The Practical Implementation of
Bayesian Model Selection

Hugh Chipman, Edward I. George and Robert E. McCulloch [*]
The University of Waterloo, The University of Texas at Austin
and The University of Chicago

April, 2001

Abstract

In principle, the Bayesian approach to model selection is straightforward. Prior probability distributions are used to describe the uncertainty surrounding all unknowns. After observing the data, the posterior distribution provides a coherent post data summary of the remaining uncertainty which is relevant for model selection. However, the practical implementation of this approach often requires carefully tailored priors and novel posterior calculation methods. In this article, we illustrate some of the fundamental practical issues that arise for two different model selection problems: the variable selection problem for the linear model and the CART model selection problem.

1   Introduction

The Bayesian approach to statistical problems is fundamentally probabilistic. A joint probability distribution is used to describe the relationships between all the unknowns and the data. Inference is then based on the conditional probability distribution of the unknowns given the observed data, the posterior distribution. Beyond the specification of the joint distribution, the Bayesian approach is automatic. Exploiting the internal consistency of the probability framework, the posterior distribution extracts the relevant information in the data and provides a complete and coherent summary of post data uncertainty. Using the posterior to solve specific inference and decision problems is then straightforward, at least in principle.

In this article, we describe applications of this Bayesian approach for model uncertainty problems where a large number of different models are under consideration for the data. The joint distribution is obtained by introducing prior distributions on all the unknowns, here the parameters of each model and the models themselves, and combining them with the models. Conditioning on the data then induces a posterior distribution of model uncertainty that can be used for model selection and other inference and decision problems. This is the essential idea and it can be very powerful. Especially appealing is its broad generality as it is based only on probabilistic considerations. However, two major challenges confront its practical implementation - the specification of the prior distributions and the calculation of the posterior. This will be our main focus.

The statistical properties of the Bayesian approach rest squarely on the specification of the prior distributions on the unknowns. But where do these prior distributions come from and what do they mean? One extreme answer to this question is the pure subjective Bayesian point of view that characterizes the prior as a wholly subjective description of initial uncertainty, rendering the posterior as a subjective post data description of uncertainty. Although logically compelling, we find this characterization to be unrealistic in complicated model selection problems where such information is typically unavailable or difficult to precisely quantify as a probability distribution. At the other extreme is the objective Bayesian point of view which seeks to find semi-automatic prior formulations or approximations when subjective information is unavailable. Such priors can serve as default inputs and make them attractive for repeated use by non-experts.

Prior specification strategies for recent Bayesian model selection implementations, including our own, have tended to fall somewhere between these two extremes. Typically, specific parametric families of proper priors are considered, thereby reducing the specification problem to that of selecting appropriate hyperparameter values. To avoid the need for subjective inputs, automatic default hyperparameter choices are often recommended. For this purpose, empirical Bayes considerations, either formal or informal, can be helpful, especially when informative choices are needed. However, subjective considerations can also be helpful, at least for roughly gauging prior location and scale and for putting small probability on implausible values. Of course, when substantial prior information is available, the Bayesian model selection implementations provide a natural environment for introducing realistic and important views.

By abandoning the pure subjective point of view, the evaluation of such Bayesian methods must ultimately involve frequentist considerations. Typically, such evaluations have taken the form of average performance over repeated simulations from hypothetical models or of cross

validations on real data. Although such evaluations are necessarily limited in scope, the Bayesian procedures have consistently performed well compared to non-Bayesian alternatives. Although more work is clearly needed on this crucial aspect, there is cause for optimism, since by the complete class theorems of decision theory, we need look no further than Bayes and generalized Bayes procedures for good frequentist performance.

The second major challenge confronting the practical application of Bayesian model selection approaches is posterior calculation or perhaps more accurately, posterior exploration. Recent advances in computing technology coupled with developments in numerical and Monte Carlo methods, most notably Markov Chain Monte Carlo (MCMC), have opened up new and promising directions for addressing this challenge. The basic idea behind MCMC here is the construction of a sampler which simulates a Markov chain that is converging to the posterior distribution. Although this provides a route to calculation of the full posterior, such chains are typically run for a relatively short time and used to search for high posterior models or to estimate posterior characteristics. However, constructing effective samplers and the use of such methods can be a delicate matter involving problem specific considerations such as model structure and the prior formulations. This very active area of research continues to hold promise for future developments.

In this introduction, we have described our overall point of view to provide context for the implementations we are about to describe. In Section 2, we describe the general Bayesian approach in more detail. In Sections 3 and 4, we illustrate the practical implementation of these general ideas to Bayesian variable selection for the linear model and Bayesian CART model selection, respectively. In Section 5, we conclude with a brief discussion of related recent implementations for Bayesian model selection.

2   The General Bayesian Approach

2.1   A Probabilistic Setup for Model Uncertainty

Suppose a set of K models $\mathcal{M} = \{M_1, \ldots, M_K\}$ are under consideration for data Y, and that under M_k, Y has density $p(Y \mid \theta_k, M_k)$ where θ_k is a vector of unknown parameters that indexes the members of M_k. (Although we refer to M_k as a model, it is more precisely a model class). The Bayesian approach proceeds by assigning a prior probability distribution $p(\theta_k \mid M_k)$ to the parameters of each model, and a prior probability $p(M_k)$ to each model. Intuitively, this complete specification can be understood as a three stage hierarchical mixture model for generating the data Y; first the model M_k is generated from $p(M_1), \ldots, p(M_K)$, second the parameter vector θ_k is generated from $p(\theta_k \mid M_k)$, and third the data Y is generated from $p(Y \mid \theta_k, M_k)$.

Letting Y_f be future observations of the same process that generated Y, this prior formulation induces a joint distribution $p(Y_f, Y, \theta_k, M_k) = p(Y_f, Y \mid \theta_k, M_k) \, p(\theta_k \mid M_k) \, p(M_k)$. Conditioning on the observed data Y, all remaining uncertainty is captured by the joint posterior distribution $p(Y_f, \theta_k, M_k | Y)$. Through conditioning and marginalization, this joint posterior can be used for a variety Bayesian inferences and decisions. For example, when the goal is exclusively prediction of Y_f, attention would focus on the predictive distribution $p(Y_f | Y)$, which is obtained by margining

out both θ_k and M_k. By averaging over the unknown models, $p(Y_f \mid Y)$ properly incorporates the model uncertainty embedded in the priors. In effect, the predictive distribution sidesteps the problem of model selection, replacing it by model averaging. However, sometimes interest focuses on selecting one of the models in $\mathcal{M}$ for the data Y, the model selection problem. One may simply want to discover a useful simple model from a large speculative class of models. Such a model might, for example, provide valuable scientific insights or perhaps a less costly method for prediction than the model average. One may instead want to test a theory represented by one of a set of carefully studied models.

In terms of the three stage hierarchical mixture formulation, the model selection problem becomes that of finding the model in $\mathcal{M}$ that actually generated the data, namely the model that was generated from $p(M_1), \ldots, p(M_K)$ in the first step. The probability that M_k was in fact this model, conditionally on having observed Y, is the posterior model probability

$$p(M_k \mid Y) = \frac{p(Y \mid M_k)p(M_k)}{\sum_k p(Y \mid M_k)p(M_k)} \tag{1}$$

where

$$p(Y \mid M_k) = \int p(Y \mid \theta_k, M_k)p(\theta_k \mid M_k)d\theta_k \tag{2}$$

is the marginal or integrated likelihood of M_k. Based on these posterior probabilities, pairwise comparison of models, say M_1 and M_2, is summarized by the posterior odds

$$\frac{p(M_1 \mid Y)}{p(M_2 \mid Y)} = \frac{p(Y \mid M_1)}{p(Y \mid M_2)} \times \frac{p(M_1)}{p(M_2)}. \tag{3}$$

This expression reveals how the data, through the Bayes factor $\frac{p(Y \mid M_1)}{p(Y \mid M_2)}$, updates the prior odds $\frac{p(M_1)}{p(M_2)}$ to yield the posterior odds.

The model posterior distribution $p(M_1 \mid Y), \ldots, p(M_K \mid Y)$ is the fundamental object of interest for model selection. Insofar as the priors $p(\theta_k \mid M_k)$ and $p(M_k)$ provide an initial representation of model uncertainty, the model posterior summarizes all the relevant information in the data Y and provides a complete post-data representation of model uncertainty. By treating $p(M_k \mid Y)$ as a measure of the "truth" of model M_k, a natural and simple strategy for model selection is to choose the most probable M_k, the one for which $p(M_k \mid Y)$ largest. Alternatively one might prefer to report a set of high posterior models along with their probabilities to convey the model uncertainty.

More formally, one can motivate selection strategies based on the posterior using a decision theoretic framework where the goal is to maximize expected utility, (Gelfand, Dey and Chang 1992 and Bernardo and Smith 1994). More precisely, let α represent the action of selecting M_k, and suppose that α is evaluated by a utility function $u(\alpha, \Delta)$, where Δ is some unknown of interest, possibly Y_f. Then, the optimal selection is that α which maximizes the expected utility

$$\int u(\alpha, \Delta)p(\Delta \mid Y)d\Delta \tag{4}$$

where the predictive distribution of Δ given Y

$$p(\Delta \mid Y) = \sum_k p(\Delta \mid M_k, Y)p(M_k \mid Y) \tag{5}$$

is a posterior weighted mixture of the conditional predictive distributions.

$$p(\Delta \mid M_k, Y) = \int p(\Delta \mid \theta_k, M_k) p(\theta_k \mid M_k, Y) d\theta_k \tag{6}$$

It is straightforward to show that if Δ identifies one of the M_k as the "true state of nature", and $u(\alpha, \Delta)$ is 0 or 1 according to whether a correct selection has been made, then selection of the highest posterior probability model will maximize expected utility. However, different selection strategies are motivated by other utility functions. For example, suppose α entails choosing $p(\Delta \mid M_k, Y)$ as a predictive distribution for a future observation Δ, and this selection is to be evaluated by the logarithmic score function $u(\alpha, \Delta) = \log p(\Delta \mid M_k, Y)$. Then, the best selection is that α which maximizes the posterior weighted logarithmic divergence

$$\sum_{k'} p(M_{k'} \mid Y) \int p(\Delta \mid M_{k'}, Y) \log \frac{p(\Delta \mid M_{k'}, Y)}{p(\Delta \mid M_k, Y)} \tag{7}$$

(San Martini and Spezzaferri 1984).

However, if the goal is strictly prediction and not model selection, then expected logarithmic utility is maximized by using the posterior weighted mixture $p(\Delta \mid Y)$ in (5). Under squared error loss, the best prediction of Δ is the overall posterior mean

$$E(\Delta \mid Y) = \sum_k E(\Delta \mid M_k, Y) p(M_k \mid Y). \tag{8}$$

Such model averaging or mixing procedures incorporate model uncertainty and have been advocated by Geisser (1993), Draper (1995), Hoeting, Madigan, Raftery & Volinsky (1999) and Clyde, Desimone and Parmigiani (1995). Note however, that if a cost of model complexity is introduced into these utilities, then model selection may dominate model averaging.

Another interesting modification of the decision theory setup is to allow for the possibility that the "true" model is not one of the M_k, a commonly held perspective in many applications. This aspect can be incorporated into a utility analysis by using the actual predictive density in place of $p(\Delta \mid Y)$. In cases where the form of the true model is completely unknown, this approach serves to motivate cross validation types of training sample approaches, (see Bernardo and Smith 1994, Berger and Pericchi 1996 and Key, Perrichi and Smith 1998).

2.2   General Considerations for Prior Selection

For a given set of models $\mathcal{M}$, the effectiveness of the Bayesian approach rests firmly on the specification of the parameter priors $p(\theta_k \mid M_k)$ and the model space prior $p(M_1), \ldots, p(M_K)$. Indeed, all of the utility results in the previous section are predicated on the assumption that the model is correct. If one takes the subjective point of view that these priors represent the statistician's prior uncertainty about all the unknowns, then the posterior would be the appropriate update of this uncertainty after the data Y has been observed. However appealing, the pure subjective point of view here has practical limitations. Because of the sheer number and complexity of unknowns in most model uncertainty problems, it is probably unrealistic to assume that such uncertainty can be meaningfully described.

The most common and practical approach to prior specification in this context is to try and construct noninformative, semi-automatic formulations, using subjective and empirical Bayes considerations where needed. Roughly speaking, one would like to specify priors that allow the posterior to accumulate probability at or near the actual model that generated the data. At the very least, such a posterior can serve as a heuristic device to identify promising models for further examination.

Beginning with considerations for choosing the model space prior $p(M_1), \ldots, p(M_K)$, a simple and popular choice is the uniform prior

$$p(M_k) \equiv 1/K \tag{9}$$

which is noninformative in the sense of favoring all models equally. Under this prior, the model posterior is proportional to the marginal likelihood, $p(M_k \mid Y) \propto p(Y \mid M_k)$, and posterior odds comparisons in (3) reduce to Bayes factor comparisons. However, the apparent noninformativeness of (9) can be deceptive. Although uniform over models, it will typically not be uniform on model characteristics such as model size. A more subtle problem occurs in setups where many models are very similar and only a few are distinct. In such cases, (9) will not assign probability uniformly to model neighborhoods and may bias the posterior away from good models. As will be seen in later sections, alternative model space priors that dilute probability within model neighborhoods can be meaningfully considered when specific model structures are taken into account.

Turning to the choice of parameter priors $p(\theta_k \mid M_k)$, direct insertion of improper noninformative priors into (1) and (2) must be ruled out because their arbitrary norming constants are problematic for posterior odds comparisons. Although one can avoid some of these difficulties with constructs such as intrinsic Bayes factors (Berger and Pericchi 1996) or fractional Bayes factors (O'Hagan 1995), many Bayesian model selection implementations, including our own, have stuck with proper parameter priors, especially in large problems. Such priors guarantee the internal coherence of the Bayesian formulation, allow for meaningful hyperparameter specifications and yield proper posterior distributions which are crucial for the MCMC posterior calculation and exploration described in the next section.

Several features are typically used to narrow down the choice of proper parameter priors. To ease the computational burden, it is very useful to choose priors under which rapidly computable closed form expressions for the marginal $p(Y \mid M_k)$ in (2) can be obtained. For exponential family models, conjugate priors serve this purpose and so have been commonly used. When such priors are not used, as is sometimes necessary outside the exponential family, computational efficiency may be obtained with the approximations of $p(Y \mid M_k)$ described in Section 2.3. In any case, it is useful to parametrize $p(\theta_k \mid M_k)$ by a small number of interpretable hyperparameters. For nested model formulations, which are obtained by setting certain parameters to zero, it is often natural to center the priors of such parameters at zero, further simplifying the specification. A crucial challenge is setting the prior dispersion. It should be large enough to avoid too much prior influence, but small enough to avoid overly diffuse specifications that tend to downweight $p(Y \mid M_k)$ through (2), resulting in too little probability on M_k. For this purpose, we have found it useful to consider subjective inputs and empirical Bayes estimates.

2.3 Extracting Information from the Posterior

Once the priors have been chosen, all the needed information for Bayesian inference and decision is implicitly contained in the posterior. In large problems, where exact calculation of (1) and (2) is not feasible, Markov Chain Monte Carlo (MCMC) can often be used to extract such information by simulating an approximate sample from the posterior. Such samples can be used to estimate posterior characteristics or to explore the posterior, searching for models with high posterior probability.

For a model characteristic η, MCMC entails simulating a Markov chain, say $\eta^{(1)}, \eta^{(2)}, \ldots$, that is converging to its posterior distribution $p(\eta|Y)$. Typically, η will be an index of the models M_k or an index of the values of (θ_k, M_k). Simulation of $\eta^{(1)}, \eta^{(2)}, \ldots$ requires a starting value $\eta^{(0)}$ and proceeds by successive simulation from a probability transition kernel $p(\eta \mid \eta^{(j)})$, see Meyn and Tweedie (1993). Two of the most useful prescriptions for constructing a kernel that generates a Markov chain converging to a given $p(\eta \mid Y)$, are the Gibbs sampler (GS) (Geman and Geman 1984, Gelfand and Smith 1990) and the Metropolis-Hastings (MH) algorithms (Metropolis 1953, Hastings 1970). Introductions to these methods can be found in Casella and George (1992) and Chib and Greenberg (1995). More general treatments that detail precise convergence conditions (essentially irreducibility and aperiodicity) can found in Besag and Green (1993), Smith and Roberts (1993) and Tierney (1994).

When $\eta \in R^p$, the GS is obtained by successive simulations from the full conditional component distributions $p(\eta_i \mid \eta_{-i})$, $i = 1, \ldots, p$, where η_{-i} denotes the most recently updated component values of η other than η_i. Such GS algorithms reduce the problem of simulating from $p(\eta \mid Y)$ to a sequence of one-dimensional simulations.

MH algorithms work by successive sampling from an essentially arbitrary probability transition kernel $q(\eta \mid \eta^{(j)})$ and imposing a random rejection step at each transition. When the dimension of $\eta^{(j)}$ remains fixed, an MH algorithm is defined by:

1. Simulate a candidate η^* from the transition kernel $q(\eta \mid \eta^{(j)})$

2. Set $\eta^{(j+1)} = \eta^*$ with probability

$$\alpha(\eta^* \mid \eta^{(j)}) = \min\left\{1, \frac{q(\eta^{(j)} \mid \eta^*)}{q(\eta^* \mid \eta^{(j)})} \frac{p(\eta^* \mid Y)}{p(\eta^{(j)} \mid Y)}\right\}$$

Otherwise set $\eta^{(j+1)} = \eta^{(j)}$,

This is a special case of the more elaborate reversible jump MH algorithms (Green 1995) which can be used when dimension of η is changing. The general availability of such MH algorithms derives from the fact that $p(\eta \mid Y)$ is only needed up to the norming constant for the calculation of α above.

The are endless possibilities for constructing Markov transition kernels $p(\eta \mid \eta^{(j)})$ that guarantee convergence to $p(\eta \mid Y)$. The GS can be applied to different groupings and reorderings of the coordinates, and these can be randomly chosen. For MH algorithms, only weak conditions restrict considerations of the choice of $q(\eta|\eta^{(j)})$ and can also be considered componentwise. The GS and MH algorithms can be combined and used together in many ways. Recently proposed variations such as tempering, importance sampling, perfect sampling and augmentation offer a

promising wealth of further possibilities for sampling the posterior. As with prior specification, the construction of effective transition kernels and how they can be exploited is meaningfully guided by problem specific considerations as will be seen in later sections. Various illustrations of the broad practical potential of MCMC are described in Gilks, Richardson, Spieglehalter (1996).

The use of MCMC to simulate the posterior distribution of a model index η is greatly facilitated when rapidly computable closed form expressions for the marginal $p(Y \mid M_k)$ in (2) are available. In such cases, $p(Y \mid \eta)p(\eta) \propto p(\eta \mid Y)$ can be used to implement GS and MH algorithms. Otherwise, one can simulate an index of the values of (θ_k, M_k) (or at least M_k and the values of parameters that cannot be eliminated analytically). When the dimension of such an index is changing, MCMC implementations for this purpose typically require more delicate design, see Carlin and Chib (1995), Dellaportas, Forster and Ntzoufras (2000), Geweke (1996), Green (1995), Kuo and Mallick (1998) and Phillips and Smith (1996).

Because of the computational advantages of having closed form expressions for $p(Y \mid M_k)$, it may be preferable to use a computable approximation for $p(Y \mid M_k)$ when exact expressions are unavailable. An effective approximation for this purpose, when $h(\theta_k) \equiv \log p(Y|\theta_k, M_k)p(\theta_k|M_k)$ is sufficiently well-behaved, is obtained by Laplace's method (see Tierney and Kadane 1986) as

$$p(Y \mid M_k) \approx (2\pi)^{d_k/2}|H(\tilde{\theta}_k)|^{1/2}p(Y \mid \tilde{\theta}_k, M_k)p(\tilde{\theta}_k \mid M_k) \tag{10}$$

where d_k is the dimension of θ_k, $\tilde{\theta}_k$ is the maximum of $h(\theta_k)$, namely the posterior mode of $p(\tilde{\theta}_k \mid Y, M_k)$, and $H(\tilde{\theta}_k)$ is minus the inverse Hessian of h evaluated at $\tilde{\theta}_k$. This is obtained by substituting the Taylor series approximation $h(\theta_k) \approx h(\tilde{\theta}_k) - \frac{1}{2}(\theta_k - \tilde{\theta}_k)'H(\tilde{\theta}_k)(\theta_k - \tilde{\theta}_k)$ for $h(\theta_k)$ in $p(M_k \mid Y) = \int \exp\{h(\theta_k)\}d\theta_k$.

When finding $\tilde{\theta}_k$ is costly, further approximation of $p(Y \mid M)$ can be obtained by

$$p(Y \mid M_k) \approx (2\pi)^{d_k/2}|H^*(\hat{\theta}_k)|^{1/2}p(Y \mid \hat{\theta}_k, M_k)p(\hat{\theta}_k \mid M_k) \tag{11}$$

where $\hat{\theta}_k$ is the maximum likelihood estimate and H^* can be H, minus the inverse Hessian of the log likelihood or Fisher's information matrix. Going one step further, by ignoring the terms in (11) that are constant in large samples, yields the BIC approximation (Schwarz 1978)

$$\log p(Y \mid M) \approx \log p(Y \mid \hat{\theta}_k, M_k) - (d_k/2)\log n \tag{12}$$

where n is the sample size. This last approximation was successfully implemented for model averaging in a survival analysis problem by Raftery, Madigan and Volinsky (1996). Although it does not explicitly depend on a parameter prior, (12) may be considered an implicit approximation to $p(Y \mid M)$ under a "unit information prior" (Kass and Wasserman 1995) or under a "normalized" Jeffreys prior (Wasserman 2000). It should be emphasized that the asymptotic justification for these successive approximations, (10), (11), (12), may not be very good in small samples, see for example, McCulloch and Rossi (1991).

3 Bayesian Variable Selection for the Linear Model

Suppose Y a variable of interest, and $X_1, \ldots, X_p$ a set of potential explanatory variables or predictors, are vectors of n observations. The problem of variable selection, or subset selection

as it often called, arises when one wants to model the relationship between Y and a subset of $X_1, \ldots, X_p$, but there is uncertainty about which subset to use. Such a situation is particularly of interest when p is large and $X_1, \ldots, X_p$ is thought to contain many redundant or irrelevant variables.

The variable selection problem is usually posed as a special case of the model selection problem, where each model under consideration corresponds to a distinct subset of $X_1, \ldots, X_p$. This problem is most familiar in the context of multiple regression where attention is restricted to normal linear models. Many of the fundamental developments in variable selection have occurred in the context of the linear model, in large part because its analytical tractability greatly facilitates insight and computational reduction, and because it provides a simple first order approximation to more complex relationships. Furthermore, many problems of interest can be posed as linear variable selection problems. For example, for the problem of nonparametric function estimation, the values of the unknown function are represented by Y, and a linear basis such as a wavelet basis or a spline basis are represented by $X_1, \ldots, X_p$. The problem of finding a parsimonious approximation to the function is then the linear variable selection problem. Finally, when the normality assumption is inappropriate, such as when Y is discrete, solutions for the linear model can be extended to alternatives such as general linear models (McCullagh and Nelder 1989).

We now proceed to consider Bayesian approaches to this important linear variable selection problem. Suppose the normal linear model is used to relate Y to the potential predictors $X_1, \ldots X_p$

$$Y \sim N_n(X\beta, \sigma^2 I) \tag{13}$$

where $X = (X_1, \ldots X_p)$, β is a $p \times 1$ vector of unknown regression coefficients, and σ^2 is an unknown positive scalar. The variable selection problem arises when there is some unknown subset of the predictors with regression coefficients so small that it would be preferable to ignore them. In Sections 3.2 and 3.4, we describe two Bayesian formulations of this problem which are distinguished by their interpretation of how small a regression coefficient must be to ignore X_i. It will be convenient throughout to index each of these 2^p possible subset choices by the vector

$$\gamma = (\gamma_1, \ldots, \gamma_p)',$$

where $\gamma_i = 0$ or 1 according to whether β_i is small or large, respectively. We use $q_\gamma \equiv \gamma'1$ to denote the size of the γth subset. Note that here, γ plays the role of model identifier M_k described in Section 2.

We will assume throughout this section that $X_1, \ldots X_p$ contains no variable that would be included in every possible model. If additional predictors $Z_1, \ldots, Z_r$ were to be included every model, then we would assume that their linear effect had been removed by replacing Y and $X_1, \ldots X_p$ with $(I - Z(Z'Z)^{-1}Z')Y$ and $(I - Z(Z'Z)^{-1}Z')X_i$, $i = 1, \ldots, p$ where $Z = (Z_1, \ldots, Z_r)$. For example, if an intercept were to be included in every model, then we would assume that Y and $X_1, \ldots X_p$ had all been centered to have mean 0. Such reductions are simple and fast, and can be motivated from a formal Bayesian perspective by integrating out the coefficients corresponding to $Z_1, \ldots, Z_r$ with respect to an improper uniform prior.

9

3.1 Model Space Priors for Variable Selection

For the specification of the model space prior, most Bayesian variable selection implementations have used independence priors of the form

$$p(\gamma) = \prod w_i^{\gamma_i}(1 - w_i)^{1-\gamma_i}, \tag{14}$$

which are easy to specify, substantially reduce computational requirements, and often yield sensible results, see, for example, Clyde, Desimone and Parmigiani (1996), George and McCulloch (1993, 1997), Raftery, Madigan and Hoeting (1997) and Smith and Kohn (1996). Under this prior, each X_i enters the model independently of the other coefficients, with probability $p(\gamma_i = 1) = 1 - p(\gamma_i = 0) = w_i$. Smaller w_i can be used to downweight X_i which are costly or of less interest.

A useful reduction of (14) has been to set $w_i \equiv w$, yielding

$$p(\gamma) = w^{q_\gamma}(1 - w)^{p-q_\gamma}, \tag{15}$$

in which case the hyperparameter w is the a priori expected proportion of $X_i's$ in the model. In particular, setting $w = 1/2$, yields the popular uniform prior

$$p(\gamma) \equiv 1/2^p, \tag{16}$$

which is often used as a representation of ignorance. However, this prior puts most of its weight near models of size $q_\gamma = p/2$ because there are more of them. Increased weight on parsimonious models, for example, could instead be obtained by setting w small. Alternatively, one could put a prior on w. For example, combined with a beta prior $w \sim Beta(\alpha, \beta)$, (15) yields

$$p(\gamma) = \frac{B(\alpha + q_\gamma, \beta + p - q_\gamma)}{B(\alpha, \beta)} \tag{17}$$

where $B(\alpha, \beta)$ is the beta function. More generally, one could simply put a prior $h(q_\gamma)$ on the model dimension and let

$$p(\gamma) = \binom{p}{q_\gamma}^{-1} h(q_\gamma), \tag{18}$$

of which (17) is a special case. Under priors of the form (18), the components of γ are exchangeable but not independent, (except for the special case (15)).

Independence and exchangeable priors on γ may be less satisfactory when the models under consideration contain dependent components such as might occur with interactions, polynomials, lagged variables or indicator variables (Chipman 1996). Common practice often rules out certain models from consideration, such as a model with an $X_1 X_2$ interaction but no X_1 or X_2 linear terms. Priors on γ can encode such preferences.

With interactions, the prior for γ can capture the dependence relation between the importance of a higher order term and those lower order terms from which it was formed. For example, suppose there are three independent main effects A, B, C and three two-factor interactions AB, AC, and BC. The importance of the interactions such as AB will often depend only on whether the main effects A and B are included in the model. This belief can be expressed by a prior for $\gamma = (\gamma_A, \ldots, \gamma_{BC})$ of the form:

$$p(\gamma) = p(\gamma_A)p(\gamma_B)p(\gamma_C)p(\gamma_{AB} \mid \gamma_A, \gamma_B)p(\gamma_{AC} \mid \gamma_A, \gamma_C)p(\gamma_{BC} \mid \gamma_B, \gamma_C). \tag{19}$$

The specification of terms like $p(\gamma_{AC}|\gamma_A, \gamma_C)$ in (19) would entail specifying four probabilities, one for each of the values of (γ_A, γ_C). Typically $p(\gamma_{AC} \mid 0, 0) < (p(\gamma_{AC} \mid 1, 0), p(\gamma_{AC} \mid 0, 1))$ $< p(\gamma_{AC} \mid 1, 1)$. Similar strategies can be considered to downweight or eliminate models with isolated high order terms in polynomial regressions or isolated high order lagged variables in ARIMA models. With indicators for a categorical predictor, it may be of interest to include either all or none of the indicators, in which case $p(\gamma) = 0$ for any γ violating this condition.

The number of possible models using interactions, polynomials, lagged variables or indicator variables grows combinatorially as the number of variables increases. In contrast to independence priors of the form (14), priors for dependent component models, such as (19), is that they concentrate mass on "plausible" models, when the number of possible models is huge. This can be crucial in applications such as screening designs, where the number of candidate predictors may exceed the number of observations (Chipman, Hamada, and Wu 1997).

Another more subtle shortcoming of independence and exchangeable priors on γ is their failure to account for similarities and differences between models due to covariate collinearity or redundancy. An interesting alternative in this regard are priors that "dilute" probability across neighborhoods of similar models, the so called dilution priors (George 1999). Consider the following simple example.

Suppose that only two uncorrelated predictors X_1 and X_2 are considered, and that they yield the following posterior probabilities:

Variables in γ	X_1	X_2	X_1, X_2
$p(\gamma \mid Y)$	0.3	0.4	0.2

Suppose now a new potential predictor X_3 is introduced, and that X_3 is very highly correlated with X_2, but not with X_1. If the model prior is elaborated in a sensible way, as is discussed below, the posterior may well look something like

Variables in γ	X_1	X_2	X_3	X_1, X_2	$\cdots$
$p(\gamma \mid Y)$	0.3	0.13	0.13	0.06	$\cdots$

The probability allocated to X_2 and X_1, X_2 has here been "diluted" across all the new models containing X_3. Such dilution seems desirable because it maintains the allocation of posterior probability across neighborhoods of similar models. The introduction of X_3 has added proxies for the models containing X_2 but not any really new models. The probability of the resulting set of equivalent models should not change, and it is dilution that prevents this from happening. Note that this dilution phenomenon would become much more pronounced when many highly correlated variables are under consideration.

The dilution phenomenon is controlled completely by the model space prior $p(\gamma)$ because $p(\gamma \mid Y) \propto p(Y \mid \gamma)p(\gamma)$ and the marginal $p(Y \mid \gamma)$ is unaffected by changes to the model space. Indeed, no dilution of neighborhood probabilities occurs under the uniform prior (16) where $p(\gamma \mid Y) \propto p(Y \mid \gamma)$. Instead the posterior probability of every γ is reduced while all pairwise posterior odds are maintained. For instance, when X_3 is introduced above and a uniform prior is used, the posterior probabilities become something like

Variables in γ	X_1	X_2	X_3	X_1, X_2	$\cdots$
$p(\gamma \mid Y)$	0.15	0.2	0.2	0.1	$\cdots$

If we continued to introduce more proxies for X_2, the probability of the X_1 only model could be made arbitrarily small and the overall probability of the X_2 like models could be made arbitrarily large, a disturbing feature if Y was strongly related to X_1 and unrelated to X_2. Note that any independence prior (14), of which (16) is a special case, will also fail to maintain probability allocation within neighborhoods of similar models, because the addition of a new X_j reduces all the model probabilities by w_j for models in which X_j is included, and by $(1 - w_i)$ for models in which X_j is excluded.

What are the advantages of dilution priors? Dilution priors avoid placing too little probability on good, but unique, models as a consequence of massing excess probability on large sets of bad, but similar, models. Thus dilution priors are desirable for model averaging over the entire posterior to avoid biasing averages such as (8) away from good models. They are also desirable for MCMC sampling of the posterior because such Markov chains gravitate towards regions of high probability. Failure to dilute the probability across clusters of many bad models would bias both model search and model averaging estimates towards those bad models. That said, it should be noted that dilution priors would not be appropriate for pairwise model comparisons because the relative strengths of two models should not depend on whether another is considered. For this purpose, Bayes factors (corresponding to selection under uniform priors) would be preferable.

3.2 Parameter Priors for Selection of Nonzero β_i

We now consider parameter prior formulations for variable selection where the goal is to ignore only those X_i for which $\beta_i = 0$ in (13). In effect, the problem then becomes that of selecting a submodel of (13) of the form

$$p(Y \mid \beta_\gamma, \sigma^2, \gamma) = N_n(X_\gamma \beta_\gamma, \sigma^2 I) \tag{20}$$

where X_γ is the n x q_γ matrix whose columns correspond to the γth subset of $X_1, \ldots, X_p$, β_γ is a $q_\gamma \times 1$ vector of unknown regression coefficients, and σ^2 is the unknown residual variance. Here, (β_γ, σ^2) plays the role of the model parameter θ_k described in Section 2.

Perhaps the most useful and commonly applied parameter prior form for this setup, especially in large problems, is the normal-inverse-gamma, which consists of a q_γ-dimensional normal prior on β_γ

$$p(\beta_\gamma \mid \sigma^2, \gamma) = N_{q_\gamma}(\bar{\beta}_\gamma, \sigma^2 \Sigma_\gamma), \tag{21}$$

coupled with an inverse gamma prior on σ^2

$$p(\sigma^2 \mid \gamma) = p(\sigma^2) = IG(\nu/2, \nu\lambda/2), \tag{22}$$

(which is equivalent to $\nu\lambda/\sigma^2 \sim \chi_\nu^2$). For example, see Clyde, DeSimone and Parmigiani (1996), Fernandez, Ley and Steel (2001), Garthwaite and Dickey (1992, 1996), George and McCulloch (1997), Kuo and Mallick (1998), Raftery, Madigan and Hoeting (1997) and Smith and Kohn

(1996). Note that the coefficient prior (21), when coupled with $p(\gamma)$, implicitly assigns a point mass at zero for coefficients in (13) that are not contained in β_γ. As such, (21) may be thought of as a point-normal prior. It should also be mentioned that in one the first Bayesian variable selection treatments of the setup (20), Mitchell and Beauchamp (1988) proposed spike-and-slab priors. The normal-inverse-gamma prior above is obtained by simply replacing their uniform slab by a normal distribution.

A valuable feature of the prior combination (21) and (22) is analytical tractability; the conditional distribution of β_γ and σ^2 given γ is conjugate for (20), so that β_γ and σ^2 can be eliminated by routine integration from $p(Y, \beta_\gamma, \sigma^2 \mid \gamma) = p(Y \mid \beta_\gamma, \sigma^2, \gamma) p(\beta_\gamma \mid \sigma^2, \gamma) p(\sigma^2 \mid \gamma)$ to yield

$$p(Y \mid \gamma) \propto |X'_\gamma X_\gamma + \Sigma_\gamma^{-1}|^{-1/2} |\Sigma_\gamma|^{-1/2} (\nu\lambda + S_\gamma^2)^{-(n+\nu)/2} \tag{23}$$

where

$$S_\gamma^2 = Y'Y - Y'X_\gamma (X'_\gamma X_\gamma + \Sigma_\gamma^{-1})^{-1} X'_\gamma Y. \tag{24}$$

As will be seen in subsequent sections, the use of these closed form expressions can substantially speed up posterior evaluation and MCMC exploration. Note that the scale of the prior (21) for β_γ depends on σ^2, and this is needed to obtain conjugacy. Integrating out σ^2 with respect to (22), the prior for β_γ conditionally only on γ is

$$p(\beta_\gamma \mid \gamma) = T_{q_\gamma}(\nu, \bar{\beta}_\gamma, \lambda\Sigma_\gamma) \tag{25}$$

the q_γ-dimensional multivariate T-distribution centered at $\bar{\beta}_\gamma$ with ν degrees of freedom and scale $\lambda\Sigma_\gamma$.

The priors (21) and (22) are determined by the hyperparameters $\bar{\beta}_\gamma, \Sigma_\gamma, \lambda, \nu,$, which must be specified for implementations. Although a good deal of progress can be made through subjective elicitation of these hyperparameter values in smaller problems when substantial expert information is available, for example see Garthwaite and Dickey (1996), we focus here on the case where such information is unavailable and the goal is roughly to assign values that "minimize" prior influence.

Beginning with the choice of λ and ν, note that (22) corresponds to the likelihood information about σ^2 provided by ν independent observations from a $N(0, \lambda)$ distribution. Thus, λ may be thought of as a prior estimate of σ^2 and ν may be thought of as the prior sample size associated with this estimate. By using the data and treating s_Y^2, the sample variance of Y, as a rough upper bound for σ^2, a simple default strategy is to choose ν small, say around 3, and λ near s_Y^2. One might also go a bit further, by treating s_{FULL}^2, the traditional unbiased estimate of σ^2 based on a saturated model, as a rough lower bound for σ^2, and then choosing λ and ν so that (22) assigns substantial probability to the interval (s_{LS}^2, s_Y^2). Similar informal approaches based on the data are considered by Clyde, Desimone and Parmigiani (1996) and Raftery, Madigan and Hoeting (1997). Alternatively, the explicit choice of λ and ν can be avoided by using $p(\sigma^2 \mid \gamma) \propto 1/\sigma^2$, the limit of (22) as $\nu \to 0$, a choice recommended by Smith and Kohn (1996) and Fernandez, Ley and Steel (2001). This prior corresponds to the uniform distribution on $\log\sigma^2$, and is invariant to scale changes in Y. Although improper, it yields proper marginals $p(Y \mid \gamma)$ when combined with (21) and so can be used formally.

Turning to (21), the usual default for the prior mean $\bar{\beta}_\gamma$ has been $\bar{\beta}_\gamma = 0$, a neutral choice reflecting indifference between positive and negative values. This specification is also consistent with standard Bayesian approaches to testing point null hypotheses, where under the alternative, the prior is typically centered at the point null value. For choosing the prior covariance matrix Σ_γ, the specification is substantially simplified by setting $\Sigma_\gamma = c\,V_\gamma$, where c is a scalar and V_γ is a preset form such as $V_\gamma = (X'_\gamma X_\gamma)^{-1}$ or $V_\gamma = I_{q_\gamma}$, the $q_\gamma \times q_\gamma$ identity matrix. Note that under such V_γ, the conditional priors (21) provide a consistent description of uncertainty in the sense that they are the conditional distributions of the nonzero components of β given γ when $\beta \sim N_p(0, c\,\sigma^2 (X'X)^{-1})$ or $\beta \sim N_p(0, c\,\sigma^2 I)$, respectively. The choice $V_\gamma = (X'_\gamma X_\gamma)^{-1}$ serves to replicate the covariance structure of the likelihood, and yields the g-prior recommended by Zellner (1986). With $V_\gamma = I_{q_\gamma}$, the components of β_γ are conditionally independent, causing (21) to weaken the likelihood covariance structure. In contrast to $V_\gamma = (X'_\gamma X_\gamma)^{-1}$, the effect of $V_\gamma = I_{q_\gamma}$ on the posterior depends on the relative scaling of the predictors. In this regard, it may be reasonable to rescale the predictors in units of standard deviation to give them a common scaling, although this may be complicated by the presence of outliers or skewed distributions.

Having fixed V_γ, the goal is then to choose c large enough so that the prior is relatively flat over the region of plausible values of β_γ, thereby reducing prior influence (Edwards, Lindman and Savage 1963). At the same time, however, it is important to avoid excessively large values of c because the prior will eventually put increasing weight on the null model as $c \to \infty$, a form of the Bartlett-Lindley paradox, Bartlett (1957). For practical purposes, a rough guide is to choose c so that (25) assigns substantial probability to the range of all plausible values for β_γ. Raftery, Madigan and Hoeting (1997), who used a combination of $V_\gamma = I_{q_\gamma}$ and $V_\gamma = (X'_\gamma X_\gamma)^{-1}$ with standardized predictors, list various desiderata along the lines of this rough guide which lead them to the choice $c = 2.85^2$. They also note that their resulting coefficient prior is relatively flat over the actual distribution of coefficients from a variety of real data sets. Smith and Kohn (1996), who used $V_\gamma = (X'_\gamma X_\gamma)^{-1}$, recommend $c = 100$ and report that performance was insensitive to values of c between 10 and 10,000. Fernandez, Ley and Steel (2001) perform a simulation evaluation of the effect of various choices for c, with $V_\gamma = (X'_\gamma X_\gamma)^{-1}$, $p(\sigma^2 | \gamma) \propto 1/\sigma^2$ and $p(\gamma) = 2^{-p}$, on the posterior probability of the true model. Noting how the effect depends on the true model and noise level, they recommend $c = \max\{p^2, n\}$.

3.3 Calibration and Empirical Bayes Variable Selection

An interesting connection between Bayesian and non-Bayesian approaches to variable selection occurs when the special case of (21) with $\bar{\beta}_\gamma = 0$ and $V_\gamma = (X'_\gamma X_\gamma)^{-1}$, namely

$$p(\beta_\gamma \mid \sigma^2, \gamma) = N_{q_\gamma}(0, c\,\sigma^2 (X'_\gamma X_\gamma)^{-1}), \tag{26}$$

is combined with

$$p(\gamma) = w^{q_\gamma}(1 - w)^{p - q_\gamma} \tag{27}$$

the simple independence prior in (15); for the moment, σ^2 is treated as known. As shown by George and Foster (2000), this prior setup can be calibrated by choices of c and w so that the same γ maximizes both the model posterior and the canonical penalized sum-of-squares criterion

$$SS_\gamma / \sigma^2 - F\,q_\gamma \tag{28}$$

14

where $SS_\gamma = \hat{\beta}'_\gamma X'_\gamma X_\gamma \hat{\beta}_\gamma$, $\hat{\beta}_\gamma \equiv (X'_\gamma X_\gamma)^{-1} X'_\gamma Y$ and F is a fixed penalty. This correspondence may be of interest because a wide variety of popular model selection criteria are obtained by maximizing (28) with particular choices of F and with σ^2 replaced by an estimate $\hat{\sigma}^2$. For example $F = 2$ yields C_p (Mallows 1973) and, approximately, AIC (Akaike 1973), $F = \log n$ yields BIC (Schwarz 1978) and $F = 2 \log p$ yields RIC (Donoho and Johnstone 1994, Foster and George 1994). The motivation for these choices are varied; C_p is motivated as an unbiased estimate of predictive risk, AIC by an expected information distance, BIC by an asymptotic Bayes factor and RIC by minimax predictive risk inflation.

The posterior correspondence with (28) is obtained by reexpressing the model posterior under (26) and (27) as

$$
\begin{aligned}
p(\gamma \mid Y) &\propto w^{q_\gamma}(1-w)^{p-q_\gamma}(1+c)^{-q_\gamma/2} \exp\left\{-\frac{Y'Y - SS_\gamma}{2\sigma^2} - \frac{SS_\gamma}{2\sigma^2(1+c)}\right\} \\
&\propto \exp\left[\frac{c}{2(1+c)}\{SS_\gamma/\sigma^2 - F(c,w)\, q_\gamma\}\right],
\end{aligned}
\tag{29}
$$

where

$$
F(c,w) = \frac{1+c}{c}\left\{2\log\frac{1-w}{w} + \log(1+c)\right\}.
\tag{30}
$$

The expression (29) reveals that, as a function of γ for fixed Y, $p(\gamma \mid Y)$ is increasing in (28) when $F = F(c,w)$. Thus, both (28) and (29) are simultaneously maximized by the same γ when c and w are chosen to satisfy $F(c,w) = F$. In this case, Bayesian model selection based on $p(\gamma \mid Y)$ is equivalent to model selection based on the criterion (28).

This correspondence between seemingly different approaches to model selection provides additional insight and interpretability for users of either approach. In particular, when c and w are such that $F(c,w) = 2, \log n$ or $2 \log p$, selecting the highest posterior model (with σ^2 set equal to $\hat{\sigma}^2$) will be equivalent to selecting the best AIC/C$_p$, BIC or RIC models, respectively. For example, $F(c,w) = 2, \log n$ and $2 \log p$ are obtained when $c \simeq 3.92, n$ and p^2 and $w = 1/2$. Similar asymptotic connections are pointed out by Fernandez, Ley and Steel (2001) when $p(\sigma^2 \mid \gamma) \propto 1/\sigma^2$ and $w = 1/2$. Because the posterior probabilities are monotone in (28) when $F = F(c,w)$, this correspondence also reveals that the MCMC methods discussed in Section 3.5 can also be used to search for large values of (28) in large problems where global maximization is not computationally feasible. Furthermore, since c and w control the expected size and proportion of the nonzero components of β, the dependence of $F(c,w)$ on c and w provides an implicit connection between the penalty F and the profile of models for which its value may be appropriate.

Ideally, the prespecified values of c and w in (26) and (27) will be consistent with the true underlying model. For example, large c will be chosen when the regresssion coefficients are large, and small w will be chosen when the proportion of nonzero coefficients are small. To avoid the difficulties of choosing such c and w when the true model is completely unknown, it may be preferable to treat c and w as unknown parameters, and use empirical Bayes estimates of c and w based on the data. Such estimates can be obtained, at least in principle, as the values of c and w that maximize the marginal likelihood under (26) and (27), namely

$$
L(c,w \mid Y) \propto \sum_\gamma p(\gamma \mid w)\, p(Y \mid \gamma, c)
$$

$$\propto \quad \sum_{\gamma} w^{q_\gamma}(1-w)^{p-q_\gamma}(1+c)^{-q_\gamma/2}\exp\left\{\frac{c\,SS_\gamma}{2\sigma^2(1+c)}\right\}. \tag{31}$$

Although this maximization is generally impractical when p is large, the likelihood (31) simplifies considerably when X is orthogonal, a setup that occurs naturally in nonparametric function estimation with orthogonal bases such as wavelets. In this case, letting $t_i = b_i v_i/\sigma$ where v_i^2 is the ith diagonal element of $X'X$ and b_i is the ith component of $\hat{\beta} = (X'X)^{-1}X'Y$, (31) reduces to

$$L(c,w\,|\,Y) \propto \prod_{i=1}^{p}\left\{(1-w)e^{-t_i^2/2} + w(1+c)^{-1/2}e^{-t_i^2/2(1+c)}\right\}. \tag{32}$$

Since many fewer terms are involved in the product in (32) than in the sum in (31), maximization of (32) is feasible by numerical methods even for moderately large p.

Replacing σ^2 by an estimate $\hat{\sigma}^2$, the estimators $\hat{c}$ and $\hat{w}$ that maximize the marginal likelihood L above can be used as prior inputs for an empirical Bayes analysis under the priors (26) and (27). In particular, (4) reveals that the γ maximizing the posterior $p(\gamma\,|\,Y,\hat{c},\hat{w})$ can be obtained as the γ that maximizes the marginal maximum likelihood criterion

$$C_{\mathrm{MML}} = SS_\gamma/\hat{\sigma}^2 - F(\hat{c},\hat{w})\,q_\gamma, \tag{33}$$

where $F(c,w)$ is given by (30). Because maximizing (31) to obtain $\hat{c}$ and $\hat{w}$ can be computationally overwhelming when p is large and X is not orthogonal, one might instead consider a computable empirical Bayes approximation, the conditional maximum likelihood criterion

$$C_{\mathrm{CML}} = SS_\gamma/\hat{\sigma}^2 - q_\gamma\left\{1 + \log_+(SS_\gamma/\hat{\sigma}^2 q_\gamma)\right\} - 2\left\{\log(p-q_\gamma)^{-(p-q_\gamma)} + \log q_\gamma^{-q_\gamma}\right\} \tag{34}$$

where $\log_+(\cdot)$ is the positive part of $\log(\cdot)$. Selecting the γ that maximizes C_{CML} provides an approximate empirical Bayes alternative to selection based on C_{MML}.

In contrast to criteria of the form (28), which penalize $SS_\gamma/\hat{\sigma}^2$ by Fq_γ, with F constant, C_{MML} uses an adaptive penalty $F(\hat{c},\hat{w})$ that is implicitly based on the estimated distribution of the regression coefficients. C_{CML} also uses an adaptive penalty, but one can be expressed by a rapidly computable closed form that can be shown to act like a combination of a modified BIC penalty $F = \log n$, which gives it same consistency property as BIC, and a modified RIC penalty $F = 2\log p$. Insofar as maximizing C_{CML} approximates maximizing C_{MML}, these interpretations at least roughly explain the behavior of the C_{MML} penalty $F(\hat{c},\hat{w})$ in (33).

George and Foster (2000) proposed the empirical Bayes criteria C_{MML} and C_{CML} and provided simulation evaluations demonstrating substantial performance advantages over the fixed penalty criteria (28); selection using C_{MML} delivers excellent performance over a much wider portion of the model space, and C_{CML} performs nearly as well. The superiority of empirical Bayes methods was confirmed in context of wavelet regression by Johnstone and Silverman (1998) and Clyde and George (1999). Johnstone and Silverman (1998) demonstrated the superiority of using maximum marginal likelihood estimates of c and w with posterior median selection criteria, and proposed EM algorithms for implementation. Clyde and George (1999) also proposed EM algorithms for implementation, extended the methods to include empirical Bayes estimates of σ^2 and considered both model selection and model averaging.

Finally, a fully Bayes analysis which integrates out c and w with respect to some noninformative prior $p(c, w)$ could be a promising alternative to empirical Bayes estimation of c and w. Indeed, the maximum marginal likelihood estimates $\hat{c}$ and $\hat{w}$ correspond to the posterior mode estimates under a Bayes formulation with independent uniform priors on c and w, a natural default choice. As such, the empirical Bayes methods can be considered as approximations to fully Bayes methods, but approximations which do not fully account for the uncertainty surrounding c and w. We are currently investigating the potential of such fully Bayes alternatives and plan to report on them elsewhere.

3.4 Parameter Priors for Selection Based on Practical Significance

A potential drawback of the point-normal prior (21) for variable selection is that with enough data, the posterior will favor the inclusion of X_i for any $\beta_i \neq 0$, no matter how small. Although this might be desirable from a purely predictive standpoint, it can also run counter to the goals of parsimony and interpretability in some problems, where it would be preferable to ignore such negligible β_i. A similar phenomenon occurs in frequentist hypothesis testing, where for large enough sample sizes, small departures from a point null become statistically significant even though they are not practically significant or meaningful.

An alternative to the point-normal prior (21) that avoids this potential drawback is the normal-normal formulation used in the SSVS procedure of George and McCulloch (1993, 1996, 1997). This formulation builds in the goal of excluding X_i from the model whenever $|\beta_i| < \delta_i$ for a given $\delta_i > 0$. The idea is that δ_i is a "threshold of practical significance" that is prespecified by the user. A simple choice might be $\delta_i = \Delta Y / \Delta X_i$, where ΔY is the size of an insignificant change in Y, and ΔX_i is the size of the maximum feasible change in X_i. To account for the cumulative effect of changes of other X's in the model, one might prefer the smaller conservative choice $\delta_i = \Delta Y / (p \Delta X_i)$. The practical potential of the SSVS formulation is nicely illustrated by Wakefield and Bennett (1996).

Under the normal-normal formulation of SSVS, the data always follow the saturated model (13) so that

$$p(Y \mid \beta, \sigma^2, \gamma) = N_n(X\beta, \sigma^2 I) \tag{35}$$

for all γ. In the general notation of Section 2, the model parameters here are the same for every model, $\theta_k \equiv (\beta, \sigma^2)$. The γth model is instead distinguished by a coefficient prior of the form

$$\pi(\beta \mid \sigma^2, \gamma) = \pi(\beta \mid \gamma) = N_p(0, D_\gamma R D_\gamma) \tag{36}$$

where R is a correlation matrix and D_γ is a diagonal matrix with diagonal elements

$$(D_\gamma)_{ii} = \begin{cases} \sqrt{v_{0i}} & \text{when} \quad \gamma_i = 0 \\ \sqrt{v_{1i}} & \text{when} \quad \gamma_i = 1 \end{cases} \tag{37}$$

Under the model space prior $p(\gamma)$, the marginal prior distribution of each component of β is here

$$p(\beta_i) = (1 - p(\gamma_i))N(0, v_{0i}) + p(\gamma_i)N(0, v_{1i}), \tag{38}$$

a scale mixture of two normal distributions.

17

Although β is independent of σ^2 in (36), the inverse Gamma prior (22) for σ^2 is still useful, as are the specification considerations for it discussed in Section 3.2. Furthermore, $R \propto (X'X)^{-1}$ and $R = I$ are natural choices for R in (36), similarly to the commonly used choices for Σ_γ in (21).

To use this normal-normal setup for variable selection, the hyperparameters v_{0i} and v_{1i} are set "small and large" respectively, so that $N(0, v_{0i})$ is concentrated and $N(0, v_{1i})$ is diffuse. The general idea is that when the data support $\gamma_i = 0$ over $\gamma_i = 1$, then β_i is probably small enough so that X_i will not be needed in the model. For a given threshold δ_i, higher posterior weighting of those γ values for which $|\beta_i| > \delta_i$ when $\gamma_i = 1$, can be achieved by choosing v_{0i} and v_{1i} such that $p(\beta_i \mid \gamma_i = 0) = N(0, v_{0i}) > p(\beta_i \mid \gamma_i = 1) = N(0, v_{1i})$ precisely on the interval $(-\delta_i, \delta_i)$. This property is obtained by any v_{0i} and v_{1i} satisfying

$$\log(v_{1i}/v_{0i})/(v_{0i}^{-1} - v_{1i}^{-1}) = \delta_{i\gamma}^2 \tag{39}$$

By choosing v_{1i} such that $N(0, v_{1i})$ is consistent with plausible values of β_i, v_{0i} can then be chosen according to (39). George and McCulloch (1997) report that computational problems and difficulties with v_{1i} too large will be avoided whenever $v_{1i}/v_{0i} \leq 10,000$, thus allowing for a wide variety of settings.

Under the normal-normal setup above, the joint distribution of β and σ^2 given γ is not conjugate for (13) because (36) excludes σ^2. This prevents analytical reduction of the full posterior $p(\beta, \sigma^2, \gamma \mid Y)$, which can severely increase the cost of posterior computations. To avoid this, one can instead consider the conjugate normal-normal formulation using

$$p(\beta \mid \sigma^2, \gamma) = N_p(0, \sigma^2 D_\gamma R D_\gamma), \tag{40}$$

which is identical to (36) except for the insertion of σ^2. Coupled with the inverse Gamma prior (22) for σ^2, the conditional distribution of β and σ^2 given γ is conjugate. This allows for the analytical margining out of β and σ^2 from $p(Y, \beta, \sigma^2 \mid \gamma) = p(Y \mid \beta, \sigma^2)p(\beta \mid \sigma^2, \gamma)p(\sigma^2 \mid \gamma)$ to yield

$$p(Y \mid \gamma) \propto |X'X + (D_\gamma R D_\gamma)^{-1}|^{-1/2} |D_\gamma R D_\gamma|^{-1/2} (\nu\lambda + S_\gamma^2)^{-(n+\nu)/2} \tag{41}$$

where

$$S_\gamma^2 = Y'Y - Y'X(X'X + (D_\gamma R D_\gamma)^{-1})^{-1}X'Y. \tag{42}$$

As will be seen in Section 3.5, this simplification confers strong advantages for posterior calculation and exploration.

Under (40), (22), and a model space prior $p(\gamma)$, the marginal distribution each component of β is

$$p(\beta_i \mid \gamma) = (1 - \gamma_i)T_1(\nu, 0, \lambda v_{0i}) + \gamma_i T_1(\nu, 0, \lambda v_{1i}), \tag{43}$$

a scale mixture of t-distributions, in contrast to the normal mixture (38). As with the nonconjugate prior, the idea is that v_{0i} and v_{1i} are to be set "small and large" respectively, so that when the data supports $\gamma_i = 0$ over $\gamma_i = 1$, then β_i is probably small enough so that X_i will not be needed in the model. However, the way in which v_{0i} and v_{1i} determine "small and large" is affected by the unknown value of σ^2, thereby making specification more difficult and less reliable than in the nonconjugate formulation. For a chosen threshold of practical significance δ_i, the

pdf $p(\beta_i \,|\, i, \gamma_i = 0) = T(\nu, 0, \lambda \upsilon_{0i})$ is larger than the pdf $p(\beta_i \,|\, i, \gamma_i = 1) = T(\nu, 0, \lambda \upsilon_{1i})$ precisely on the interval $(-\delta_i, \delta_i)$, when υ_{0i} and υ_{1i} satisfy

$$(\upsilon_{0i}/\upsilon_{1i})^{\nu/(\nu+1)} = [\upsilon_{0i} + \delta_i^2/(\nu\lambda)]/[\upsilon_{1i} + \delta_i^2/(\nu\lambda)] \tag{44}$$

By choosing υ_{1i} such that $T(\nu, 0, \lambda \upsilon_{1i})$ is consistent with plausible values of β_i, υ_{0i} can then be chosen according to (44).

Another potentially valuable specification of the conjugate normal-normal formulation can be used to address the problem of outlier detection, which can be framed as a variable selection problem by including indicator variables for the observations as potential predictors. For such indicator variables, the choice $\upsilon_{0i}^* = 1$ and $\upsilon_{1i}^* = K > 0$ yields the well-known additive outlier formulation, see, for example, Petit and Smith (1985). Furthermore, when used in combination with the previous settings for ordinary predictors, the conjugate prior provides a hierarchical formulation for simultaneous variable selection and outlier detection. This has also been considered by Smith and Kohn (1996). A related treatment has been considered by Hoeting, Raftery and Madigan (1996).

3.5 Posterior Calculation and Exploration for Variable Selection

3.5.1 Closed Form Expressions for $p(Y \,|\, \gamma)$

A valuable feature of the previous conjugate prior formulations is that they allow for analytical margining out of β and σ^2 from $p(Y, \beta, \sigma^2 \,|\, \gamma)$ to yield the closed form expressions in (23) and (41) which are proportional to $p(Y \,|\, \gamma)$. Thus, when the model prior $p(\gamma)$ is computable, this can be used to obtain a computable, closed form expression $g(\gamma)$ satisfying

$$g(\gamma) \propto p(Y \,|\, \gamma)p(\gamma) \propto p(\gamma \,|\, Y). \tag{45}$$

The availability of such $g(\gamma)$ can greatly facilitate posterior calculation and estimation. Furthermore, it turns out that for certain formulations, the value of $g(\gamma)$ can be rapidly updated as γ is changed by a single component. As will be seen, such rapid updating schemes can be used to speed up algorithms for evaluating and exploring the posterior $p(\gamma \,|\, Y)$.

Consider first the conjugate point-normal formulation (21) and (22) for which $p(Y \,|\, \gamma)$ proportional to (23) can be obtained. When $\Sigma_\gamma = c\,(X_\gamma' X_\gamma)^{-1}$, a function $g(\gamma)$ satisfying (45) can be expressed as

$$g(\gamma) = (1 + c)^{-q_\gamma/2}(\nu\lambda + Y'Y - (1 + 1/c)^{-1}W'W)^{-(n+\nu)/2}p(\gamma) \tag{46}$$

where $W = T'^{-1}X_\gamma'Y$ for upper triangular T such that $T'T = X_\gamma' X_\gamma$ for (obtainable by the Cholesky decomposition). As noted by Smith and Kohn (1996), the algorithm of Dongarra, Moler, Bunch and Stewart (1979) provides fast updating of T, and hence W and $g(\gamma)$, when γ is changed one component at a time. This algorithm requires $O(q_\gamma^2)$ operations per update, where γ is the changed value.

Now consider the conjugate normal-normal formulation (40) and (22) for which $p(Y \,|\, \gamma)$ proportional to (41) can be obtained. When $R = I$ holds, a function $g(\gamma)$ satisfying (45) can be

19

expressed as

$$g(\gamma) = (\prod_{i=1}^{p} T_{ii}^2 \left[(1 - \gamma_i)v_{0\gamma_{(i)}}^* + \gamma_i v_{1\gamma_{(i)}}^*\right])^{-1/2}(\nu\lambda + Y'Y - W'W)^{-(n+\nu)/2}p(\gamma) \qquad (47)$$

where $W = T'^{-1}\tilde{X}'\tilde{Y}$ for upper triangular T such that $T'T = \tilde{X}'\tilde{X}$ (obtainable by the Cholesky decomposition). As noted by George and McCulloch (1997), the Chambers (1971) algorithm provides fast updating of T, and hence W and $g(\gamma)$, when γ is changed one component at a time. This algorithm requires $O(p^2)$ operations per update.

The availability of these computable, closed form expressions for $g(\gamma) \propto p(\gamma \mid Y)$ enables exhaustive calculation of $p(\gamma \mid Y)$ in moderately sized problems. In general, this simply entails calculating $g(\gamma)$ for every γ value and then summing over all γ values to obtain the normalization constant. However, when one of the above fast updating schemes can be used, this calculation can be substantially speeded up by sequential calculation of the 2^p $g(\gamma)$ values where consecutive γ differ by just one component. Such an ordering is provided by the Gray Code, George and McCulloch (1997). After computing T, W and $g(\gamma)$ for an initial γ value, subsequent values of T, W and $g(\gamma)$ can be obtained with the appropriate fast updating scheme by proceeding in the Gray Code order. Using this approach, this exhaustive calculation is feasible for p less than about 25.

3.5.2 MCMC Methods for Variable Selection

MCMC methods have become a principal tool for posterior evaluation and exploration in Bayesian variable selection problems. Such methods are used to simulate a sequence

$$\gamma^{(1)}, \gamma^{(2)}, \ldots \qquad (48)$$

that converges (in distribution) to $p(\gamma \mid Y)$. In formulations where analytical simplification of $p(\beta, \sigma^2, \gamma \mid Y)$ is unavailable, (48) can be obtained as a subsequence of a simulated Markov chain of the form

$$\beta^{(1)}, \sigma^{(1)}, \gamma^{(1)}, \beta^{(2)}, \sigma^{(2)}, \gamma^{(2)}, \ldots \qquad (49)$$

that converges to $p(\beta, \sigma^2, \gamma \mid Y)$. However, in conjugate formulations where β and σ^2 can be analytically eliminated form the posterior, the availability of $g(\gamma) \propto p(\gamma \mid Y)$ allows for the flexible construction of MCMC algorithms that simulate (48) directly as a Markov chain. Such chains are often more useful, in terms of both computational and convergence speed.

In problems where the number of potential predictors p is very small, and $g(\gamma) \propto p(\gamma \mid Y)$ is unavailable, the sequence (48) may be used to evaluate the entire posterior $p(\gamma \mid Y)$. Indeed, empirical frequencies and other functions of the γ values will be consistent estimates of their values under $p(\gamma \mid Y)$. In large problems where exhaustive calculation of all 2^p values of $p(\gamma \mid Y)$ is not feasible, the sequence (48) may still provide useful information. Even when the length of the sequence (48) is much smaller than 2^p, it may be possible to identify at least some of the high probability γ, since those γ are expected to appear more frequently. In this sense, these MCMC methods can be used to stochastically search for high probability models.

In the next two subsections, we describe various MCMC algorithms which may be useful for simulating (48). These algorithms are obtained as variants of the Gibbs sampler (GS) and Metropolis-Hastings (MH) algorithms described in Section 2.3.

3.5.3 Gibbs Sampling Algorithms

Under the nonconjugate normal-normal formulation (36) and (22) for SSVS, the posterior $p(\beta, \sigma^2, \gamma \mid Y)$ is p-dimensional for all γ. Thus, a simple GS that simulates the full parameter sequence (49) is obtained by successive simulation from the full conditionals

$$p(\beta \mid \sigma^2, \gamma, Y)$$

$$p(\sigma^2 \mid \beta, \gamma, Y) = p(\sigma^2 \mid \beta, Y) \tag{50}$$

$$p(\gamma_i \mid \beta, \sigma^2, \gamma_{(i)}, Y) = p(\gamma_i \mid \beta, \gamma_{(i)}), \ i = 1, \ldots, p$$

where at each step, these distributions are conditioned on the most recently generated parameter values. These conditionals are standard distributions which can be simulated quickly and efficiently by routine methods.

For conjugate formulations where $g(\gamma)$ is available, a variety of MCMC algorithms for generating (48) directly as a Markov chain, can be conveniently obtained by applying the GS with $g(\gamma)$. The simplest such implementation is obtained by generating each γ value componentwise from the full conditionals,

$$\gamma_i \mid \gamma_{(i)}, Y \quad i = 1, 2, \ldots, p, \tag{51}$$

$(\gamma_{(i)} = (\gamma_1, \gamma_2, \ldots, \gamma_{i-1}, \gamma_{i+1}, \ldots, \gamma_p))$ where the γ_i may be drawn in any fixed or random order. By margining out β and σ^2 in advance, the sequence (48) obtained by this algorithm should converge faster than the nonconjugate Gibbs approach, rendering it more effective on a per iteration basis for learning about $p(\gamma \mid Y)$, see Liu, Wong and Kong (1994).

The generation of the components in (51) in conjunction with $g(\gamma)$ can be obtained trivially as a sequence of Bernoulli draws. Furthermore, if $g(\gamma)$ allows for fast updating as in (46) or (47), the required sequence of Bernoulli probabilities can be computed faster and more efficiently. To see this, note that the Bernoulli probabilities are simple functions of the ratio

$$\frac{p(\gamma_i = 1, \gamma_{(i)} \mid Y)}{p(\gamma_i = 0, \gamma_{(i)} \mid Y)} = \frac{g(\gamma_i = 1, \gamma_{(i)})}{g(\gamma_i = 0, \gamma_{(i)})}. \tag{52}$$

At each step of the iterative simulation from (51), one of the values of $g(\gamma)$ in (52) will be available from the previous component simulation. Since γ has been varied by only a single component, the other value of $g(\gamma)$ can then be obtained by using the appropriate updating scheme. Under the point-normal prior (21) with $\Sigma_\gamma = c (X'_\gamma X_\gamma)^{-1}$, the fast updating of (46) requires $O(q_\gamma^2)$ operations, whereas under the conjugate normal-normal prior formulation (40) with $R = I$ fast updating of (47) requires $O(p^2)$ operations. Thus, GS algorithms in the former case can be substantially faster when $p(\gamma \mid Y)$ is concentrated on those γ for which q_γ is small, namely the parsimonious models. This advantage could be pronounced in large problems with many useless predictors.

Simple variants of the componentwise GS can be obtained by generating the components in a different fixed or random order. Note that in any such generation, it is not necessary to generate each and every component once before repeating a coordinate. Another variant of the GS can be obtained by drawing the components of γ in groups, rather than one at a time. Let $\{I_k\}$, $k = 1, 2, \ldots, m$ be a partition of $\{1, 2, \ldots, p\}$ so that, $I_k \subseteq \{1, 2, \ldots, p\}$, $\cup I_k = \{1, 2, \ldots, p\}$ and $I_{k_1} \cap I_{k_2} = \emptyset$ for $k_1 \neq k_2$. Let $\gamma_{I_k} = \{\gamma_i \mid i \in I_k\}$ and $\gamma_{(I_k)} = \{\gamma_i \mid i \notin I_k\}$. The grouped GS generates (48) by iterative simulation from

$$\gamma_{I_k} \mid \gamma_{(I_k)}, Y \quad k = 1, 2, \ldots, m. \tag{53}$$

Fast updating of $g(\gamma)$, when available, can also be used to speed up this simulation by computing the conditional probabilities of each γ_{I_k} in Gray Code order. The potential advantage of such a grouped GS is improved convergence of (48). This might be achieved by choosing the partition so that strongly correlated γ_i are contained in the same I_k, thereby reducing the dependence between draws in the simulation. Intuitively, clusters of such correlated γ_i should correspond to clusters of correlated X_i which, in practice, might be identified by clustering procedures. As before, variants of the grouped GS can be obtained by generating the γ_{I_k} in a different fixed or random order.

3.5.4 Metropolis-Hastings Algorithms

The availability of $g(\gamma) \propto p(\gamma \mid Y)$ also facilitates the use of MH algorithms for direct simulation of (48). By restricting attention to the set of γ values, a discrete space, the simple MH form described in Section 2.3 can be used. Because $g(\gamma)/g(\gamma') = p(\gamma|Y)/p(\gamma'|Y)$, such MH algorithms are here of the form:

1. Simulate a candidate γ^* from a transition kernel $q(\gamma^* \mid \gamma^{(j)})$.

2. Set $\gamma^{(j+1)} = \gamma^*$ with probability

$$\alpha(\gamma^* \mid \gamma^{(j)}) = \min\left\{\frac{q(\gamma^{(j)} \mid \gamma^*)}{q(\gamma^* \mid \gamma^{(j)})} \frac{g(\gamma^*)}{g(\gamma^{(j)})}, 1\right\}. \tag{54}$$

Otherwise, $\gamma^{(j+1)} = \gamma^{(j)}$.

When available, fast updating schemes for $g(\gamma)$ can be exploited. Just as for the Gibbs sampler, the MH algorthims under the point-normal formulations (21) with $\Sigma_\gamma = c\,(X'_\gamma X_\gamma)^{-1}$ will be the fastest scheme when $p(\gamma \mid Y)$ is concentrated on those γ for which q_γ is small.

A special class of MH algorithms, the Metropolis algorithms, are obtained from the class of transition kernels $q(\gamma^1 \mid \gamma^0)$ which are symmetric in γ^1 and γ^0. For this class, the form of (54) simplifies to

$$\alpha^M(\gamma^* \mid \gamma^{(j)}) = \min\left\{\frac{g(\gamma^*)}{g(\gamma^{(j)})}, 1\right\}. \tag{55}$$

Perhaps the simplest symmetric transition kernel is

$$q(\gamma^1 \mid \gamma^0) = 1/p \quad \text{if} \quad \sum_1^p |\gamma_i^0 - \gamma_i^1| = 1. \tag{56}$$

This yields the Metropolis algorithm

1. Simulate a candidate γ^* by randomly changing one component of $\gamma^{(j)}$.

2. Set $\gamma^{(j+1)} = \gamma^*$ with probability $\alpha^M(\gamma^* \mid \gamma^{(j)})$. Otherwise, $\gamma^{(j+1)} = \gamma^{(j)}$.

This algorithm was proposed in a related model selection context by Madigan and York (1995) who called it MC^3. It was used by Raftery, Madigan and Hoeting (1997) for model averaging, and was proposed for the SSVS prior formulation by Clyde and Parmigiani (1994).

The transition kernel (56) is a special case of the class of symmetric transition kernels of the form

$$q(\gamma^1 \mid \gamma^0) = q_d \quad \text{if} \quad \sum_1^p |\gamma_i^0 - \gamma_i^1| = d. \tag{57}$$

Such transition kernels yield Metropolis algorithms of the form

1. Simulate a candidate γ^* by randomly changing d components of $\gamma^{(j)}$ with probability q_d.

2. Set $\gamma^{(j+1)} = \gamma^*$ with probability $\alpha^M(\gamma^* \mid \gamma^{(j)})$. Otherwise, $\gamma^{(j+1)} = \gamma^{(j)}$.

Here q_d is the probability that γ^* will have d new components. By allocating some weight to q_d for larger d, the resulting algorithm will occasionally make big jumps to different γ values. In contrast to the algorithm obtained by (56) which only moves locally, such algorithms require more computation per iteration.

Finally, it may also be of interest to consider asymmetric transition kernels such as

$$q(\gamma^1 \mid \gamma^0) = q_d \quad \text{if} \quad \sum_1^p (\gamma_i^0 - \gamma_i^1) = d. \tag{58}$$

Here q_d is the probability of generating a candidate value γ^* which corresponds to a model with d more variables $\gamma^{(j)}$. When $d < 0$, γ^* will represent a more parsimonious model than $\gamma^{(j)}$. By suitable weighting of the q_d probabilities, such Metropolis-Hastings algorithms can be made to explore the posterior in the region of more parsimonious models.

3.5.5 Extracting Information from the Output

In nonconjugate setups, where $g(\gamma)$ is unavailable, inference about posterior characteristics based on (48) ultimately rely on the empirical frequency estimates the visited γ values. Although such estimates of posterior characteristics will be consistent, they may be unreliable, especially if the size of the simulated sample is small in comparison to 2^p or if there is substantial dependence between draws. The use of empirical frequencies to identify high probability γ values for selection can be similarly problematic.

However, the situation changes dramatically in conjugate setups where $g(\gamma) \propto p(\gamma \mid Y)$ is available. To begin with, $g(\gamma)$ provides the relative probability of any two values γ^0 and γ^1 via $g(\gamma^0) / g(\gamma^1)$ and so can be used to definitively identify the higher probability γ in the sequence (48) of simulated values. Only minimal additional effort is required to obtain such calculations since $g(\gamma)$ must be calculated for each of the visited γ values in the execution of the MCMC algorithms described in Sections 3.5.3 and 3.5.4.

The availability of $g(\gamma)$ can also be used to obtain estimators of the normalizing constant C

$$p(\gamma \mid Y) = C \, g(\gamma) \tag{59}$$

23

based on the MCMC output (48) , say $\gamma^{(1)}, \dots, \gamma^{(K)}$. Let A be a preselected subset of γ values and let $g(A) = \sum_{\gamma \in A} g(\gamma)$ so that $p(A \mid Y) = C\,g(A)$. Then, a consistent estimator of C is

$$\hat{C} = \frac{1}{g(A)K} \sum_{k=1}^{K} I_A(\gamma^{(k)}) \tag{60}$$

where $I_A(\,)$ is the indicator of the set A, George and McCulloch (1997). Note that if (48) were an uncorrelated sequence, then $\operatorname{Var}(\hat{C}) = (C^2/K)(1 - p(A \mid Y))/p(A \mid Y)$ suggesting that the variance of (60) is decreasing as $p(A \mid Y)$ increases. It is also desirable to choose A such that $I_A(\gamma^{(k)})$ can be easily evaluated. George and McCulloch (1997) obtain very good results by setting A to be those γ values visited by a preliminary simulation of (48). Peng (1998) has extended and generalized these ideas to obtain estimators of C that improve on (60).

Inserting $\hat{C}$ into (59) yields improved estimates of the probability of individual γ values,

$$\hat{p}(\gamma \mid Y) = \hat{C}\,g(\gamma), \tag{61}$$

as well as an estimate of the total visited probability,

$$\hat{p}(B \mid Y) = \hat{C}\,g(B), \tag{62}$$

where B is the set of visited γ values. Such $\hat{p}(B \mid Y)$ can provide valuable information about when to stop a MCMC simulation. Another useful quantity is

$$|(\hat{C}/C) - 1|. \tag{63}$$

Since $\hat{p}(\gamma \mid Y)/p(\gamma \mid Y) \equiv \hat{C}/C$, (63) measures of the uniform accuracy of the probability estimates. It also measures the total probability discrepancy since $\sum_{\gamma} |\hat{p}(\gamma \mid Y) - p(\gamma \mid Y)| = |\hat{C} - C| \sum_{\gamma} g(\gamma) = |(\hat{C}/C) - 1|$.

The simulated values (48) can also play an important role in model averaging. For example, suppose one wanted to predict a quantity of interest Δ by the posterior mean

$$E(\Delta \mid Y) = \sum_{\text{all } \gamma} E(\Delta \mid \gamma, Y) p(\gamma \mid Y). \tag{64}$$

When p is too large for exhaustive enumeration and $p(\gamma \mid Y)$ cannot be computed, (64) is unavailable and is typically approximated by something of the form

$$\hat{E}(\Delta \mid Y) = \sum_{\gamma \in S} E(\Delta \mid \gamma, Y) \hat{p}(\gamma \mid Y, S) \tag{65}$$

where S is a manageable subset of models and $\hat{p}(\gamma \mid Y, S)$ is a probability distribution over S. (In some cases, $E(\Delta \mid \gamma, Y)$ will also be approximated).

Using the Markov chain sample for S, a natural choice for (65) is

$$\hat{E}_f(\Delta \mid Y) = \sum_{\gamma \in S} E(\Delta \mid \gamma, Y) \hat{p}_f(\gamma \mid Y, S) \tag{66}$$

where $\hat{p}_f(\gamma \mid Y, S)$ is the relative frequency of γ in S, George (1999). Indeed, (66) will be a consistent estimator of $E(\Delta \mid Y)$. However, here too, it appears that when $g(\gamma)$ is available, one can do better by using

$$\hat{E}_g(\Delta \mid Y) = \sum_{\gamma \in S} E(\Delta \mid \gamma, Y) \hat{p}_g(\gamma \mid Y, S) \tag{67}$$

where

$$\hat{p}_g(\gamma \mid Y, S) = g(\gamma)/g(S) \tag{68}$$

is just the renormalized value of $g(\gamma)$. For example, when S is an iid sample from $p(\gamma \mid Y)$, (67) increasingly approximates the best unbiased estimator of $E(\Delta \mid Y)$ as the sample size increases. To see this, note that when S is an iid sample, $\hat{E}_f(\Delta \mid Y)$ is unbiased for $E(\Delta \mid Y)$. Since S (together with g) is sufficient, the Rao-Blackwellized estimator $E(\hat{E}_f(\Delta \mid Y) \mid S)$ is best unbiased. But as the sample size increases, $E(\hat{E}_f(\Delta \mid Y) \mid S) \to \hat{E}_g(\Delta \mid Y)$.

4   Bayesian CART Model Selection

For our second illustration of Bayesian model selection implementations, we consider the problem of selecting a classification and regression tree (CART) model for the relationship between a variable y and a vector of potential predictors $x = (x_1, \ldots, x_p)$. An alternative to linear regression, CART models provide a more flexible specification of the conditional distribution of y given x. This specification consists of a partition of the x space, and a set of distinct distributions for y within the subsets of the partition. The partition is accomplished by a binary tree T that recursively partitions the x space with internal node splitting rules of the form $\{x \in A\}$ or $\{x \notin A\}$. By moving from the root node through to the terminal nodes, each observation is assigned to a terminal node of T which then associates the observation with a distribution for y.

Although any distribution may be considered for the terminal node distributions, it is convenient to specify these as members of a single parametric family $p(y \mid \theta)$ and to assume all observations of y are conditionally independent given the parameter values. In this case, a CART model is identified by the tree T and the parameter values $\Theta = (\theta_1, \ldots, \theta_b)$ of the distributions at each of the b terminal nodes of T. Note that T here plays the role of M_k of model identifier as described in Section 2. The model is called a regression tree model or a classification tree model according to whether y is quantitative or qualitative, respectively. For regression trees, two simple and useful specifications for the terminal node distributions are the mean shift normal model

$$p(y \mid \theta_i) = N(\mu_i, \sigma^2), \ i = 1, \ldots, b, \tag{69}$$

where $\theta_i = (\mu_i, \sigma)$, and the mean-variance shift normal model

$$p(y \mid \theta_i) = N(\mu_i, \sigma_i^2), \ i = 1, \ldots, b, \tag{70}$$

where $\theta_i = (\mu_i, \sigma_i)$. For classification trees where y belongs to one of K categories, say $C_1, \ldots, C_K$, a natural choice for terminal node distributions are the simple multinomials

$$p(y \mid \theta_i) = \prod_{k=1}^{K} p_{ik}^{I(y \in C_k)} \quad i = 1, \ldots, b, \tag{71}$$

where $\theta_i = p_i \equiv (p_{i1}, \ldots, p_{iK})$, $p_{ik} \geq 0$ and $\sum_k p_{ik} = 1$. Here $p(y \in C_k) = p_{ik}$ at the ith terminal node of T.

As illustration, Figure 1 depicts a regression tree model where $y \sim N(\theta, 2^2)$ and $x = (x_1, x_2)$. x_1 is a quantitative predictor taking values in [0,10], and x_2 is a qualitative predictor with

$$X_2 \in \{C, D\} \qquad X_2 \in \{A, B\}$$

$$X_1 \leq 3 \qquad X_1 > 3 \qquad X_1 \leq 5 \qquad X_1 > 5$$

$$\theta_1 = 1 \qquad \theta_4 = 8 \qquad \theta_5 = 2$$

$$X_1 \leq 7 \qquad X_1 > 7$$
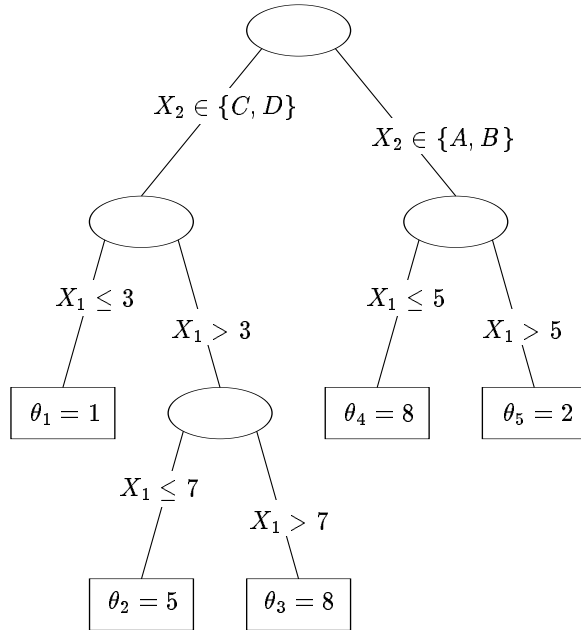
$$\theta_2 = 5 \qquad \theta_3 = 8$$

Figure 1: A regression tree where $y \sim N(\theta, 2^2)$ and $x = (x_1, x_2)$.

categories $\{A, B, C, D\}$. The binary tree has 9 nodes of which $b = 5$ are terminal nodes that partition the x space into 5 subsets. The splitting rules are displayed at each internal node. For example, the leftmost terminal node corresponds to $x_1 \leq 3.0$ and $x_2 \in \{C, D\}$. The θ_i value identifying the mean of y given x is displayed at each terminal node. Note that in contrast to a linear model, θ_i decreases in x_1 when $x_2 \in \{A, B\}$, but increases in x_1 when $x_2 \in \{C, D\}$.

The two basic components of the Bayesian approach to CART model selection are prior specification and posterior exploration. Prior specification over CART models entails putting a prior on the tree space and priors on the parameters of the terminal node distributions. The CART model likelihoods are then used to update the prior to yield a posterior distribution that can be used for model selection. Although straightforward in principle, practical implementations require subtle and delicate attention to details. Prior formulation must be interpretable and computationally manageable. Hyperparameter specification can be usefully guided by overall location and scale measures of the data. A feature of this approach is that the prior specification can be used to downweight undesirable model characteristics such as tree complexity or to express a preference for certain predictor variables. Although the entire posterior cannot be computed in nontrivial problems, posterior guided MH algorithms can still be used to search for good tree models. However, the algorithms require repeated restarting or other modifications because of the multimodal nature of the posterior. As the search proceeds, selection based on marginal likelihood rather than posterior probability is preferable because of the dilution properties of the prior. Alternatively, a posterior weighted average of the visited models can be

easily obtained.

CART modeling was popularized in the statistical community by the seminal book of Breiman, Friedman, Olshen and Stone (1984), which proposed a nonparametric approach for tree selection based on a greedy algorithm named CART. A concise description of this approach, which seeks to partition the x space into regions where the distribution of y is 'homogeneous', and its implementation in S appears in Clark and Pregibon (1992). Bayesian approaches to CART are enabled by elaborating the CART tree formulation to include parametric terminal node distributions, effectively turning it into a statistical model and providing a likelihood. Conventional greedy search algorithms are also replaced by the MCMC algorithms that provide a broader search over the tree model space.

The Bayesian CART model selection implementations described here were proposed by Chipman, George and McCulloch (1998) and Denison, Mallick and Smith (1998a), hereafter referred to as CGM and DMS, respectively. An earlier Bayesian approach to classification tree modeling was proposed by Buntine (1992) which, compared to CGM and DMS, uses similar priors for terminal node distributions, but different priors on the space of trees, and deterministic, rather than stochastic, algorithms for model search. Priors for tree models based on Minimum Encoding ideas were proposed by Quinlan and Rivest (1989) and Wallace and Patrick (1993). Oliver and Hand (1995) discuss and provide an empirical comparison of a variety of pruning and Bayesian model averaging approaches based on CART. Paass and Kindermann (1997) applied a simpler version of the CGM approach and obtained results which uniformly dominated a wide variety of competing methods. Other alternatives to greedy search methods include Sutton (1991) and Lutsko and Kuijpers (1994) who use simulated annealing, Jordan and Jacobs (1994) who use the EM algorithm, Breiman (1996), who averages trees based on bootstrap samples, and Tibshirani and Knight (1995) who select trees based on bootstrap samples.

4.1 Prior Formulations for Bayesian CART

Since a CART model is identified by (Θ, T), a Bayesian analysis of the problem proceeds by specifying priors on the parameters of the terminal node distributions of each tree $p(\Theta \mid T)$ and a prior distribution $p(T)$ over the set of trees. Because the prior for T does not depend on the form of the terminal node distributions, $p(T)$ can be generally considered for both regression trees and classification trees.

4.1.1 Tree Prior Specification

A CART model tree T partitions the x space and consists of both the binary tree structure and the set of splitting rules associated with the internal nodes. A general formulation approach for $p(T)$ proposed by CGM, is to specify $p(T)$ implicitly by the following tree-generating stochastic process which "grows" trees from a single root tree by randomly "splitting" terminal nodes:

1. Begin by setting T to be the trivial tree consisting of a single root (and terminal) node denoted η.

2. Split the terminal node η with probability $p_\eta = \alpha(1 + d_\eta)^{-\beta}$ where d_η is the depth of the node η, and $\alpha \in (0, 1)$ and $\beta \geq 0$ are prechosen control parameters.

27

3. If the node splits, randomly assign it a splitting rule as follows: First choose x_i uniformly from the set of available predictors. If x_i is quantitative, assign a splitting rule of the form $\{x_i \leq s\}$ vs $\{x_i > s\}$ where s is chosen uniformly from the available observed values of x_i. If x_i is qualitative, assign a splitting rule of the form $\{x_i \in C\}$ vs $\{x_i \notin C\}$ where C is chosen uniformly from the set of subsets of available categories of x_i. Next assign left and right children nodes to the split node, and apply steps 2 and 3 to the newly created tree with η equal to the new left and the right children (if nontrivial splitting rules are available).

By available in step 3, we mean those predictors, split values and category subsets that would not lead to empty terminal nodes. For example, if a binary predictor was used in a splitting rule, it would no longer be available for splitting rules at nodes below it. Each realization of such a process can simply be considered as a random draw from $p(T)$. Furthermore, this specification allows for straightforward evaluation of $p(T)$ for any T, and can be effectively coupled with the MH search algorithms described in Section 4.2.1.

Although other useful forms can easily be considered for the splitting probability in step 2 above, the choice of $p_\eta = \alpha(1 + d_\eta)^{-\beta}$ is simple, interpretable, easy to compute and dependent only on the depth d_η of the node η. The parameters α and β control the size and shape of the binary tree produced by the process. To see how, consider the simple specification, $p_\eta \equiv \alpha < 1$ when $\beta = 0$. In this case the probability of any particular binary tree with b terminal nodes (ignoring the constraints of splitting rule assignments in step 3) is just $\alpha^{b-1}(1-\alpha)^b$, a natural generalization of the geometric distribution. (A binary tree with b terminal nodes will always have exactly $(b-1)$ internal nodes). Setting α small will tend to yield smaller trees and is a simple convenient way to control the size of trees generated by growing process.

The choice of $\beta = 0$ above assigns equal probability to all binary trees with b terminal nodes regardless of their shape. Indeed any prior that is only a function of b will do this; for example, DMS recommend this with a truncated Poisson distribution on b. However, for increasing $\beta > 0$, p_η is a decreasing function of d_η making deeper nodes less likely to split. The resulting prior $p(T)$ puts higher probability on "bushy" trees, those whose terminal nodes do not vary too much in depth. Choosing α and β in practice can guided by looking at the implicit marginal distributions of characteristics such as b. Such marginals can be easily simulated and graphed.

Turning to the splitting rule assignments, step 3 of the tree growing process represents the prior information that at each node, available predictors are equally likely to be effective, and that for each predictor, available split values or category subsets are equally likely to be effective. This specification is invariant to monotone transformations of the quantitative predictors, and is uniform on the observed quantiles of a quantitative predictor with no repeated values. However, it is not uniform over all possible splitting rules because it assigns lower probability to splitting rules based on predictors with more potential split values or category subsets. This feature is necessary to maintain equal probability on predictor choices, and is essentially yields the dilution property discussed in Sections 2.2 and 3.1. Predictors with more potential split values will give rise to more trees. By downweighting the splitting rules of such predictors, $p(T)$ serves to dilute probability within neighborhoods of similar trees.

Although the uniform choices for $p(T)$ above seem to be reasonable defaults, non-uniform choices may also be of interest. For example, it may be preferable to place higher probability on predictors that are thought to be more important. A preference for models with fewer variables could be expressed by putting greater mass on variables already assigned to ancestral nodes. For the choice of split value, tapered distribution at the extremes would increase the tendency to split more towards the interior range of a variable. One might also consider the distribution of split values to be uniform on the available range of the predictor and so could weight the available observed values accordingly. For the choice of category subset, one might put extra weight on subsets thought to be more important.

As a practical matter, note that all of the choices above consider only the observed predictor values as possible split points. This induces a discrete distribution on the set of splitting rules, and hence the support of $p(T)$ will be a finite set of trees in any application. This is not really a restriction since it allows for all possible partitions of any given data set. The alternative of putting a continuous distribution on the range of the predictors would needlessly increase the computational requirements of posterior search while providing no gain in generality. Finally, we note that the dependence of $p(T)$ on the observed x values is typical of default prior formulations, as was the case for some of the coefficient prior covariance choices discussed in Sections 3.2 and 3.4.

4.1.2 Parameter Prior Specifications

As discussed in Section 2.3, the computational burden of posterior calculation and exploration is substantially reduced when the marginal likelihood, here $p(Y \mid T)$, can be obtained in closed form. Because of the large size of the space of CART models, this computational consideration is key in choosing the prior $p(\Theta \mid T)$ for the parameters of the terminal node distributions. For this purpose, we recommend the conjugate prior forms below for the parameters of the models (69)-(71). For each of these priors, Θ can be analytically margined out via (2), namely

$$p(Y \mid T) = \int p(Y \mid \Theta, T) p(\Theta \mid T) d\Theta, \tag{72}$$

where Y here denotes the observed values of y.

For regression trees with the mean-shift model normal model (69), perhaps the simplest prior specification for $p(\Theta \mid T)$ is the standard normal-inverse-gamma form where $\mu_1, \ldots, \mu_b$ are iid given σ and T with

$$p(\mu_i \mid \sigma, T) = N(\bar{\mu}, \sigma^2/a) \tag{73}$$

and

$$p(\sigma^2 \mid T) = p(\sigma^2) = IG(\nu/2, \nu\lambda/2). \tag{74}$$

Under this prior, standard analytical simplification yields

$$p(Y \mid T) \propto \frac{c \, a^{b/2}}{\prod_{i=1}^{b} (n_i + a)^{1/2}} \left(\sum_{i=1}^{b} (s_i + t_i) + \nu\lambda \right)^{-(n+\nu)/2} \tag{75}$$

where c is a constant which does not depend on T, s_i is $(n_i - 1)$ times the sample variance of the ith terminal node Y values, $t_i = \frac{n_i a}{n_i + a} (\bar{y}_i - \bar{\mu})^2$, and $\bar{y}_i$ is the sample mean of the ith terminal node Y values.

29

In practice, the observed Y can be used to guide the choice of hyperparameter values for $(\nu, \lambda, \bar{\mu}, a)$. Considerations similar to those discussed for Bayesian variable selection in Section 3.2 are also useful here. To begin with, because the mean-shift model attempts to explain the variation of Y, it is reasonable to expect that σ^2 will be smaller than s_Y^2, the sample variance of Y. Similarly, it is reasonable to expect that σ^2 will be larger than a pooled variance estimate obtained from a deliberate overfitting of the data by a greedy algorithm, say s_G^2. Using these values as guides, ν and λ would then be chosen so that the prior for σ^2 assigns substantial probability to the interval (s_G^2, s_Y^2). Once ν and λ have been chosen, $\bar{\mu}$ and a would be selected so that the prior for μ is spread out over the range of Y values.

For the more flexible mean-variance shift model (70) where σ_i can also vary across the terminal nodes, the normal-inverse-gamma form is easily extended to

$$p(\mu_i \mid \sigma_i, T) = N(\bar{\mu}, \sigma_i^2/a) \tag{76}$$

and

$$p(\sigma_i^2 \mid T) = p(\sigma_i^2) = IG(\nu/2, \nu\lambda/2), \tag{77}$$

with the pairs $(\mu_1, \sigma_1), \ldots, (\mu_b, \sigma_b)$ independently distributed given T. Under this prior, analytical simplification is still straightforward, and yields

$$p(Y \mid T) \propto \prod_{i=1}^{b} \pi^{-n_i/2}(\lambda\nu)^{\nu/2} \frac{\sqrt{a}}{\sqrt{n_i + a}} \frac{\Gamma((n_i + \nu)/2)}{\Gamma(\nu/2)}(s_i + t_i + \nu\lambda)^{-(n_i+\nu)/2} \tag{78}$$

where s_i and t_i are as above. Interestingly, the MCMC computations discussed in the next section are facilitated by the factorization of this marginal likelihood across nodes, in contrast to the marginal likelihood (75) for the equal variance model.

Here too, the observed Y can be used to guide the choice of hyperparameter values for $(\nu, \lambda, \bar{\mu}, a)$. The same ideas above may be used with an additional consideration. In some cases, the mean-variance shift model may explain variance shifts much more so than mean shifts. To handle this possibility, it may be better to choose ν and λ so that σ_Y^2 is more toward the center rather than the right tail of the prior for σ^2. We might also tighten up our prior for μ about the average y value. In any case, it can be useful to explore the consequences of several different prior choices.

For classification trees with the simple multinomial model (71), a useful conjugate prior specification for $\Theta = (p_1, \ldots, p_b)$ is the standard Dirichlet distribution of dimension $K - 1$ with parameter $\alpha = (\alpha_1, \ldots, \alpha_K)$, $\alpha_k > 0$, where $p_1, \ldots, p_b$ are iid given T with

$$p(p_i \mid T) = \text{Dirichlet}(p_i \mid \alpha) \propto p_{i1}^{\alpha_1 - 1} \cdots p_{iK}^{\alpha_K - 1}. \tag{79}$$

When $K = 2$ this reduces to the familiar Beta prior. Under this prior, standard analytical simplification yields

$$p(Y \mid T) \propto \left(\frac{\Gamma\left(\sum_k \alpha_k\right)}{\prod_k \Gamma(\alpha_k)} \right)^b \prod_{i=1}^{b} \frac{\prod_k \Gamma(n_{ik} + \alpha_k)}{\Gamma\left(n_i + \sum_k \alpha_k\right)} \tag{80}$$

where n_{ik} is the number of ith terminal node Y values in category C_k, $n_i = \sum_k n_{ik}$ and $k = 1, ..., K$ over the sums and products above. For a given tree, $p(Y \mid T)$ will be larger when

nodes are assigned more homogeneous values of y. To see this, note that assignments for which the Y values at the same node are similar will lead to more disparate values of $n_{i1}, \ldots, n_{iK}$, which in turn will lead to larger values of $p(Y \mid T)$.

The natural default choice for α is the vector $(1, \ldots, 1)$ for which the Dirichlet prior (79) is the uniform. However, by setting certain α_k to be larger for certain categories, $p(Y \mid T)$ will become more sensitive to misclassification at those categories. This would be desirable when classification into those categories is most important.

One detail of analytical simplifications yielding integrated likelihoods (75), (78) or (80) merits attention. Independence of parameters across terminal nodes means that integration can be carried out separately for each node. Normalizing constants in integrals for each node that would usually be discarded (for example $a^{b/2}$ in (75)) need to be kept, since the number of terminal nodes, b, varies across trees. This means that these normalizing constants will be exponentiated to a different power for trees of different size.

All the prior specifications above assume that given the tree T, the parameters in the terminal nodes are independent. When terminal nodes share many common parents, it may be desirable to introduce dependence between their θ_i values. Chipman, George, and McCulloch (2000) introduce such a dependence for the regression tree model, resulting in a Bayesian analogue of the tree shrinkage methods considered by Hastie and Pregibon (1990) and Leblanc and Tibshirani (1998).

4.2 Stochastic Search of the CART Model Posterior

Combining any of the closed form expressions (75), (78) or (80) for $p(Y \mid T)$ with $p(T)$ yields a closed form expression $g(T)$ satisfying

$$g(T) \propto p(Y \mid T)p(T) \propto p(T \mid Y). \tag{81}$$

Analogous to benefits of the availability $g(\gamma)$ in (45) for Bayesian variable selection, the availability of $g(T)$ confers great advantages for posterior computation and exploration in Bayesian CART model selection.

Exhaustive evaluation of $g(T)$ over all T will not be feasible, except in trivially small problems, because of the sheer number of possible trees. This not only prevents exact calculation of the norming constant, but also makes it nearly impossible to determine exactly which trees have largest posterior probability. In spite of these limitations, MH algorithms can still be used to explore the posterior. Such algorithms simulate a Markov chain sequence of trees

$$T^0, T^1, T^2, \ldots \tag{82}$$

which are converging in distribution to the posterior $p(T|Y)$. Because such a simulated sequence will tend to gravitate towards regions of higher posterior probability, the simulation can be used to stochastically search for high posterior probability trees. We now proceed to describe the details of such algorithms and their effective implementation.

4.2.1 Metropolis-Hastings Search Algorithms

By restricting attention to a finite set of CART trees T, as discussed in the last paragraph of Section 4.1.1, the simple MH form described in Section 2.3 can be used for direct simulation of the Markov chain (82). Because $g(T)/g(T') = p(T \mid Y)/p(T' \mid Y)$, such MH algorithms are obtained as follows. Starting with an initial tree T^0, iteratively simulate the transitions from T^j to T^{j+1} by the two steps:

1. Simulate a candidate T^* from the transition kernel $q(T \mid T^j)$.

2. Set $T^{j+1} = T^*$ with probability

$$\alpha(T^* \mid T^j) = \min \left\{ \frac{q(T^j \mid T^*)}{q(T^* \mid T^j)} \frac{g(T^*)}{g(T^j)}, 1 \right\}. \tag{83}$$

Otherwise, set $T^{j+1} = T^j$.

The key to making this an effective MH algorithm is the choice of transition kernel $q(T \mid T^j)$. A useful strategy in this regard is to construct $q(T \mid T^j)$ as a mixture of simple local moves from one tree to another - moves that have a chance of increasing posterior probability. In particular, CGM use the following $q(T \mid T^j)$, which generates T from T^j by randomly choosing among four steps:

- GROW: Randomly pick a terminal node. Split it into two new ones by randomly assigning it a splitting rule using the same random splitting rule assignment used to determine $p(T)$.

- PRUNE: Randomly pick a parent of two terminal nodes and turn it into a terminal node by collapsing the nodes below it.

- CHANGE: Randomly pick an internal node, and randomly reassign it using the same random splitting rule assignment used to determine $p(T)$.

- SWAP: Randomly pick a parent-child pair which are both internal nodes. Swap their splitting rules unless the other child has the identical rule. In that case, swap the splitting rule of the parent with that of both children.

In executing the GROW, CHANGE and SWAP steps, attention is restricted to splitting rule assignments that do not force the tree have an empty terminal node. CGM also recommend further restricting attention to splitting rule assignments which yield trees with at least a small number (such as five) observations at every terminal node. A similar $q(T \mid T^j)$, without the SWAP step, was proposed by DMS. An interesting general approach for constructing such moves was proposed by Knight, Kustra and Tibshirani (1998).

The transition kernel $q(T \mid T^j)$ above has some appealing features. To begin with, every step from T to T^* has a counterpart that moves from T^* to T. Indeed, the GROW and PRUNE steps are counterparts of one another, and the CHANGE and SWAP steps are their own counterparts. This feature guarantees the irreducibility of the algorithm, which is needed for convergence. It also makes it easy to calculate the ratio $q(T^j \mid T^*)/q(T^* \mid T^j)$ in (83). Note that other reversible moves may be much more difficult to implement because their counterparts are impractical to construct. For example, pruning off more than a pair of terminal nodes would require a complicated and computationally expensive reverse step. Another computational feature occurs

in the GROW and PRUNE steps, where there is substantial cancelation between g and q in the calculation of (83) because the splitting rule assignment for the prior is used.

4.2.2 Running the MH Algorithm for Stochastic Search

The MH algorithm described in the previous section can be used to search for desirable trees. To perform an effective search it is necessary to understand its behavior as it moves through the space of trees. By virtue of the fact that its limiting distribution is $p(T \mid Y)$, it will spend more time visiting tree regions where $p(T \mid Y)$ is large. However, our experience in assorted problems (see the examples in CGM) has been that the algorithm quickly gravitates towards such regions and then stabilizes, moving locally in that region for a long time. Evidently, this is a consequence of a transition kernel that makes local moves over a sharply peaked multimodal posterior. Once a tree has reasonable fit, the chain is unlikely to move away from a sharp local mode by small steps. Because the algorithm is convergent, we know it will eventually move from mode to mode and traverse the entire space of trees. However, the long waiting times between such moves and the large size of the space of trees make it impractical to search effectively with long runs of the algorithm. Although different move types might be implemented, we believe that any MH algorithm for CART models will have difficulty moving between local modes.

To avoid wasting time waiting for mode to mode moves, our search strategy has been to repeatedly restart the algorithm. At each restart, the algorithm tends to move quickly in a direction of higher posterior probability and eventually stabilize around a local mode. At that point the algorithm ceases to provide new information, and so we intervene in order to find another local mode more quickly. Although the algorithm can be restarted from any particular tree, we have found it very productive to repeatedly restart at the trivial single node tree. Such restarts have led to a wide variety of different trees, apparently due to large initial variation of the algorithm. However, we have also found it productive to restart the algorithm at other trees such as previously visited intermediate trees or trees found by other heuristic methods. For example, CGM demonstrate that restarting the algorithm at trees found by bootstrap bumping (Tibshirani and Knight 1996) leads to further improvements over the start points.

A practical implication of restarting the chain is that the number of restarts must be traded off against the length of the chains. Longer chains may more thoroughly explore a local region of the model space, while more restarts could cover the space of models more completely. In our experience, a preliminary run with a small number of restarts can aid in deciding these two parameters of the run. If the marginal likelihood stops increasing before the end of each run, lengthening runs may be less profitable than increasing the number of restarts.

It may also be useful to consider the slower "burn in" modification of the algorithm proposed by DMS. Rather than let their MH algorithm move quickly to a mode, DMS intervene, forcing the algorithm to move around small trees with around 6 or fewer nodes, before letting it move on. This interesting strategy can take advantage of the fact that the problems caused by the sharply peaked multimodal posterior are less acute when small trees are constructed. Indeed, when trees remain small, the change or swap steps are more likely to be permissible (since there are fewer children to be incompatible with), and help move around the model space. Although this "burn in" strategy will slow down the algorithm, it may be a worthwhile tradeoff if it

sufficiently increases the probability of finding better models.

4.2.3 Selecting the "Best" Trees

As many trees are visited by each run of the algorithm, a method is needed to identify those trees which are of most interest. Because $g(T) \propto p(T \mid Y)$ is available for each visited tree, one might consider selecting those trees with largest posterior probability. However, this can be problematic because of the dilution property of $p(T)$ discussed in Section 4.1.1. Consider the following simple example. Suppose we were considering all possible trees with two terminal nodes and a single rule. Suppose further that we had only two possible predictors, a binary variable with a single available splitting rule, and a multilevel variable with 100 possible splits. If the marginal likelihood $p(Y \mid T)$ was the same for all 101 rules, then the posterior would have a sharp mode on the binary variable because the prior assigns small probability to each of the 100 candidate splits for the multilevel predictor, and much larger probability to the single rule on the binary predictor. Selection via posterior probabilities is problematic because the relative sizes of posterior modes does not capture the fact that the total posterior probability allocated to the 100 trees splitting on the multilevel variable is the same as that allocated to the single binary tree.

It should be emphasized that the dilution property is not a failure of the prior. By using it, the posterior properly allocates high probability to tree neighborhoods which are collectively supported by the data. This serves to guide the algorithm towards such regions. The difficulty is that relative sizes of posterior modes do not capture the relative allocation of probability to such regions, and so can lead to misleading comparisons of single trees. Note also that dilution is not a limitation for model averaging. Indeed, one could approximate the overall posterior mean by the average of the visited trees using weights proportional to $p(Y \mid T)p(T)$. Such model averages provide a Bayesian alternative to the tree model averaging proposed by see Breiman (1996) and Oliver and Hand (1995).

A natural criterion for tree model selection, which avoids the difficulties described above, is to use the marginal likelihood $p(Y \mid T)$. As illustrated in CGM, a useful tool in this regard is a plot of the largest observed values of $p(Y \mid T)$ against the number of terminal nodes of T, an analogue of the C_p plot (Mallows 1973). This allows the user to directly gauge the value of adding additional nodes while removing the influence of $p(T)$. In the same spirit, we have also found it useful to consider other commonly used tree selection criteria such as residual sums of squares for regression trees and misclassification rates for classification trees.

After choosing a selection criterion, a remaining issue is what to do when many different models are found, all of which fit the data well. Indeed, our experience with stochastic search in applications has been to find a large number of good tree models, distinguished only by small differences in marginal likelihood. To deal with such output, in Chipman, George and McCulloch (1998b, 2001a), we have proposed clustering methods for organizing such multiple model output. We found such clustering to reveal a few distinct neighborhoods of similar models. In such cases, it may be better to select a few representative models rather than a single "best" model.

5 Much More to Come

Because of its broad generality, the formulation for Bayesian model uncertainty can be applied to a wide variety of problems. The two examples that we have discussed at length, Bayesian variable selection for the linear model and Bayesian CART model selection, illustrate some of the main ideas that have been used to obtain effective practical implementations. However, there have been many other recent examples. To get an idea of the extent of recent activity, consider the following partial list of some of the highlights just within the regression framework.

To begin with, the Bayesian variable selection formulation for the linear model has been extended to the multivariate regression setting by Brown, Vannucci and Fearn (1998). It has been applied and extended to nonparametric spline regression by Denison, Mallick and Smith (1998bc), Gustafson (2000), Holmes and Mallick (2001), Liang, Truong and Wong (2000), Smith and Kohn (1996, 1997), Smith, Wong and Kohn (1998); and to nonparametric wavelet regression by Abramovich, Sapatinas and Silverman (1998), Chipman, Kolaczyk and McCulloch (1997), Clyde and George (1999,2000), Clyde, Parmigiani and Vidakovic (1998), Holmes and Mallick (2000) and Johnstone and Silverman (1998). Related Bayesian approaches for generalized linear models and time series models have been put forward by Chen, Ibrahim and Yiannoutsos (1999), Clyde (1999), George, McCulloch and Tsay (1995), Ibrahim and Chen (1999), Mallick and Gelfand (1994), Raftery (1996), Raftery, Madigan and Volinsky (1996), Raftery and Richardson (1996), Shively, Kohn and Wood (1999), Troughton and Godsill (1997) and Wood and Kohn (1998); for loglinear models by Dellaportas and Foster (1999) and Albert (1996); and to graphical model selection by Giuduci and Green (1999) and Madigan and York (1995). Bayesian CART has been extended to Bayesian treed modeling by Chipman, George and McCulloch (2001); an related Bayesian partitioning approach has been proposed by Holmes, Denison and Mallick (2000). Alternative recent Bayesian methods for the regression setup include the predictive criteria of Laud and Ibrahim (1995), the information distance approach of Goutis and Robert (1998) and the utility approach of Brown, Fearn and Vannucci (1999) based on the early work of Lindley (1968). An excellent summary of many of the above articles and additional contributions to Bayesian model selection can be found in Ntzoufras (1999).

Spurred on by applications to new model classes, refinements in prior formulations and advances in computing methods and technology, implementations of Bayesian approaches to model uncertainty are widening in scope and becoming increasingly prevalent. With the involvement of a growing number of researchers, the cross-fertilization of ideas is further accelerating developments. As we see it, there is much more to come.

References

Abramovich, F., Sapatinas, T., and Silverman, B.W. (1998). Wavelet thresholding via a Bayesian approach. *Journal of the Royal Statistics Society, Series B*, **60**, 725–749.

Albert, J.H. (1996). The Bayesian Selection of Log-linear Models. *Canadian Journal of Statistics*, 24, 327–347.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle.

In *2nd International Symposium on Information Theory*, Ed. B.N. Petrov and F. Csaki, pp. 267-81. Budapest: Akademia Kiado.

Bartlett, M. (1957). A comment on D. V. Lindley's statistical paradox. *Biometrika*, 44, 533-534.

Bernardo, J. M., and Smith, A.F.M. (1994). *Bayesian Theory*, Wiley, New York.

Besag, J and Green, P.J. (1993). Spatial statistics and Bayesian computation (with discussion) *Journal of the Royal Statistical Society, Series B*, 55 25–37.

Berger, J.O. and Pericchi, L.R. (1996). The intrinsic Bayes factor for model selection and prediction. *J. Amer. Statist. Assoc.*, 91, 109-122.

Breiman, L., Friedman, J. Olshen, R. and Stone, C. (1984). *Classification and Regression Trees*, Wadsworth.

Breiman, L (1996). Bagging predictors, *Machine Learning*, 24, 123–140.

Brown, P.J., Vannucci,M., and Fearn, T. (1998). Multivariate Bayesian variable selection and prediction. *Journal of the Royal Statistical Society, Ser. B*, 60, 627-642.

Brown, P.J., Fearn, T. and Vannucci, M. (1999). The choice of variables in multivariate regression: a non-conjugate Bayesian decision theory approach. *Biometrika*, **86**, 635–648.

Buntine, W. (1992). Learning Classification Trees. *Statistics and Computing*, 2, 63–73.

Carlin, B.P. and Chib, S. (1995). Bayesian Model Choice via Markov Chain Monte Carlo, *Journal of the Royal Statistical Society, Series B*, 77, 473-484.

Casella, G. and George, E.I. (1992). Explaining the Gibbs Sampler, *The American Statistician*, 46, 167-174.

Chambers, J.M. (1971). Regression Updating, *Journal of the American Statistical Association*, 66, 744-748.

Chen, M.H., Ibrahim, J.G. and Yiannoutsos, C. (1999). Prior Elicitation, Variable Selection and Bayesian Computation for Logistic Regression Models. *Journal of the Royal Statistical Society B*, 61, 223–243.

Chib, S., and Greenberg, E. (1995). Understanding the Metropolis-Hastings Algorithm, *The American Statistician*, 49, 327-335.

Chipman, H. (1996). Bayesian Variable Selection with Related Predictors, *The Canadian Journal of Statistics* 24, 17–36.

Chipman, H. A., George, E. I., and McCulloch, R. E. (1998a) Bayesian CART Model Search (with discussion). *Journal of the American Statistical Association*, 93, 935–960.

Chipman, H. A., George, E. I. and McCulloch, R. E. (1998b) Making Sense of a Forest of Trees, *Proceedings of the 30th Symposium on the Interface*, S. Weisberg, Ed., Interface Foundation of North America.

Chipman, H., George, E. I., and McCulloch, R. E. (2000) Hierarchical Priors for Bayesian CART Shrinkage, *Statistics and Computing*, 10, 17–24.

Chipman, H., George, E.I. and McCulloch, R.E. (2001a) Managing Multiple Models, *Artificial Intelligence and Statistics 2001*, (eds. Tommi Jaakkola, Thomas Richardson). pp 11–18, Morgan Kaufmann, San Francisco, CA.

Chipman, H., George, E. I., and McCulloch, R. E. (2001b) Bayesian Treed Models, *Machine Learning* (to appear).

Chipman, H., Hamada, M. and Wu, C. F. J., (1997) A Bayesian Variable Selection Approach for Analyzing Designed Experiments with Complex Aliasing, *Technometrics*, 39, 372–381.

Chipman, H., Kolaczyk, E., and McCulloch, R. (1997). Adaptive Bayesian wavelet shrinkage. *Journal of the American Statistical Association*, **92**, 1413–1421.

Clark, L., and Pregibon, D. (1992). Tree-Based Models. In *Statistical models in S*, J. Chambers and T. Hastie, Eds., Wadsworth.

Clyde, M.A. (1999). Bayesian Model Averaging and Model Search Strategies (with discussion). In *Bayesian Statistics 6* (J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith, eds.) 157–185, Oxford University Press.

Clyde, M.A., DeSimone, H., and Parmigiani, G. (1996). Prediction via orthogonalized model mixing, *Journal of the American Statistical Association*, 91 1197-1208.

Clyde, M. and George, E.I. (1999) Empirical Bayes estimation in wavelet nonparametric Regression. In *Bayesian Inference in Wavelet Based Models*, Ed. P. Muller and B. Vidakovic, pp. 309-22. New York: Springer-Verlag.

Clyde, M. and George, E.I. (2000). Flexible empirical Bayes estimation for wavelets. *J. R. Statist. Soc.* B **62**, 681-698.

Clyde, M.A., and Parmigiani, G. (1994). Bayesian Variable Selection and Prediction with Mixtures, *Journal of Biopharmaceutical Statistics*.

Clyde, M., Parmigiani, G., Vidakovic, B. (1998). Multiple shrinkage and subset selection in wavelets. *Biometrika*, **85**, 391–402.

Denison, D., Mallick, B. and Smith, A.F.M. (1998a) A Bayesian CART Algorithm, *Biometrika*, 85, 363-377.

Denison, D.G.T., Mallick, B.K. and Smith, A.F.M. (1998b) Automatic Bayesian Curve Fitting, *J. Roy. Statst. Soc. B* 60 333–350.

Denison, D. G. T., Mallick, B. K. and Smith, A. F. M. (1998c). Bayesian MARS, *Statistics and Computing*, 8, 337–346.

Dellaportas, P. and Foster, J.J. (1999) Markov Chain Monte Carlo Model Determination for Hierarchical and Graphical Log-Linear Models, *Biometrika*, **86**, 615–634.

Dellaportas, P. Forster, J.J. and Ntzoufras, I. (2000) On Bayesian Model and Variable Selection Using MCMC. *Statistics and Computing* (to appear).

Donoho, D.L., & Johnstone, I.M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika* **81**, 425-56.

Draper, D. (1995) Assessment and Propagation of Model Uncertainty (with discusssion). *J. R. Statist. Soc.* B **57**, 45–98.

Dongarra, J.J., Moler C.B., Bunch, J.R. and Stewart, G.W. (1979). Linpack Users' Guide, Philadelphia: Siam.

Edwards, W Lindman H. and Savage, L.J. (1963). Bayesian Statistical Inference for Psychological Research, *Psychological Review*, 70 193-242.

Fernandez, C., Ley, E. and Steel, M.F.J. (2001). Benchmark Priors for Bayesian Model Averaging, *Journal of Econometrics*, 100, 381-427.

Foster, D.P., & George, E.I. (1994). The risk inflation criterion for multiple regression. *Ann. Statist.* **22**, 1947-75.

Garthwaite,P. H. and Dickey, J.M. (1992) Elicitation of Prior Distributions for Variable-Selection Problems in Regression, *Annals of Statistics* 20 4 1697-1719.

Garthwaite, P. H. & Dickey, J.M. (1996). Quantifying and using expert opinion for variable-selection problems in regression (with discussion). *Chemomet. Intel. Lab. Syst.* **35**, 1-34.

Gelfand, A. E., Dey, D. K., and Chang, H. (1992). Model Determination Using Predictive Distributions With Implementations via Sampling-Based Methods, in *Bayesian Statistics 4*, eds. J. M. Bernardo et al., Oxford U.K.: Oxford University Press, pp. 147–167.

Gelfand, A., and Smith, A. F. M. (1990). Sampling-Based Approaches to Calculating Marginal Densities, *Journal of the American Statistical Association*, 85, 398–409.

George, E.I. (1998). Bayesian Model Selection, *Encyclopedia of Statistical Sciences, Update Volume 3*, (eds. S. Kotz, C. Read and D. Banks). pp 39-46, Wiley, N.Y.

George, E.I. (1999). Discussion of Bayesian Model Averaging and Model Search Strategies by M.A. Clyde. *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. (2000) Calibration and Empirical Bayes Variable Selection. *Biometrika* 87 731-748.

George, E.I. and McCulloch, R.E. (1993). Variable Selection via Gibbs Sampling, *Journal of the American Statistical Society*, 88, 881-889.

George, E.I. and McCulloch, R.E. (1996). Stochastic Search Variable Selection, In *Markov Chain Monte Carlo in Practice*, (eds W.R. Gilks, S. Richardson and D.J. Spiegelhalter), pp. 203-214. London: Chapman & Hall.

George, E.I., & McCulloch, R.E. (1997). Approaches for Bayesian variable selection. *Statist. Sinica* **7**, 339-73.

George, E.I., McCulloch, R.E. and Tsay, R. (1995) Two approaches to Bayesian Model selection with applications. In *Bayesian Statistics and Econometrics: Essays in Honor of Arnold Zellner*, D. Berry, K. Chaloner, and J. Geweke (eds.), Wiley, New York, 339-348.

Geisser, S. (1993). *Predictive Inference: an Introduction.* London: Chapman & Hall.

Geman, S. and Geman, D. (1984) Stochastic Relaxation, Gibbs Distributions and the Bayesian Restoration of Images, *IEEE Trans. Pattn. Anal. Mach. Intell.* 6, 721–741.

Geweke, J. (1996). Variable Selection and Model Comparison in Regression, em Bayesian Statistics 5, ed. J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith, Oxford Press, 609-620.

Gilks, W.R., Richardson S. and Spiegelhalter, D.J. (1996) *Markov Chain Monte Carlo in Practice*, London: Chapman & Hall.

Giuduci, P. and Green, P.J. (1999) Decomposable graphical Gausssian Model Determination, *Biometrika*, **86**, 785–802.

Goutis, C. and Robert, C.P. (1998). Model Choice in Generalized Linear Models: A Bayesian Approach via Kullback-Leibler Projections. *Biometrika*, 82, 711–732.

Green, P. (1995). Reversible Jump Markov chain Monte Carlo computation and Bayesian model determination, *Biometrika*, 82, 711-732.

Gustafson, P. (2000). Bayesian Regression Modelling with Interactions and Smooth Effects. *Journal of the American Statistical Association* 95, 795–806.

Hastie, T., and Pregibon, D. (1990). Shrinking Trees, AT&T Bell Laboratories Technical Report.

Hastings, W.K. (1970) Monte Carlo Sampling Methods using Markov Chain and their Applications. *Biometrika*, 57, 97–109.

Hoeting, J. A., Raftery, A.E. and Madigan, D. (1996). A Method for Simultaneous Variable Selection and Outlier Identification in Linear Regression, *Computational Statistics and Data Analysis.*

Hoeting, J. A., Madigan, D., Raftery, A.E., and Volinsky, C.T. (1999) Bayesian Model Averaging: A Tutorial (with discussion). *Statistical Science*, **14**:4, 382–417. (Corrected version available at http://www.stat.washington.edu/www/research/ online/hoeting1999.pdf).

Holmes C.C., Denison, D.G.T. and Mallick, B.K. (2000). Bayesian prediction via partitioning, Technical Report, Dept of Mathematics, Imperial College, London.

Holmes, C.C. and Mallick, B.K. (2000). Bayesian wavelet networks for nonparametric regression, *IEEE Trans. Neur. Netwrks.* 10 1217–1233.

Holmes, C.C. and Mallick, B.K. (2001). Bayesian Regression with Mulivariate Linear Splines, *J. Roy. Statst. Soc. B* 63 3–18.

Ibrahim, J.G. and Chen, M.H. (1999). Prior Elicitation and Variable Selection for Generalized Mixed Models. *Generalized Linear Models: A Bayesian Perspective* (D. K. Dey, S. Ghosh, and B. Mallick, eds.). New York: Marcel Dekker, (to appear).

Johnstone, I.M. & Silverman, B.W. (1998). Empirical Bayes approaches to mixture problems and wavelet regression. Tech Report. University of Bristol.

Jordan, M.I. and Jacobs, R.A. (1994). Mixtures of Experts and the EM Algorithm, *Neural Computation*, 6, 181-214.

Key, J.T., Pericchi, L.R. and Smith, A.F.M. (1998). Choosing among models when none of them are true. In *Proceedings of the Workshop on Model Selection, Special Issue of Rassegna di Metodi Statistici ed Applicazioni*, Ed. W. Racugno, Bologna: Pitagora Editrice.

Kass, R. E. and Wasserman, L. (1995). A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *J. Am. Statist. Assoc.* **90**, 928-34.

Leblanc, M. and Tibshirani, R. (1998). Monotone Shrinkage of Trees, *Journal of Computational and Graphical Statistics*, 7, 417–433.

Laud, P.W. and Ibrahim, J.G. (1995). Predictive Model Selection. *Journal of the Royal Statistical Society B*, 57, 247–262.

Liang F., Truong, Y.K. and Wong, W.H. (2000). Automatic Bayesian Model Averaging for Linear Regression and Applications in Bayesian Curve Fitting. Technical Report, Dept of Statistics, UCLA.

Liu, J.S., Wong, W.H., and Kong, A. (1994). Covariance Structure and Convergence Rate of the Gibbs Sampler with applications to the comparisons of estimators and augmentation schemes, *Biometrika*, 81, 27-40.

Lindley, D.V. (1968). The Choice of Variables in Regression. *Journal of the Royal Statistical Society B*, 30, 31–66.

Lutsko, J. F. and Kuijpers, B. (1994). Simulated Annealing in the Construction of Near-Optimal Decision Trees, in *Selecting Models from Data: AI and Statistics IV*, P. Cheeseman and R. W. Oldford, Eds., p 453–462.

Knight, K., Kustra, R. and Tibshirani, R. (1998). Discussion of "Bayesian CART Model Search by Chipman, H. A., George, E. I., and McCulloch, R. E." *Journal of the American Statistical Association*, 93, 950–957.

Kuo, L. and Mallick, B. (1998). Variable Selection for Regression Models. *Sankhya B*, 60, 65–81.

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

Mallick, B.K. and Gelfand, A.E. (1994). Generalized Linear Models with Unknown Number of Components. *Biometrika*, 81, 237–245.

Mallows, C. L. (1973), Some Comments on C_p. *Technometrics*, 15, 661–676.

McCullagh, P. and Nelder, J.A. (1989). *Generalized Liner Models, 2nd Edition*, Chapman and Hall, N.Y.

McCulloch, R.E. and Rossi P. (1991). A Bayesian Approach to Testing the Arbitrage Pricing Theory,
em Journal of Econometrics, 49, 141–168.

Meyn, S.P. and Tweedie, R.L. (1993). *Markov Chains and Stochastic Stability*, Springer-Verlag, New York.

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H. Teller,E. (1953) Equations of State Calculations by Fast Computing Machines. *J. Chem. Phys.*, 21 1087–1091.

Mitchell, T.J. and Beauchamp, J.J. (1988). Bayesian variable selection in linear regression (with discussion), *Journal of the American Statistical Association*, 83, 1023-1036.

Ntzoufras, I. (1999). Aspects of Bayesian Model and Variable Selection Using MCMC, Ph.D. Dissertation, Dept. of Statistics, Athens University of Economics and Business, Athens, Greece.

Oliver, J.J. and Hand, D.J. (1995). On Pruning and Averaging Decision Trees, *Proceedings of the International Machine Learning Conference*, 430–437.

O'Hagan, A. (1995). Fractional Bayes factors for model comparison (with discussion). *J. R. Statist. Soc.* B **57**, 99-138. Proceedings of the International Machine Learning Conference, 430–437.

Paass, G. and Kindermann, J. (1997). Describing the Uncertainty of Bayesian Predictions by Using Ensembles of Models and its Application, *1997 Real World Computing Symposium*, Real World Computing Partnership, Tsukuba Research Center, Tsukuba, Japan, p. 118–125.

Petit, L.I., and Smith, A. F. M. (1985). Outliers and influential observations in linear models, In *Bayesian Statisticcs 2*, (J.M. Bernardo, M.H. DeGroot, D.V. Lindley and A.F.M. Smith, eds). Amsterdam: North-Holland, 473-494.

Phillips, D. B., and Smith, A. F. M. (1995). Bayesian Model Comparison Via Jump Diffusions, In *Practical Markov Chain Monte Carlo in Practice*, (eds W.R. Gilks, S. Richardson and D.J. Spiegelhalter), pp. 215-239. London: Chapman & Hall.

Peng, L. (1998). Normalizing Constant Estimation for Discrete Distribution Simulation, Ph.D. Dissertation, Dept MSIS, University of Texas, Austin, TX.

Quinlan, J.R. and Rivest, R.L. (1989). Inferring decision trees using the minimum description length principle, *Information and Computation*, 80:227-248.

Raftery, A.E. (1996). Approximate Bayes Factors and Accounting for Model Uncertainty in Generalized Linear Models. *Biometrika*, 83, 251–266.

Raftery, A. E., Madigan, D. and Hoeting, J. A. (1997). Bayesian Model Averaging for Linear Regression Models. *Journal of the American Statistical Association*, 92, 179–191.

Raftery, A.E., Madigan, D. M., and Volinsky C.T. (1995). Accounting for Model Uncertainty in Survival Analysis Improves Predictive Performance (with discussion). In *Bayesian Statistics 5*, ed. J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith, Oxford Press, 323-350.

Raftery, A.E. and Richardson, S. (1996). Model Selection for Generalized Linear Models via GLIB: Application to Nutrition and Breast Cancer. *Bayesian Biostatistics* (D.A. Berry and D.K. Strangl, eds.). New York: Marcel Dekker, 321–353.

San Martini, A. and Spezzaferri, F. (1984). A predictive model selection criterion *J. Royal Statist. Soc., Ser. B*, 46 296-303.

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6**, 461-4.

Shively, T.S., Kohn, R. and Wood, S. (1999). Variable Selection and Function Estimation in Additive Nonparametric Regression Using a Data-Based Prior (with discussion), *Journal of the American Statistical Association*, **94**, 777–806.

Smith, A.F.M and Roberts, G.O. (1993). Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society, Series B*, 55 3-24.

Smith, M. and Kohn, R.(1996). Nonparametric regression using Bayesian variable selection, *Journal of Econometrics*, 75, 317-344.

Smith, M. and Kohn, R. (1997). A Bayesian Approach to Nonparametric Bivariate Regression, *Journal of the American Statistical Association*, **92**, 1522–1535.

Smith, M., Wong, C.M. and Kohn, R. (1998). Additive Nonparametric Regression with Auto-correlated Errors. *Journal of the Royal Statistical Society B*, 60, 311–331.

Sutton, C. (1991). Improving Classification Trees with Simulated Annealing, *Proceedings of the 23rd Symposium on the Interface*, E. Keramidas, Ed., Interface Foundation of North America.

Tibshirani, R. and Knight, K. (1999). Model Search by Bootstrap "Bumping" , *Journal of Compuatational and Graphical Statistics*, 8, 671–686.

Tierney, L. (1994). Markov Chains for Exploring Posterior Distributions (with discussion), *Annals of Statistics*, 22, 1674-1762.

Tierney, L. and Kadane, J.B. (1986). Accurate Approximations for Posterior Moments and Marginal Densities, *Journal of the American Statistical Association* 81 82-86.

Troughton, P.T. and Godsill, S.J. (1997). Bayesian Model Selection for time series using Markov Chain Monte Carlo. *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*.

Wakefield, J.C. and Bennett, J.E. (1996). The Bayesian modelling of covariates for population pharmacokinetic models *Journal of the American Statistical Society*, 91, 917-928.

Wallace, C.C. and Patrick, J.D. (1993). Coding decision trees, *Machine Learning*, 11, 7-22.

Wasserman, L. (2000). Bayesian Model Selection and Model Averaging. *Journal of Mathematical Psychology*, **44**, 92-107.

Wood, S. and Kohn, R. (1998). A Bayesian Approach to Robust Binary Nonparametric Regression, *Journal of the American Statistical Association*, **93**, 203–213.

Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finietti*, Ed. P.K. Goel and A. Zellner, pp 233-43. Amsterdam:North-Holland.