Model Selection
IMS Lecture Notes - Monograph Series (2001) Volume 38

The Practical Implementation of Bayesian Model Selection

Hugh Chipman, Edward I. George and Robert E. McCulloch

The University of Waterloo, The University of Pennsylvania
and *The University of Chicago*

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.

[0]Hugh Chipman is Associate Professor of Statistics, Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, ON N2L 3G1, Canada; email: hachipman@uwaterloo.ca. Edward I. George is Professor of Statistics, Department of Statistics, The Wharton School of the University of Pennsylvania, 3620 Locust Walk, Philadelphia, PA 19104-6302, U.S.A; email: edgeorge@wharton.upenn.edu. Robert E. McCulloch is Professor of Statistics, Graduate School of Business, University of Chicago, 1101 East 58th Street, Chicago, IL, U.S.A; email: Robert.McCulloch@gsb.uchicago.edu. This work was supported by NSF grant DMS-98.03756 and Texas ARP grant 003658.690.

Contents

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 then combining them with the distributions for the data. 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 $p(Y_f, \theta_k, M_k \mid Y)$. 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 $p(Y_f \mid Y)$, 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

$$p(M_k \mid Y) = \frac{p(Y \mid M_k)p(M_k)}{\sum_k p(Y \mid M_k)p(M_k)} \tag{2.1}$$

where

$$p(Y \mid M_k) = \int p(Y \mid \theta_k, M_k)p(\theta_k \mid M_k)d\theta_k \tag{2.2}$$

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

$$\frac{p(M_1 \mid Y)}{p(M_2 \mid Y)} = \frac{p(Y \mid M_1)}{p(Y \mid M_2)} \times \frac{p(M_1)}{p(M_2)}. \tag{2.3}$$

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 $p(M_1|Y), \ldots, p(M_K|Y)$ 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

$$\int u(\alpha, \Delta) p(\Delta \mid Y) d\Delta \tag{2.4}$$

where the predictive distribution of Δ given Y

$$p(\Delta \mid Y) = \sum_k p(\Delta \mid M_k, Y) p(M_k \mid Y) \tag{2.5}$$

is a posterior weighted mixture of the conditional predictive distributions.

$$p(\Delta \mid M_k, Y) = \int p(\Delta \mid \theta_k, M_k) p(\theta_k \mid M_k, Y) d\theta_k \tag{2.6}$$

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

$$\sum_{k'} p(M_{k'} \mid Y) \int p(\Delta \mid M_{k'}, Y) \log \frac{p(\Delta \mid M_{k'}, Y)}{p(\Delta \mid M_k, Y)} \tag{2.7}$$

(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 (2.5). Under squared error loss, the best prediction of Δ is the overall posterior mean

$$E(\Delta \mid Y) = \sum_k E(\Delta \mid M_k, Y) p(M_k \mid Y). \tag{2.8}$$

Such model averaging or mixing procedures incorporate model uncertainty and have been advocated by Geisser (1993), Draper (1995), Hoeting, Madigan, Raftery and 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 $p(\theta_k|M_k)$ 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 this specification 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

$$p(M_k) \equiv 1/K \tag{2.9}$$

which is noninformative in the sense of favoring all models equally. Under this prior, the model posterior is proportional to the marginal likelihood, $p(M_k|Y) \propto p(Y|M_k)$, and posterior odds comparisons in (2.3) reduce to Bayes factor comparisons. However, the apparent noninformativeness of (2.9) 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, (2.9) 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 (2.1) and (2.2) 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 (2.2) 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 $p(Y|M_k)$ 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 (2.2), 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 (2.1) and (2.2) 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 $p(\eta \mid Y)$. 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 $q(\eta \mid \eta^{(j)})$ 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, and 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 (2.2) 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 $h(\theta_k) \equiv \log p(Y \mid \theta_k, M_k)p(\theta_k \mid M_k)$ is sufficiently well-behaved, is obtained by Laplace's method (see Tierney and Kadane 1986) as

$$p(Y \mid M_k) \approx (2\pi)^{d_k/2} |H(\tilde{\theta}_k)|^{1/2} p(Y \mid \tilde{\theta}_k, M_k) p(\tilde{\theta}_k \mid M_k) \tag{2.10}$$

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

$$p(Y \mid M_k) \approx (2\pi)^{d_k/2} |H^*(\hat{\theta}_k)|^{1/2} p(Y \mid \hat{\theta}_k, M_k) p(\hat{\theta}_k \mid M_k) \tag{2.11}$$

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 (2.11) that are constant in large samples, yields the BIC approximation (Schwarz 1978)

$$\log p(Y \mid M) \approx \log p(Y \mid \hat{\theta}_k, M_k) - (d_k/2) \log n \tag{2.12}$$

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, (2.12) 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, (2.10), (2.11), (2.12), 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 $X_1, \ldots X_p$

$$Y \sim N_n(X\beta, \sigma^2 I) \tag{3.1}$$

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.

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

$$p(\gamma) = \prod w_i^{\gamma_i}(1 - w_i)^{1-\gamma_i}, \tag{3.2}$$

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 (3.2) has been to set $w_i \equiv w$, yielding

$$p(\gamma) = w^{q_\gamma}(1-w)^{p-q_\gamma}, \tag{3.3}$$

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

$$p(\gamma) \equiv 1/2^p, \tag{3.4}$$

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 $w \sim Beta(\alpha, \beta)$, (3.3) yields

$$p(\gamma) = \frac{B(\alpha + q_\gamma, \beta + p - q_\gamma)}{B(\alpha, \beta)} \tag{3.5}$$

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

$$p(\gamma) = \binom{p}{q_\gamma}^{-1} h(q_\gamma), \tag{3.6}$$

of which (3.5) is a special case. Under priors of the form (3.6), the components of γ are exchangeable but not independent, (except for the special case (3.3)).

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_1X_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:

$$p(\gamma) = p(\gamma_A)p(\gamma_B)p(\gamma_C)p(\gamma_{AB} \mid \gamma_A, \gamma_B)p(\gamma_{AC} \mid \gamma_A, \gamma_C)p(\gamma_{BC} \mid \gamma_B, \gamma_C). \tag{3.7}$$

The specification of terms like $p(\gamma_{AC}|\gamma_A, \gamma_C)$ in (3.7) would entail specifying four probabilities, one for each of the values of (γ_A, γ_C). Typically $p(\gamma_{AC}|0,0) < (p(\gamma_{AC}|1,0)$, // $p(\gamma_{AC} \mid 0,1)) < p(\gamma_{AC} \mid 1,1)$. 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 (3.2), priors for dependent component models, such as (3.7), 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_1 remains unchanged, whereas the probability allocated to X_2 and X_1, X_2 has 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 (3.4) where $p(\gamma|Y) \propto p(Y|\gamma)$. 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 (3.2), of which (3.4) 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 (2.8) 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 (3.1). In effect, the problem then becomes that of selecting a submodel of (3.1) of the form

$$p(Y \mid \beta_\gamma, \sigma^2, \gamma) = N_n(X_\gamma \beta_\gamma, \sigma^2 I) \tag{3.8}$$

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

$$p(\beta_\gamma \mid \sigma^2, \gamma) = N_{q_\gamma}(\bar{\beta}_\gamma, \sigma^2 \Sigma_\gamma), \tag{3.9}$$

coupled with an inverse gamma prior on σ^2

$$p(\sigma^2 \mid \gamma) = p(\sigma^2) = IG(\nu/2, \nu\lambda/2), \tag{3.10}$$

(which is equivalent to $\nu\lambda/\sigma^2 \sim \chi^2_\nu$). 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 (3.9), when coupled with $p(\gamma)$, implicitly assigns a point mass at zero for coefficients in (3.1) that are not contained in β_γ. As such, (3.9) 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 (3.8), 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 (3.9) and (3.10) is analytical tractability; the conditional distribution of β_γ and σ^2 given γ is conjugate for (3.8), 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

$$p(Y \mid \gamma) \propto |X'_\gamma X_\gamma + \Sigma_\gamma^{-1}|^{-1/2} |\Sigma_\gamma|^{-1/2} (\nu\lambda + S^2_\gamma)^{-(n+\nu)/2} \tag{3.11}$$

where

$$S^2_\gamma = Y'Y - Y'X_\gamma (X'_\gamma X_\gamma + \Sigma_\gamma^{-1})^{-1} X'_\gamma Y. \tag{3.12}$$

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 (3.9) for β_γ depends on σ^2, and this is needed to obtain conjugacy. Integrating out σ^2 with respect to (3.10), the prior for β_γ conditionally only on γ is

$$p(\beta_\gamma \mid \gamma) = T_{q_\gamma}(\nu, \bar{\beta}_\gamma, \lambda\Sigma_\gamma) \tag{3.13}$$

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

The priors (3.9) and (3.10) 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 (3.10) 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 (3.10) assigns substantial probability to the interval (s_{FULL}^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 (3.10) 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 (3.9) and so can be used formally.

Turning to (3.9), 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 (3.9) 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 (3.9) 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 (3.13) 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 $V_\gamma = (X_\gamma' X_\gamma)^{-1}$, $p(\sigma^2 \mid \gamma) \propto 1/\sigma^2$ 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 (3.9) with $\bar{\beta}_\gamma = 0$ and $V_\gamma = (X_\gamma' X_\gamma)^{-1}$, namely

$$p(\beta_\gamma \mid \sigma^2, \gamma) = N_{q_\gamma}(0, c\,\sigma^2 (X_\gamma' X_\gamma)^{-1}), \tag{3.14}$$

is combined with

$$p(\gamma) = w^{q_\gamma}(1-w)^{p-q_\gamma} \tag{3.15}$$

the simple independence prior in (3.3); 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

$$SS_\gamma/\sigma^2 - F\, q_\gamma \tag{3.16}$$

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 (3.16) 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 (3.16) is obtained by reexpressing the model posterior under (3.14) and (3.15) as

$$\begin{aligned} p(\gamma \mid Y) &\propto w^{q_\gamma} (1-w)^{p-q_\gamma} (1+c)^{-q_\gamma/2} \exp\left\{ -\frac{Y'Y - SS_\gamma}{2\sigma^2} - \frac{SS_\gamma}{2\sigma^2(1+c)} \right\} \\ &\propto \exp\left[\frac{c}{2(1+c)} \{ SS_\gamma/\sigma^2 - F(c,w)\, q_\gamma \} \right], \end{aligned} \tag{3.17}$$

where

$$F(c,w) = \frac{1+c}{c} \left\{ 2 \log \frac{1-w}{w} + \log(1+c) \right\}. \tag{3.18}$$

The expression (3.17) reveals that, as a function of γ for fixed Y, $p(\gamma \mid Y)$ is increasing in (3.16) when $F = F(c,w)$. Thus, both (3.16) and (3.17) 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 (3.16).

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 respectively, all with $w = 1/2$. 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 (3.16) 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 (3.16) 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 (3.14) and (3.15) will be consistent with

the true underlying model. For example, large c will be chosen when the regression 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 (3.14) and (3.15), namely

$$\begin{aligned} L(c, w \mid Y) &\propto \sum_{\gamma} p(\gamma \mid w)\, p(Y \mid \gamma, c) \\ &\propto \sum_{\gamma} w^{q_\gamma}(1-w)^{p-q_\gamma}(1+c)^{-q_\gamma/2} \exp\left\{\frac{c\, SS_\gamma}{2\sigma^2(1+c)}\right\}. \end{aligned} \tag{3.19}$$

Although this maximization is generally impractical when p is large, the likelihood (3.19) 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 $\hat{\beta} = (X'X)^{-1}X'Y$, (3.19) reduces to

$$L(c, w \mid Y) \propto \prod_{i=1}^{p} \left\{(1-w)e^{-t_i^2/2} + w(1+c)^{-1/2} e^{-t_i^2/2(1+c)}\right\}. \tag{3.20}$$

Since many fewer terms are involved in the product in (3.20) than in the sum in (3.19), maximization of (3.20) 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 (3.14) and (3.15). In particular, (3.17) reveals that the γ maximizing the posterior $p(\gamma \mid Y, \hat{c}, \hat{w})$ can be obtained as the γ that maximizes the marginal maximum likelihood criterion

$$C_{\text{MML}} = SS_\gamma/\hat{\sigma}^2 - F(\hat{c}, \hat{w})\, q_\gamma, \tag{3.21}$$

where $F(c, w)$ is given by (3.18). Because maximizing (3.19) 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

$$C_{\text{CML}} = SS_\gamma/\hat{\sigma}^2 - q_\gamma \left\{1 + \log_+(SS_\gamma/\hat{\sigma}^2 q_\gamma)\right\} - 2\left\{\log(p-q_\gamma)^{-(p-q_\gamma)} + \log q_\gamma^{-q_\gamma}\right\} \tag{3.22}$$

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 (3.16), 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 (3.21).

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 (3.16); 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 (3.9) 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 (3.9), which avoids this potential drawback, is the normal-normal formulation used in the stochastic search variable selection (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 (3.1) so that

$$p(Y \mid \beta, \sigma^2, \gamma) = N_n(X\beta, \sigma^2 I) \tag{3.23}$$

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

$$\pi(\beta \mid \sigma^2, \gamma) = \pi(\beta \mid \gamma) = N_p(0, D_\gamma R D_\gamma) \tag{3.24}$$

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

$$(D_\gamma)_{ii} = \begin{cases} \sqrt{\upsilon_{0i}} & \text{when} \quad \gamma_i = 0 \\ \sqrt{\upsilon_{1i}} & \text{when} \quad \gamma_i = 1 \end{cases} \tag{3.25}$$

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

$$p(\beta_i) = (1 - p(\gamma_i))N(0, \upsilon_{0i}) + p(\gamma_i)N(0, \upsilon_{1i}), \tag{3.26}$$

a scale mixture of two normal distributions.

Although β is independent of σ^2 in (3.24), the inverse Gamma prior (3.10) 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 (3.24), similarly to the commonly used choices for Σ_γ in (3.9).

To use this normal-normal setup for variable selection, the hyperparameters υ_{0i} and υ_{1i} are set "small and large" respectively, so that $N(0, \upsilon_{0i})$ is concentrated and $N(0, \upsilon_{1i})$ 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 υ_{0i} and υ_{1i} such that $p(\beta_i \mid \gamma_i = 0) = N(0, \upsilon_{0i}) > p(\beta_i \mid \gamma_i = 1) = N(0, \upsilon_{1i})$ precisely on the interval $(-\delta_i, \delta_i)$. This property is obtained by any υ_{0i} and υ_{1i} satisfying

$$\log(\upsilon_{1i}/\upsilon_{0i})/(\upsilon_{0i}^{-1} - \upsilon_{1i}^{-1}) = \delta_{i\gamma}^2 \tag{3.27}$$

By choosing υ_{1i} such that $N(0, \upsilon_{1i})$ is consistent with plausible values of β_i, υ_{0i} can then be chosen according to (3.27). George and McCulloch (1997) report that computational problems and difficulties with υ_{1i} too large will be avoided whenever $\upsilon_{1i}/\upsilon_{0i} \leq 10,000$, 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 (3.1) because (3.24) 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

$$p(\beta \mid \sigma^2, \gamma) = N_p(0, \sigma^2 D_\gamma R D_\gamma), \tag{3.28}$$

which is identical to (3.24) except for the insertion of σ^2. Coupled with the inverse Gamma prior (3.10) 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

$$p(Y \mid \gamma) \propto |X'X + (D_\gamma R D_\gamma)^{-1}|^{-1/2} |D_\gamma R D_\gamma|^{-1/2} (\nu\lambda + S_\gamma^2)^{-(n+\nu)/2} \tag{3.29}$$

where

$$S_\gamma^2 = Y'Y - Y'X(X'X + (D_\gamma R D_\gamma)^{-1})^{-1} X'Y. \tag{3.30}$$

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

Under (3.28), (3.10), and a model space prior $p(\gamma)$, the marginal distribution each component of β is

$$p(\beta_i \mid \gamma) = (1 - \gamma_i) T_1(\nu, 0, \lambda\upsilon_{0i}) + \gamma_i T_1(\nu, 0, \lambda\upsilon_{1i}), \tag{3.31}$$

a scale mixture of t-distributions, in contrast to the normal mixture (3.26). As with the nonconjugate prior, the idea is that υ_{0i} and υ_{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 υ_{0i} and υ_{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 $p(\beta_i \mid i, \gamma_i = 0) = T(\nu, 0, \lambda\upsilon_{0i})$ is larger than the pdf $p(\beta_i \mid i, \gamma_i = 1) = T(\nu, 0, \lambda\upsilon_{1i})$ precisely on the interval $(-\delta_i, \delta_i)$, when υ_{0i} and υ_{1i} satisfy

$$(\upsilon_{0i}/\upsilon_{1i})^{\nu/(\nu+1)} = [\upsilon_{0i} + \delta_i^2/(\nu\lambda)]/[\upsilon_{1i} + \delta_i^2/(\nu\lambda)] \tag{3.32}$$

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 (3.32).

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 $v_{0i}^* = 1$ and $v_{1i}^* = K > 0$ 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 $p(Y \mid \gamma)$

A valuable feature of the previous conjugate prior formulations is that they allow for analytical margining out of β and σ^2 from $p(Y, \beta, \sigma^2 \mid \gamma)$ to yield the closed form expressions in (3.11) and (3.29) which are proportional to $p(Y \mid \gamma)$. Thus, when the model prior $p(\gamma)$ is computable, this can be used to obtain a computable, closed form expression $g(\gamma)$ satisfying

$$g(\gamma) \propto p(Y \mid \gamma) p(\gamma) \propto p(\gamma \mid Y). \tag{3.33}$$

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 $p(\gamma \mid Y)$.

Consider first the conjugate point-normal formulation (3.9) and (3.10) for which $p(Y \mid \gamma)$ proportional to (3.11) can be obtained. When $\Sigma_\gamma = c\,(X_\gamma' X_\gamma)^{-1}$, a function $g(\gamma)$ satisfying (3.33) can be expressed as

$$g(\gamma) = (1+c)^{-q_\gamma/2} (\nu\lambda + Y'Y - (1+1/c)^{-1} W'W)^{-(n+\nu)/2} p(\gamma) \tag{3.34}$$

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 (3.28) and (3.10) for which $p(Y \mid \gamma)$ proportional to (3.29) can be obtained. When $R = I$ holds, a function $g(\gamma)$

satisfying (3.33) can be expressed as

$$g(\gamma) = (\prod_{i=1}^{p} T_{ii}^2 \left[(1-\gamma_i)v^*_{0\gamma_{(i)}} + \gamma_i v^*_{1\gamma_{(i)}}\right])^{-1/2} (\nu\lambda + Y'Y - W'W)^{-(n+\nu)/2} p(\gamma) \quad (3.35)$$

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

$$\gamma^{(1)}, \gamma^{(2)}, \ldots \quad (3.36)$$

that converges (in distribution) to $p(\gamma|Y)$. In formulations where analytical simplification of $p(\beta, \sigma^2, \gamma \mid Y)$ is unavailable, (3.36) can be obtained as a subsequence of a simulated Markov chain of the form

$$\beta^{(1)}, \sigma^{(1)}, \gamma^{(1)}, \beta^{(2)}, \sigma^{(2)}, \gamma^{(2)}, \ldots \quad (3.37)$$

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 (3.36) 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 (3.36) 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 (3.36) may still provide useful information. Even when the length of the sequence (3.36) 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 (3.36). 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 (3.24) and (3.10) 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 (3.37) is obtained by successive simulation from the full conditionals

$$\begin{gathered} p(\beta \mid \sigma^2, \gamma, Y) \\ p(\sigma^2 \mid \beta, \gamma, Y) = p(\sigma^2 \mid \beta, Y) \\ p(\gamma_i \mid \beta, \sigma^2, \gamma_{(i)}, Y) = p(\gamma_i \mid \beta, \gamma_{(i)}),\ i = 1, \ldots, p \end{gathered} \tag{3.38}$$

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 (3.36) 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,

$$\gamma_i \mid \gamma_{(i)}, Y \quad i = 1, 2, \ldots, p, \tag{3.39}$$

($\gamma_{(i)} = (\gamma_1, \gamma_2, \ldots, \gamma_{i-1}, \gamma_{i+1}, \ldots, \gamma_p)$) where the γ_i may be drawn in any fixed or random order. By margining out β and σ^2 in advance, the sequence (3.36) 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 (3.39) 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 (3.34) or (3.35), 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

$$\frac{p(\gamma_i = 1, \gamma_{(i)} \mid Y)}{p(\gamma_i = 0, \gamma_{(i)} \mid Y)} = \frac{g(\gamma_i = 1, \gamma_{(i)})}{g(\gamma_i = 0, \gamma_{(i)})}. \tag{3.40}$$

At each step of the iterative simulation from (3.39), one of the values of $g(\gamma)$ in (3.40) 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 (3.9) with $\Sigma_\gamma = c\,(X'_\gamma X_\gamma)^{-1}$, the fast updating of (3.34) requires $O(q_\gamma^2)$ operations, whereas under the conjugate normal-normal prior formulation (3.28) with $R = I$ fast updating of (3.35) 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 $\gamma_{I_k} = \{\gamma_i \,|\, i \in I_k\}$ and $\gamma_{(I_k)} = \{\gamma_i \mid i \notin I_k\}$. The grouped GS generates (3.36) by iterative simulation from

$$\gamma_{I_k} \mid \gamma_{(I_k)}, Y \quad k = 1, 2, \ldots, m. \tag{3.41}$$

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 (3.36). 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 $g(\gamma) \propto p(\gamma|Y)$ also facilitates the use of MH algorithms for direct simulation of (3.36). 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 $g(\gamma)/g(\gamma') = p(\gamma \mid Y)/p(\gamma' \mid Y)$, 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

$$\alpha(\gamma^* \mid \gamma^{(j)}) = \min\left\{\frac{q(\gamma^{(j)} \mid \gamma^*)}{q(\gamma^* \mid \gamma^{(j)})}\frac{g(\gamma^*)}{g(\gamma^{(j)})}, 1\right\}. \tag{3.42}$$

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 (3.9) 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 (3.42) simplifies to

$$\alpha^M(\gamma^* \mid \gamma^{(j)}) = \min\left\{\frac{g(\gamma^*)}{g(\gamma^{(j)})}, 1\right\}. \tag{3.43}$$

Perhaps the simplest symmetric transition kernel is

$$q(\gamma^1 \mid \gamma^0) = 1/p \quad \text{if} \quad \sum_1^p |\gamma_i^0 - \gamma_i^1| = 1. \tag{3.44}$$

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 (3.44) is a special case of the class of symmetric transition kernels of the form

$$q(\gamma^1 \mid \gamma^0) = q_d \quad \text{if} \quad \sum_1^p |\gamma_i^0 - \gamma_i^1| = d. \tag{3.45}$$

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 (3.44) 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

$$q(\gamma^1 \mid \gamma^0) = q_d \text{ if } \sum_1^p (\gamma_i^0 - \gamma_i^1) = d. \tag{3.46}$$

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 (3.36) 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 $g(\gamma) \propto p(\gamma|Y)$ 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 (3.36) 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

$$p(\gamma \mid Y) = C\, g(\gamma) \tag{3.47}$$

based on the MCMC output (3.36) , say $\gamma^{(1)}, \ldots, \gamma^{(K)}$. 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

$$\hat{C} = \frac{1}{g(A)K} \sum_{k=1}^{K} I_A(\gamma^{(k)}) \tag{3.48}$$

where $I_A(\,)$ is the indicator of the set A, George and McCulloch (1997). Note that if (3.36) were an uncorrelated sequence, then $\mathrm{Var}(\hat{C}) = (C^2/K)\frac{1-p(A \mid Y)}{p(A \mid Y)}$ suggesting that

the variance of (3.48) 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 (3.36). Peng (1998) has extended and generalized these ideas to obtain estimators of C that improve on (3.48).

Inserting $\hat{C}$ into (3.47) yields improved estimates of the probability of individual γ values

$$\hat{p}(\gamma \mid Y) = \hat{C}\, g(\gamma), \tag{3.49}$$

as well as an estimate of the total visited probability

$$\hat{p}(B \mid Y) = \hat{C}\, g(B), \tag{3.50}$$

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. Since $\hat{p}(\gamma \mid Y)/p(\gamma \mid Y) \equiv \hat{C}/C$, the uniform accuracy of the probability estimates (3.49) is

$$|(\hat{C}/C) \; - \; 1|. \tag{3.51}$$

This quantity is also 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 (3.36) 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

$$E(\Delta \mid Y) = \sum_{\text{all } \gamma} E(\Delta \mid \gamma, Y) p(\gamma \mid Y). \tag{3.52}$$

When p is too large for exhaustive enumeration and $p(\gamma \mid Y)$ cannot be computed, (3.52) is unavailable and is typically approximated by something of the form

$$\hat{E}(\Delta \mid Y) = \sum_{\gamma \in S} E(\Delta \mid \gamma, Y) \hat{p}(\gamma \mid Y, S) \tag{3.53}$$

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 (3.53) is

$$\hat{E}_f(\Delta \mid Y) = \sum_{\gamma \in S} E(\Delta \mid \gamma, Y) \hat{p}_f(\gamma \mid Y, S) \tag{3.54}$$

where $\hat{p}_f(\gamma \mid Y, S)$ is the relative frequency of γ in S, George (1999). Indeed, (3.54) 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

$$\hat{E}_g(\Delta \mid Y) = \sum_{\gamma \in S} E(\Delta \mid \gamma, Y) \hat{p}_g(\gamma \mid Y, S) \tag{3.55}$$

where

$$\hat{p}_g(\gamma \mid Y, S) = g(\gamma)/g(S) \tag{3.56}$$

is just the renormalized value of $g(\gamma)$. For example, when S is an iid sample from $p(\gamma|Y)$, (3.55) 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, $E(\hat{E}_f(\Delta \mid Y) \mid S) \rightarrow \hat{E}_g(\Delta \mid Y)$.

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 $p(y|\theta)$ 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

$$p(y \mid \theta_i) = N(\mu_i, \sigma^2),\ i = 1, \ldots, b, \tag{4.1}$$

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

$$p(y \mid \theta_i) = N(\mu_i, \sigma_i^2),\ i = 1, \ldots, b, \tag{4.2}$$

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

$$p(y \mid \theta_i) = \prod_{k=1}^{K} p_{ik}^{I(y \in C_k)} \quad i = 1, \ldots, b, \tag{4.3}$$

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 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 modelling was popularized in the statistical community by the seminal book of Breiman, Friedman, Olshen and Stone (1984). Earlier work by Kass (1980) and Hawkins and Kass (1982) developed tree models in a statistical framework. There has also been substantial research on trees in the field of machine learning, for example the C4.5 algorithm and its predecessor, ID3 (Quinlan 1986, 1993). Here, we focus on the method of Breiman et al. (1984), which proposed a nonparametric approach for tree selection based on a greedy algorithm named CART. A concise description of this approach, which

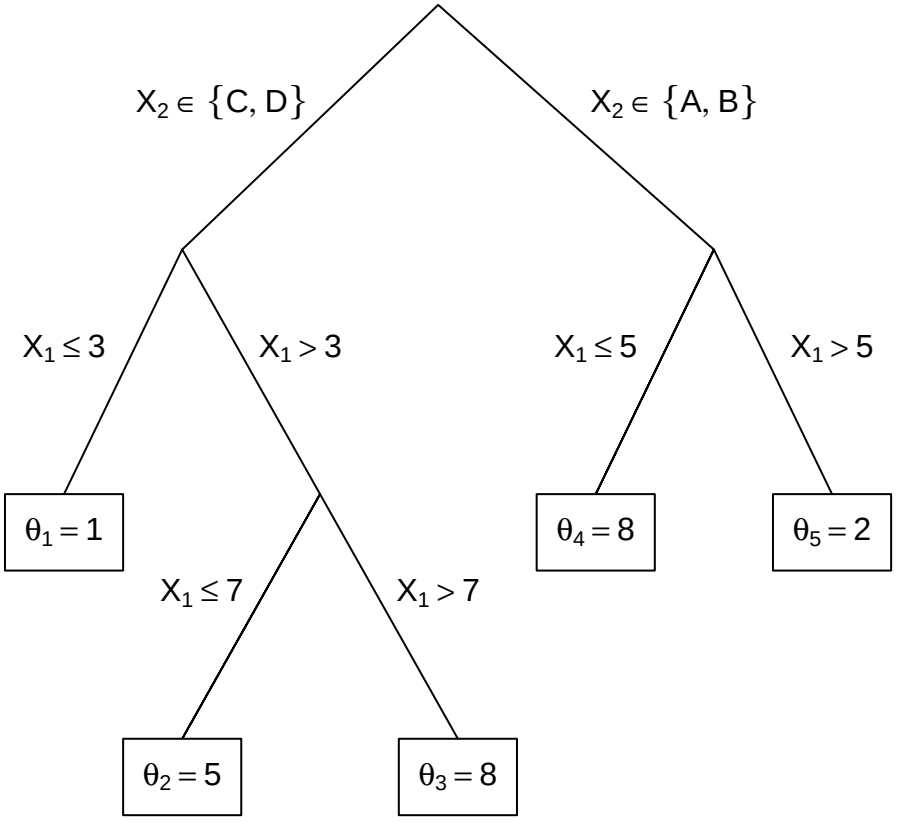

Figure 1: *A regression tree where* $y \sim N(\theta, 2^2)$ *and* $x = (x_1, x_2)$.

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 modelling 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 (1999) 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 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.

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 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 (4.1)-(4.3). For each of these priors, Θ can be analytically margined out via (2.2), namely

$$p(Y \mid T) = \int p(Y \mid \Theta, T) p(\Theta \mid T) d\Theta, \tag{4.4}$$

where Y here denotes the observed values of y.

For regression trees with the mean-shift normal model (4.1), perhaps the simplest prior specification for $p(\Theta|T)$ is the standard normal-inverse-gamma form where $\mu_1, \ldots, \mu_b$ are iid given σ and T with

$$p(\mu_i \mid \sigma, T) = N(\bar{\mu}, \sigma^2/a) \tag{4.5}$$

and

$$p(\sigma^2 \mid T) = p(\sigma^2) = IG(\nu/2, \nu\lambda/2). \tag{4.6}$$

Under this prior, standard analytical simplification yields

$$p(Y \mid T) \propto \frac{c\, a^{b/2}}{\prod_{i=1}^{b}(n_i + a)^{1/2}} \left(\sum_{i=1}^{b} (s_i + t_i) + \nu\lambda \right)^{-(n+\nu)/2} \tag{4.7}$$

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.

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 (4.2) where σ_i can also vary across the terminal nodes, the normal-inverse-gamma form is easily extended to

$$p(\mu_i \mid \sigma_i, T) = N(\bar{\mu}, \sigma_i^2/a) \tag{4.8}$$

and

$$p(\sigma_i^2 \mid T) = p(\sigma_i^2) = IG(\nu/2, \nu\lambda/2), \tag{4.9}$$

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

$$p(Y \mid T) \propto \prod_{i=1}^{b} \pi^{-n_i/2} (\lambda\nu)^{\nu/2} \frac{\sqrt{a}}{\sqrt{n_i + a}} \frac{\Gamma((n_i + \nu)/2)}{\Gamma(\nu/2)} (s_i + t_i + \nu\lambda)^{-(n_i+\nu)/2} \tag{4.10}$$

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 (4.7) 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 (4.3), 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

$$p(p_i \mid T) = \text{Dirichlet}(p_i \mid \alpha) \propto p_{i1}^{\alpha_1 - 1} \cdots p_{iK}^{\alpha_K - 1}. \tag{4.11}$$

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

$$p(Y \mid T) \propto \left(\frac{\Gamma\left(\sum_k \alpha_k\right)}{\prod_k \Gamma(\alpha_k)} \right)^b \prod_{i=1}^{b} \frac{\prod_k \Gamma(n_{ik} + \alpha_k)}{\Gamma\left(n_i + \sum_k \alpha_k\right)} \tag{4.12}$$

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 the Y values within each node are more homogeneous. 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 (4.11) 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 (4.7), (4.10) or (4.12) 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 (4.7)) 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 (4.7), (4.10) or (4.12) for $p(Y \mid T)$ with $p(T)$ yields a closed form expression $g(T)$ satisfying

$$g(T) \propto p(Y \mid T)p(T) \propto p(T \mid Y). \tag{4.13}$$

Analogous to benefits of the availability $g(\gamma)$ in (3.33) 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

$$T^0, T^1, T^2, \ldots \tag{4.14}$$

which are converging in distribution to the posterior $p(T \mid Y)$. 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 trees, 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 (4.14). 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

$$\alpha(T^* \mid T^j) = \min \left\{ \frac{q(T^j \mid T^*)}{q(T^* \mid T^j)} \frac{g(T^*)}{g(T^j)}, 1 \right\}. \tag{4.15}$$

   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 (4.15). 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 cancellation between g and q in the calculation of (4.15) 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 $p(T|Y)$, it will spend more time visiting tree regions where $p(T|Y)$ 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 1999) 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 multiple models. 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 (2000), 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 modelling 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. *J. Roy. Statist. Soc. Ser. B* **60**, 725-749.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *2nd Internat. Symp. Inform. Theory* (B.N. Petrov and F. Csaki, eds.) 267-81, Akademia Kiado, Budapest.

Albert, J.H. (1996). The Bayesian Selection of Log-linear Models. *Canad. J. Statist.* **24**, 327-347.

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.

Berger, J.O. and Pericchi, L.R. (1996). The intrinsic Bayes factor for model selection

and prediction. *J. Amer. Statist. Asso.* **91**, 109-122.

Besag, J and Green, P.J. (1993). Spatial statistics and Bayesian computation (with discussion) *J. Roy. Statist. Soc. Ser. B* **55** 25-37.

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

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

Brown, P.J., Vannucci,M., and Fearn, T. (1998). Multivariate Bayesian variable selection and prediction. *J. Roy. Statist. Soc. 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. *Statist. Comput.* **2**, 63-73.

Carlin, B.P. and Chib, S. (1995). Bayesian model choice via Markov Chain Monte Carlo. *J. Roy. Statist. Soc. Ser. 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, *J. Amer. Statist. Asso.* **66**, 744-748.

Chen, M.H., Ibrahim, J.G. and Yiannoutsos, C. (1999). Prior elicitation, variable selection and Bayesian computation for logistic regression models. *J. Roy. Statist. Soc. Ser. 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. *Canad. J. Statist.* **24**, 17-36.

Chipman, H. A., George, E. I., and McCulloch, R. E. (1998a). Bayesian CART model search (with discussion). *J. Amer. Statist. Asso.* **93**, 935-960.

Chipman, H. A., George, E. I. and McCulloch, R. E. (1998b). Making sense of a forest of trees. *Computing Science and Statistics, Proc. 30th Symp. Interface* (S. Weisberg, Ed.) 84-92, Interface Foundation of North America, Fairfax, VA.

Chipman, H., George, E. I., and McCulloch, R. E. (2000). Hierarchical priors for Bayesian CART shrinkage. *Statist. Comput.* **10**, 17-24.

Chipman, H., George, E.I. and McCulloch, R.E. (2001a). Managing multiple models. *Artificial Intelligence and Statistics 2001*, (Tommi Jaakkola and Thomas Richardson, eds.) 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. *J. Amer. Statist. Asso.* **92**, 1413-1421.

Clark, L. and Pregibon, D. (1992). Tree-Based Models. In *Statistical Models in S* (J. Chambers and T. Hastie, Eds.) 377-419, 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 Univ. Press.

Clyde, M.A., DeSimone, H., and Parmigiani, G. (1996). Prediction via orthogonalized model mixing. *J. Amer. Statist. Asso.* **91**, 1197-1208.

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

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

Clyde, M.A. and Parmigiani, G. (1994). Bayesian variable selection and prediction with mixtures, *J. Biopharmaceutical Statist.*

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

Denison, D.G.T., Mallick, B.K. 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. Statist. Soc. Ser. B* **60**, 333-350.

Denison, D.G.T., Mallick, B.K. and Smith, A.F.M. (1998c). Bayesian MARS. *Statist. Comput.* **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. *Statist. Comput.* To appear.

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

Donoho, D.L. and 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. Roy. Statist. Soc. Ser. B* **57**, 45-98.

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. *J. Econometrics* **100**, 381-427.

Foster, D.P. and 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, *Ann. Statist.* **20**, 1697-1719.

Garthwaite, P. H. and 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* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.) 147–167, Oxford Univ. Press.

Gelfand, A., and Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities, *J. Amer. Statist. Asso.* **85**, 398-409.

George, E.I. (1998). Bayesian model selection. *Encyclopedia of Statistical Sciences, Update Volume 3* (S. Kotz, C. Read and D. Banks, eds.) 39-46, Wiley, New York.

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.) 175-177, Oxford Univ. 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. *J. Amer. Statist. Asso.* **88**, 881-889.

George, E.I. and McCulloch, R.E. (1996). Stochastic search variable selection. In *Markov Chain Monte Carlo in Practice* (W.R. Gilks, S. Richardson and D.J. Spiegelhalter, eds.) 203-214, Chapman and Hall, London.

George, E.I. and 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.) 339-348, Wiley, New York.

Geisser, S. (1993). *Predictive Inference: An Introduction*. Chapman and Hall, London.

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. In *Bayesian Statistics 5* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.) 609-620, Oxford Univ. Press.

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

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. *J. Amer. Statist. Asso.* **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.

Hawkins, D. M. and Kass, G. V. (1982). Automatic interaction detection. In *Topic in Applied Multivariate Analysis* (D. M. Hawkins, ed.) 267-302, Cambridge Univ. Press.

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* **22**, 251-270.

Hoeting, J. A., Madigan, D., Raftery, A.E., and Volinsky, C.T. (1999). Bayesian Model averaging: A tutorial (with discussion). *Statist. Sci.* **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 multivariate linear splines. *J. Roy. Statist. Soc. Ser. B* **63**, 3-18.

Ibrahim, J.G. and Chen, M.-H. (2000). Prior Elicitation and Variable Selection for Generalized Linear Mixed Models. In *Generalized Linear Models: A Bayesian Perspective*, (Dey, D.K., Ghosh, S.K. and Mallick, B.K. eds.) 41-53, Marcel Dekker, New York.

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

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

Kass, G. V. (1980). An exploratory technique for investigating large quantities of categorical data. *Appl. Statist.* **29**, 119-127.

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

Key, J.T., Pericchi, L.R. and Smith, A.F.M. (1998). Choosing among models when none of them are true. In *Proc. Workshop Model Selection, Special Issue of Rassegna di Metodi Statistici ed Applicazioni* (W. Racugno, ed.) 333-361, Pitagora Editrice, Bologna.

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." *J. Amer. Statist. Asso.* **93**, 950-957.

Kuo, L. and Mallick, B. (1998). Variable selection for regression models. *Sankhyā Ser. B* **60**, 65-81.

Laud, P.W. and Ibrahim, J.G. (1995). Predictive model selection. *J. Roy. Statist. Soc. Ser. B* **57**, 247-262.

Leblanc, M. and Tibshirani, R. (1998). Monotone shrinkage of trees. *J. Comput. Graph. Statist.* **7**, 417-433.

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, Univ. of California, Los Angeles.

Lindley, D.V. (1968). The choice of variables in regression. *J. Roy. Statist. Soc. Ser. B* **30**, 31-66.

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.

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.) 453-462.

Madigan, D. and York, J. (1995). Bayesian graphical models for discrete data. *Internat. Statist. Rev.* **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 Ed.* Chapman and Hall, New York.

McCulloch, R.E. and Rossi P. (1991). A Bayesian approach to testing the arbitrage pricing theory. *J. Econometrics* **49**, 141-168.

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H. and Teller,E. (1953). Equations of state calculations by fast computing machines. *J. Chem. Phys.* **21**, 1087-1091.

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

Mitchell, T.J. and Beauchamp, J.J. (1988). Bayesian variable selection in linear regression (with discussion). *J. Amer. Statist. Asso.* **83**, 1023-1036.

Ntzoufras, I. (1999). Aspects of Bayesian model and variable selection using MCMC. Ph.D. dissertation, Dept. of Statistics, Athens Univ. of Economics and Business, Athens, Greece.

O'Hagan, A. (1995). Fractional Bayes factors for model comparison (with discussion). *Jour. of the Roy. Statist. Soc. Ser. B* **57**, 99-138.

Oliver, J.J. and Hand, D.J. (1995). On pruning and averaging decision trees. *Proc. Internat. Machine Learning Conf.* 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 Comput. Symp.* 118-125, Real World Computing Partnership, Tsukuba Research Center, Tsukuba, Japan.

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

Peng, L. (1998). Normalizing constant estimation for discrete distribution simulation.

Ph.D. dissertation, Dept. MSIS, Univ. of Texas, Austin.

Phillips, D. B., and Smith, A. F. M. (1995). Bayesian model comparison via jump diffusions, In *Practical Markov Chain Monte Carlo in Practice* (W.R. Gilks, S. Richardson and D.J. Spiegelhalter, eds.) 215-239, Chapman and Hall, London.

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning* **1**, 81-106.

Quinlan, J. R. (1993). *C4.5: Tools for Machine Learning*, Morgan Kauffman, San Mateo, CA.

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. *J. Amer. Statist. Asso.* **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* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.) 323-350, Oxford Univ. Press.

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.) 321-353, Marcel Dekker, New York.

San Martini, A. and Spezzaferri, F. (1984). A predictive model selection criterion *J. Roy. 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). *J. Amer. Statist. Asso.* **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. *J. Roy. Statist. Soc. Ser. 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. *J. Amer. Statist. Asso.* **92**, 1522-1535.

Smith, M., Wong, C.M. and Kohn, R. (1998). Additive nonparametric regression with autocorrelated errors. *J. Roy. Statist. Soc. Ser. B* **60**, 311-331.

Sutton, C. (1991). Improving classification trees with simulated annealing. *Computing Science and Statistics, Proc. 23rd Symp. Interface* (E. Keramidas, ed.) 396-402, Interface Foundation of North America.

Tibshirani, R. and Knight, K. (1999). Model search by bootstrap "bumping". *J. Comput. Graph. Statist.* **8**, 671-686.

Tierney, L. (1994). Markov chains for exploring posterior distributions (with discussion). *Ann. Statist.* **22**, 1674-1762.

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

Troughton, P.T. and Godsill, S.J. (1997). Bayesian model selection for time series using Markov Chain Monte Carlo. *Proc. IEEE Internat. Conf. Acoustics, Speech and Signal Processing*, 3733-3736.

Wakefield, J.C. and Bennett, J.E. (1996). The Bayesian modelling of covariates for population pharmacokinetic models. *J. Amer. Statist. Asso.* **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. *J. Math. Psychology* **44**, 92-107.

Wood, S. and Kohn, R. (1998). A Bayesian approach to robust binary nonparametric regression. *J. Amer. Statist. Asso.* **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* (P.K. Goel and A. Zellner, eds.) 233-43, North-Holland, Amsterdam.

DISCUSSION

M. Clyde

Duke University

I would like to begin by thanking the authors for their many contributions to Bayesian model selection and for providing an excellent summary of the growing body of literature on Bayesian model selection and model uncertainty. The development of computational tools such as the Gibbs sampler and Markov chain Monte Carlo approaches, has led to an explosion in Bayesian approaches for model selection. On the surface, Bayesian model averaging (BMA) and model selection is straightforward to implement: one specifies the distribution of the data, and the prior probabilities of models and model specific parameters; Bayes theorem provides the rest. As the authors point out the two major challenges confronting the practical implementation of Bayesian model selection are choosing prior distributions and calculating posterior distributions. In my experience, I have found that this is especially true in high dimensional problems, such as wavelet regression or non-parametric models, where subjective elicitation of prior distributions is practically infeasible and enumeration of the number of potential models is intractable (Clyde et al. 1998; Clyde and George 2000).

Choice of Prior Distributions

The specification of prior distributions is often broken down into two parts: (1) elicitation of distributions for parameters specific to each model, such as the distribution for regression coefficients in linear models, $p(\beta|\gamma, \sigma^2)$, and (2) selection of a prior distribution over models $p(\gamma)$. For high dimensional problems one cannot specify the prior probability of each γ directly, and practical implementations of Bayesian selection have usually made prior assumptions that the presence or absence of a variable is independent of the presence or absence of other variables. As a special case of this, the uniform prior distribution over models is appealing in that posterior probabilities of models depend only on the likelihood. However, this prior distribution may not be sensible for model averaging when there are groups of similar variables and does not provide the proper "dilution" of posterior mass over similar models (see Clyde 1999; Hoeting et al. 1999), and discussion therein (George 1999; 1999a). In this regard, uniform and independent prior distributions must be used carefully with highly correlated explanatory variables and special consideration should be given to constructing the model space. Even with

[0]Merlise Clyde is Associate Professor, Institute of Statistics and Decision Sciences, Duke University, Durham NC 27708-0251, U.S.A. email: clyde@stat.Duke.EDU.

uniform priors on models, the posterior distribution over models naturally penalizes adding redundant variables, however, this may not be enough to lead to the proper rate of dilution with nearly redundant variables. One approach to constructing dilution prior distributions is to start with a uniform prior over models and use imaginary training data to construct a posterior distribution for γ based on the training data; this posterior would become the prior distribution for γ for the observed data. Selection of the training data and hyperparameters are non-trivial issues, however, this approach would likely provide better dilution properties than starting with an independent prior. Clearly, construction of prior distributions for models that capture similarities among variables in a sensible way is a difficult task and one that needs more exploration.

For (1), by far the most common choice is a normal prior distribution for β, such as in the conjugate setup for point prior selection models (section 3.2 CGM), where $\beta_\gamma \sim N(0, \sigma^2\Sigma_\gamma)$. Again, as one cannot typically specify a separate prior distribution for β under each model, any practical implementation for Bayesian model uncertainty usually resorts to structured families of prior distributions. Another important consideration is whether prior specifications for β are "compatible" across models (Dawid and Lauritzen 2000) . For example, suppose that Model 1 contains variables 1 and 2, Model 2 contains variables 2 and 3, and Model 3 includes only variable 2. With apologies for possible abuse of notation, let β_2 denote the coefficient for variable 2 in each model. With completely arbitrary choices for Σ_γ, under Model 1 the variance for β_2 given that $\beta_1 = 0$ may not be the same as the variance for β_2 given that $\beta_3 = 0$ derived under Model 2, and both may differ from the variance for β_2 under the prior distribution given Model 3.

To avoid this incompatibility, choices for Σ_γ are often obtained from conditional specifications (i.e. conditional on $\beta_i = 0$ for $\gamma_i = 0$) derived from a prior distribution for β under the full model. For example, Zellner's g-prior (Zellner 1986) is commonly used, which leads to $\Sigma = c(X'X)^{-1}$ for the full model and $\Sigma_\gamma = c(X_\gamma' X_\gamma)^{-1}$ for model γ.

While in many cases conditioning leads to sensible choices, the result may depend on the choice of parameterization, which can lead to a Borel paradox (Kass and Raftery 1995; Dawid and Lauritzen 2000). Other structured families may be induced by marginalization or projections, leading to possibly different prior distributions. While structured prior distributions may reduce the number of hyperparameters that must be selected, i.e. to one parameter c, how to specify c is still an issue.

The choice of c requires careful consideration, as it appears in the marginal likelihoods of the data and hence the posterior model probabilities, and in my experience, can be influential. In the situation where little prior information is available, being too "non-informative" about β (taking c large) can have the un-intended consequence of favoring the null model *a posteriori* (Kass and Raftery 1995). While "default" prior distributions

(both proper and improper) can be calibrated based on information criteria such as AIC (Akaike Information Criterion - Akaike 1973), BIC (Bayes Information Criterion – Schwarz 1978), or RIC (Risk Inflation Criterion – Foster and George 1994), there remains the question of which one to use (Clyde 2000; George and Foster 2000; Fernandez et al. 1998); such decisions may relate more to utilities for model selection rather than prior beliefs (although it may be impossible to separate the two issues). Based on simulation studies, Fernandez et al.(1998) recommend RIC-like prior distributions when $n < p^2$ and BIC-like prior distributions otherwise. In wavelet regression, where $p = n$, there are cases where priors calibrated based on BIC have better predictive performance than prior distributions calibrated using RIC, and vice versa. From simulation studies and asymptotic arguments, it is clear that there is no one default choice for c that will "perform" well for all contingencies (Fernandez et al. 1998; Hansen and Yu 1999).

Empirical Bayes approaches provide an adaptive choice for c. One class of problems where BMA has had outstanding success is in non-parametric regression using wavelet bases. In nonparametric wavelet regression where subjective information is typically not available, empirical Bayes methods for estimating the hyperparameters have (empirically) led to improved predictive performance over a number of fixed hyperparameter specifications as well as default choices such as AIC, BIC, and RIC (Clyde and George 1999, 2000; George and Foster 2000; Hansen and Yu 1999) for both model selection and BMA. Because of the orthogonality in the design matrix under discrete wavelet transformations, EB estimates can be easily obtained using EM algorithms (Clyde and George 1999, 2000; Johnstone and Silverman 1998) and allow for fast analytic expressions for Bayesian model averaging and model selection despite the high dimension of the parameter space ($p = n$) and model space (2^n), bypassing MCMC sampling altogether.

George and Foster (2000) have explored EB approaches to hyperparameter selection for linear models with correlated predictors. EM algorithms for obtaining estimates of c and ω, as in Clyde and George (2000), can be adapted to the non-orthogonal case with unknown σ^2 using the conjugate point prior formulation and independent model space priors (equation 3.2 in CGM). For the EM algorithm both model indicators γ and σ^2 are treated as latent data and imputed using current estimates of c and $\omega = (\omega_1, \ldots, \omega_p)$, where ω_j is the prior probability that variable j is included under the independence prior. At iteration i, this leads to

$$\hat{\gamma}^{(i)} = \frac{p(Y|\gamma, c^{(i)})p(\gamma|\omega^{(i)})}{\sum_{\gamma'} p(Y|\gamma, c^{(i)})p(\gamma|\omega^{(i)})} \quad (1)$$

$$S_\gamma^{2(i)} = Y'Y - \frac{c^{(i)}}{1+c^{(i)}}\mathrm{SSR}(\gamma) \quad (2)$$

where $SSR(\gamma)$ is the usual regression sum of squares and $S_\gamma^{2(i)}$ is a Bayesian version of

residual sum of squares using the current estimate $c^{(i)}$. Values of c and ω that maximize the posterior distribution for c and ω given the observed data and current values of the latent data are

$$\hat{\omega}_j^{(i+1)} = \sum_{\gamma \text{ such that } \gamma_j=1} \hat{\gamma}^{(i)} \tag{3}$$

$$\hat{c}^{(i+1)} = \max\left\{0, \left(\sum_{\gamma} \hat{\gamma}^{(i)} \frac{\mathrm{SSR}(\gamma)/(p\sum_j \hat{\omega}_j^{(i+1)})}{(\lambda\nu + S_\gamma^{2(i)})/(n+\nu)}\right) - 1\right\} \tag{4}$$

where ν and λ are hyperparameters in the inverse gamma prior distribution for σ^2 (CGM equation 3.10). These steps are iterated until estimates converge. EB estimates based on a common ω for all variables are also easy to obtain. For ridge-regression independent priors with $\Sigma_\gamma = cI$ or other prior distributions for β, estimates for ω_j have the same form, but estimates for c are slightly more complicated and require numerical optimization.

The ratio in the expression for $\hat{c}$ has the form of a generalized F-ratio, which is weighted by estimates of posterior model probabilities. The EM algorithm highlights an immediate difficulty with a common c for all models, as one or two highly significant coefficients may influence the EB estimate of c. For example, the intercept may be centered far from 0, and may have an absolute t-ratio much larger than the t-ratios of other coefficients. As the intercept is in all models, it contributes to all of the $SSR(\gamma)$ terms, which has the effect of increasing c as the absolute t-ratio increases. Since the same c appears in the prior variance of all other coefficients, if c becomes too large in order to account for the size of the intercept, we risk having the null model being favored (Bartlett's paradox (Kass and Raftery 1995)). While one could use a different prior distribution for the intercept (even a non-informative prior distribution, which would correspond to centering all variables), the problem may still arise among the other variables if there are many moderate to large coefficients, and a few that have extreme standardized values. Implicit in the formulation based on a common c is that the non-zero standardized coefficients follow a normal distribution with a common variance. As such, this model cannot accommodate one or a few extremely large standardized coefficients without increasing the odds that the remaining coefficients are zero. Using a heavy-tailed prior distribution for β may result in more robust EB estimates of c (Clyde and George 2000). Other possible solutions including adding additional structure into the prior that would allow for different groups of coefficients with a different c in each group. In the context of wavelet regression, coefficients are grouped based on the multi-resolution wavelet decomposition; in other problems there may not be any natural a priori groupings. Related to EB methods is the minimum description length (MDL) approach to model selection, which effectively uses a different c_γ estimated from the data

for each model (Hansen and Yu 1999). While EB methods have led to improvements in performance, part of the success depends on careful construction of the model/prior. Some of the problems discussed above highlight possible restrictions of the normal prior.

Unlike the EM estimates for orthogonal regression, the EB approach with correlated predictors involves a summation over all models, which is clearly impractical in large problems. As in the inference problem, one can base EB estimation on a sample of models. This approach has worked well in moderate sized problems, where leaps and bounds (Furnival and Wilson 1974) was used to select a subset of models; these were then used to construct the EB prior distribution, and then estimates under BMA with the estimated prior. For larger problems, leaps and bounds may not be practical, feasible, or suitable (such as CART models), and models must be sampled using MCMC or other methods. How to scale the EB/EM approach up for larger problems where models must be sampled is an interesting problem.

In situations where there is uncertainty regarding a parameter, the Bayesian approach is to represent that prior uncertainty via a distribution. In other words, why not add another level to the hierarchy and specify a prior distribution on c rather than using a fixed value? While clearly feasible using Gibbs sampling and MCMC methods, analytic calculation of marginal likelihoods is no longer an option. Empirical Bayes (EB) estimation of c often provides a practical compromise between the fully hierarchical Bayes model and Bayes procedures where c is fixed in advance. The EB approach plugs in the modal c into $g(\gamma)$ which ignores uncertainty regarding c, while a fully Bayes approach would integrate over c to obtain the marginal likelihood. As the latter does not exist in closed form, Monte Carlo frequencies of models provide consistent estimates of posterior model probabilities. However, in large dimensional problems where frequencies of most models may be only 0 or 1, it is not clear that Monte Carlo frequencies of models $p_f(\gamma|Y, S)$ from implementing MCMC for the fully Bayesian approach are superior to using renormalized marginal likelihoods evaluated at the EB estimate of c. When the EB estimate of c corresponds to the posterior mode for c, renormalized marginal likelihoods $g(\gamma)$ evaluated at the EB estimate of c are closely related to Laplace approximations (Tierney and Kadane 1986) for integrating the posterior with respect to c (the Laplace approximation would involve a term with the determinant of the negative Hessian of the log posterior). A hybrid approach where MCMC samplers are used to identify/sample models from the fully hierarchical Bayes model, but one evaluates posterior model probabilities for the unique models using Laplace approximations may provide better estimates that account for uncertainty in c.

Implementing Sampling of Models

In the variable selection problem for linear regression, marginal likelihoods are available in closed form (at least for nice conjugate prior distributions); for generalized linear models and many other models, Laplace's method of integration can provide accurate approximations to marginal distributions. The next major problem is that the model space is often too large to allow enumeration of all models, and beyond 20-25 variables, estimation of posterior model probabilities, model selection, and BMA must be based on a sample of models.

Deterministic search for models using branch and bounds or leaps and bounds algorithms (Furnival and Wilson 1974) is efficient for problems with typically fewer than 30 variables. For larger problems, such as in non-parametric models or generalized additive models, these methods are too expensive computationally or do not explore a large enough region of the model space, producing poor fits (Hanson and Kooperberg 1999). While Gibbs and MCMC sampling have worked well in high dimensional orthogonal problems, Wong et al. (1997) found that in high dimensional problems such as non-parametric regression using non-orthogonal basis functions that Gibbs samplers were unsuitable, from both a computational efficiency standpoint as well as for numerical reasons, as the sampler tended to get stuck in local modes. Their proposed focused sampler "focuses" on variables that are more "active" at each iteration, and in simulation studies provided better MSE performance than other classical non-parametric approaches or Bayesian approaches using Gibbs or reversible jump MCMC sampling.

Recently, Holmes and Mallick (1998) adapted perfect sampling (Propp and Wilson 1996) to the context of orthogonal regression. While more computationally intensive per iteration, this may prove to be more efficient for estimation than Gibbs sampling or MH algorithms in problems where the method is applicable. Whether perfect sampling can be used with non-orthogonal designs is still open.

With the exception of deterministic search, most methods for sampling models rely on algorithms that sample models with replacement. In cases where $g(\gamma)$ is known, model probabilities may be estimated using renormalized model probabilities (Clyde et al. 1996). As no additional information is provided under resampling models, algorithms based on sampling models without replacement may be more efficient. Under random sampling without replacement (with equal probabilities), the estimates of model probabilities (CGM equation 3.56) are ratios of Horvitz-Thompson estimators (Horvitz and Thompson 1952) and are simulation consistent. Current work (joint with M. Littman) involves designing adaptive algorithms for sampling without replacement where sampling probabilities are sequentially updated. This appears to be a promising direction for implementation of Bayesian model selection and model averaging.

Summary

CGM have provided practical implementations for Bayesian model selection in linear and generalized linear models, non-parametric regression, and CART models, as well as spurred advances in research so that it is feasible to account for model uncertainty in a wide variety of problems. Demand for methods for larger problems seems to outpace growth in computing resources, and there is a growing need for Bayesian model selection methods that "scale up" as the dimension of the problem increases. Guidance in prior selection is also critical, as in many cases prior information is limited. For example, current experiments using gene-array technology result in high dimensional design matrices, $p > 7000$, however, the sample size may only be on the order of 10-100 (Spang et al. 2000). Identifying which genes (corresponding to columns of X) are associated with outcomes (response to treatment, disease status, etc) is a challenging problem for Bayesian model selection, from both a computational standpoint, as well as the choice of prior distributions.

ADDITIONAL REFERENCES

Clyde, M. (2000). Model uncertainty and health effect studies for particulate matter. To appear in *Environmentrics.*

Dawid, A. and Lauritzen, S. (2000). Compatible prior distributions. Technical Report.

Fernandez, C., Ley, E., and Steel, M.F. (1998). Benchmark priors for Bayesian model averaging. Technical report, Dept. of Econometrics, Tillburg Univ., Netherlands.

Furnival, G.M. and Wilson, Robert W.J. (1974). Regression by Leaps and Bounds. *Technometrics,* **16**, 499-511.

George, E.I. (1999a). Discussion of "Bayesian Model Averaging: A Tutorial" by Hoeting, J.A., Madigan, D., Raftery, A.E., and Volinsky, C.T. *Statist. Sci.* **14**, 401-404.

Hansen, M. and Yu, B. (1999). Model selection and the principle of minimum description. Technical Report *http://cm.bell-labs.com/who/cocteau/papers.*

Hanson, M. and Kooperberg, C. (1999). Spline adaptation in extended linear models. Technical Report *http://cm.bell-labs.com/who/cocteau/papers.*

Hoeting, H.A., Madigan, D., Raftery, A.E., and Volinsky, C.T. (1999). Bayesian model averaging: A tutorial (with discussion). *Statist. Sci.* **14**, 382-417.

Holmes, C. and Mallick, B.K. (1998). Perfect simulation for orthogonal model mixing. Technical Report, Dept. of Math., Imperial College, London.

Horvitz, D. and Thompson, D. (1952). A generalization of sampling without replacement from a finite universe. *J. Amer. Statist. Asso.* **47**, 663-685.

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

Propp, J. and Wilson, D. (1996). Exact sampling with coupled Markov chains and applications to statistical mechanics. *Random Structures and Algorithms,* **9**, 223-252.

Spang, R., Zuzan, H., West, M., Nevins, J, Blanchette, C., and Marks, J.R. (2000). Prediction and uncertainty in the analysis of gene expression profiles. Discussion paper, ISDS, Duke Univ.

Wong, F., Hansen,M.H., Kohn, R., and Smith, M. (1997). Focused Sampling and its Application to Nonparametric and Robust Regression. Bell Labs Technical Report. Technical Report *http://cm.bell-labs.com/who/cocteau/papers.*

Dean P. Foster and Robert A. Stine

University of Pennsylvania

We want to draw attention to three ideas in the paper of Chipman, George and McCulloch (henceforth CGM). The first is the importance of an adaptive variable selection criterion. The second is the development of priors for interaction terms. Our perspective is information theoretic rather than Bayesian, so we briefly review this alternative perspective. Finally, we want to call attention to the practical importance of having a fully automatic procedure. To convey the need for automatic procedures, we discuss the role of variable selection in developing a model for credit risk from the information in a large database.

Adaptive variable selection

A method for variable selection should be *adaptive*. By this, we mean that the prior, particularly $p(\gamma)$, should adapt to the complexity of the model that matches the data rather than impose an external presumption of the number of variables in the model. One may argue that in reasonable problems the modeler should have a good idea how many predictors are going to be useful. It can appear that a well-informed modeler does not need an adaptive prior and can use simpler, more rigid alternatives that reflect knowledge of the substantive context. While domain knowledge is truly useful, it does

[0]Dean P. Foster and Robert P. Stine are Associate Professors, Department of Statistics, The Wharton School of the University of Pennsylvania, Philadelphia, PA 19104-6302, U.S.A; emails: foster@diskworld.wharton.upenn.edu and stine@wharton.upenn.edu.

not follow that such knowledge conveys how many predictors belong in a model. The problem is made most transparent in the following admittedly artificial setting.

A small error in the choice of the basis in an orthogonal regression can lead to a proliferation in the number of required predictors. Suppose that we wish to predict future values of a highly periodic sequence, one dominated by a single sinusoid with frequency ω. If we approach this as a problem in variable selection and use the common Fourier basis to define the collection of predictors, the number of predictors is influenced by how close the frequency of the dominant cycle comes to a Fourier frequency. Fourier frequencies are of the form $\omega_j = 2\pi j/n$, indicating sinusoids that complete precisely j cycles during our n observations. If it so happens that $\omega = \omega_k$, then our model will likely need but one sinusoid to model the response. If ω is not of this form, however, our model will require many sinusoids from the Fourier basis to fit the data well. For example, with $n = 256$ and $\omega = 2\pi 5.5/n$, it takes 8 sinusoids at Fourier frequencies to capture 90% of the variation in this signal. The optimal basis would need but one sinusoid. Adaptive thresholding — the empirical Bayes approach — is forgiving of such errors, whereas dogmatic methods that anticipate, say, a single sinusoid are not.

Information theory and the choice of priors

A difficult choice in the use of Bayesian models for variable selection is the choice of a prior, particularly a prior for the subspace identifying the predictors. We have found coding ideas drawn from information theory useful in this regard, particularly the ideas related to Rissanen's minimum description length (*MDL*). The concreteness of coding offers appealing metaphors for picking among priors that produce surprisingly different selection criteria. In the Bayesian setting, calibration also offers a framework for contrasting the range of variable selection criteria.

The problem we consider from information theory is compression. This problem is simple to state. An *encoder* observes a sequence of n random variables $Y = (Y_1, \ldots, Y_n)$, and his objective is to send a *message* conveying these observations to a *decoder* using as few bits as possible. In this context, a *model* is a completely specified probability distribution, a distribution that is shared by the encoder and decoder. Given that both encoder and decoder share a model $P(Y)$ for the data, the optimal message length (here, the so-called "idealized length" since we ignore fractional bits and the infinite precision of real numbers) is

$$\ell(Y) = \log_2 \frac{1}{P(Y)} \text{ bits.}$$

If the model is a good representation for the data, then $P(Y)$ is large and the resulting message length is small. Since the encoder and decoder share the model $P(Y)$ they can

use a technique known as arithmetic coding to realize this procedure. But what model should they use?

Common statistical models like the linear model are parametric models P_{θ_q}, indexed by a q-dimensional parameter θ_q. For example, suppose that the data Y are generated by the Gaussian linear model

$$Y = \theta_1 X_1 + \theta_2 X_2 + \cdots + \theta_q X_q + \epsilon, \quad \epsilon \sim N(0, \sigma^2) \; .$$

To keep the analysis straightforward, we will assume σ^2 is known (see Barron, Rissanen and Yu 1998, for the general case). Given this model, the shortest code for the data is obtained by maximizing the probability of Y, namely using maximum likelihood (i.e., least squares) to estimate θ_q and obtain a message with length

$$\ell_{\hat{\theta}_q}(Y) = \frac{\log_2 e}{2\sigma^2} RSS(\hat{\theta}_q) + \frac{n}{2} \log_2(2\pi\sigma^2) \; ,$$

where $RSS(\hat{\theta}_q)$ is the residual sum of squares. This code length is not realizable, however, since $P_{\hat{\theta}_q}$ is *not* a model in our sense. The normal density for Y with parameters $\hat{\theta}_q = \theta_q(Y)$ integrates to more than one,

$$C_{n,q} = \int_Y P_{\theta_q(Y)}(Y) dY > 1.$$

Once normalized with the help of some benign constraints that make the integral finite but do not interfere with variable selection (see, e.g., Rissanen 1986), the code length associated with the model $P_{\hat{\theta}_q}/C_{n,q}$ is

$$\ell_q(Y) = \log_2 C_{n,q} + \ell_{\hat{\theta}_q}(Y) \; . \tag{1}$$

The need for such normalization reminds us that coding does not allow improper priors; improper priors generate codes of infinite length. We can think of the first summand in (1) as the length of a code for the parameters $\hat{\theta}_q$ (thus defining a prior for θ_q) and the second as a code for the compressed data. This perspective reveals how coding guards against over-fitting: adding parameters to the model will increase $C_{n,q}$ while reducing the length for the data.

So far, so good, but we have not addressed the problem of variable selection. Suppose that both the encoder and decoder have available a collection of p possible predictors to use in this q-variable regression. Which predictors should form the code? In this expanded context, our code at this point is incomplete since it includes $\hat{\theta}_q$, but does not identify the q predictors. It is easy to find a remedy: simply prefix the message with the p bits in γ. Since codes imply probabilities, the use of p bits to encode γ implies a

prior, p_1 say, for these indicators. This prior is the iid Bernoulli model with probability $\Pr(\gamma_i = 1) = \frac{1}{2}$ for which the optimal code length for γ is indeed p,

$$\log_2 1/p_1(\gamma) = \log_2 2^p = p \; .$$

Since adding a predictor does not affect the length of this prefix – it's always p bits – we add the predictor X_{q+1} if the gain in data compression (represented by the reduction in RSS) compensates for the increase in the normalizing constant $C_{n,q}$. Using a so-called two-part code to approximate the code length (1), we have shown (Foster and Stine 1996) that this approach leads to a thresholding rule. For orthogonal predictors, this criterion amounts to choosing those predictors whose z-statistic $z_j = \hat{\theta}_j/\mathrm{SE}(\hat{\theta}_j)$ exceeds a threshold near 2. Such a procedure resembles the frequentist selection procedure *AIC*, which uses a threshold of $\sqrt{2}$ in this context.

Now suppose that p is rather large. Using the p bits to represent γ seems foolish if we believe but one or two predictors are likely to be useful. If indeed few predictors are useful, we obtain a shorter message by instead forming a prefix from the *indices* of the those $\gamma_j = 1$. Each index now costs us about $\log_2 p$ bits and implies a different prior for γ. This prior, p_2 say, is again iid Bernoulli, but with small probability $\Pr(\gamma_i = 1) = 1/p$; the optimal code length for γ under p_2 is

$$\log_2 1/p_2(\gamma) = q \log p - (p - q) \log(1 - q/p) \approx q \log p \; ,$$

for $q = \sum_j \gamma_j \ll p$. Notice how this code assigns a higher cost to adding a predictor to the model. Adding a predictor does not affect the length of the prefix given by $p_1(\gamma)$. With $p_2(\gamma)$ as the prior, however, adding a predictor adds both a coefficient as well as its index to the message. The prefix grows by an additional $\log_2 p$ bits. For orthogonal regression and two-part codes, this approach implies a threshold for z_j near $\sqrt{2 \log p}$ which also happens to correspond roughly to another frequentist procedure. This is the well-known Bonferroni method which retains predictors whose p-value is less than α/p for some $0 \leq \alpha \leq 1$.

Both of these codes have some appeal and correspond to frequentist methods as well as Bayesian priors, but neither is adaptive. The prior for the first code with a fixed p-bit prefix expects half of the predictors to be useful. The second has a prior that expects only one predictor to enter the model. As codes, both are flawed. The gold standard in coding is to compress the data down to the limit implied by the entropy of the underlying process, whatever that process may be. Both $p_1(\gamma)$ and $p_2(\gamma)$ only approach that limit when they happen to be right (e.g., when in fact only one predictor is needed in the model). Alternatively, so-called universal codes exist for representing binary sequences, and these compress γ almost as well as if the underlying probability

were known. Assuming that the elements γ_j are iid (one can drop this condition as well), a universal code represents γ_q using about $pH(q/p)$ bits, where H denotes the Boolean entropy function

$$H(u) = u \log_2 \frac{1}{u} + (1-u) \log_2 \frac{1}{1-u}, \quad 0 \le u \le 1 \; .$$

Universal codes are *adaptive* in that they perform well for all values of q/p, doing almost as well as either of the previous codes when they happen to be right, but much better in other cases. Returning to the setting of an orthogonal regression, a universal code also implies a threshold for adding a predictor. The threshold in this case now depends on how many predictors are in the model. One adds the predictor X_j to a model that already has q predictors if its absolute z-statistic $|\hat{\theta}_j/\mathrm{SE}(\hat{\theta}_j)| > \sqrt{2 \log p/q}$. This is essentially the empirical Bayes selection rule discussed by CGM in Section 3.3. The threshold decreases as the model grows, adapting to the evidence in favor of a larger collection of predictors. Again, this procedure is analogous to a frequentist method, namely step-up testing as described, for example, in Benjamini and Hochberg (1995).

Coding also suggests novel priors for other situations when the elements of γ are not so "anonymous". For example, consider the treatment of interaction terms. In the application we discuss in the next section, we violate the principle of marginality and treat interactions in a non-hierarchical fashion. That is, we treat them just like any other coefficient. Since we start with 350 predictors, the addition of interactions raises the number of possible variables to about $p = 67,000$. Since they heavily outnumber the linear terms, interactions dominate the predictors selected for our model. Coding the model differently leads to a different prior. For example, consider a variation on the second method for encoding γ by giving the index of the predictor. One could modify this code to handle interactions by appending a single bit to all indices for interactions. This one bit would indicate whether the model included the underlying linear terms as well. In this way, the indices for X_j, X_k and $X_j * X_k$ could be coded in $1 + \log_2 p$ bits rather than $3 \log_2 p$ bits, making it much easier for the selection criterion to add the linear terms.

An application of automatic, adaptive selection

Methods for automatic variable selection matter most in problems that confront the statistician with many possibly relevant predictors. If the available data set holds, say, 1000 observations but only 10 predictors, then variable selection is not going to be very important. The fitted model with all 10 of these predictors is going to do about as well as anything. As the number of predictors increases, however, there comes a point where an

automatic method is necessary. What constitutes a large number of possible predictors? Probably several thousand or more.

Such problems are not simply imaginary scenarios and are the common fodder for "data mining". Here is one example of such a problem, one that we discuss in detail in Foster and Stine (2000). The objective is to anticipate the onset of bankruptcy for credit card holders. The available data set holds records of credit card usage for a collection of some 250,000 accounts. For each account, we know a variety of demographic characteristics, such as place and type of residence of the card holder. When combined with several months of past credit history and indicators of missing data, we have more than 350 possible predictors. The presence of missing data adds further features, and indeed we have found the absence of certain data to be predictive of credit risk. Though impressive at first, the challenge of choosing from among 350 features is nonetheless small by data mining standards. For predicting bankruptcy, we have found interactions between pairs or even triples to be very useful. Considering pairwise interactions swells the number of predictors to over 67,000. It would be interesting to learn how to apply a Gibbs sampler to such problems with so many possible features.

Though challenging for any methodology, problems of this size make it clear that we must have automated methods for setting prior distributions. To handle 67,000 predictors, we use adaptive thresholding and stepwise regression. Beginning from the model with no predictors, we identify the first predictor X_{j_1} that by itself explains the most variation in the response. We add this predictor to the model if its t-statistic $t_{11} = \hat{\beta}_{j_1,1}/se(\hat{\beta}_{j_1,1})$ (in absolute value) exceeds the threshold $\sqrt{2\log p}$. If $\hat{\beta}_{j_1,1}$ meets this criterion, we continue and find the second predictor X_{j_2} that when combined with X_{j_1} explains the most variation in Y. Rather than compare the associated t-statistic $t_{j_2,2}$ to the initial threshold, we reduce the threshold to $\sqrt{2\log p/2}$, making it now easier for X_{j_2} to enter the model. This process continues, greedily adding predictors so long as the t-statistic for each exceeds the declining threshold,

$$\text{Step } q\text{: Add predictor } X_{j_q} \quad \Longleftrightarrow \quad |t_{j_q,q}| > \sqrt{2\log p/q}$$

Benjamini and Hochberg (1995) use essentially the same procedure in multiple testing where it is known as step-up testing. This methodology works in this credit modelling in that it finds structure without over-fitting. Figure 1 shows a plot of the residual sum of squares as a function of model size; as usual, RSS decreases with p. The plot also shows the cross-validation sum of squares (CVSS) computed by predicting an independent sample. The validation sample for these calculations has about 2,400,000 observations; we scaled the CVSS to roughly match the scale of the RSS. Our search identified 39 significant predictors, and each of these — with the exception of the small "bump" — improves the out-of-sample performance of the model. Although the CVSS curve is flat

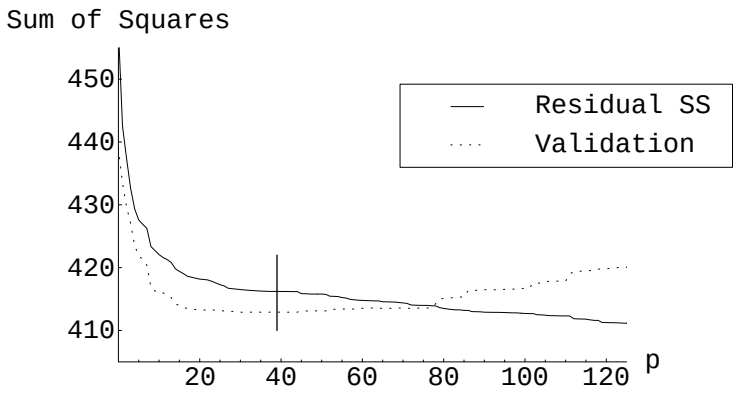

Figure 2: *Residual and cross-validation sums of squares for predicting bankruptcy.*

near $p = 39$, it does not show the rapid increase typically associated with over-fitting.

Gibbs sampling and the practical Bayesian methods discussed by CGM offer an interesting alternative to our search and selection procedure. They have established the foundations, and the next challenge would seem to be the investigation of how their search procedure performs in the setting of real applications such as this. Greedy selection methods such as stepwise regression have been well-studied and probably do not find the best set of predictors. Once X_{j_1} becomes the first predictor, it must be in the model. Such a strategy is clearly optimal for orthogonal predictors, but can be 'tricked' by collinearity. Nonetheless, stepwise regression is fast and comparisons have shown it to be competitive with all possible subsets regression (e.g., see the discussion in Miller 1990). Is greedy good enough, or should one pursue other ways of exploring the space of models via Gibbs sampling?

ADDITIONAL REFERENCES

Barron, A., Rissanen, J. and Yu, B. (1998). The minimum description length principle in coding and modelling. *IEEE Trans. Info. Theory* **44**, 2743-2760.

Benjamini, Y. and Hoohberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B* **57**, 289-300.

Foster, D.P. and Stine, R.A. (1996). Variable selection via information theory. Technical Report, Northwestern Univ., Chicago.

Foster, D.P. and Stine, R.A. (2000). Variable selection in data mining: Building a predictive model for bankruptcy. Unpublished Manuscript.

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

Rissanen, J. (1986). Stochastic complexity and modelling. *Ann. Statist.* **14**, 1080-1100.

REJOINDER

Hugh Chipman, Edward I. George and Robert E. McCulloch

First of all, we would like to thank the discussants, Merlise Clyde, Dean Foster and Robert Stine, for their generous discussions. They have each made profound contributions to model selection and this comes through in their insightful remarks. Although there is some overlap in the underlying issues they raise, they have done so from different vantage points. For this reason, we have organized our responses around each of their discussions separately.

Clyde

Clyde raises key issues surrounding prior selection. For choosing model space priors for the linear model with redundant variables, she confirms the need to move away from uniform and independence priors towards dilution priors. This is especially true in high dimensional problems where independence priors will allocate most of their probability to neighborhoods of redundant models. Clyde's suggestion to use an imaginary training data to construct a dilution prior is a very interesting idea. Along similar lines, we have considered dilution priors for the linear model where $p(\gamma)$ is defined as the probability that $Y^* \sim N_n(0, I)$ is "closer" to the span of X_γ than the span of any other $X_{\gamma'}$. Here Y^* can be thought of as imaginary training data reflecting ignorance about the direction of Y. Further investigation of the construction of effective dilution priors for linear models is certainly needed.

Clyde comments on the typical choices of Σ_γ for the coefficient prior $p(\beta_\gamma \mid \sigma^2, \gamma) = N_{q_\gamma}(\bar{\beta}_\gamma, \sigma^2\Sigma_\gamma)$ in (3.9), namely $\Sigma_\gamma = c\,(X_\gamma' X_\gamma)^{-1}$ and $\Sigma_\gamma = c\,I_{q_\gamma}$. In a sense, these choices are the two extremes, $c\,(X_\gamma' X_\gamma)^{-1}$ serves to reinforce the likelihood covariance while $c\,I_{q_\gamma}$ serves to break it apart. As we point out, the coefficient priors under such Σ_γ, are the natural conditional distributions of the nonzero components of β given γ when $\beta \sim N_p(0, c\,\sigma^2(X'X)^{-1})$ and $\beta \sim N_p(0, c\,\sigma^2 I)$, respectively. The joint prior $p(\beta_\gamma, \gamma \mid \sigma^2)$ then corresponds to a reweighting of the conditional distributions according to the chosen model space prior $p(\gamma)$. With respect to such joint priors, the conditional distributions are indeed compatible in the sense of Dawid and Lauritzen (2000). Although not strictly

necessary, we find such compatible specifications to provide an appealingly coherent description of prior information.

We agree with Clyde that the choice of c can be absolutely crucial. As the calibration result in (3.17) shows, different values of c, and hence different selection criteria such as AIC, BIC and RIC, are appropriate for different states of nature. For larger models with many small nonzero coefficients, smaller values of c are more appropriate, whereas for parsimonious models with a few large coefficients, larger values of c are better. Of course, when such information about the actual model is unavailable, as is typically the case, the adaptive empirical Bayes methods serve to insure against poor choices. It is especially appealing that by avoiding the need to specify hyperparameter values, empirical Bayes methods are automatic, a valuable feature for large complicated problems. Similar features are also offered by fully Bayes methods that margin out the hyperparameters with respect to hyperpriors. The challenge for the implementation of effective fully Bayes methods is the selection of hyperpriors that offer strong performance across the model space while avoiding the computational difficulties described by Clyde.

Clyde points out an important limitation of using empirical Bayes methods with conditional priors of the form $p(\beta_\gamma \mid \sigma^2, \gamma) = N_{q_\gamma}(\bar{\beta}_\gamma, \sigma^2 c\, V_\gamma)$. When the actual model has many moderate sized coefficients and but a few very large coefficients, the few large coefficients will tend to inflate the implicit estimate of c, causing the moderate sized coefficients to be ignored as noise. In addition to the heavy-tailed and the grouped prior formulations she describes for mitigating such situations, one might also consider elaborating the priors by adding a shape hyperparameter.

Finally, Clyde discusses the growing need for fast Bayesian computational methods that "scale up" for very large high dimensional problems. In this regard, it may be useful to combine heuristic strategies with Bayesian methods. For example, George and McCulloch (1997) combined globally greedy strategies with local MCMC search in applying Bayesian variable selection to build tracking portfolios. In our response to Foster and Stine below, we further elaborate on the potential of greedy algorithms for such purposes.

Foster and Stine

Foster and Stine begin by emphasizing the need for adaptive procedures. We completely agree. The adaptive empirical Bayes methods described in Section 3.3 offer improved performance across the model space while automatically avoiding the need for hyperparameter specification. For more complicated settings, adaptivity can be obtained by informal empirical Bayes approaches that use the data to gauge hyperparameter values,

such as those we described for the inverse gamma distribution in Sections 3.2 and 4.1.2. In the sinusoid modelling example of Foster and Stine, a simple adaptive resolution is obtained by a Bayesian treatment with a prior on ω_k. This nicely illustrates the fundamental adaptive nature of Bayesian analysis. By using priors rather than fixed arbitrary values to describe the uncertainty surrounding the unknown characteristics in a statistical problem, Bayesian methods are automatically adaptive. We attribute the adaptivity of empirical Bayes methods to their implicit approximation of a fully Bayes approach.

Foster and Stine go on to discuss some revealing analogies between strategies for minimum length coding and formulations for Bayesian model selection. The key idea is that the probability model for the data, namely the complete Bayesian formulation, also serves to generate the coding strategy. Choosing the probability model that best predicts the data is tantamount to choosing the optimal coding strategy. Foster and Stine note that improper priors are unacceptable because they generate infinite codes. This is consistent with our strong preference for proper priors for model selection. They point out the potential inefficiencies of Bernoulli model prior codes for variable selection, and use them to motivate a universal code that adapts to the appropriate model size. This is directly analogous to our observation in Section 3.3 that different hyperparameter choices for the Bernoulli model prior (3.15) correspond to different model sizes, and that an empirical Bayes hyperparameter estimate adapts to the appropriate model size. It should be the case that the universal prior corresponds to a fully Bayes prior that is approximated by the empirical Bayes procedure. Finally, their coding scheme for interactions is interesting and clearly effective for parsimonious models. Such a coding scheme would seem to correspond to a hierarchical prior that puts a Bernoulli $1/p$ prior on each potential triple - two linear terms and their interaction - and a conditionally uniform prior on the elements of the triple.

The credit risk example given by Foster and Stine raises several interesting issues. It illustrates that with this large dataset, an automatic stepwise search algorithm can achieve promising results. Figure 1 shows how their adaptive threshold criterion guards against overfitting, although the cross validation results seem also to suggest that a smaller number of terms, around 20, is adequate for prediction. Another automatic adaptive alternative to consider here would be a stepwise search based on the empirical Bayes criterion C_{CML} in (3.22). It would also be interesting to investigate the potential of one of the Bayesian variable selection approaches using the hierarchical priors described in Section 3.1 to account for potential relationships between linear and interaction terms. As opposed to treating all potential predictors independently, such priors tend to concentrate prior mass in a smaller, more manageable region of the model space.

For example, Chipman, Hamada and Wu (1997) considered an 18 run designed ex-

periment with 8 predictors used in a blood-glucose experiment. The non-orthogonal design made it possible to consider a total of 113 terms, including quadratic terms and interactions. They found that independence priors of the form (3.2) led to such a diffuse posterior that, in 10,000 steps of a Gibbs sampling run, the most frequently visited model was visited only 3 times. On the other hand, hierarchical priors like (3.7) raised posterior mass on the most probable model to around 15%. In the same problem stepwise methods were unable to find all the different models identified by stochastic search. In effect, priors that account for interactions (or other structure, such as correlations between predictors which can lead to the dilution problem discussed in Section 3.1) can narrow the posterior to models which are considered more "plausible". We note, however, that the credit risk example is much larger than this example, and because the number of observations there is much larger than the number of predictors, such a hierarchical prior may have only a minor effect.

The credit risk example is also a clear illustration of the everlasting potential of greedy search algorithms on very large problems. At the very least, greedy algorithms can provide a "baseline" against which MCMC stochastic search results can be compared and then thrown out if an improvement is not found. Furthermore, greedy algorithms can provide a fast way to get rough estimates of hyperparameter values, and can be used directly for posterior search. Greedy algorithms also offer interesting possibilities for enhancement of stochastic search. At the most basic level, the models identified by greedy algorithms can be used as starting points for stochastic searches. Stochastic search algorithms can also be made more greedy, for example, by exponentiating the probabilities in the accept/reject step of the MH algorithms. The use of a wide variety of search algorithms, including MCMC stochastic search, can only increase the chances of finding better models.