2
Spiders	1=present	0.29	0.000	0.027	2
Time observed (days)	41–4556	2001			
Status	0=censored 1=died	0.40			

7.1.2 Results

The PBC data set provides an opportunity to compare BMA with model selection methods in the presence of moderate censoring. The model chosen by a stepwise (backward elimination) procedure, starting with the variables in Table 1, included the following variables: age, edema, bilirubin, albumin, urine copper, and prothrombin time (which is the FH model with the inclusion of urine copper). BMA was performed using the leaps-and-bounds approach described in Section 4.3. The model with the highest approximate posterior probability was the same as the stepwise model. Nonetheless, this model represents only 17% of the total posterior probability, indicating that there is a fair amount of model uncertainty. The FH model places sixth in the table with a posterior model probability of only 5%. Table 2 lists the models with the highest posterior probabilities. Inference about independent variables is expressed in terms of the posterior effect probabilities.

Table 1 contains the posterior means, standard deviations and posterior effect probabilities, $\mathrm{P}(\beta \neq 0|D)$, for the coefficient associated with each variable. Note that these parameter estimates and standard deviations directly incorporate model uncertainty. For instance, the averaged posterior distribution associated with the independent variable SGOT has 78% of

Table 2: PBC Example: Results for the Full Data Set. PMP denotes the posterior model probability. Only the 10 models with the highest PMP values are shown.

Model No.	Age	Edema	Bili	Albu	UCopp	SGOT	Prothromb	Hist	PMP	Log Lik
1	•	•	•	•	•		•		.17	−174.4
2	•	•	•	•	•		•	•	.07	−172.6
3	•	•	•	•	•			•	.07	−172.5
4	•		•	•	•		•		.06	−172.2
5*	•	•	•	•			•		.05	−172.0
6	•	•	•	•	•				.05	−172.0
7	•	•	•	•	•	•	•		.04	−171.7
8	•	•	•	•		•	•		.04	−171.4
9	•	•	•	•		•	•	•	.04	−171.3
10	•	•	•	•	•	•	•	•	.03	−170.9
$\mathrm{Pr}_{\mathrm{MA}}[\beta_i \neq 0]$	1.00	0.84	1.00	1.00	0.72	0.22	0.78	0.34		

*Model Selected by FH.

its mass at zero. This shrinks the estimate towards zero, not unlike other shrinkage estimates such as ridge regression. In addition, this tends to increase the standard deviation of the estimate, to take account of model uncertainty.

Figure 3 shows the posterior effect probabilities, plotted against the corresponding p-value from the stepwise variable selection model. Overall, the posterior effect probabilities imply weaker evidence for effects than do the p-values, which do not take model uncertainty into effect. Comparison of p-values is often used as a measure of evidence (as in the standard interpretation of $p < .05$ and $p < .01$ as significant and highly significant), even though they should not necessarily be interpreted this way. In fact, p-values arguably overstate the evidence for an effect even when there is no model uncertainty (Edwards *et al.*, 1963; Berger and Delampady, 1987; Berger and Sellke, 1987).

For the three variables, albumin, age and bilirubin (which is highly significant and not shown in Figure 3), the posterior effect probabilities and the p-values agree that there is very strong evidence for an effect ($p < 0.001$ and $P(\beta \neq 0|D) > 99\%$). For the five variables in Table 3, however, the two approaches lead to qualitatively different conclusions. Each p-value overstates the evidence for an effect. For the first three of the variables, the p-value suggests that the effect is "highly significant" ($p < 0.01$), while the posterior effect probability indicates that the evidence is positive but not strong. For the other two variables (histology and SGOT), the p-values are "marginally significant" ($p < 0.10$), but the posterior effect probabilities actually indicate (weak) evidence *against* an effect.

For the remaining 7 variables (the clump of points in the lower right corner of the figure), p-values and posterior effect probabilities agree in saying that there is little or no evidence

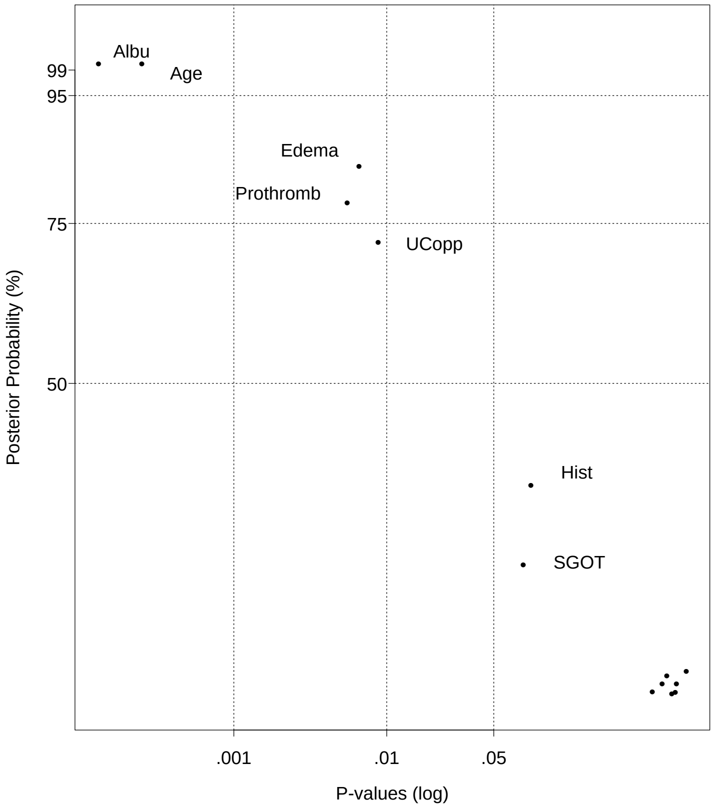

Figure 3: PBC Example: Posterior Effect Probabilities From BMA Versus p-values from the Stepwise Variable Selection Model.

for an effect. However, posterior effect probabilities enable one to make one distinction that p-values cannot. One may fail to reject the null hypothesis of "no effect" because either (a) there are not enough data to detect an effect, or (b) the data provide evidence *for* the null hypothesis. P-values cannot distinguish between these two situations, but posterior effect probabilities can. Thus, for example, for SGOT, $P(\beta \neq 0|D) = 22\%$, so that the data are indecisive, while for the treatment effect of DPCA, $P(\beta \neq 0|D) = 2\%$, indicating evidence *for* the null hypothesis of no effect. The posterior probability of "no effect" can be viewed as an approximation to the posterior probability of the effect being "small", namely $P(|\beta| < \varepsilon)$, provided that ε is at most about one-half of a standard error (Berger and Delampady, 1987).

7.1.3 Predictive Performance

For assessing predictive performance, we randomly split the data into two parts such that an equal number of events (61 deaths) occurred in each part. We compare the results for BMA with those for stepwise model selection and for the single model with the highest posterior

Table 3: PBC Example: A Comparison of Some p-values from the Stepwise Selection Model to the Posterior Effect Probabilities from BMA.

Var	p-value	$P(\beta \neq 0 \mid D)$ (%)
Edema	.007**	84
Prothrombin	.006**	78
Urine Copper	.009**	72
Histology	.09*	34
SGOT	.08	22

Table 4: PBC Example: Partial Predictive Scores for Model Selection Techniques and BMA. FH denotes the model selected by Fleming and Harrington.

Method	PPS
Top PMP Model	221.6
Stepwise	220.7
FH Model	222.8
BMA	217.1

model probability. Table 4 shows the partial predictive scores (PPS) for the competing methods. The PPS is an approximation to the predictive log score in Equations (17) and (18). A smaller PPS indicates better predictive performance. The top model and stepwise model may be different than those in the analysis for the full data since they are built using only half the data.

The difference in PPS of 3.6 can be viewed as an increase in predictive performance *per event* by a factor of $\exp(3.6/61) = 1.06$ or by about 6%. This means that BMA predicts who is at risk 6% more effectively than a method which picks the model with the highest posterior model probability (as well as 10% better than the Fleming and Harrington model and 2% more effectively than a stepwise method). We also performed this analysis on 20 different splits of the data, and over the 20 splits BMA was an average of 2.7 points better (5% per event) than both the top PMP model and the stepwise model.

In practice, categorizing patients into discrete risk categories such as high, medium, or low risk may prove more practical than numerical prediction. To assess the performance of a single model with respect to this goal we proceed as follows:

Table 5: PBC Example: Classification for Predictive Discrimination.

		BMA			Stepwise		
		Survived	Died	% Died	Survived	Died	% Died
Risk group	Low	34	3	8%	41	3	7%
	Med	47	15	24%	36	15	29 %
	High	10	43	81%	14	43	75%

	Top PMP	
Survived	Died	% Died
42	4	9%
31	11	26%
18	46	72%

1. Fit the model to the build data (the subset of data from which the models are selected) to get estimated coefficients $\hat{\beta}$.

2. Calculate risk scores $(\mathbf{x_i}^T\hat{\beta})$ for each subject in the build data.

3. Define low, medium and high risk groups for the model by the empirical (1/3) and (2/3) quantiles of the risk scores.

4. Calculate risk scores for the test data and assign each subject to a risk group.

5. Observe the actual survival status of those assigned to the three groups.

To assess BMA in this manner, we replace the first steps above with

1′. Fit each model $M_1, \ldots, M_K$ in $\mathcal{A}$ to get estimated coefficients $\hat{\beta}_k$.

2′. Calculate risk scores $(\mathbf{x_i}^T\hat{\beta}_k)$ under each model in $\mathcal{A}$ for each person in the build data. A person's risk score under BMA is the weighted average of these: $\sum_{k=1}^{K}(\mathbf{x_i}^T\hat{\beta}_k)\,\mathrm{pr}(M_k \mid D^B)$.

A method is better if it consistently assigns higher risks to the people who actually died. Table 5 shows the classification of the 152 people in the test data, and whether or not those people died in the study period. The people assigned to the high risk group by BMA had a higher death rate than did those assigned high risk by other methods; similarly those assigned to the low and medium risk groups by BMA had a lower total death rate.

In summary, we found that BMA improves predictive performance for the PBC study as measured both by PPS and predictive discrimination. The BMA results also provide additional evidence that the p-values for the model selected using stepwise variable selection overstate confidence at least in part because they ignore model uncertainty.

7.2 Example 2: Predicting Percent Body Fat

7.2.1 Overview

Percent body fat is now commonly used as an indicator of fitness or potential health problems (Lohman, 1992, p. 1). Percent body fat can be measured in a variety of ways including underwater weighing, skinfold calipers, and bioelectric impedance (Katch and McArdle, 1993). One drawback with these methods is that they require specialized equipment or expertise on the part of the person taking the measurements. As a result, simpler methods for measuring body fat have been developed. One such approach is to predict percent body fat using basic body measurements such as height and weight. This approach is non-invasive and requires little training or instrumentation. The drawback of this approach is a potential loss in accuracy in estimating body fat.

The goal of the analysis described here is to predict body fat using 13 simple body measurements in a multiple regression model. We consider body fat measurements for 252 men. The data were originally referenced in an abstract by Penrose *et al.* (1985) and are listed in Johnson (1996). For each subject, percentage of body fat, age, weight, height, and ten body circumference measurements were recorded (Table 6). We omitted one subject (observation 42) whose height was apparently erroneously listed as 29.5 inches.

The response in the regression model is percent body fat. Percent body fat was determined using body density, the ratio of body mass to body volume. Body volume was measured using an underwater weighing technique (Katch and McArdle, 1993, p. 242-244). Body density was then used to estimate percent body fat using Brozek's equation (Brozek *et al.*, 1963),

$$\%\text{ body fat} = 457/\text{Density} - 414.2. \tag{19}$$

For more details on the derivation of Equation (19) see Johnson (1996) and Brozek *et al.* (1963). Percent body fat for the subjects in this study ranged from 0 to 45% with a mean of 18.9% and standard deviation of 7.8%. One subject was quite lean and thus the percentage body fat (as computed using Brozek's equation) was negative. The body fat for this individual was truncated to 0%.

Regression results for the full model are given in Table 7. For this model, standard diagnostic checking did not reveal any gross violations of the assumptions underlying normal linear regression (Weisberg, 1985).

The standard approach to this analysis is to choose a single best subset of predictors using one of the many variable selection methods available. Since a model with fewer predictors

Table 6: Body Fat Example: Summary statistics for full data set. Abdomen circumference was measured at the umbilicus and level with the iliac crest. Wrist circumference (cm) was measured distal to the styloid processes.

Predictor number	Predictor	mean	s.d.	min	max
X_1	Age (years)	45	13	21	81
X_2	Weight (pounds)	179	29	118	363
X_3	Height (inches)	70	3	64	78
X_4	Neck circumference (cm)	38	2	31	51
X_5	Chest circumference (cm)	101	8	79	136
X_6	Abdomen circumference (cm)	93	11	69	148
X_7	Hip circumference (cm)	100	7	85	148
X_8	Thigh circumference (cm)	59	5	47	87
X_9	Knee circumference (cm)	39	2	33	49
X_{10}	Ankle circumference (cm)	23	2	19	34
X_{11}	Extended biceps circumference	32	3	25	45
X_{12}	Forearm circumference (cm)	29	2	21	35
X_{13}	Wrist circumference (cm)	18	1	16	21

than the full model may be selected, one advantage to this approach is that the number of measurements that are required to estimate body fat may be reduced. An alternative to this approach is to do Bayesian model averaging. BMA will require that all 13 measurements are taken. However, if BMA produces better predictions than the single model approach, then it may be worthwhile to take these additional measurements.

We will compare Bayesian model averaging to single models selected using several standard variable selection techniques to determine whether there are advantages to accounting for model uncertainty for these data. In what follows, we first analyze the full data set and then we split the data set into two parts, using one portion of the data to do BMA and select models using standard techniques and the other portion to assess performance. We compare the predictive performance of BMA to that of individual models selected using standard techniques.

7.2.2 Results

There are 13 candidate predictors of body fat and so potentially $2^{13} = 8192$ different sets of predictors, or linear regression models. For the Bayesian approach, all possible combinations of predictors were assumed to be equally likely *a priori*. To implement the Bayesian approach,

Table 7: Body Fat Example: Least Squares Regression Results From the Full Model. Residual Standard Error = 4, $R^2 = 0.75$, N=251, F-statistic = 53.62 on 13 and 237 df, p-value < 0.0001.

Predictor		Coef	Std Error	t-statistic	$p-$value
Intercept		-17.80	20.60	-0.86	0.39
X_1	age	0.06	0.03	1.89	0.06
X_2	weight	-0.09	0.06	-1.50	0.14
X_3	height	-0.04	0.17	-0.23	0.82
X_4	neck	-0.43	0.22	-1.96	0.05
X_5	chest	-0.02	0.10	-0.19	0.85
X_6	abdomen	0.89	0.08	10.62	<0.01
X_7	hip	-0.20	0.14	-1.44	0.15
X_8	thigh	0.24	0.14	1.74	0.08
X_9	knee	-0.02	0.23	-0.09	0.93
X_{10}	ankle	0.17	0.21	0.81	0.42
X_{11}	biceps	0.16	0.16	0.98	0.33
X_{12}	forearm	0.43	0.18	2.32	0.02
X_{13}	wrist	-1.47	0.50	-2.97	<0.01

we computed the posterior model probability for all possible models using the diffuse (but proper) prior distributions derived by Raftery *et al.* (1997). For larger problems where it is more difficult to compute the posterior model probability for all possible models, one can use MC^3 or the leaps and bounds algorithm to approximate BMA (see Section 3.1).

Table 8 shows the posterior effect probabilities, $\Pr(\beta_i \neq 0|D)$, obtained by summing the posterior model probabilities across models for each predictor. Two predictors, abdomen circumference and weight, appear in the models that account for a very high percentage of the total model probability. Five predictors have posterior effect probabilities smaller than 10% including age, height, and chest, ankle and knee circumference. The top three predictors by $\Pr(\beta_i \neq 0|D)$, weight, and abdomen and wrist circumference, appear in the model with the highest posterior model probability (Table 9).

The BMA results indicate considerable model uncertainty, with the model with the highest posterior model probability (PMP) accounting for only 14% of the total posterior probability (Table 9). The top 10 models by PMP account for 57% of the total posterior probability.

We compare the Bayesian results with models that might be selected using standard techniques. We chose three popular variable selection techniques, Efroymson's stepwise

Table 8: Body Fat Example: Comparison of BMA Results to Model Selected Using Standard Model Selection Methods. Stepwise, minimum Mallow's C_p, and maximum adjusted R^2 all selected the same model. The predictors are sorted by $\Pr(\beta_i \neq 0|D)$ which is expressed as a percentage. The results given here are based on standardized data (columns have means equal to 0 and variances equal to 1).

		Bayesian Model Averaging			**Stepwise Model**
Predictor		Mean $\beta\|D$	SD $\beta\|D$	$\Pr(\beta \neq 0\|D)$	p-value
X_6	abdomen	1.2687	0.08	100	<0.01
X_2	weight	–0.4642	0.15	97	0.03
X_{13}	wrist	–0.0924	0.08	62	<0.01
X_{12}	forearm	0.0390	0.06	35	0.01
X_4	neck	–0.0231	0.06	19	0.05
X_{11}	biceps	0.0179	0.05	17	
X_8	thigh	0.0176	0.05	15	0.02
X_7	hip	–0.0196	0.07	13	0.12
X_5	chest	0.0004	0.02	6	
X_1	age	0.0029	0.02	5	0.05
X_9	knee	0.0020	0.02	5	
X_3	height	–0.0015	0.01	4	
X_{10}	ankle	0.0011	0.01	4	

method (Miller, 1990), minimum Mallow's C_p, and maximum adjusted R^2 (Weisberg, 1985). Efroymson's stepwise method is like forward selection except that when a new variable is added to the subset, partial correlations are considered to see if any of the variables currently in the subset should be dropped. Similar hybrid methods are found in most standard statistical computer packages. For the stepwise procedure we used a 5% significance level which means that the significance levels for the F-to-enter and F-to-delete values were equal to 5%. Shortcomings of stepwise regression, Mallow's C_p, and adjusted R^2 are well known (see, for example, Weisberg, 1985).

All three standard model selection methods selected the same eight-predictor model (Table 8). There is clear agreement among the frequentist and BMA methods that the predictors abdomen circumference, weight, and wrist circumference are important predictors of percent body fat. If a cut-off of $\alpha = 0.05$ is chosen for interpretation of significant predictors, the p-values for the predictors for the single model selected using standard techniques are small for age, and forearm, neck and thigh circumference as compared to the posterior effect probabilities for those predictors computed from the BMA results. Based on these results, one

Table 9: Body Fat Example: 10 Models with Highest Posterior Model Probability (PMP).

X_2	X_4	X_6	X_8	X_{11}	X_{12}	X_{13}	PMP
•		•				•	.14
•		•			•	•	.14
•		•					.12
•		•		•		•	.05
•		•	•				.03
•	•	•					.03
•		•			•		.02
•		•		•			.02
•	•	•			•	•	.02
•	•	•			•		.02

could argue that, as in Example 1, the p-values overstate the evidence for an effect.

The posterior distribution for the coefficient of predictor 13 (wrist circumference), based on the BMA results, is shown in Figure 4. The BMA posterior distribution for β_{13} is a mixture of non-central Student's t distributions. The spike in the plot of the posterior distribution corresponds to P($\beta_{13} = 0|D$)=.38. This is an artifact of our approach as we consider models with a predictor fully removed from the model. This is in contrast to the practice of setting the predictor close to 0 with high probability as in George and McCulloch (1993).

7.2.3 Predictive Performance

As in Example 1, we use the predictive ability of the selected models for future observations to measure the effectiveness of a model selection strategy. Our objective is to compare the quality of the predictions based on BMA to the quality of predictions based on any single model that an analyst might reasonably have selected.

To measure performance we split the complete data set into two subsets. We used the split of the data that was used by the original researchers for model building (Penrose *et al.*, 1985). The first 142 observations were used to do BMA and apply the model selection procedures and the remaining 109 observations were used to evaluate performance.

Predictive coverage was measured using the proportion of observations in the performance set that fall in the corresponding 90% prediction interval. The prediction interval is based on the posterior predictive distribution for individual models and a mixture of these posterior predictive distributions for BMA. The predictive coverage for BMA is 90.8% while

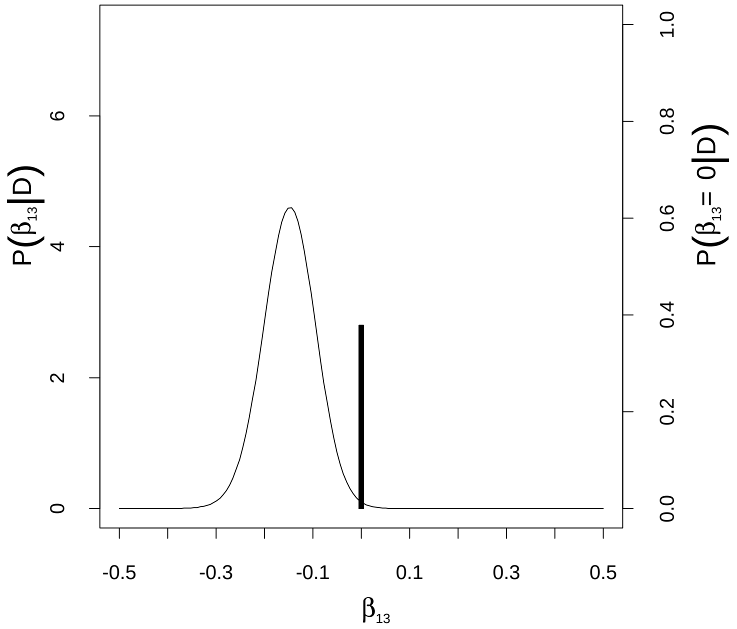

Figure 4: Body Fat Example: BMA Posterior Distribution for β_{13}, the Coefficient for Wrist Circumference. The spike corresponds to P($\beta_{13} = 0|D$)=0.38. The vertical axis on the left corresponds to the posterior distribution for β_{13} and the vertical axis on the right corresponds to the posterior distribution for β_{13} equal to 0. The density is scaled so that the maximum of the density is equal to P($\beta_{13} \neq 0|D$) on the right axis.

the predictive coverage for each of the individual models selected using standard techniques is less than 85%. For different random splits of this data set, the algorithms often selected different models, but BMA typically had superior predictive coverage as compared to the predictive coverage of the individual models.

Conditioning on a single selected model ignores model uncertainty which, in turn, can lead to the underestimation of uncertainty when making inferences about quantities of interest. For these data, the underestimation of model uncertainty for single selected models can lead to predictive coverage that is less than the stated coverage level.

8 Discussion

8.1 Choosing the Class of Models for BMA

In the examples we have discussed, the model structure was chosen to start with (e.g. linear regression), and then BMA averaged either over a reduced set of models supported by the data (e.g. subsets of predictors selected using Occam's Window) or over the entire class of models (e.g. all possible subsets of predictors). Several authors have suggested alternative

Table 10: Body Fat Example: Performance Comparison. Predictive coverage % is the percentage of observations in the performance set that fall in the 90% prediction interval. For BMA, the top 2500 models, accounting for 99.99% of the posterior model probability, were used to estimate predictive coverage.

Method	Model									Predictive coverage %
BMA	model averaging									90.8
Stepwise & C_p	X_1	X_2		X_6			X_{10}	X_{12}	X_{13}	84.4
Adjusted R^2	X_1	X_2	X_4	X_6	X_7	X_8	X_{10}	X_{12}	X_{13}	83.5

approaches to choosing the class of models for BMA.

Draper (1995) suggested finding a good model and then averaging over an expanded class of models "near" the good model (see also Besag *et al.*, 1995, Section 5.6). Within a single model structure, this approach is similar to the Madigan and Raftery (1994) suggestion that one average over a small set of models supported by the data. Draper also discusses the possibility of averaging over models with different error structures, e.g., averaging over models with different link functions in generalized linear models.

8.2 Other Approaches to Model Averaging

We have focused here on Bayesian solutions to the model uncertainty problem. Little has been written about frequentist solutions to the problem. Perhaps the most obvious frequentist solution is to bootstrap the entire data analysis, including model selection. However, Freedman *et al.* (1988) have shown that this does not necessarily give a satisfactory solution to the problem.

George (1986a; 1986b; 1986c) proposes a minimax multiple shrinkage Stein estimator of a multivariate normal mean under squared error loss. When the prior distributions are finite normal mixtures, these minimax multiple shrinkage estimates are empirical Bayes and formal Bayes estimates. George shows that this approach can easily be extended to estimate the coefficients in multiple regression in which case this method essentially provides minimax model averaging estimates of the regression coefficients.

Buckland *et al.* (1997) suggested several *ad hoc* non-Bayesian approaches to accounting for model uncertainty. They suggested using Akaike's information criterion (AIC) (Akaike,

1973) to approximate the model weights. This approach is similar to the BIC approximating strategies described above in terms of implementation, but not in terms of the underlying rationale; the results also tend to be quite different. Kass and Raftery (1995) discussed the relative merits of AIC and BIC in this context. To estimate model uncertainty, Buckland *et al.* (1997) suggested several bootstrapping methods. For a simulated example, they found coverage to be well below the nominal level if model uncertainty is ignored, and to be very accurate when model uncertainty is taken into account when forming the intervals.

Computational Learning Theory (COLT) provides a large body of theoretical work on predictive performance of non-Bayesian model mixing (see, for example, Kearns *et al.*, 1994, Chan *et al.*, 1996, and the references therein). Related literature discusses algorithms such as stacking (Wolpert, 1992), boosting (Freund, 1995), and bagging (Breiman, 1996) (see also Rao and Tibshirani (1997)). Note that while Bayesian model averaging researchers focus primarily on properties of predictive *distributions* such as predictive calibration and coverage of predictive intervals, Neural Network, Machine Learning, and COLT researchers generally focus on point prediction, often in the context of supervised learning.

8.3 Perspectives on Modeling

Bernardo and Smith (1994, p 383-385) drew the distinction between model selection when one knows the entire class of models to be entertained in advance, and the situation where the model class is not fully known in advance, but rather is determined and defined iteratively as the analysis and scientific investigation proceed. They referred to the former situation as the "$\mathcal{M}$-closed perspective", and to the latter as the "$\mathcal{M}$-open perspective". They argued that, while the $\mathcal{M}$-closed situation does arise in practice, usually in rather formally constrained situations, the $\mathcal{M}$-open perspective often provides a better approximation to the scientific inference problem.

At first sight, it appears as if the Bayesian model averaging approach on which we have concentrated is relevant solely within the $\mathcal{M}$-closed perspective, because it consists of averaging over a class of models that is specified in advance, at least in principle. However, we believe that the basic principles of Bayesian model averaging also apply, perhaps with even greater force, to the $\mathcal{M}$-open situation. This is because in the $\mathcal{M}$-open situation, with its open and less constrained search for better models, model uncertainty may be even greater than in the $\mathcal{M}$-closed case, and so it may be more important for well-calibrated inference to take account of it.

The Occam's Window approach of Madigan and Raftery (1994) can be viewed as an implementation of the $\mathcal{M}$-open perspective, since the model class (and not just the inferences) used for model averaging is effectively updated as new variables and data become available. This is because, as new variables and models are discovered that provide better predictions, they are included in the Bayesian model averaging. Similarly, when new and superior models are discovered, older models that do not predict as well relative to the new ones are excluded from the Bayesian model averaging in the Occam's window approach, whereas in the original ("$\mathcal{M}$-closed") Bayesian model averaging, all models ever considered continue to be included in the model averaging, even if they have been effectively discredited.

8.4 Conclusion

We have argued here that it can be important to take account of model uncertainty, or uncertainty about statistical structure, when making inferences. A coherent and conceptually simple way to do this is Bayesian model averaging, and we have outlined several practical implementation strategies for this, as well as pointing to some freely available software. We have provided implementation details for four classes of models: linear regression models, generalized linear models, survival analysis and graphical models.

In theory, BMA provides better average predictive performance than any single model that could be selected, and this theoretical result has now been supported in practice in a range of applications involving different model classes and types of data. BMA also provides inference about parameters that takes account of this sometimes important source of uncertainty, and in our examples we have found that BMA-based confidence intervals are better calibrated than single-model based confidence intervals; the latter tend to be too narrow.

One common criticism of model averaging is that the results may be too complicated to present easily. However, if desired, presentation can focus on the posterior effect probabilities, which are easy to understand, arguably more so than $p-$values. An alternative is to focus on a single "best" model when presenting results, using the BMA analysis as a formally justified form of sensitivity analysis and a basis for inference about parameters of interest that takes account of model uncertainty. Model averaging also avoids the problem of having to defend the choice of model, thus simplifying presentation. Indeed, model averaging results are robust to model choice (but not necessarily robust to model class, as discussed in Section 8.1). Model averaging also allows users to incorporate several competing models in the estimation process; thus model averaging may offer a committee of scientists a better estimation method than the traditional approach of trying to get the committee to agree on one best model.

Another potential concern is that model averaging tends to produce higher estimates of variance than do estimates that ignore model uncertainty. Why would practitioners use model averaging when they are less likely to get significant results? The simple answer is that model averaging is more correct, because it takes account of a source of uncertainty that analyses based on model selection ignore. The implication is that standard analyses probably tend to find significant results too often. Also, if results are significant under model averaging, then conclusions are more robust than those that depend upon the particular model that has been selected.

There are many open research questions related to BMA. These include the choice of prior distribution, investigation of the performance of BMA when the true model is not in the model class, the performance and tuning of Occam's window and similar approaches as against the more computationally demanding full BMA, the development of BMA methodology for model classes not considered here, and the development of more efficient computational approaches. As more examples of the dangers of ignoring model uncertainty are publicized, as computing power continues to expand, and as the size of databases, the numbers of variables, and hence the numbers of possible models increase, we predict that accounting for model uncertainty will become an integral part of statistical modeling.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In Petrox, B. and Caski, F., editors, *Second International Symposium on Information Theory*, page 267.

Barnard, G. A. (1963). New methods of quality control. *J. Roy. Statist. Soc. Ser. A*, 126:255.

Bates, J. M. and Granger, C. W. J. (1969). The combination of forecasts. *Operational Research Quarterly*, 20:451–468.

Berger, J. O. and Delampady, M. (1987). Testing precise hypotheses. *Statis. Sci.*, 2:317–352.

Berger, J. O. and Sellke, T. (1987). Testing a point null hypothesis (with discussion). *J. Amer. Statist. Assoc.*, 82:112–122.

Bernardo, J. and Smith, A. (1994). *Bayesian Theory*. Wiley: Chichester.

Besag, J. E., Green, P., Higdon, D., and Mengerson, K. (1995). Bayesian computation and stochastic systems. *Statis. Sci.*, 10:3–66.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, 26:123–140.

Breiman, L. and Friedman, J. H. (1985). Estimating optimal transformations for multiple regression and correlation (with discussion). *J. Amer. Statist. Assoc.*, 80:580–619.

Brozek, J., Grande, F., Anderson, J., and Keys, A. (1963). Densitometric analysis of body composition: Revision of some quantitative assumptions. *Annals of the New York Academy of Sciences*, 110:113–140.

Buckland, S. T., Burnham, K. P., and Augustin, N. H. (1997). Model selection: An integral part of inference. *Biometrics*, 53:275–290.

Buntine, W. (1992). Learning classification trees. *Statistics and Computing*, 2:63–73.

Carlin, B. P. and Chib, S. (1993). Bayesian model choice via Markov chain Monte Carlo. *J. Roy. Statist. Soc. Ser. B*, 55:473–484.

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.

Chan, P. K. and Stolfo, S. J. (1997). On the accuracy of meta-learning for scalable data mining. *Journal of Intelligent Integration of Information*, 8(1):5–28.

Chatfield, C. (1995). Model uncertainty, data mining, and statistical inference (with discussion). *J. Roy. Statist. Soc. Ser. A*, 158:419–466.

Chib, S. and Greenberg, E. (1995). Understanding the Metropolis-Hastings algorithm. *The American Statistician*, 40(4):327–335.

Clemen, R. T. (1989). Combining forecasts: a review and annotated bibliography. *International Journal of Forecasting*, 5:559–583.

Clyde, M., DeSimone, H., and Parmigiani, G. (1996). Prediction via orthoganalized model mixing. *J. Amer. Statist. Assoc.*, 91:1197–1208.

Cox, D. R. (1972). Regression models and life tables (with discussion). *J. Roy. Statist. Soc. Ser. B*, 34:187–220.

Dawid, A. P. (1984). Statistical theory–the prequential approach. *J. Roy. Statist. Soc. Ser. A*, 147:278–292.

Dickinson, J. P. (1973). Some statistical results on the combination of forecasts. *Operational Research Quarterly*, 24:253–260.

Dijkstra, T. K. (1988). *On Model Uncertainty and its statistical implications*. Springer, Berlin.

Draper, D. (1995). Assessment and propagation of model uncertainty. *J. Roy. Statist. Soc. Ser. B*, 57:45–97.

Draper, D., Gaver, D. P., Goel, P. K., Greenhouse, J. B., Hedges, L. V., Morris, C. N., Tucker, J., and Waternaux, C. (1993). *Combining information: National Research Council Panel on Statistical Issues and Opportunities for Research in the Combination of Information.* National Academy Press, Washington, D.C.

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. Technical Report Rand Note N-2683-RC, The RAND Corporation, Santa Monica, California.

Edwards, W., Lindman, H., and Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review*, 70:193–242.

Fernández, C., Ley, E., and Steel, M. F. (1997). Statistical modeling of fishing activities in the north atlantic. Technical report, Department of Econometrics, Tilburg Univeristy, the Netherlands.

Fernández, C., Ley, E., and Steel, M. F. (1998). Benchmark priors for Bayesian model averaging. Technical report, Department of Econometrics, Tilburg Univeristy, the Netherlands.

Fleming, T. R. and Harrington, D. H. (1991). *Counting Processes and Survival Analysis.* Wiley, New York.

Freedman, D. A., Navidi, W., and Peters, S. C. (1988). On the impact of variable selection in fitting regression equations. In Dijkstra, T. K., editor, *On Model Uncertainty and its statistical implications*, pages 1–16. Springer, Berlin.

Freund, Y. (1995). Boosting a weak learning algorithm by majority. *Information and Computation*, 121:256–285.

Fried, L. P., Borhani, N. O., Enright, P., Furberg, C. D., Gardin, J. M., Kronmal, R. A., Kuller, L. H., Manolio, T. A., Mittelmark, M. B., Newman, A., O'Leary, D. H., Psaty, B., Rautaharju, P., Tracy, R. P., and Weiler, P. G. (1991). The cardiovascular health study: Design and rationale. *Annals of Epidemiology*, 1:263–276.

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

Geisser, S. (1980). Discussion on sampling and Bayes' inference in scientific modeling and robustness (by G. E. P. Box). *J. Roy. Statist. Soc. Ser. A*, 143:416–417.

George, E. I. (1986a). Combining minimax shrinkage estimators. *J. Amer. Statist. Assoc.*, 81:437–445.

George, E. I. (1986b). A formal Bayes multiple shrinkage estimator. *Communications in Statistics: Part A - Theory and Methods (Special issue on Stein-type multivariate estimation*, 15(7):2099–2114.

George, E. I. (1986c). Minimax multiple shrinkage estimation. *Annals of Statistics*, 14:188–205.

George, E. I. (1999). Bayesian model selection. In *Encyclopedia of Statistical Sciences Update*, volume 3. Wiley, New York, to appear.

George, E. I. and McCulloch, R. (1993). Variable selection via Gibbs sampling. *J. Amer. Statist. Assoc.*, 88:881–889.

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

Good, I. J. (1952). Rational decisions. *J. Roy. Statist. Soc. Ser. B*, 14:107–114.

Grambsch, P. M., Dickson, E. R., Kaplan, M., LeSage, G., Fleming, T. R., and Langworthy, A. L. (1989). Extramural cross-validation of the Mayo primary biliary cirrhosis survival model establishes its generalizability. *Hepatology*, 10:846–850.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82:711–732.

Heckerman, D., Geiger, D., and Chickering, D. M. (1994). Learning Bayesian networks: the combination of knowledge and statistical data. In de Mantaras, B. L. and Poole, D., editors, *Uncertainty in Artificial Intelligence, Proceedings of the Tenth Conference*, pages 293–301. Morgan Kaufman: San Francisco.

Hodges, J. S. (1987). Uncertainty, policy analysis, and statistics. *Statis. Sci.*, 2:259–291.

Hoeting, J. A. (1994). *Accounting for Model Uncertainty in Linear Regression.* PhD thesis, Univ. of Washington.

Hoeting, J. A., Raftery, A. E., and Madigan, D. (1996). A method for simultaneous variable selection and outlier identification in linear regression. *Journal of Computational Statistics*, 22:251–271.

Hoeting, J. A., Raftery, A. E., and Madigan, D. (1999). Bayesian simultaneous variable and transformation selection in linear regression. Technical Report 9905, Department of Statistics, Colorado State Univ., available at www.stat.colostate.edu.

Ibrahim, J. G. and Laud, P. W. (1994). A predictive approach to the analysis of designed experiments. *J. Amer. Statist. Assoc.*, 89:309–319.

Johnson, R. W. (1996). Fitting percentage of body fat to simple body measurements. *Journal of Statistics Education*, 4.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *J. Amer. Statist. Assoc.*, 90:773–795.

Kass, R. E. and Wasserman, L. (1995). A reference Bayesian test for nested hypothese with large samples. *J. Amer. Statist. Assoc.*, 90:928–934.

Katch, F. and McArdle, W. (1993). *Nutrition, Weight Control, and Exercise.* Williams and Wilkins, Philadelphia, 4th edition.

Kearns, M. J., Schapire, R. E., and Sellie, L. M. (1994). Toward efficient agnostic learning. *Machine Learning*, 17:115–142.

Kincaid, D. and Cheney, W. (1991). *Numerical Analysis.* Brooks/Cole, Pacific Grove, California.

Kuk, A. Y. C. (1984). All subsets regression in a proportional hazards model. *Biometrika*, 71:587–592.

Kwok, S. and Carter, C. (1990). Multiple decision trees. In Shachter, R., Levitt, T., Kanal, L., and Lemmer, J., editors, *Uncertainty in Artificial Intelligence 4*, pages 323–349. North Holland.

Lauritzen, S. L. (1996). *Graphical Models.* Clarendon Press, Oxford.

Lauritzen, S. L., Thiesson, B., and Spiegelhalter, D. J. (1994). Diagnostic systems created by model selection methods - a case study. In Cheeseman, P. and Oldford, W., editors, *Uncertainty in Artificial Intelligence 4*, pages 143–152. Springer Verlag.

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

Leamer, E. E. (1978). *Specification Searches.* Wiley, New York.

Lohman, T. (1992). *Advance in Body Composition Assessment, Current Issues in Exercise Science (Monograph Number 3).* Human Kinetics Publishers, Champaign, IL.

Madigan, D., Andersson, S. A., Perlman, M., and Volinsky, C. T. (1996a). Bayesian model averaging and model selection for Markov equivalence classes of acyclic digraphs. *Communications in Statistics: Theory and Methods*, 25:2493–2520.

Madigan, D., Andersson, S. A., Perlman, M. D., and Volinsky, C. T. (1996b). Bayesian model averaging and model selection for Markov equivalence classes of acyclic digraphs. *Communications in Statistics - Theory and Methods*, 25:2493–2519.

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

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

Madigan, D. and Raftery, A. E. (1994). Model selection and accounting for model uncertainty in graphical models using Occam's window. *J. Amer. Statist. Assoc.*, 89:1535–1546.

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

Madigan, D. and York, J. (1995). Bayesian graphical models for discrete data. *Int. Statist. Rev.*, 63:215–232.

Markus, B. H., Dickson, E. R., Grambsch, P. M., Fleming, T. R., Mazzaferro, V., Klintmalm, G., Weisner, R. H., Van Thiel, D. H., and Starzl, T. E. (1989). Efficacy of liver transplantation in patients with primary biliary cirrhosis. *New England Journal of Medicine*, 320:1709–1713.

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

McCullagh, P. and Nelder, J. (1989). *Generalized Linear Models*. Chapman & Hall, London, 2nd edition.

Miller, A. J. (1990). *Subset Selection in Regression*. Chapman and Hall.

Penrose, K., Nelson, A., and Fisher, A. (1985). Generalized body composition prediction equation for men using simple measurement techniques (abstract). *Medicine and Science in Sports and Exercise*, 17:189.

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. (1993). Bayesian model selection in structural equation models. In Bollen, K. and Long, J., editors, *Testing Structural Equation Models*, number 163–180. Sage, Newbury Park, CA.

Raftery, A. E. (1995). Bayesian model selection in social research (with discussion). In Marsden, P. V., editor, *Sociological Methodology 1995*, pages 111–195. Blackwells Publishers, Cambridge, Mass.

Raftery, A. E. (1996). Approximate Bayes factors and accounting for model uncertainty in generalised linear models. *Biometrika*, 83:251–266.

Raftery, A. E., Madigan, D., and Hoeting, J. (1997). Bayesian model averaging for linear regression models. *J. Amer. Statist. Assoc.*, 92:179–191.

Raftery, A. E., Madigan, D., and Volinsky, C. T. (1996). Accounting for model uncertainty in survival analysis improves predictive performance (with discussion). In Bernardo, J., Berger, J., Dawid, A., and Smith, A., editors, *Bayesian Statistics 5*, pages 323–349. Oxford University Press.

Rao, J. S. and Tibshirani, R. (1997). The out-of-bootstrap method for model averaging and selection. Technical report, Department of Statistics, Univ. of Toronto.

Regal, 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.

Roberts, H. V. (1965). Probabilistic prediction. *J. Amer. Statist. Assoc.*, 60:50–62.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6:461–46.

Smith, A. F. M. and Roberts, G. O. (1993). Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods (with discussion). *J. Roy. Statist. Soc. Ser. B*, 55:3 – 23.

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

Spiegelhalter, D. J., Dawid, A., Lauritzen, S., and Cowell, R. (1993). Bayesian analysis in expert systems (with discussion). *Statis. Sci.*, 8:219–283.

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

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. *J. Roy. Statist. Soc. Ser. B*, 55:829–836.

Thompson, E. A. and Wijsman, E. M. (1990). Monte Carlo methods for the genetic analysis of complex traits. Technical Report 193, Department of Statistics, Univ. of Washington.

Tierney, L. and Kadane, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *J. Amer. Statist. Assoc.*, 81:82–86.

Volinsky, C. T. (1997). *Bayesian Model Averaging for Censored Survival Models.* PhD thesis, Univ. of Washington.

Volinsky, C. T., Madigan, D., Raftery, A. E., and Kronmal, R. A. (1997). Bayesian Model Averaging in proportional hazard models: Assessing the risk of a stroke. *Applied Statistics*, 46(3):433– 448.

Weisberg, S. (1985). *Applied Linear Regression.* Wiley, New York, 2nd edition.

Wolpert, D. H. (1992). Stacked generalization. *Neural Networks*, 5:241–259.

York, J., Madigan, D., Heuch, I., and Lie, R. T. (1995). Estimating a proportion of birth defects by double sampling: a Bayesian approach incorporating covariates and model uncertainty. *Applied Statistics*, 44:227–242.