Statistical Science
1999, Vol. 14, No. 4, 382–417

Bayesian Model Averaging: A Tutorial

Jennifer A. Hoeting, David Madigan, Adrian E. Raftery and Chris T. Volinsky

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.

Key words and phrases: Bayesian model averaging, Bayesian graphical models, learning; model uncertainty, Markov chain Monte Carlo.

CONTENTS

Jennifer A. Hoeting is Assistant Professor, Department of Statistics, Colorado State University, Fort Collins, Colorado, 80523. David Madigan is with ATT Labs Research, 180 Park Avenue, P.O. Box 971, Florham Park, New Jersey 07932-0000. Adrian E. Raftery is Professor of Statistics and Sociology, Department of Statistics, University of Washington, Box 354322, Seattle, Washington 98195-4322. Chris T. Volinsky is with ATT Labs Research, 180 Park Ave., P.O. Box 971, Florham Park, New Jersey 07932-0000.

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, page 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

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

This is an average of the posterior distributions under each of the models considered, weighted by their posterior model probability. In (1), $M_1, \ldots, M_K$ are the models considered. The posterior probability for model M_k is given by

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

where

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

is the integrated likelihood of model M_k, θ_k is the vector of parameters of model M_k (e.g., for regression $\theta = (\beta, \sigma^2)$), $\text{pr}(\theta_k \mid M_k)$ is the prior density of θ_k under model M_k, $\text{pr}(D \mid \theta_k, M_k)$ is the likelihood, and $\text{pr}(M_k)$ 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 $\mathscr{M}$, the set of all models being considered.

The posterior mean and variance of Δ are as follows:

$$E[\Delta \mid D] = \sum_{k=0}^{K} \hat{\Delta}_k \text{pr}(M_k \mid D),$$

$$\text{Var}[\Delta \mid D] = \sum_{k=0}^{K} \left(\text{Var}[\Delta \mid D, M_k] + \hat{\Delta}_k^2\right)\text{pr}(M_k \mid D) - E[\Delta \mid D]^2,$$

where $\hat{\Delta}_k = E[\Delta \mid D, M_k]$ (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 $\mathscr{M}$. 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, discussed in the sections of this paper as noted:

- 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 $\text{pr}(M_k)$, 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 1970s 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 (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

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

should be excluded from (1) where C is chosen by the data analyst. Their second, optional, principle, appealing to Occam's razor, led them to exclude complex models which receive less support from the data than their simpler counterparts. More formally, they also exclude from (1) models belonging to:

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

and (1) is replaced by

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

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

This greatly reduces the number of models in the sum in (1) and now all that is required is a search strategy to identify the models in $\mathscr{A}$. 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 $\text{pr}(M_0 \mid D)/\text{pr}(M_1 \mid D)$. 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 1/20 and 1 for O_L and O_R, respectively (see Figure 1). Raftery, Madigan and Volinsky (1996) show that adopting 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.

Another way to search for the models in $\mathscr{A}$ is suggested by Volinsky, Madigan, Raftery and Kronmal (1997). They use the "leaps and bounds" algorithm (Furnival and Wilson, 1974) to rapidly identify models to be used in the summation of (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 $\mathscr{M}$ denote the space of models under consideration. One can construct a Markov chain $\{M(t)\}$, $t = 1, 2, \ldots$ with state space $\mathscr{M}$ and equilibrium distribution $\text{pr}(M_i \mid D)$ and simulate this Markov chain to obtain observations $M(1), \ldots, M(N)$. Then for any

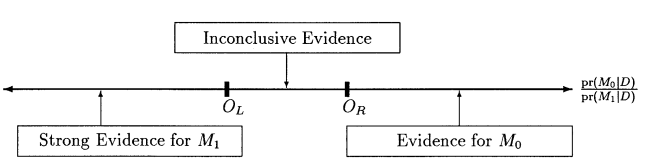

FIG. 1. *Occam's window: interpreting the posterior odds.*

function $g(M_i)$ defined on $\mathscr{M}$, 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,

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

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

To construct the Markov chain, define a neighborhood $\text{nbd}(M)$ for each $M \in \mathscr{M}$. 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 $M' \notin \text{nbd}(M)$ and $q(M \to M')$ nonzero for all $M' \in \text{nbd}(M)$. If the chain is currently in state M, proceed by drawing M' from $q(M \to M')$. M' is accepted with probability

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

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, Andersson, Perlman and Volinsky (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, Madigan, Heuch and Lie (1995) incorporated missing data and a latent variable into their MC^3 scheme. For linear models, Raftery, Madigan and Hoeting (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, DeSimone and Parmigiani (1996) introduced an importance sampling strategy based on a reexpression 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, Green, Higdon and Mengerson (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, Madigan and Hoeting, 1997), closed form integrals for the marginal likelihood, (3), are available. The Laplace method (Tierney and Kadane, 1986) can provide an excellent approximation to $\mathrm{pr}(D \mid M_k)$; 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, Madigan and Volinsky (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

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

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, for example, 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, Madigan and Hoeting (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, Ley and Steel (1997, 1998) offer an alternative prior structure aiming at a more automatic choice of hyperparameters.

Hoeting, Raftery and Madigan (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 $Y^{(\rho)} = X\beta + \varepsilon$ where $\varepsilon \sim N(0, \sigma^2 I)$ and

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

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 $(-1, 0, 0.5, 1)$, 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 $Y = X\beta + \varepsilon$ 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

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

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, 94). 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,

$$B_{10} = \text{pr}(D \mid M_1)/\text{pr}(D \mid M_0),$$

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

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

where $\alpha_k = \text{pr}(M_k)/\text{pr}(M_0)$ 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 $X_i = (x_{i1}, \ldots, x_{ip})$ is a corresponding vector of independent variables, for $i = 1, \ldots, n$. A generalized linear model M_1 is defined by specifying $\text{pr}(Y_i \mid X_i, \beta)$ in such a way that $E[Y_i \mid X_i] = \mu_i$, $\text{Var}[Y_i \mid X_i] = \sigma^2 v(\mu_i)$ and $g(\mu_i) = X_i\beta$, where $\beta = (\beta_1, \ldots, \beta_p)^T$; 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 $x_{i1} = 1$ $(i = 1, \ldots, n)$. 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 $\beta_j = 0$ $(j = 2, \ldots, p)$, 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:

$$\begin{aligned} p(D \mid M_k) \approx{}& (2\pi)^{p_k/2} \mid \Psi \mid^{1/2} \\ & \cdot \text{pr}(D \mid \tilde{\beta}_k, M_k)\text{pr}(\tilde{\beta}_k \mid M_k), \end{aligned} \tag{10}$$

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 $h(\beta_k) = \log\{\text{pr}(D \mid \beta_k, M_k)\text{pr}(\beta_k \mid M_k)\}$, 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 (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 $E[\beta_k \mid M_k] = \omega_k$ and $\mathrm{Var}[\beta_k \mid M_k] = W_k$. Then approximating the posterior mode, $\tilde{\beta}_k$, by a single step of the Newton–Raphson algorithm (e.g., Kincaid and Cheney, 1991, page 26) starting from the MLE, $\hat{\beta}_k$, and substituting the result into (10) yields the approximation

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

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

$$\begin{aligned} E_k = {} & 2\lambda_k(\hat{\beta}_k) + \lambda_k'(\hat{\beta}_k)^T(F_k + G_k)^{-1} \\ & \cdot \{2 - F_k(F_k + G_k)^{-1}\}\lambda_k'(\hat{\beta}_k) \\ & - \log \mid F_k + G_k \mid + p_k \log(2\pi), \end{aligned}$$

where F_k is the expected Fisher information matrix, $G_k = W_k^{-1}$, $\lambda_k(\beta_k) = \log \mathrm{pr}(\beta_k \mid M_k)$ 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 $O(n^{-1/2})$. 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 $O(n^{-1})$.

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, Madigan and Volinsky (1996) and Volinsky et al. (1997), VMRK hereafter, adopted a number of approximations. In particular, VMRK used the MLE approximation,

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

and the Laplace approximation,

$$\begin{aligned} \log \mathrm{pr}(D \mid M_k) \approx {} & \log \mathrm{pr}(D \mid \hat{\beta}_k, M_k) \\ & - d_k \log n, \end{aligned} \tag{13}$$

where d_k is the dimension of β_k. This is the Bayesian information criterion (BIC) approximation. In (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 $\hat{\beta}_k$, $\mathrm{var}(\hat{\beta}_k)$, and R_k^2 for each model M_k returned. Lawless and Singhal (1978) and Kuk (1984) provided a modified algorithm for nonnormal 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 ($\mathscr{A}$) plus many models not in $\mathscr{A}$. VMRK used the approximate likelihood ratio test to reduce the remaining subset of models to those most likely to be in $\mathscr{A}$. 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 (4). (VMRK set $C' = C^2$ and almost no models in $\mathscr{A}$ 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 $\mathscr{A}$.

For the models in $\mathscr{A}$, VMRK calculated posterior model probabilities by normalizing over the model set, as in (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 nonzero ("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 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 $\mathscr{X}_v$. 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 nondescendants 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

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

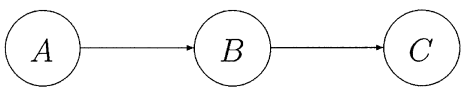

FIG. 2. *A simple discrete graphical model.*

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, Geiger and Chickering (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 reexpression of the usual Bayes factor comparing two models, M_0 and M_1:

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

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 (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, Dawid, Lauritzen and Cowell (1993) and Lauritzen, Thiesson and 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.

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

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

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 (1995) 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 (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, page 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,

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

and measures the predictive performance of BMA with

$$-\sum_{d \in D^T} \log \left\{ \sum_{M \in \mathscr{A}} \text{pr}(d \mid M, D^B) \text{pr}(M \mid D^B) \right\}. \tag{18}$$

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. Other applications of BMA can be found in a number of works (Chatfield, 1995; Draper, 1995; Fernández, Ley and Steel, 1997; Hoeting, Raftery and Madigan, 1999; Hoeting, Raftery and Madigan, 1996; Madigan, Andersson, Perlman and Volinsky, 1996b; Madigan and Raftery, 1994; Raftery, Madigan and Hoeting, 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, Madigan and Volinsky (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.

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. Table 2 lists the models with the highest posterior probabilities. 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%. 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, $P(\beta \neq 0 \mid D)$, 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 its mass at zero. This shrinks

TABLE 1
PBC example: summary statistics and BMA estimates

Variable	Range	Mean	Mean $\beta \mid D$	SD $\beta \mid D$	$P(\beta \neq 0 \mid D)$
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
	0.5 = 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			

TABLE 2
PBC example: results for the full data set[1]

Model no.	Age	Edema	Bili	Albu	UCopp	SGOT	Prothromb	Hist	PMP	Log lik
1	•	•	•	•	•		•		0.17	−174.4
2	•	•	•	•	•		•	•	0.07	−172.6
3	•	•	•	•	•			•	0.07	−172.5
4	•		•	•	•		•		0.06	−172.2
5[2]	•	•	•	•			•		0.05	−172.0
6	•	•	•	•	•				0.05	−172.0
7	•	•	•	•	•	•	•		0.04	−171.7
8	•	•	•	•		•	•		0.04	−171.4
9	•	•	•	•		•	•	•	0.04	−171.3
10	•	•	•	•	•	•	•	•	0.03	−170.9
$\Pr_{MA}[\beta_i \neq 0]$	1.00	0.84	1.00	1.00	0.72	0.22	0.78	0.34		

[1] PMP denotes the posterior model probability. Only the 10 models with the highest PMP values are shown.
[2] Model selected by FH.

the estimate toward 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 $p < 0.05$ and $p < 0.01$ 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, Lindman and Savage, 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 $P(\beta \neq 0 \mid D) > 99\%$]. For the five variables in Table 3, however, the two approaches lead to qualitatively different conclu-

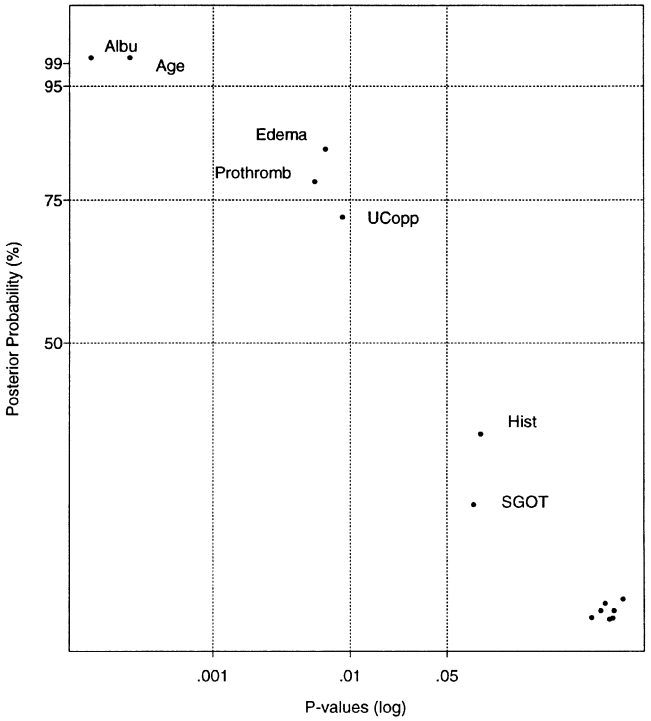

FIG. 3. *PBC example: posterior effect probabilities from BMA versus p-values from the stepwise variable selection model.*

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	0.007**	84
Prothrombin	0.006**	78
Urine copper	0.009**	72
Histology	0.09*	34
SGOT	0.08	22

sions. 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 seven variables (the clump of points in the lower right corner of Figure 3), p-values and posterior effect probabilities agree in saying that there is little or no evidence 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, $P(\beta \neq 0 \mid D) = 22\%$, so that the data are indecisive, while for the treatment effect of DPCA, $P(\beta \neq 0 \mid D) = 2\%$, 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 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 (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

TABLE 4
PBC example: partial predictive scores for model selection techniques and BMA[1]

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

[1] FH denotes the model selected by Fleming and Harrington.

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%

may prove more practical than numerical prediction. To assess the performance of a single model with respect to this goal we proceed as follows:

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 $\mathscr{A}$ to get estimated coefficients $\hat{\beta}_k$.
2′. Calculate risk scores ($\mathbf{x_i}^T\hat{\beta}_k$) under each model in $\mathscr{A}$ 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, page 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 noninvasive 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, Nelson and Fisher (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, pages 242–244). Body density was then used to estimate percent body fat using Brozek's equation

TABLE 6
Body fat example: summary statistics for full data set[1]

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

[1]Abdomen circumference was measured at the umbilicus and level with the iliac crest. Wrist circumference (cm) was measured distal to the styloid processes.

TABLE 7
Body fat example: least squares regression results from the full model[1]

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

[1]Residual standard error = 4, $R^2 = 0.75$, $N = 251$, F-statistic = 53.62 on 13 and 237 df, p-value <0.0001.

(Brozek, Grande, Anderson and Keys, 1963),

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

For more details on the derivation of (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 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 be 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, we computed the posterior model probability for all possible models using the diffuse (but proper) prior distributions derived by Raftery, Madigan and Hoeting (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, $P(\beta_i \neq 0 \mid D)$, 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 $P(\beta_i \neq 0 \mid D)$, 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

TABLE 8
Body fat example: comparison of BMA results to model selected using standard model selection methods[1]

Predictor		Bayesian model averaging: Mean $\beta \mid D$	Bayesian model averaging: SD $\beta \mid D$	Bayesian model averaging: $P(\beta \neq 0 \mid D)$	Stepwise model 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	

[1]Stepwise, minimum Mallow's C_p, and maximum adjusted R^2 all selected the same model. The predictors are sorted by $P(\beta_i \neq 0 \mid D)$ 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).

TABLE 9
Body fat example: Ten models with highest posterior model probability (PMP)

X_2	X_4	X_6	X_8	X_{11}	X_{12}	X_{13}	PMP
•		•				•	0.14
•		•			•	•	0.14
•		•					0.12
•		•		•		•	0.05
•		•	•				0.03
•	•	•					0.03
•		•			•		0.02
•		•		•			0.02
•	•	•			•	•	0.02
•	•	•			•		0.02

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 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, e.g., 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 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 noncentral Student's t distributions. The spike in the plot of the posterior distribution corresponds to $P(\beta_{13} = 0 \mid D) = 0.38$. 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).

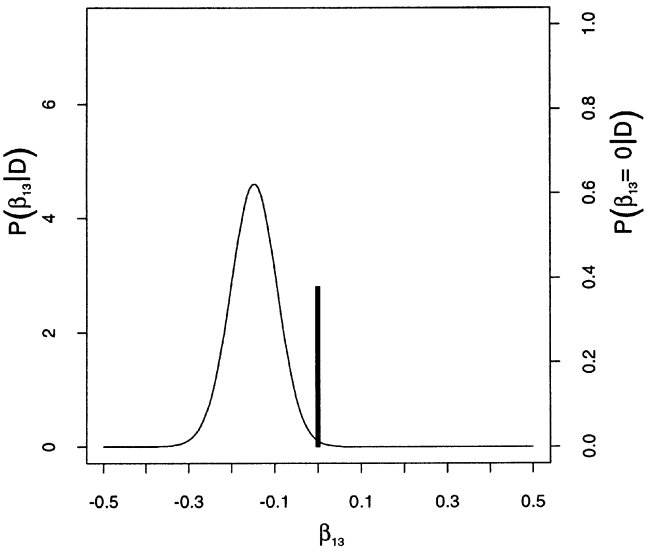

FIG. 4. *Body fat example: BMA posterior distribution for β_{13}, the coefficient for wrist circumference. The spike corresponds to $P(\beta_{13} = 0 \mid D) = 0.38$. The vertical axis on the left corresponds to the posterior distribution for β_{13} and the vertical axis on the right corresponds to the posterior distribution for β_{13} equal to 0. The density is scaled so that the maximum of the density is equal to $P(\beta_{13} \neq 0 \mid D)$ on the right axis.*

TABLE 10
Body fat example: performance comparison[1]

Method	Model	Predictive coverage %
BMA	Model averaging	90.8
Stepwise and 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

[1]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.

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, Nelson and Fisher, 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 (Table 10). 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 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 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, for example, 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, Navidi and Peters (1988) have shown that this does not necessarily give a satisfactory solution to the problem.

George (1986a, b, c) 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, Burnham and Augustin (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, Burnham and Augustin (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, e.g., Kearns, Schapire and Sellie 1994; Chan and Stolfo, 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, 1977). 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, pages 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.

ACKNOWLEDGMENT

Research supported in part by NSF Grant DMS-98-06243 and the Office of Naval Research (N00014-96-1-1092). We thank Leon Gleser (Executive Editor) and two anonymous referees for valuable comments.

REFERENCES

AKAIKE, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory* (B. Petrox and F. Caski, eds.) 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. *Statist. 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. *Statist. 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. *Ann. New York Acad. Sci.* **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. *Statist. Comput.* **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. *Canad. J. Statist.* **19** 399–405.

CHAN, P. K. and STOLFO, S. J. (1996). On the accuracy of meta-learning for scalable data mining. *J. Intelligent Integration of Information* **8** 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. *Amer. Statist.* **40** 327–335.

CLEMEN, R. T. (1989). Combining forecasts: a review and annotated bibliography. *Internat. J. 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, DC.

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, 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, Dept. Econometrics, Tilburg Univ., The Netherlands.

FERNÁNDEZ, C., LEY, E. and STEEL, M. F. (1998). Benchmark priors for Bayesian model averaging. Technical report, Dept. Econometrics, Tilburg Univ., 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 *On Model Uncertainty and Its Statistical Implications* (T. K. Dijkstra, ed.) 1–16. Springer, Berlin.

FREUND, Y. (1995). Boosting a weak learning algorithm by majority. *Inform. and Comput.* **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 GEPB). *J. Roy. Statist. Soc. Ser. A* **143** 416–417.

GEORGE, E. and MCCULLOCH, R. (1993). Variable selection via Gibbs sampling. *J. Amer. Statist. Assoc.* **88** 881–889.

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. *Commun. Statist. Theory Methods* (*Special issue on Stein-type multivariate estimation*) **15** 2099–2114.

GEORGE, E. I. (1986c). Minimax multiple shrinkage estimation. *Ann. Statist.* **14** 188–205.

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

GOOD, I. J. (1950). Probability and the weighing of evidence. 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 *Uncertainty in Artificial Intelligence, Proceedings of the Tenth Conference* (B. L. de Mantaras and D. Poole, eds.) 293–301. Morgan Kaufman, San Francisco.

HODGES, J. S. (1987). Uncertainty, policy analysis, and statistics. *Statist. Sci.* **2** 259–291.

HOETING, J. A. (1994). Accounting for model uncertainty in linear regression. Ph.D. dissertation, Univ. Washington, Seattle.

HOETING, J. A., RAFTERY, A. E. and MADIGAN, D. (1996). A method for simultaneous variable selection and outlier identification in linear regression. *J. Comput. Statist.* **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, Dept. 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. *J. 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 hypotheses with large samples. *J. Amer. Statist. Assoc.* **90** 928–934.

KATCH, F. and MCARDLE, W. (1993). *Nutrition, Weight Control, and Exercise*, 4th ed. Williams and Wilkins, Philadelphia.

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, CA.

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 *Uncertainty in Artificial Intelligence* (R. Shachter, T. Levitt, L. Kanal and J. Lemmer, eds.) **4** 323–349. North-Holland, Amsterdam.

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 *Uncertainty in Artificial Intelligence* (P. Cheeseman and W. Oldford, eds.) **4** 143–152. Springer Berlin.

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*. 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. *Comm. Statist. Theory 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. *Comm. Statist. Theory Methods* **25** 2493–2519.

MADIGAN, D., GAVRIN, J. and RAFTERY, A. E. (1995). Eliciting prior information to enhance the predictive performance of Bayesian graphical models. *Comm. Statist. Theory 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. 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 *Selecting Models from Data: Artificial Intelligence and Statistics* (P. Cheeseman and W. Oldford, eds.) **4** 91–100. Springer, Berlin.

MADIGAN, D. and YORK, J. (1995). Bayesian graphical models for discrete data. *Internat. 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 J. 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*, 2nd ed. Chapman & Hall, London.

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

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 *Testing Structural Equation Models* (K. Bollen and J. Long, eds.) 163–180. Sage, Newbury Park, CA.

RAFTERY, A. E. (1995). Bayesian model selection in social research (with discussion). In *Sociological Methodology 1995* (P. V. Marsden, ed.) 111–195. Blackwell, Cambridge, MA.

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 *Bayesian Statistics 5* (J. Bernardo, J. Berger, A. Dawid and A. Smith, eds.) 323–349. Oxford Univ. Press.

RAO, J. S. and TIBSHIRANI, R. (1997). The out-of-bootstrap method for model averaging and selection. Technical report, Dept. Statistics, Univ. 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. *Ann. Statist.* **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). *Statist. 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, Dept. Statistics, Univ. Washington, Seattle.

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. Ph.D. dissertation, Univ. Washington, Seattle.

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. *J. Roy. Statist. Soc. Ser. C* **46** 433–448.

WEISBERG, S. (1985). *Applied Linear Regression*, 2nd ed. Wiley, New York.

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. *J. Roy. Statist. Soc. Ser. C* **44** 227–242.

Comment

M. Clyde

I would like to begin by thanking the authors for providing an excellent historical perspective and summarization of literature on BMA. They provided insightful examples in a wide variety of applications, demonstrated that BMA provides much better predictive performance than using a single model and have developed useful software so that BMA can be put into practice for a wide class of models. I would like to comment first on some interesting connections between some of the algorithms used for implementing BMA and then discuss issues related to the choice of prior distributions for BMA.

Merlise Clyde is Professor, Institute of Statistics and Decision Sciences, 219A Old Chemistry, Box 90251, Duke University, Durham, North Carolina 27708–0251 (e-mail: clyde@stat.duke.edu).

1. IMPLEMENTING MODEL AVERAGING

On the surface, model averaging is straightforward to implement: one needs the marginal distribution of the data, the prior probabilities of models and the posterior distribution of the quantity of interest conditional on each model. In linear regression, these components are available in closed form (at least for nice prior distributions); for generalized linear models and many other models, Laplace's method of integration can provide accurate approximations to marginal distributions. One problem is that, in many applications, the model space is too

large to allow enumeration of all models, and beyond 20–25 variables (George and McCulloch, 1997) estimation of posterior model probabilities and BMA must be based on a sample of models.

Deterministic search for models using branch and bounds or leaps and bounds algorithms (Furnival and Wilson, 1974) is efficient for problems with typically fewer than 30 variables. For larger problems, such as in non-parametric models or generalized additive models, these methods are too expensive computationally or do not explore a large enough region of the model space, producing poor fits (Hanson and Kooperberg, 1999). Markov chain Monte Carlo (MCMC) methods provide a stochastic method of obtaining samples from the posterior distributions $f(M_k \mid \mathbf{Y})$ and $f(\boldsymbol{\beta}_{M_k} \mid M_k, \mathbf{Y})$ and many of the algorithms that the authors mention can be viewed as special cases of reversible jump MCMC algorithms.

1.1 Reversible Jump MCMC

The reversible jump algorithm (Green, 1995) is applicable when the models have parameter spaces of different dimensions and can be described as follows. If the current state of the chain is $(\boldsymbol{\beta}_M, M)$, then:

1. Propose a jump to a new model M^* of dimension $p^*_{M^*}$ with probability $j(M^* \mid M)$ given the current model M.
2. Generate a vector $\mathbf{u}$ from a continuous distribution $q(\mathbf{u} \mid \boldsymbol{\beta}_M, M, M^*)$.
3. Set $(\boldsymbol{\beta}^*_{M^*}, \mathbf{u}^*) = g_{M,M^*}(\boldsymbol{\beta}_M, \mathbf{u})$ where g is a bijection between $(\boldsymbol{\beta}_M, \mathbf{u})$ and $(\boldsymbol{\beta}^*_{M^*}, \mathbf{u}^*)$ and the lengths of $\mathbf{u}$ and $\mathbf{u}^*$ satisfy $p_M + \dim(\mathbf{u}) = p^*_{M^*} + \dim(\mathbf{u}^*)$, where p_M and p_{M^*} are the dimensions of M and M^*, respectively.
4. Accept the proposed move to $(\boldsymbol{\beta}^*_{M^*}, M^*)$ with probability

$$\alpha = \min\Bigg\{1, f(\mathbf{Y} \mid \boldsymbol{\beta}^*_{M^*}, M^*) f(\boldsymbol{\beta}^*_{M^*} \mid M^*) f(M^*) \times j(M \mid M^*) q(\mathbf{u}^* \mid \boldsymbol{\beta}^*_{M^*}, M^*, M) \times \big[f(\mathbf{Y} \mid \boldsymbol{\beta}_M, M) f(\boldsymbol{\beta}_M \mid M) f(M) \times j(M^* \mid M) q(\mathbf{u} \mid \boldsymbol{\beta}_M, M, M^*)\big] \times \left|\frac{\partial g_{M,M^*}(\boldsymbol{\beta}_M, \mathbf{u})}{\partial(\boldsymbol{\beta}_M, \mathbf{u})}\right|\Bigg\}.$$

In special cases, the posterior distribution of $\boldsymbol{\beta}_{M^*} \mid M^*$ is available in closed form with known normalizing constants. If one takes $q(\mathbf{u} \mid \boldsymbol{\beta}_M, M, M^*)$ to be the posterior distribution of $\boldsymbol{\beta}_{M^*} \mid M^*$, then the Jacobian term is identically one (i.e., $\mathbf{u} = \boldsymbol{\beta}^*_{M^*}$, and $\mathbf{u}^* = \boldsymbol{\beta}_M$), and the acceptance probability simplifies to

$$\alpha = \min\left\{1, \frac{f(\mathbf{Y} \mid M^*) f(M^*) j(M \mid M^*)}{f(\mathbf{Y} \mid M) f(M) j(M^* \mid M)}\right\}$$

so that there is no need to generate $\mathbf{u}$ or $\boldsymbol{\beta}_M$.

In linear models with conjugate prior distributions, both the integrated SSVS (George and McCulloch, 1997) and MC^3 (Raftery, Madigan and Hoeting, 1997) algorithms can be viewed as special cases of the above reversible jump algorithm that differ only in their choice of proposal distributions on the model space. In RMHs implementation of MC^3, the model M^* is determined by picking one of the p variables at random and either deleting (if it is currently in the model) or adding it (if it is currently out of the model); thus $j(M^* \mid M) = 1/p$ for models that differ from M in one coordinate and 0 otherwise. The reverse jump probability is the same and cancels from the acceptance probability ratio, so that M^* is accepted with probability

$$\alpha = \min\left\{1, \frac{f(\mathbf{Y} \mid M^*) f(M^*)}{f(\mathbf{Y} \mid M) f(M)}\right\} = \min\left\{1, \frac{P(M^* \mid \mathbf{Y})}{P(M \mid \mathbf{Y})}\right\},$$

which depends only on the ratio of the marginal likelihood of the data and the prior probabilities of models. The integrated version of SSVS is a Gibbs sampler over the model space. The corresponding jump proposal is the full conditional distribution of the indicator variable that the jth variable is included in the model, given $\mathbf{Y}$ and the remaining indicator variables for variables currently included in M. The posterior and proposal distributions cancel in the acceptance ratio, so that M^* is accepted with probability 1. Other choices of j lead to other Metropolis–Hastings algorithms.

Dellaportas and Forster (1996) and Godsill (1998) discuss relationships among other popular algorithms for sampling models and reversible jump MCMC algorithms in classes of problems such as graphical models. In problems where the posterior distribution for $\boldsymbol{\beta}$ is not known in closed form, $\mathbf{u}$ may be generated from a standard proposal distribution for $\boldsymbol{\beta}_{M^*}$. While the simple "birth" or "death" proposals for adding or deleting variables are easy to implement, other choices of j may lead to algorithms that can move more rapidly through the model space. Clyde and DeSimone-Sasinowska (1997) and Clyde (1999a) used approximate posterior probabilities of variable inclusion for the proposal distribution over the model space to target the more important variables, rather than proposing all variables be added or deleted with equal probability $1/p$. One of the key issues in designing

a sampler for BMA is to achieve rapid mixing and coverage of the model space.

One class of problems where BMA has had outstanding success is in nonparametric regression using wavelet bases. In wavelet regression it is common to treat the error variance, σ^2, as fixed, substituting a robust estimate based on the finest level wavelet coefficients. Because the columns of the design matrix are orthogonal, the posterior distribution of M given σ^2 (under conjugate priors) can be represented as a product of independent distributions, and SSVS provides i.i.d. draws from the posterior distribution. However, for many quantities of interest such as posterior means and variances, posterior expectations can be computed analytically despite the high dimension of the parameter space ($p = n$), thus avoiding sampling models altogether (Clyde, Parmigiani and Vidakovic, 1998). Sampling models and σ^2 in conjunction with the use of Rao–Blackwellized estimators does appear to be more efficient in terms of mean squared error, when there is substantial uncertainty in the error variance (i.e., small sample sizes or low signal-to-noise ratio) or important prior information. Recently, Holmes and Mallick (1998) adapted perfect sampling (Propp and Wilson, 1996) to the context of orthogonal regression. While more computationally intensive per iteration, this may prove to be more efficient for estimation than SSVS or MC^3 in problems where the method is applicable and sampling is necessary.

While Gibbs and MCMC sampling has worked well in high-dimensional orthogonal problems, Wong, Hansen, Kohn and Smith (1997) found in high-dimensional problems such as nonparametric regression using nonorthogonal basis functions that Gibbs samplers were unsuitable, from both a computational efficiency standpoint as well as for numerical reasons, because the sampler tends to get stuck in local modes. Their proposed sampler "focuses" on variables that are more "active" at each iteration and in simulation studies provided better MSE performance than other classical nonparametric approaches or Bayesian approaches using Gibbs or reversible jump (Holmes and Mallick, 1997) sampling.

With the exception of a deterministic search, most methods for implementing BMA rely on algorithms that sample models with replacement and use ergodic averages to compute expectations, as in (7). In problems, such as linear models, where posterior model probabilities are known up to the normalizing constant, it may be more efficient to devise estimators using renormalized posterior model probabilities (Clyde, DeSimone and Parmigiani, 1996; Clyde, 1999a) and to devise algorithms based on sampling models without replacement. Based on current work with M. Littman, this appears to be a promising direction for implementation of BMA.

While many recent developments have greatly advanced the class of problems that can be handled using BMA, implementing BMA in high-dimensional problems with correlated variables, such as nonparametric regression, is still a challenge from both a computational standpoint and the choice of prior distributions.

2. PRIOR SPECIFICATION

In applications, I have found that specifying the prior distributions on both the parameters and model space to be perhaps the most difficult aspect of BMA. While the authors discussed prior distributions on the model space, the hyperparameters in the prior distributions for the parameters within each model also require careful consideration, as they appear in the marginal likelihoods and hence the posterior model probabilities. In many problems, subjective elicitation of prior hyperparameters is extremely difficult. Even if subjective elicitation is possible, it may be desirable to also use default prior distributions as part of a broader sensitivity analysis. Proper, but diffuse, prior distributions for $\boldsymbol{\beta}$ may have unintended influences on posterior model probabilities and deserve careful attention. While default prior distributions (both proper and improper) can be calibrated based on information criteria such as AIC, BIC or RIC (Clyde, 1999b; George and Foster, 1997; Fernandez et al. 1998), no one prior distribution emerges as the clear choice for BMA (although, in general, I have found that BMA based on BIC and RIC priors out-performs BMA using AIC calibrated prior distributions). Based on simulation studies, Fernandez, Ley and Steel (1998) recommend RIC-like prior distributions when $n < p^2$ and BIC-like prior distributions otherwise. In wavelets, where $p = n$, there are cases where priors calibrated based on BIC have better predictive performance than prior distributions calibrated using RIC, and vice versa. In problems such as wavelets where subjective information is not available, empirical Bayes (EB) methods for estimating the hyperparameters ensure that the prior distribution is not at odds with the data and have (empirically) led to improved predictive performance over a number of fixed hyperparameter specifications as well as default choices such as AIC, BIC, and RIC (Clyde and George, 1998, 1999; George and Foster, 1997; Hanson and Yu, 1999) for both model selection and BMA.

In research on the health effects of airborne particulate matter (PM), I have found that results using BMA may be very sensitive to the choice of prior hyperparameters (Clyde, 1999b). While model averaged estimates of relative risks conditioning on models that included particulate matter did not differ greatly using AIC, RIC or BIC calibrated prior distributions, the probability of no health effect attributed to PM was highly sensitive to the choice of prior distribution. As subjective prior distributions may be viewed with scepticism by some consumers of PM research, EB methods for generalized linear models may be a useful alternative.

While posterior probabilities, $P(\beta \neq 0 \mid \mathbf{Y})$, of no effect may be more natural than p-values, care must be taken in their interpretation. Besides being sensitive to prior hyperparameters (an indication that the evidence in the data may be weak), one needs to consider what other models are also under consideration. While Occam's window may be an effective means for accounting for model uncertainty by focusing on a few models, I am concerned that it leads to biased posterior effect probabilities, as marginal posterior probabilities for important variables are often inflated, while small posterior effect probabilities are underestimated, because of averaging over a restricted class of models. Perhaps an alternative is to use the models in Occam's window to communicate uncertainty, using displays such as in Clyde (1999b), which focus on the "best" models for illustrating model uncertainty, but provide estimates based on averaging over all (sampled) models.

Posterior effect probabilities depend on the variables (and their correlation structure) under consideration in the model space. In health effect–pollution studies, one may also want to consider other copollutants, such as ozone, NO_2, CO and SO_2, in addition to PM. As pollution levels may be highly correlated (perhaps due to latent variables), they may effectively divide up the posterior mass based on "equivalent models." The posterior effect probabilities for each individual pollutant, based on marginal probabilities, may be very small [the "dilution" (George, 1999) depends on the number of copollutants], even though the overall probability of a pollution effect (at least one pollutant variable is included in the model) could be near 1. Of increasing interest in PM research is whether the chemical composition of PM can explain health effects. As measurements of chemical constituents of PM become available, the number of potential variables will increase dramatically, leading to the potential dilution of effect probabilities.

The application of model averaging in the $\mathcal{M}$-open perspective leads to additional questions. For example, how should we specify prior distributions in a consistent manner as we add a number of possibly highly correlated variables? Are estimates of posterior effect probabilities or other quantities stable as we add more, possibly irrelevant, variables to the model? Using Occam's window to discredit models from future consideration may lead to a situation where none of the models in Occam's window contain a particular variable. Do we now disregard this variable from further analyses, even though it may, in conjunction with the new variables, lead to a better model than those found so far? If we decide to include this variable only in models that contain the new variables, is our assignment of prior distributions coherent? While updating posterior distributions sequentially, as new data are collected, is coherent, is applying Occam's window sequentially a coherent strategy and what are the implied prior distributions? While model averaging is a natural way to incorporate model uncertainty in the $\mathcal{M}$-open perspective, the choice of prior distributions over both parameter and model spaces becomes even more difficult.

Comment

David Draper

1. MODEL UNCERTAINTY, YES, DISCRETE MODEL AVERAGING, MAYBE

This paper offers a good review of one approach to dealing with statistical model uncertainty, an important topic and one which has only begun to come into focus for us as a profession in this decade (largely because of the availability of Markov chain Monte Carlo computing methods). The authors—who together might be said to have founded the Seattle school of model uncertainty—are to be commended for taking this issue forward so vigorously over the past five years. I have eight comments on the paper, some general and some specific to the body-fat example (Jennifer Hoeting kindly sent me the data, which are well worth looking at; the data set, and a full description of it, may be obtained by emailing the message send jse/v4n1/datasets.johnson to archive@jse.stat.ncsu.edu).

1. In the Bayesian approach that makes the most sense to me personally, the de Finetti (1974, /1975) predictive point of view, a model is a joint probability distribution $p(y_1, \ldots, y_n \mid \mathscr{A})$ for observables y_i, in which probability is an expression of your personal uncertainty and $\mathscr{A}$ is the set of assumptions (implicit and explicit) on which your uncertainty assessment is based. From this vantage point, how can we talk about posterior model probabilities, as the authors (hereafter HMRV) do in their equation (2)? Wouldn't we be talking about probabilities of probabilities?

   Both crucially and reasonably, de Finetti emphasized starting with exchangeability judgments in constructing $p(y_1, \ldots, y_n \mid \mathscr{A})$, and the point seems to be that (a) what HMRV call a model arises from these judgments and (b) HMRV's "model uncertainty" in de Finetti's language is uncertainty about the level of aggregation or conditioning at which (your uncertainty about) the data may reasonably be regarded as exchangeable. In the simplest possible setting, for instance, of binary observables, initially with no covariates, de Finetti's (1931) celebrated representation theorem says that $\mathscr{E}_1 = \{y_i \text{ exchangeable}\}$ is functionally equivalent to the simple hierarchical model $\{\theta \sim p(\theta), (y_i \mid \theta) \sim_{\text{iid}} \text{Bernoulli}(\theta)\}$, where $\theta = P(y_i = 1)$. Now suppose I also observe a binary covariate x, and I am uncertain about whether to assume $\mathscr{E}_1$ or $\mathscr{E}_2 = \{y_i \text{ conditionally exchangeable given } x_i\}$, by which I mean that $(y_i \mid x_i = 1)$ and $(y_i \mid x_i = 0)$ are separately exchangeable but $y = (y_1, \ldots, y_n)$ is not. Then in predicting y_{n+1}, say, I could either calculate $p(y_{n+1} \mid y, \mathscr{E}_1)$, or $p(y_{n+1} \mid y, [x_1, \ldots, x_{n+1}], \mathscr{E}_2)$, or I could expand the model hierarchically by adding uncertainty about $\mathscr{E}_j$ at the top of the hierarchy, which would be implemented via equations like HMRV's (1–3). Thus "model" uncertainty, or "structural" uncertainty (Draper 1995; this paper, Section 8.4), or uncertainty about exchangeability, is something that I think de Finetti would agree can be discussed probabilistically (as long as we don't take the "models" to be anything other than sets of judgments that help in predicting observables). (It is probably also worth mentioning regarding HMRV's equations (1–3) that Δ needs to have the same meaning in all "models" for the equations to be straightforwardly interpretable; the coefficient of x_1 in a regression of y on x_1 is a different beast than the coefficient of x_1 in a regression of y on x_1 and x_2.)
2. At the beginning of Section 5 HMRV say that "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." This sounds unassailable, but reality is actually a bit more slippery, as was first pointed out to me by John Tukey. Suppose there are two models (a) which get counted separately in the list of possible models but (b) which are functionally (almost) equivalent as far as making predictions is concerned; then "assuming that they are equally likely in the prior" amounts to giving the single model, of which there are essentially two slightly different versions, twice as

David Draper is Professor of Statistics, Department of Mathematical Sciences, University of Bath, Claverton Down, Bath BA2 7AY, England (e-mail: d.draper@maths.bath.ac.uk).

much actual prior weight. An example of this in regression problems arises when there is a high degree of collinearity among the predictors, as occurs in the body fat data: for instance, the correlations between the (abdomen, hip) and (abdomen, chest) pairs of variables are +0.94 and +0.91, respectively. Consider the simple version of HMRV's set-up with an outcome y, predictors x_1 and x_2 and a third predictor x_3 which is perfectly correlated with x_2. A researcher who wishes to use the HMRV "flat-prior" idea and who deals with the collinearity by ignoring x_3 altogether would place prior weight 1/4 on each of the models in $\mathcal{M}_1 = \{$no predictors; x_1; x_2; $(x_1, x_2)\}$, but another investigator who tries to include x_3 in the model in a naive manner (a) will initially put prior weight 1/8 on each of the models in $\mathcal{M}_2 = \{$no predictors; x_1; x_2; x_3; (x_1, x_2); (x_1, x_3); (x_2, x_3); $(x_1, x_2, x_3)\}$; (b) will be forced to drop the last two models in $\mathcal{M}_2$ as unfittable; (c) will be left with weight 1/6 on each of the models in $\mathcal{M}_3 = \{$no predictors; x_1; x_2; x_3; (x_1, x_2); $(x_1, x_3)\}$ and (d) will thus effectively put respective weights (1/6, 1/6, 1/3, 1/3) on the models in $\mathcal{M}_1$. This sort of thing would presumably have little effect on HMRV's calculations in the body-fat example, but would be more important in settings (for instance, in economics and sociology; see, e.g., Western, 1996) where the number p of predictors is large relative to the sample size n and there is a high degree of collinearity among many of the x_j. What role should collinearity play in the HMRV approach?

3. In the body fat example it is interesting to think about why the R^2 value is only 75% for the (Gaussian maximum likelihood) model in HMRV's Table 7 (which includes all the predictors entered linearly), because the variables in the authors' Table 6 provide a fairly exhaustive catalogue of body measurements that might be relevant to fat level. One possibility is nonlinear terms in the available predictors, and indeed HMRV neglect to mention that there is a noticeable interaction between the abdomen and weight variables. While I am on that subject, is anything new required in the authors' approach to deal with interactions, or to HMRV are they just additional predictors? What about the data-analytic idea (e.g., Mosteller and Tukey, 1977) that it is always good form to make sure the main effects for x_1 and x_2 (say) are in any model which includes the interaction x_1x_2?. I do not mean to downplay the importance of model uncertainty, but it is arguable from a purely scientific point of view that a better use of our time is figuring out what other variables should be included in the body fat model, such as (self-reported) exercise level, that might be unreliable but would be better than nothing, to get the residual SD down considerably below 4 percentage points. After all, even an R^2 of 75% means that we have only been able to drive the approximate posterior predictive standard deviation (SD) for a given individual down from its value in a model with no predictors, namely 7.7% (the overall SD of the body fat variable, on a scale from 0% to 45%), to 4.0% with all 13 predictors; in other words, even taking account of all available x's the actual body fat of a man whose predicted value is 19% (the mean) could easily be anywhere between 11% (a lean athlete) and 27% (the 85th percentile of the observed distribution). As statisticians we tend to feel that we have done a good job in a regression problem when all of the technical assumptions of the model look plausible given the behavior of the residuals, even if the unexplained variation in y is large; but as scientists we should not be happy until this variation is small enough for the model to be of substantial practical use in, for example, decision making.
4. Question: What is the Bayesian justification for out-of-sample predictive validation, as with the splitting of the data into build and test subsets in HMRV's Section 6? After all, isn't keeping score on the quality of your predictions of the test data an inherently frequentist activity? Here are two answers to this question: (1) Who says there is anything wrong with a Bayesian-frequentist fusion? The two paradigms have both been around for 350 years, since the earliest attempts to attach meaning to probabilistic concepts (e.g., Hacking 1975), and if the two schools were like prize fighters punching it out for centuries it is clear that both boxers are still standing, which I take to be empirical proof of a theorem saying there must be elements of merit in both viewpoints. A fusion of the form {reason in a Bayesian way when formulating your inferences and predictions, and think frequentistly when evaluating their quality} seems to me to bring the advantages of Bayesian uncertainty assessment and predictive calibration together, and is the approach I try to use in both my research and consulting. See Good (1983) for more thoughts on the topic of Bayes/non-Bayes synthesis. (2) BMA is really about model selection, because even averaging over a lot of models is a form of model choice. Model selection should

best be viewed as a decision problem (Key, Perocchi and Smith, 1999; Draper 1999a): to choose a model you have to say to what purpose the model will be put, for how else will you know whether your model is good enough? If you base the utility function in your decision formulation on the quality of predictions for future data not yet collected, then the expectation in the usual maximization-of-expected-utility prescription is over that future data; having not yet seen that data, the only way to evaluate such an expectation is to assume that the present data is exchangeable with the future and use some of it to proxy for data that hasn't arrived yet—hence the build and test framework. This idea can be formalized (Gelfand, Dey and Chang, 1992; Draper and Fouskakis, 1999).

5. What characteristics of a statistical example predict when BMA will lead to large gains? The only obvious answer I know is the ratio n/p of observations to predictors (with tens of thousands of observations and only dozens of predictors to evaluate, intuitively the price paid for shopping around in the data for a model should be small). Are the authors aware of any other simple answers to this question?

   As an instance of the n/p effect, in regression-style problems like the cirrhosis example where p is in the low dozens and n is in the hundreds, the effect of model averaging on the predictive scale can be modest. HMRV are stretching a bit when they say, in this example, that "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"; this can be seen by attaching uncertainty bands to the estimates in Table 5. Over the single random split into build and test data reported in that table, and assuming (at least approximate) independence of the 152 yes/no classifications aggregated in the table, death rates in the high risk group, with binomial standard errors, are 81% ± 5%, 75% ± 6% and 72% ± 6% for the BMA, stepwise, and top PMP methods, and combining the low and medium risk groups yields 18% ± 4%, 19% ± 4% and 17% ± 4% for the three methods, respectively, hardly a rousing victory for BMA. It is probable that by averaging over many random build–test splits a "statistically significant" difference would emerge, but the predictive advantage of BMA in this example is not large in practical terms.

6. Following on from item (4) above, now that the topic of model choice is on the table, why are we doing variable selection in regression at all? People who think that you have to choose a subset of the predictors typically appeal to vague concepts like "parsimony," while neglecting to mention that the "full model" containing all the predictors may well have better out-of-sample predictive performance than many models based on subsets of the x_j. With the body-fat data, for instance, on the same build–test split used by HMRV, the model that uses all 13 predictors in the authors' Table 7 (fitted by least squares–Gaussian maximum likelihood) has actual coverage of nominal 90% predictive intervals of (95.0 ± 1.8)% and (86.4 ± 3.3)% in the build and test data subsets, respectively; this out-of-sample figure is better than any of the standard variable-selection methods tried by HMRV (though not better than BMA in this example). To make a connection with item (5) above, I generated a data set 10 times as big but with the same mean and covariance structure as the body-fat data; with 2,510 total observations the actual coverage of nominal 90% intervals within the 1,420 data values used to fit the model was (90.6 ± 0.8)%, and on the other 1,090 observations it was (89.2 ± 0.9)%. Thus with only 251 data points and 13 predictors, the "full model" overfits the cases used for estimation and underfits the out-of-sample cases, but this effect disappears with large n for fixed p (the rate at which this occurs could be studied systematically as a function of n and p). (I put "full model" in quotes because the concept of a full model is unclear when things like quadratics and interactions in the available predictors are considered.)

   There is another sense in which the "full model" is hard to beat: one can create a rather accurate approximation to the output of the complex, and computationally intensive, HMRV regression machinery in the following closed-form Luddite manner. (1) Convert y and all of the x_j to standard units, by subtracting off their means and dividing by their SDs, obtaining y^* and x_j^* (say). This goes some distance toward putting the predictors on a common scale. (2) Use least squares–Gaussian maximum likelihood to regress y^* on all [or almost all $(*)$] the x_j^*, resolving collinearity problems by $(*)$ simply dropping out of the model altogether any x's that are highly correlated with other x's (when in doubt, drop the x in a pair of such predictors that is more weakly correlated with y. This

TABLE 1
A comparison of the HMRV and PSC approaches to judging when the effect of a predictor on the outcome is large, in the body fat example with caliper $c = 1/6$.

Method	X_6	X_2	X_{13}	X_{12}	X_4	X_{11}	X_8	X_7	X_5	X_1	X_9	X_3	X_{10}
HMRV $P(\beta_j \neq 0 \mid D)$	100	97	62	35	19	17	15	13	6	5	5	4	4
PSC $P(\lvert\beta_j\rvert \geq c \mid D)$	100	96	59	16	23	6	30	0*	11	8	2	0	0

Note: Table entries are percentages.
*This variable was dropped in the PSC approach due to collinearity.

doesn't handle collinearity problems in which three or more of the x_j are close to linearly determined, but I am trying for simplicity here). (3) People who claim that a (standardized) regression coefficient β_j is zero can't really mean that; no continuous quantity is ever precisely zero. What they presumably mean is that β_j is close enough to zero that the effect of x_j on y is close to negligible from a practical significance point of view. Therefore, introduce the idea of a *practical significance caliper* (PSC) c and assert that β_j is *practically* nonzero if $P(\lvert\beta_j\rvert \geq c \mid \text{data})$ is large. Table 1 reports a comparison of this approach with that of HMRV on the body-fat data; $c = 1/6$ is a caliper value that produces reasonably good agreement between the two methods (in other problems in which I have tried both approaches, values of c^{-1} in the vicinity of 4–6 have produced virtually identical agreement; see the last sentence in the authors' Section 7.1.2 for a reason why). This means that the HMRV approach with this data set is essentially equivalent to a rather conservative rule of the form {a predictor x has a "significant" effect if changing it by six or more SDs is associated with a 1-SD or larger change in y}.

7. The authors' Figure 4 really bothers me: thinking scientifically, both wrist circumference and body fat are being modeled as continuous quantities in HMRV's approach, and my uncertainty about "the effect of wrist circumference on body fat" (now there's a phrase that deserves to be put into quotes, if any phrase ever did) is surely continuous as well. So where does the spike at zero come from? The authors acknowledge that this is an "artifact" of their approach, and I will now argue that it is an unnecessary artifact.

   HMRV's method averages together a lot of models, in each of which the β for a given variable is either effectively left at its value from least squares–Gaussian maximum likelihood fitting (since they use flat priors on the parameters conditional on a given model's structure) or dragged all the way to zero. But by now we have almost 30 years of good empirical Bayes research (e.g., Sclove, Morris, Radhakrishna, 1972; Copas, 1983) to show that modeling approaches that shrink a given coefficient part of the way back to zero can dominate all-or-nothing methods. One way forward is via hierarchical modeling of the form

$$
\begin{aligned}
&(\gamma, \tau^2, \sigma^2) \sim p(\gamma, \tau^2, \sigma^2) \text{ (e.g., diffuse)},\\
(1)\quad &(\beta \mid Z, \gamma, \tau^2) \sim N_{p+1}(Z\gamma, \tau^2 I_{p+1}),\\
&(y \mid X, \beta, \sigma^2) \sim N_n(X\beta, \sigma^2 I_n),
\end{aligned}
$$

   where y is the n-vector of outcomes (in standard units), X is the $n \times (p+1)$ design matrix from the "full model" with all of the predictors standardized; β is the $(p+1)$-vector of regression coefficients and Z is a vector or matrix quantifying prior information about the signs and relative magnitude of the "effects of the x_j on y" (see Greenland, 1993 for empirical Bayes fitting of a model like (1) with dichotomous outcomes in epidemiological examples). Elsewhere (Browne, 1995; Draper, 1999b) Bill Browne and I show that this model can lead to out-of-sample predictive performance at least as good as that of BMA, and it has the advantage of scientific credibility in that the posterior distribution for any given β_j depends naturally on substantive prior information and is continuous. Figure 1 compares the posterior distribution for β_{13}, the coefficient in HMRV's Figure 4, under three sets of modeling assumptions: the "full model" (with a diffuse prior on the β vector), BMA, and model (1) above using a Z vector, $(+1, -4, 0, -2, 0, +10, -2, +2, 0, 0, +2, +1, -1)$, in the order of variables in the authors' Table 6 obtained from discussions with physicians in the Bath area. Given the physicians' judgments, hierarchical modeling shrinks β_{13} toward zero, as does the HMRV approach, but model (1) does so smoothly and in a way that is controlled by substantive prior information. This

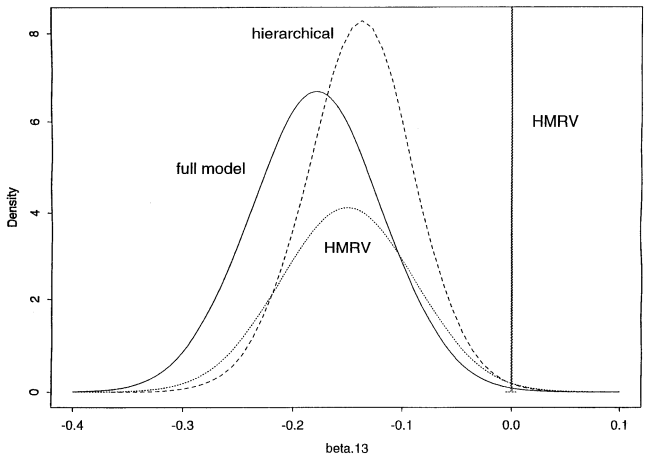

FIG. 1. *Three posterior distributions for β_{13}, the coefficient for wrist circumference in the body fat example of Section* 7.2.

is the motivation for the title of my comments: I am with the authors 100% on the importance of model uncertainty, but I believe that discrete model averaging should be reserved for problems (such as the oil price example in Draper, 1995) where structural uncertainty is truly discrete.

8. Finally, and continuing with the theme that model selection should be done decision-theoretically, in HMRV's Section 7.2 the goal is to predict body fat using 13 physiological variables or a subset thereof. Consider a typical use of the resulting equation in a doctor's office or a mobile clinic set up in a shopping mall. Predictor x_j takes t_j seconds of time to measure per person, where the t_j for the variables in the authors' Table 6 might vary from a low of 3 to 4 seconds (to elicit and record the subject's age) to a high of 25 to 30 seconds (to accurately measure and record the subject's weight). These amounts of time translate directly either into money (how much does it cost to keep the mobile clinic running?) or to numbers of subjects who can be seen in a given unit of time such as a day—in other words, they are costs. The potential benefit of including a predictor is the increase in predictive accuracy resulting from that predictor being in the equation versus not being there. These benefits can also be expressed in monetary terms, for instance by quantifying the value to the subject of being correctly classified as either having or not having a body fat value sufficiently high to warrant clinical intervention. Analyses based either on traditional variable-selection methods or on HMRV's formulation of BMA are based on benefit-only calculations that examine predictive accuracy, but it seems to me (and to a Ph.D. student, Dimitris Fouskakis, who is working with me) that cost–benefit calculations which weigh predictive accuracy against data collection cost are more relevant to the real-world problem of model choice in the class of generalized linear models. In work that we are writing up now (Draper and Fouskakis, 1999) we make this cost–benefit tradeoff explicit in the context of a problem in health policy analysis, and we show that sharply different subsets of predictors are chosen when costs are considered along with predictive quality. The BMA analysis in Table 8 is interesting, but how does it relate to the manner in which the body-fat model will actually be used in the world?

I hope it is clear from these remarks that I differ from the authors principally in matters of implementation and interpretation; there can be no disagreement about the importance of model uncertainty itself.

Comment

E. I. George

I would like to begin by congratulating the authors of this paper, who have done a masterful job of pulling together and synthesizing a large and growing body of work on a very promising approach to modeling and prediction. They have explained the essential features of Bayesian model averaging (BMA) with unusual clarity, stressing its many advantages and paying attention to the crucial issues of implementation and interpretation. It is by understanding what BMA offers, as well as its limitations, that statisticians will be able to exploit its real potential.

E. I. George is Professor, MSIS Department, CBA 5.202, University of Texas, Austin, Texas 78712–1175 (e-mail: ed.george@bus.utexas.edu).

1. THE PREDICTIVE SUCCESS OF BMA

My own experience with BMA has been similar to that of the authors. In problems where model uncertainty is present, I have found BMA to consistently yield predictive performance improvements over single selected models. I have also found this phenomenon to persist under a wide variety of prior choices. Many colleagues have also reported similar experiences.

Although puzzling at first glance, a simplistic explanation for the predictive success of BMA stems from the observation that BMA predictions are weighted averages of single model predictions. If the individual predictions are roughly unbiased estimates of the same quantity, then averaging will tend to reduce unwanted variation. (Perhaps a more sellable name for BMA would be Bayesian prediction averaging.) Going a bit further, it can be argued that under model uncertainty, selecting a single prediction is tantamount to using a randomized rule (where the randomization takes place over the model space). If prediction is evaluated by a convex loss function, as it usually is, then by Jensen's inequality, an appropriately weighted average prediction will tend to improve on the randomized rule. Geometrically, averaging pulls the prediction inside the convex hull of the individual predictions, a sensible strategy under a convex loss function unless there is systemic bias in all the predictions. However, there is more to it than that, because BMA weights each single model prediction by the corresponding posterior model probability. Thus, BMA uses the data to adaptively increase the weights on those predictions whose models are more supported by the data.

The key idea which motivates BMA comes from posing the model uncertainty problem within the Bayesian formalism. By using individual model prior probabilities to describe model uncertainty, the class of models under consideration is replaced by a single large mixture model. Under this mixture model, a single model is drawn from the prior, the prior parameters are then drawn from the corresponding parameter priors and finally the data is drawn from the identified model. For the problem of prediction under logarithmic scoring and many other loss functions including squared error loss, BMA* arises naturally as the Bayes rule which minimizes posterior expected loss. (I have used BMA* here to denote BMA which averages over all the models, as opposed to the approximations discussed below.) The authors implicitly refer to this fact when they point out that BMA* provides better average predictive performance than any of the single models under consideration, because they use "average" to mean posterior expectation under the mixture model. Indeed, under the mixture posterior, BMA* is superior to any other procedure.

It is tempting to criticize BMA* because it does not offer better average predictive performance than a correctly specified single model. However, this fact is irrelevant when model uncertainty is present because specification of the correct model with certainty is then an unavailable procedure. In most practical applications, the probability of selecting the correct model is less than 1, and a mixture model elaboration seems appropriate. The real difficulty is that the mixture probabilities are unknown (except for Bayesian purists), and this is where the prior specification problem comes into play. Fortunately, for the prediction problem, BMA appears to be robust and offer improvements over a wide variety of model space priors. This is in part because the posterior model probabilities tend to be strongly adaptive. As long as the model space prior is not extremely asymmetric, averaging will tend to improve predictions for the reasons I alluded to above.

It should also be mentioned that BMA is well suited to yield predictive improvements over single selected models when the entire model class is misspecified. In a sense, the mixture model elaboration is an expansion of the model space to include adaptive convex combinations of models. By incorporating a richer class of models, BMA can better approximate models outside the model class.

2. CONSIDERATIONS FOR PRIOR SPECIFICATION

BMA implementation requires prior specification on the individual parameter spaces and on the overall model space. The parameter space priors determine the integrated likelihood in (3) which controls both the individual predictive distributions and the adaptivity of the posterior weights in (1). It is especially important in this context to use robust parameter priors which are relatively flat over the range of plausible parameter values. I suspect the MLE and Laplace approximations discussed in Section 4 implicitly correspond to using such robust priors, and I wonder if the authors agree. Failure to use robust parameter priors can lead to unstable, sharply adaptive posterior weights which denigrate, rather than improve, predictive performance. At the other extreme, to obtain frequentist guarantees of robustness, it may be necessary to use improper priors or even pseudo marginal distributions as was done to obtain the minimax multiple shrinkage estimators in George (1986a, b, c, 1987). However, by going outside the proper prior realm, norming constants

become arbitrary, a drawback that complicates the specification and interpretation of the prior model probabilities.

The specification of the prior model probabilities can also be a delicate matter. The simplest, and seemingly appropriate choice when no prior information is available, is the uniform prior $\text{pr}(M_i) \equiv 1/2^p$. Although the authors did not state their model space priors in the Section 7 examples, (unless I missed it), I suspect this is the prior they used. We also initially favored this prior in George and McCulloch (1993), but then realized that the more general independence prior (16) offered the flexibility of putting more weight on parsimonious models and of differential weighting across the variables, as described by the authors. Unfortunately, it is usually difficult to have the kind of prior information to differentiate among the possible choices. To mitigate some of these specification difficulties, I have recently found it useful to use empirical Bayes methods to estimate prior hyperparameters for both the parameter and model priors. In George and Foster (1997) and Clyde and George (1999a, b), we found that such empirical Bayes methods consistently yielded predictive improvements over fixed hyperparameter choices.

Another more subtle problem for model space prior specification, discussed in George (1999), is a posterior phenomenon which I call dilution. Consider, for example, a regression problem where many of the covariates were so highly correlated that large subsets of models were essentially equivalent. If a uniform prior were put on the model space, then excessive prior probability would be allocated to these subsets, at the expense of some of the unique models. The predictive potential of BMA would be compromised if the only good models were unique and different from the rest. One way of avoiding this problem would be to use prior specifications which dilute the probability within subsets of similar models. Such priors could maintain the probability assigned to model neighborhoods when redundant models were added to the model space. An example of priors which have this dilution property are the tree priors proposed in Chipman, George and McCulloch (1998). I am currently developing dilution priors for multiple regression and will report on these elsewhere.

3. AVERAGING OVER SUBSETS

As the authors make clear, BMA*, which averages over all models, may not be practical in even moderately sized problems. It is then necessary to consider approximations which average over a selected subset of these models. Because the fundamental BMA* quantities of interest are posterior expectations, the approximation problem is just the classic problem of using sample averages to estimate population means. In this regard, I like using MCMC methods such as MC^3 which, in the spirit of random sampling, attempt to select a representative subset of models. Ergodic properties of the MCMC sampler carry over directly to the sample averages.

An aspect that I would add to the authors' discussion of MCMC methods is that one can do much better than simple averages such as (7) in this context. In many problems $\text{pr}(D \mid M_k)$ can be either computed exactly or approximated very well. When S is the selected subset, one can then compute $\text{pr}(M_k \mid D, S)$ similarly to (2) and use conditional expectations such as (6) to estimate posterior quantities. For example, one could use

$$E(\Delta \mid D, S) = \sum_{M_k \in S} \hat{\Delta}_k \ \text{pr}(M_k \mid D, S)$$

to estimate $E[\Delta \mid D]$ in Section 1. Under iid sampling, such estimates are nearly best unbiased (George, 1999) and appear to very substantially improve on simple averages (George and McCulloch, 1997).

For the purpose of approximating BMA*, I am less sanguine about Occam's window, which is fundamentally a heuristic search algorithm. By restricting attention to the "best" models, the subset of models selected by Occam's Window are unlikely to be representative, and may severely bias the approximation away from BMA*. For example, suppose substantial posterior probability was diluted over a large subset of similar models, as discussed earlier. Although MCMC methods would tend to sample such subsets, they would be entirely missed by Occam's Window. A possible correction for this problem might be to base selection on a uniform prior, i.e. Bayes factors, but then use a dilution prior for the averaging. However, in spite of its limitations as an approximation to BMA*, the heuristics which motivate Occam's Window are intuitively very appealing. Perhaps it would simply be appropriate to treat and interpret BMA under Occam's Window as a conditional Bayes procedure.

4. INFERENCE WITH BMA

In their examples, the authors carefully illustrate the inferential potential of BMA. Compared to standard frequentist inferences, the Bayesian answers are much more natural because they directly address questions of real interest. A statement like $\text{P}(\beta \neq 0 \mid D) = 50\%$ seems more understandable and relevant than a statement like "the p-value

is .05 conditionally on a selected model." However, in spite of their appeal, the posterior probabilities must be cautiously interpreted because of their dependence on a complicated prior formulation which is in many ways arbitrary. I worry that consumers of such analyses may be misled by understanding $P(\beta \neq 0 \mid D) = 80\%$ in a frequentist sense. I wonder what guidance the authors would give to such a consumer.

Sounding another cautionary note, I believe there is a temptation to use the finding of predictive improvement to justify both the prior formulation and the derived Bayesian inference. One has to be wary of such justification because BMA yields predictive improvements under such a wide variety of priors. On the other hand, useful inferential justification can be obtained by checks for predictive calibration such as the cross validation assessment of predictive coverage which appears in Section 7.2.3. It would have also been useful to perform prior sensitivity analyses, to see how the Bayesian inferential statements in Section 7 held up as the priors were varied.

Finally, I would like to point out the obvious: that BMA inferences are necessarily conditional on the selected model class. BMA addresses model uncertainty within the selected class, but it does not address the uncertainty of model class selection. Even if one combined model classes or averaged over model classes, there would always be model classes left out. Furthermore, continual expansion of the model space would at some point begin to include many redundant or irrelevant models, and would begin to diminish predictive performance. Although BMA is a wonderful approach to the problem of accounting for model uncertainty, it can never completely avoid the selection dilemma.

Rejoinder

Jennifer A. Hoeting, David Madigan, Adrian E. Raftery and Chris T. Volinsky

Bayesian model averaging is a very active area, and Merlise Clyde, David Draper and Ed George are three of the researchers who have contributed the most to the area in the past five years. We are very grateful to them for their discussions.

In addition to making penetrating and useful points about our paper, they have provided additional insights here that take the field forward. George's intuitive explanation of the success of BMA in predictive performance provides a useful and different perspective. Granger and Newbold (1977) had already found that weighted averages of time series forecasts were better than the component forecasts that made them up, and that this result was robust to the weights used; George's comment also suggests a possible explanation of their empirical result. Clyde's demonstration that MC^3 and SSVS can be viewed as special cases of reversible jump MCMC is a striking unifying result. We would also like to call attention to Clyde's (1999b) appealing displays of BMA results, which are a real contribution to the field.

1. PARAMETER PRIORS

Clyde and George both discuss the specification of priors for the model parameters, and this is a vital part of the technology. As George points out, it is often desirable to have priors that are relatively uninformative and also robust in the sense that conclusions are qualitatively insensitive to reasonable changes in the priors. For Bayesian *estimation*, this is often achieved using improper priors, but, as George comments, for BMA this leads to arbitrary norming constants that complicate matters. Our preference is to avoid improper priors in this context, and instead to use priors that are sufficiently spread out but remain proper. There are now several approaches to finding such priors that seem to work well in specific model classes.

The technology is now available to implement BMA with fairly general priors. For many models, the Laplace method can give accurate approximations to integrated likelihoods (Raftery, 1996a). If estimation is carried out using MCMC, good methods are now available for calculating integrated likelihoods from MCMC output (e.g., Raftery, 1996b; DiCiccio et al., 1997; Oh, 1999).

For BMA, it is desirable that the prior on the parameters be spread out enough that it is relatively flat over the region of parameter space where the likelihood is substantial (i.e., that we be in the "stable estimation" situation described by Edwards, Lindman and Savage, 1963). It is also desirable that the prior not be much more spread out than is necessary to achieve this. This is because the integrated likelihood for a model declines roughly as σ^{-d} as

σ becomes large, where σ is the prior standard deviation and d is the number of free parameters in the model. Thus highly spread out priors tend to over-penalize larger models.

This suggests the use of data-dependent proper priors. While this sounds at first like a contradiction in terms, in fact it can be viewed as an approximation to the (subjective) prior of someone who knows just a little about the matter at hand. Wasserman (1998) has shown that data-dependent priors can have optimality properties, leading to better performance than any data-independent prior in some situations.

One proposal along these lines is Raftery's (1996a) reference set of proper priors for generalized linear models; this is calculated automatically by the `glib` software (http://lib.stat.cmu.edu/S/glib). This specifies data-dependent priors that are as concentrated as possible while remaining in the stable estimation situation and have minimal impact on ratios of posterior model probabilities when both nested and nonnested models are being compared. A similar idea was implemented for linear regression models by Raftery, Madigan and Hoeting (1997).

A second such proposal is the unit information prior (UIP), which is a multivariate normal prior centered at the maximum likelihood estimate with variance matrix equal to the inverse of the mean observed Fisher information in one observation. Under regularity conditions, this yields the simple BIC approximation given by equation (13) in our paper (Kass and Wasserman, 1995; Raftery, 1995).

The unit information prior, and hence BIC, have been criticized as being too conservative (i.e., too likely to favor simple models). Cox (1995) suggested that the prior standard deviation should decrease with sample size. Weakliem (1999) gave sociological examples where the UIP is clearly too spread out, and Viallefont et al. (1998) have shown how a more informative prior can lead to better performance of BMA in the analysis of epidemiological case-control studies. The UIP is a proper prior but seems to provide a conservative solution. This suggests that if BMA based on BIC favors an "effect," we can feel on solid ground in asserting that the data provide evidence for its existence (Raftery, 1999). Thus BMA results based on BIC could be routinely reported as a baseline reference analysis, along with results from other priors if available.

A third approach is to allow the data to estimate the prior variance of the parameters. Lindley and Smith (1972) showed that this is essentially what ridge regression does for linear regression, and Volinsky (1997) pointed out that ridge regression has consistently outperformed other estimation methods in simulation studies. Volinsky (1997) proposed combining BMA and ridge regression by using a "ridge regression prior" in BMA. This is closely related to empirical Bayes BMA, which Clyde and George (1999) have shown to work well for wavelets, a special case of orthogonal regression. Clyde, Raftery, Walsh and Volinsky (2000) show that this good performance of empirical Bayes BMA extends to (nonorthogonal) linear regression.

2. PRIOR MODEL PROBABILITIES

In our examples, we have used prior model probabilities that are the same for each model. In the by now relatively extensive experience that we describe and cite in our paper, we have found this to yield good performance. Usually, the (integrated) likelihood on model space seems to be well enough behaved and concentrated enough that the results are insensitive to moderate changes away from 1/2 in the π_j in our equation (16); $\pi_j = 1/2$ corresponds to the uniform prior. The uniform prior has the advantage of being simple, transparent and easy to explain to clients.

Nevertheless, as George points out, it may be possible to obtain better performance with other priors. Madigan, Gavrin and Raftery (1995) showed how elicited informative priors can lead to improved predictive performance for BMA in a clinical context. Also, Clyde and George have both shown in their writings how empirical Bayes estimation of the π_j hyperparameters in (16) can yield improvements.

An issue raised by all three discussants is "dilution," which arises in regression when independent variables are highly correlated. Suppose X_1 and X_2 are highly correlated regressors, and that both (individually) are highly predictive of Y. BMA typically considers four models: M_0, the null model, $M_1 : \{X_1\}$, $M_2 : \{X_2\}$ and $M_3 : \{X_1, X_2\}$. With very high correlations, $\mathrm{pr}(M_1 \mid D) + \mathrm{pr}(M_2 \mid D) + \mathrm{pr}(M_3 \mid D)$ would be close to 1, and $\mathrm{pr}(M_1 \mid D) \approx \mathrm{pr}(M_2 \mid D)$. The discussants view this as undesirable.

However, there are two different situations here. The first is when X_1 and X_2 correspond to substantively different mechanisms, and it is reasonable to postulate that one of the mechanisms might be operating and the other not. For example, suppose Y is one's occupational attainment as an adult, measured on a socioeconomic status scale, X_1 is one's mother's education and X_2 is one's (parents') family income (Featherman and Hauser, 1977). X_1 and X_2 are highly correlated, but the mechanisms by which they might impact Y are quite different, so all four models are plausible a priori. The posterior model probabilities are saying that at least one of X_1 and

X_2 has an effect on Y, but that the data cannot tell us whether the effect is due to one, the other, or both. This seems like a reasonable summary of what the data are saying.

Clyde makes the important point that simple inspection of marginal posterior probabilities of parameters being nonzero might obscure this message. It would seem useful to supplement displays like our Table 1 with diagnostics showing groups of variables of which only one usually appears at a time.

The second situation is when there is a single mechanism (e.g., pollution causes deaths, as in Clyde and DeSimone-Sasinowska, 1997), but several of the X_j's are measures of the same mechanism. Then regression itself can be misleading, and so, a fortiori, can BMA. A solution to this seems to be to recognize that the independent variable of interest is really a latent construct (e.g., "pollution") with multiple indicators, and to use something like a LISREL-type model (Bollen, 1989). BMA and Bayesian model selection can still be applied in this context (e.g., Hauser and Kuo, 1998).

3. OCCAM'S WINDOW

Madigan and Raftery (1994) had two motivations for introducing Occam's window. The first is that it represents scientific practice in that models that have been clearly discredited do get discarded in scientific research; by this argument Occam's window is preferable in principle to full BMA, or BMA*, as George calls it.

The second motivation is that Occam's window might provide a good approximation to BMA*. Clyde and George cast doubt on this argument, and indeed we know of no formal theoretical support for it. However, while the biases that Clyde and George mention may exist in principle, they seem small in practice. We have found Occam's window to provide a surprisingly good approximation to the posterior effect probabilities from BMA* in all the applications we have worked on. (Of course, Occam's window overestimates the actual posterior model probabilities of the models it includes, but it preserves ratios of these probabilities.) For example, Table 1 shows the posterior effect probabilities from the crime data analyzed by Raftery, Madigan and Hoeting (1997) from both Occam's window and BMA*. The differences are small (and not always in the direction that Clyde suggests), and they are typical of what we have seen in many applications.

Clyde raises some excellent questions about the implementation of Occam's window when there are many highly correlated variables. When the high correlation arises from the fact that several variables are being used as measurements of the same underlying latent construct, then using a LISREL-type model to account for this explicitly would remove the problem with Occam's window, as well as being more faithful to the science. When the correlated variables do represent different mechanisms, Occam's window is likely to include models that contain some but not all of these variables, and it could be argued that this is a reasonable representation of the uncertainty given the data.

4. INTERPRETATION

Draper raises the perennial issue of whether BMA really amounts to combining "apples and oranges," and hence is invalid. This arises in the regression context, where it is often suggested that the coefficient of X_1 in the regression of Y on X_1 alone is inherently different from the coefficient of X_1 in the regression of Y on X_1 and X_2, and hence they should never be combined.

One way of thinking about this is that, in the regression context, BMA can be viewed as standard Bayesian inference for just one model, the full model in which all variables are included. The twist is that the prior allows for the possibility that some of the coefficients might be equal to zero (or, essentially equivalently, close to zero). This is desirable statistically, as setting unnecessary parameters to zero can lead to better estimates. It also often represents scientists' views. Once we recast the way we think of BMA this way, in terms of just one model, the "apples and oranges" problem disappears.

As Draper points out, in our equations (1)–(3), Δ needs to have the same meaning in all models. But precisely what does this mean? A sufficient condition would seem to be that Δ be an observable quantity that could be predicted. This would include many regression coefficients if we allow quantities of interest that could be observed "asymptotically," that is, almost exactly given a very large new data set of size n_{new}, exchangeable with the one at hand. For example, if X_1 and X_2 are regressors for Y and can take only the values 0 and 1, and if $\bar{Y}_{ij}^{\text{new}}$ is the mean of the values of Y in the new data set for which $X_1 = i$ and $X_2 = j$, then

$$\beta_1 = \tfrac{1}{2}\left(\bar{Y}_{21}^{\text{new}} - \bar{Y}_{11}^{\text{new}}\right) + \tfrac{1}{2}\left(\bar{Y}_{22}^{\text{new}} - \bar{Y}_{12}^{\text{new}}\right) + O_p\left(n_{\text{new}}^{-1/2}\right),$$

for both the models $M_1 : \{X_1\}$ and $M_2 : \{X_1, X_2\}$. Thus making inference about β_1 from BMA based on M_1 and M_2 would seem valid, because β_1 is (asymptotically) an observable quantity, and its posterior distribution is also a predictive distribution.

TABLE 1
Posterior effect probabilities (%) for crime data

Predictor	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Occam's window	73	2	99	64	36	0	0	12	53	0	43	1	100	83	0
BMA* (via MC^3)	79	17	98	72	50	6	7	23	62	11	45	30	100	83	22

Source: Raftery, Madigan and Hoeting (1997)

Based on this reasoning, BMA seems likely to be valid for many regression coefficients. However, its validity for nonlinear effects and interactions is more problematic. Restricting BMA to quantities that can be interpreted (perhaps asymptotically) as observables seems a good way to stay out of trouble.

5. OTHER ISSUES

Clyde and Draper point out the importance of model checking, and we strongly agree. *All* the models considered might be poor, and then combining them will not do much good. Model diagnostics such as residual analysis and posterior predictive checks (Gelman et al., 1996) applied to the best models, are useful for diagnosing this situation. Our view is that such diagnostics are useful in an exploratory sense for suggesting when the models may be inadequate and how they should be improved. However, the tests that come with such diagnostics are usually based on *P*-values or nearly equivalent quantities, and often multiple such tests are carried out, at least implicitly. Thus, we feel that such tests should not be used in a formal manner to "reject" the models under consideration. Rather, they should be used to suggest better models, and these models should then be compared with the ones first thought of using Bayes factors and BMA (Kass and Raftery, 1995).

Draper says that model choice is a decision problem, and that the use to which the model is to be put should be taken into account explicitly in the model selection process. This is true, of course, but in practice it seems rather difficult to implement. This was first advocated by Kadane and Dickey (1980) but has not been done much in practice, perhaps because specifying utilities and carrying out the full utility maximization is burdensome, and also introduces a whole new set of sensitivity concerns. We do agree with Draper's suggestion that the analysis of the body fat data would be enhanced by a cost–benefit analysis which took account of both predictive accuracy and data collection costs.

In practical decision-making contexts, the choice of statistical model is often not the question of primary interest, and the real decision to be made is something else. Then the issue is decision-making in the presence of model uncertainty, and BMA provides a solution to this. In equation (1) of our article, let Δ be the utility of a course of action, and choose the action for which $E[\Delta \mid D]$ is maximized.

Draper does not like our Figure 4. However, we see it as a way of depicting on the same graph the answers to two separate questions: is wrist circumference associated with body fat after controlling for the other variables? and if so, how strong is the association? The posterior distribution of β_{13} has two components corresponding to these two questions. The answer to the first question is "no" (i.e., the effect is zero or small) with probability 38%, represented by the solid bar in Figure 4. The answer to the second question is summarized by the continuous curve. Figure 4 shows double shrinkage, with both discrete and continuous components. The posterior distribution of β_{13}, given that $\beta_{13} \neq 0$, is shrunk *continuously* towards zero via its prior distribution. Then the posterior is further shrunk (discretely this time) by taking account of the probability that $\beta_{13} = 0$. The displays in Clyde (1999b) convey essentially the same information, and some may find them more appealing than our Figure 4.

Draper suggests the use of a practical significance caliper and points out that for one choice, this gives similar results to BMA. Of course the big question here is how the caliper is chosen. BMA can itself be viewed as a significance caliper, where the choice of caliper is based on the data. Draper's Table 1 is encouraging for BMA, because it suggests that BMA does coincide with practical significance. It has often been observed that *P* values are at odds with "practical" significance, leading to strong distinctions being made in textbooks between statistical and practical significance. This seems rather unsatisfactory for our discipline: if statistical and practical significance do not at least approximately coincide, what is the use of statistical testing? We have found that BMA often gives results closer to the practical significance judgments of practitioners than do *P*-values.

ADDITIONAL REFERENCES

BOLLEN, K. A. (1989). *Structural Equations with Latent Variables.* Wiley, New York.

BROWNE, W. J. (1995). *Applications of Hierarchical Modelling*. M.Sc. dissertation, Dept. Mathematical Sciences, Univ. Bath, UK.

CHIPMAN, H., GEORGE, E. I. and MCCULLOCH, R. E. (1998). Bayesian CART model search (with discussion). *J. Amer. Statist. Assoc.* **93** 935–960.

CLYDE, M. (1999a). Bayesian model averaging and model search strategies (with discussion). In *Bayesian Statistics 6*. (J. M. Bernardo, A. P. Dawid, J. O. Berger and A. F. M. Smith, eds) 157–185. Oxford Univ. Press.

CLYDE, M. (1999b). Model uncertainty and health effect studies for particulate matter. ISDS Discussion Paper 99–28. Available at www.isds.duke.edu.

CLYDE, M. and DESIMONE-SASINOWSKA, H. (1997). Accounting for model uncertainty in Poisson regression models: does particulate matter particularly matter? ISDS Discussion Paper 97–06. Available at www.isds.duke.edu.

CLYDE, M. and GEORGE., E. I. (1998). Flexible empirical Bayes estimation for wavelets. ISDS Discussion Paper 98–21. Available at www.isds.duke.edu.

CLYDE, M. and GEORGE., E. I. (1999). Empirical Bayes estimation in wavelet nonparametric regression. In *Bayesian Inference in Wavelet-Based Models* (P. Muller and B. Vidakovic, eds.) 309–322. Springer, Berlin.

CLYDE, M. and GEORGE, E. I. (1999a). Empirical Bayes estimation in wavelet nonparametric regression. In *Bayesian Inference in Wavelet Based Models* (P. Muller and B. Vidakovic, eds.) Springer, Berlin. To appear.

CLYDE, M. and GEORGE, E. I. (1999b). Flexible empirical Bayes estimation for wavelets. Technical Report, ISDS, Duke Univ.

CLYDE, M., PARMIGIANI, G. and VIDAKOVIC, B. (1998). Multiple shrinkage and subset selection in wavelets. *Biometrika* **85** 391–402.

CLYDE, M., RAFTERY, A. E., WALSH, D. and VOLINSKY, C. T. (2000). Technical report. Available at www.stat.washington.edu/tech.reports.

COPAS, J. B. (1983). Regression, prediction, and shrinkage (with discussion). *J. Roy. Statist. Soc. Ser. B* **45** 311–354.

COX, D. R. (1995). The relation between theory and application in statistics (disc: P228-261). *Test* **4** 207–227.

DE FINETTI, B. (1931). Funzioni caratteristica di un fenomeno aleatorio. *Atti Acad. Naz. Lincei* **4** 86–133.

DE FINETTI, B. (1974, 1975). *Theory of Probability* **1** and **2**. (Trans. by A. F. M. Smith and A. Machi). Wiley, New York.

DELLAPORTAS, P. and FORSTER, J. J. (1996). Markov chain Monte Carlo model determination for hierarchical and graphical log-linear models. Technical Report, Faculty of Mathematics, Southampton Univ. UK.

DICICCIO, T. J., KASS, R. E., RAFTERY, A. E. and WASSERMAN, L. (1997). Computing Bayes factors by combining simulation and asymptotic approximations. *J. Amer. Statist. Assoc.* **92** 903–915.

DRAPER, D. (1999a). Discussion of "Decision models in screening for breast cancer" by G. Parmigiani. In *Bayesian Statistics 6* (J. M. Bernardo, J. Berger, P. Dawid and A. F. M. Smith eds.) 541–543 Oxford Univ. Press.

DRAPER, D. (1999b). Hierarchical modeling, variable selection, and utility. Technical Report, Dept. Mathematical Sciences, Univ. Bath, UK.

DRAPER, D. and FOUSKAKIS, D. (1999). Stochastic optimization methods for cost-effective quality assessment in health. Unpublished manuscript.

FEATHERMAN, D. and HAUSER, R. (1977). *Opportunity and Change.* Academic Press, New York.

GELFAND, A. E., DEY, D. K. and CHANG, H. (1992). Model determination using predictive distributions, with implementation via sampling-based methods (with discussion). In *Bayesian Statistics 4* (J. M. Bernardo, J. O. Berger, A. P. Dawid, A. F. M. Smith, eds.) 147–167. Oxford Univ. Press.

GELMAN, A., MENG, X.-L. and STERN, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statist. Sinica* **6** 733–760.

GEORGE, E. I. (1987). Multiple shrinkage generalizations of the James–Stein estimator. In *Contributions to the Theory and Applications of Statistics* (*A Volume in Honor of Herbert Solomon*) (A. E. Gelfand, ed.) 397–428. Academic Press, New York.

GEORGE, E. I. (1999). Discussion of "Model averaging and model search strategies" by M. Clyde. In *Bayesian Statistics 6* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.) 157–185. Oxford University Press.

GEORGE, E. I. (1999). Discussion of "Model averaging and model search by M. Clyde." In *Bayesian Statistics 6* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.) Oxford University Press.

GEORGE, E. I. and FOSTER, D. P. (1997). Calibration and empirical Bayes variable selection. Technical Report, Dept. MSIS, Univ. Texas, Austin.

GEORGE, E. I., and MCCULLOCH, R. E. (1997). Approaches for Bayesian variable selection. *Statist. Sinica* **7** 339–373.

GODSILL, S. (1998). On the relationship between MCMC model uncertainty methods. Technical report Univ. Cambridge.

GOOD, I. J. (1983). *Good Thinking: The Foundations of Probability and Its Applications*. Univ. Minnesota Press, Minneapolis.

GRANGER, C. W. J. and NEWBOLD, P. (1976). The use of R^2 to determine the appropriate transformation of regression variables. *J. Econometrics* **4** 205–210.

GREENLAND, S. (1993). Methods for epidemiologic analyses of multiple exposures—a review and comparative study of maximum-likelihood, preliminary testing, and empirical Bayes regression. *Statistics in Medicine* **12** 717–736.

HACKING, I. (1975). *The Emergence of Probability*. Cambridge University Press.

HANSON, M. and KOOPERBERG, C. (1999). Spline adaptation in extended linear models. Bell Labs Technical Report. Available at cm.bell-labs.com/who/cocteau/papers.

HANSON, M. and YU, B. (1999). Model selection and the principle of minimum description. Bell Labs Technical Report. Available at cm.bell-labs.com/who/cocteau/papers.

HAUSER, R. and KUO, H. (1998). Does the gender composition of sibships affect women's educational attainment? *Journal of Human Resources* **33** 644–657.

HOLMES. C. C. and MALLICK, B. K. (1997). Bayesian radial basis functions of unknown dimension. Dept. Mathematics technical report, Imperial College, London.

HOLMES, C. C. and MALLICK, B. K. (1998). Perfect simulation for orthogonal model mixing. Dept. Mathematics technical report, Imperial College, London.

KADANE, J. B. and DICKEY, J. M. (1980). Bayesian decision theory and the simplification of models. In *Evaluation of Econometric Models* (J. Kmenta and J. Ramsey, eds.) Academic Press, New York.

KEY, J. T., PERICCHI, L. R. and SMITH, A. F. M. (1999). Bayesian model choice: what and why? (with discussion). In *Bayesian Statistics 6* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.) 343–370. Oxford Univ. Press.

LINDLEY, D. V. and SMITH, A. F. M. (1972). Bayes estimates for the linear model (with discussion). *J. Roy. Statist. Soc. Ser. B* **34** 1–41.

MOSTELLER, F. and TUKEY, J. W. (1977). *Data Analysis and Regression*. Addison-Wesley, Reading, MA.

OH, M.-S. (1999). Estimation of posterior density functions from a posterior sample. *Comput. Statist. Data Anal.* **29** 411–427.

PROPP, J. G. and WILSON, D. B. (1996). Exact sampling with coupled Markov cahins and applications to statistical mechanics. *Random Structures Algorithms* **9** 223–252.

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

RAFTERY, A. E. (1996b). Hypothesis testing and model selection. In *Markov Chain Monte Carlo in Practice* (W. R. Gilks and D. Spiegelhalter, eds.) 163–188. Chapman and Hall, London.

RAFTERY, A. E. (1999). Bayes factors and BIC: Comment on "A Critique of the Bayesian information criterion for model selection." *Sociological Methods and Research* **27** 411–427.

SCLOVE, S. L., MORRIS, C. N. and RADHAKRISHNA, R. (1972). Nonoptimality of preliminary-test estimators for the mean of a multivariate normal distribution. *Ann. Math. Statist.* **43** 1481–1490.

VIALLEFONT, V., RAFTERY, A. E. and RICHARDSON, S. (1998). Variable selection and Bayesian Model Averaging in case-control studies. Technical Report 343, Dept. Statistics, Univ. Washington.

WASSERMAN, L. (1998). Asymptotic inference for mixture models using data dependent priors. Technical Report 677, Dept. Statistics, Carnegie-Mellon Univ.

WEAKLIEM, D. L. (1999). A critique of the Bayesian information criterion for model selection. *Sociological Methods and Research* **27** 359–297.

WESTERN, B. (1996). Vague theory and model uncertainty in macrosociology. *Sociological Methodology* **26** 165–192.

WONG, F., HANSEN, M. H., KOHN, R. and SMITH, M. (1997). Focused sampling and its application to nonparametric and robust regression. Bell Labs technical report. Available at cm.bell-labs.com/who/cocteau/papers.