Statistica Sinica **7**(1997), 339-373

APPROACHES FOR BAYESIAN VARIABLE SELECTION

Edward I. George and Robert E. McCulloch

University of Texas at Austin and University of Chicago

Abstract: This paper describes and compares various hierarchical mixture prior formulations of variable selection uncertainty in normal linear regression models. These include the nonconjugate SSVS formulation of George and McCulloch (1993), as well as conjugate formulations which allow for analytical simplification. Hyperparameter settings which base selection on practical significance, and the implications of using mixtures with point priors are discussed. Computational methods for posterior evaluation and exploration are considered. Rapid updating methods are seen to provide feasible methods for exhaustive evaluation using Gray Code sequencing in moderately sized problems, and fast Markov Chain Monte Carlo exploration in large problems. Estimation of normalization constants is seen to provide improved posterior estimates of individual model probabilities and the total visited probability. Various procedures are illustrated on simulated sample problems and on a real problem concerning the construction of financial index tracking portfolios.

Key words and phrases: Conjugate prior, Gibbs sampling, Gray Code, hierarchical models, Markov chain Monte Carlo, Metropolis-Hastings algorithms, normal mixtures, normalization constant, regression, simulation.

1. Introduction

In the context of building a multiple regression model, we consider the following canonical variable selection problem. Given a dependent variable Y and a set of p potential regressors $X_1, \ldots, X_p$, the problem is to find the "best" model of the form $Y = X_1^*\beta_1^* + \cdots + X_q^*\beta_q^* + \epsilon$ where $X_1^*, \ldots, X_q^*$ is a "selected" subset of $X_1, \ldots, X_p$.

A Bayes procedure for identifying "promising" subsets of predictors was proposed by George and McCulloch (1993). This procedure, called SSVS (Stochastic Search Variable Selection), entails the specification of a hierarchical Bayes mixture prior which uses the data to assign larger posterior probability to the more promising models. To avoid the overwhelming burden of calculating the posterior probabilities of all 2^p models, SSVS uses the Gibbs sampler to simulate a sample from the posterior distribution. Because high probability models are more likely to appear quickly, the Gibbs sampler can sometimes identify such models with relatively short runs. Effectively, the Gibbs sampler is used to search for promising models rather than compute the entire posterior. The key to the potential of SSVS is the fast and efficient simulation of the Gibbs sampler.

In this paper we describe, compare, and apply a variety of approaches to Bayesian variable selection which include SSVS as a special case. These approaches all use hierarchical mixture priors to describe the uncertainty present in variable selection problems. Hyperparameter settings which base selection on practical significance, and the implications of using mixtures with point priors are discussed. Conjugate versions of these priors are shown to yield posterior expressions which can sometimes be sequentially computed using efficient updating schemes. When p is moderate (less than about 25), performing such sequential updating in a Gray Code order yields a feasible approach for exhaustive evaluation of all 2^p posterior probabilities. For larger values of p, Markov chain Monte Carlo (MCMC) methods, such as the Gibbs sampler or the Metropolis-Hastings algorithms, can exploit such updating schemes to rapidly search for high probability models. Estimation of normalization constants is seen to provide improved posterior estimates of individual model probabilities and the total visited probability. Nonconjugate and conjugate MCMC implementations are compared on three simulated sample problems.

Bayesian variable selection has been the subject of substantial research in recent years. Some of the papers which propose procedures related to those discussed here include Carlin and Chib (1995), Chipman (1995), Clyde and Parmigiani (1994), Clyde, DeSimone, and Parmigiani (1996), George and McCulloch (1993, 1995), George, McCulloch and Tsay (1995), Geweke (1996), Hoeting, Raftery and Madigan (1995), Kuo and Mallick (1994), Meehan, Dempster and Brown (1994), Mitchell and Beauchamp (1988), Phillips and Smith (1995), Raftery, Madigan, and Hoeting (1993), Raftery, Madigan, and Volinsky (1995), Smith and Kohn (1995), and Wakefield and Bennett (1996).

This paper is structured as follows. Section 2 describes a general hierarchical Bayes mixture formulation for variable selection. Section 3 treats nonconjugate implementations of this formulation which include SSVS. Section 4 treats conjugate implementations of the this formulation which allow for substantial analytical simplification and fast algorithms for posterior evaluation and exploration. Section 5 presents performance assessments of three basic MCMC implementations. Finally, Section 6 presents a real application to the construction of financial index tracking portfolios.

2. A Hierarchical Mixture Model for Variable Selection

We begin by describing a general hierarchical mixture model which forms the basis for the various methods considered in this paper. First of all, the standard normal linear model is used to describe the relationship between the observed dependent variable and the set of all potential predictors $X_1, \ldots, X_p$, namely

$$f(Y|\beta, \sigma) = N_n(X\beta, \sigma^2 I), \tag{1}$$

where Y is $n \times 1$, $X = [X_1, \ldots, X_p]$ is an $n \times p$ matrix, β is a $p \times 1$ vector of unknown regression coefficients, and σ 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. Throughout this paper, we index each of these possible 2^p subset choices by the vector

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

where $\gamma_i = 0$ or 1 if β_i is small or large, respectively. The size of the γth subset is denoted as $q_\gamma \equiv \gamma' 1$. Since the appropriate value of γ is unknown, we model the uncertainty underlying variable selection by a mixture prior $\pi(\beta, \sigma, \gamma) = \pi(\beta|\sigma, \gamma)\pi(\sigma|\gamma)\pi(\gamma)$ which can be conditionally specified as follows:

The γth subset model is described by modeling β as a realization from a multivariate normal prior

$$\pi(\beta|\sigma, \gamma) = N_p(0, \Upsilon_{(\sigma,\gamma)}), \tag{2}$$

where the ith diagonal element of $\Upsilon_{(\sigma,\gamma)}$ is appropriately set to be small or large according to whether $\gamma_i = 0$ or 1, respectively. The specification of $\Upsilon_{(\sigma,\gamma)}$ determines the essential properties of the hierarchical prior. The consequences for variable selection of this specification are the main focus of subsequent sections of this paper.

The residual variance σ^2 for the γth model is conveniently modeled as a realization from an inverse gamma prior

$$\pi(\sigma^2|\gamma) = \mathrm{IG}(\nu/2, \nu\lambda_\gamma/2) \tag{3}$$

which is equivalent to $\nu\lambda_\gamma/\sigma^2 \sim \chi^2_\nu$. Although setting λ_γ constant has led to reasonable results in our experience, it might be desirable have λ_γ decrease with the size of the selected subset, q_γ. For specification purposes, λ_γ 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. In the absence of prior information about σ^2, we recommend choosing $\lambda_\gamma \equiv s^2_{LS}$, where s^2_{LS} is the classical least squares estimate of σ^2 based on a saturated model, and then choosing ν so that $\pi(\sigma^2|\gamma)$ assigns substantial probability to the interval (s^2_{LS}, s^2_Y), where s^2_Y is the sample variance of Y.

Although γ itself can be modeled as a realization from any (nontrivial) prior $\pi(\gamma)$ on the 2^p possible values of γ, priors of the form

$$\pi(\gamma) = \prod w_i^{\gamma_i}(1 - w_i)^{(1-\gamma_i)}, \tag{4}$$

such as $\pi(\gamma) \equiv 1/2^p$, are easy to specify, substantially reduce computational requirements, and often yield sensible results. We interpret $\pi(\gamma_i = 1) = 1 - \pi(\gamma_i =$

$0) = w_i$ as the prior probability that β_i is large enough to justify including X_i in the model. The prior (4) can also be used to put increased weight on parsimonious models by setting the w_i small. Under (4), the components of γ are apriori independent. Treatments of alternative priors with dependent components for this setup have been considered by Chipman (1995) and Geweke (1996).

For this hierarchical setup, the marginal posterior distribution $\pi(\gamma|Y)$ contains the relevant information for variable selection. Based on the data Y, the posterior $\pi(\gamma|Y)$ updates the prior probabilities on each of the 2^p possible values of γ. Identifying each γ with a submodel via $(\gamma_i = 1) \Leftrightarrow (X_i$ is included), those γ with higher posterior probability $\pi(\gamma|Y)$ identify the more "promising" submodels, that is those supported most by the data and the prior distribution.

The practical value of this hierarchical Bayes formulation for Bayesian variable selection depends on two key issues. First, the hyperparameters of the prior, especially $\Upsilon_{(\sigma,\gamma)}$, must be chosen so that, based on the data, the posterior $\pi(\gamma|Y)$ will assign higher probability to the predictor subsets of ultimate interest. Secondly, it is necessary to be able to compute $\pi(\gamma|Y)$ at least to the extent where high probability values of γ can be identified. In the following sections we address both of these issues for various hyperparameter settings.

3. Nonconjugate Hierarchical Setups

In this section, we consider the special case of the Section 2 model where $\Upsilon_{(\sigma,\gamma)}$ in (2) is of the form

$$\pi(\beta|\sigma,\gamma) = \pi(\beta|\gamma) = N_p(0, D_\gamma R_\gamma D_\gamma), \tag{5}$$

where D_γ is a diagonal matrix and R_γ is a correlation matrix. Although any covariance matrix can be written in the form $D_\gamma R_\gamma D_\gamma$, this parametrization is useful for specification purposes. We denote the ith diagonal element of D_γ^2 by

$$(D_\gamma^2)_{ii} = \begin{cases} \upsilon_{0\gamma_{(i)}} & \text{when} \quad \gamma_i = 0, \\ \upsilon_{1\gamma_{(i)}} & \text{when} \quad \gamma_i = 1, \end{cases} \tag{6}$$

where $\gamma_{(i)} = (\gamma_1, \ldots, \gamma_{i-1}, \gamma_{i+1}, \ldots, \gamma_p)$. Note that $\upsilon_{0\gamma_{(i)}}$ and $\upsilon_{1\gamma_{(i)}}$ can depend on the entire subset specified by γ. The SSVS procedure of George and McCulloch (1993, 1995) is based on the special case where $\upsilon_{0\gamma_{(i)}} \equiv \upsilon_{0i}$ and $\upsilon_{1\gamma_{(i)}} \equiv \upsilon_{1i}$ are constant for all $\gamma_{(i)}$, and $R_\gamma \equiv R$ does not depend on γ.

The joint distribution of β and σ given γ is not of the conjugate form because (5) does not depend on σ. Indeed, β and σ are here independent given γ. Throughout the paper we refer to the prior $\pi(\beta, \sigma, \gamma)$ using (5) as "the nonconjugate prior".

3.1. Nonconjugate hyperparameter settings

Under (5), each component of β is modeled as having come from a scale mixture of two normal distributions which, conditionally on $\gamma_{(i)}$, may be represented by

$$\pi(\beta_i|\gamma) = (1-\gamma_i)N(0, v_{0\gamma_{(i)}}) + \gamma_i N(0, v_{1\gamma_{(i)}}). \tag{7}$$

To use this hierarchical mixture setup for variable selection, the hyperparameters $v_{0\gamma_{(i)}}$ and $v_{1\gamma_{(i)}}$ are set "small and large" respectively, so that $N(0, v_{0\gamma_{(i)}})$ is concentrated and $N(0, v_{1\gamma_{(i)}})$ is diffuse as in Figure 1 below. The general idea is 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.

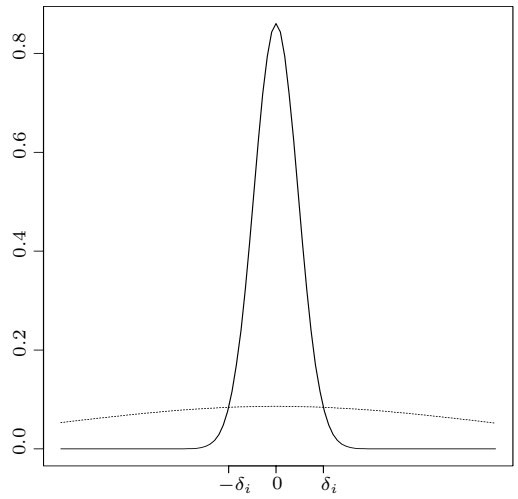

Figure 1. $N(0, v_{1\gamma_{(i)}})$ and $N(0, v_{0\gamma_{(i)}})$ densities. Intersection at $\delta_{i\gamma}$.

Several strategies may be considered for choosing $v_{0\gamma_{(i)}}$ and $v_{1\gamma_{(i)}}$. To begin with, such a choice can be based on considerations of "practical significance", as follows. Suppose a value $\delta_{i\gamma} > 0$ could be chosen such that if $|\beta_i| < \delta_{i\gamma}$ in the γth model, it would be preferable to exclude X_i. Such a $\delta_{i\gamma}$ could be considered the "threshold of practical significance". A simple choice, which does not depend on γ, might be $\delta_{i\gamma} \equiv \delta_i = \Delta Y/\Delta X_i$, 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. Alternatively, to account for the cumulative effect of changes of other X's in the model, one might use the smaller choice $\delta_{i\gamma} = \Delta Y/(q_\gamma \Delta X_i)$.

As described in George and McCulloch (1993, 1995), when such $\delta_{i\gamma}$ can be chosen, higher posterior weighting of those γ values for which $|\beta_i| > \delta_{i\gamma}$ when $\gamma_i = 1$, can be achieved by choosing $v_{0\gamma_{(i)}}$ and $v_{1\gamma_{(i)}}$ such that the pdf $\pi(\beta_i|\gamma_{(i)}, \gamma_i =$

$0) = N(0, v_{0\gamma_{(i)}})$ is larger than the pdf $\pi(\beta_i|\gamma_{(i)}, \gamma_i = 1) = N(0, v_{1\gamma_{(i)}})$ precisely on the interval $(-\delta_{i\gamma}, \delta_{i\gamma})$ (see Figure 1). This property is obtained by any $v_{0\gamma_{(i)}}$ and $v_{1\gamma_{(i)}}$ satisfying

$$\log(v_{1\gamma_{(i)}}/v_{0\gamma_{(i)}})/(v_{0\gamma_{(i)}}^{-1} - v_{1\gamma_{(i)}}^{-1}) = \delta_{i\gamma}^2. \tag{8}$$

We recommend choosing such $v_{0\gamma_{(i)}}$ and $v_{1\gamma_{(i)}}$ so that the $N(0, v_{1\gamma_{(i)}})$ distribution is consistent with prior beliefs about plausible values of β_i under γ. However, as described in Section 3.2, computational problems can arise when $v_{1\gamma_{(i)}}/v_{0\gamma_{(i)}}$ is set too large. In our experience, such computational problems will be avoided whenever $v_{1\gamma_{(i)}}/v_{0\gamma_{(i)}} \leq 10000$, thus allowing for a wide variety of settings.

Incorporation of a threshold of practical significance above requires choosing $v_{0\gamma_{(i)}} > 0$ for all i. In doing so, the distribution (5) will be p-dimensional for all γ. In this case, the prior distribution is allowing for the possibility that submodels are only approximations to the"true" model. This would be appropriate for the common data analysis situation where linear models are used to approximate more complicated relationships between the variables.

When a threshold of practical significance $\delta_{i\gamma}$ cannot be meaningfully specified, one might, instead, consider setting $v_{0\gamma_{(i)}} \equiv 0$ and setting $v_{1\gamma_{(i)}}$ to be consistent with reasonable values of β_i, a setup considered by Geweke (1996). Under this setting, (7) becomes

$$\pi(\beta_i|\gamma) = (1 - \gamma_i)I_0 + \gamma_i N(0, v_{1\gamma_{(i)}}), \tag{9}$$

where I_0 is a point mass at 0. For this choice, $\delta_{i\gamma} \equiv 0$, corresponding to the preference that any $\beta_i \neq 0$ be included in the model. This criterion will select β_i on the basis of how well they can be distinguished from 0 rather than their absolute size. Indeed, when $v_{0\gamma_{(i)}} = 0$ and $\beta_i \neq 0$, the marginal Bayes factor $\pi(\gamma_i = 1|Y)/\pi(\gamma_i = 0|Y)$ will be large with increasing probability as the amount of data increases. Thus, any nonzero β_i, no matter how small, will be included in the model with enough data.

Note that when $v_{0\gamma_{(i)}} \equiv 0$, $\pi(\beta|\gamma)$ in (5) will be a singular q_γ-dimensional distribution. A useful alternative representation in this case, is $\pi(\beta|\gamma) = \pi(\beta_\gamma|\gamma)$ $\pi(\beta_{\bar{\gamma}}|\gamma)$ where β_γ and $\beta_{\bar{\gamma}}$ are subvectors of β such that

$$\pi(\beta_\gamma|\gamma) = N_{q_\gamma}(0, D_{1\gamma}R_{1\gamma}D_{1\gamma}) \text{ and } \pi(\beta_{\bar{\gamma}} = 0|\gamma) = 1. \tag{10}$$

In addition to $v_{0\gamma_{(i)}}$ and $v_{1\gamma_{(i)}}$, the specification of $\pi(\beta|\gamma)$ in (5) requires the choice of a prior correlation matrix R_γ. For the simple choice $R_\gamma \equiv I$, the components of β are apriori independent. Other natural choices which replicate the correlation structure of the least squares estimates are $R_\gamma \propto (X'X)^{-1}$ for

$v_{0\gamma_{(i)}} > 0$ and $R_{1\gamma} \propto (X'_\gamma X_\gamma)^{-1}$ for $v_{0\gamma_{(i)}} \equiv 0$ where X_γ is the $n \times q_\gamma$ matrix whose columns correspond to the components of β_γ.

3.2. Nonconjugate MCMC exploration of the posterior

Although analytical simplification of $\pi(\beta, \sigma, \gamma|Y)$ is intractable under the nonconjugate hierarchical setup, MCMC (Markov chain Monte Carlo) methods such as the Gibbs sampler or Metropolis-Hastings algorithms (see Smith and Roberts (1993) for an overview and references) can be used to explore the posterior $\pi(\gamma|Y)$. Applied to the complete posterior $\pi(\beta, \sigma, \gamma|Y)$, such methods simulate a Markov chain

$$\beta^{(1)}, \sigma^{(1)}, \gamma^{(1)}, \beta^{(2)}, \sigma^{(2)}, \gamma^{(2)}, \ldots, \tag{11}$$

which converges in distribution to $\pi(\beta, \sigma, \gamma|Y)$. The embedded subsequence

$$\gamma^{(1)}, \gamma^{(2)}, \ldots, \tag{12}$$

thus converges to $\gamma \sim \pi(\gamma|Y)$.

In problems where the number of potential predictors p is small, the sequence (12) can be used to evaluate the entire posterior $\pi(\gamma|Y)$. In large problems, where thorough evaluation is not feasible, the sequence (12) may still provide useful information. In many cases, the γ values of interest, namely those with high probability, will appear most frequently and quickly, making them easier to identify. Even when the length of the sequence (12) is much smaller than 2^p, it may thus be possible to identify at least some of the high probability values. In such situations, MCMC methods can at least be used to search for high probability γ values.

The SSVS procedure of George and McCulloch (1993) is based on using the Gibbs sampler to simulate the full parameter sequence (11) when $v_{0\gamma_{(i)}} > 0$. This simply entails successive simulation from the full conditionals

$$\begin{gathered} \pi(\beta|\sigma, \gamma, Y) \\ \pi(\sigma|\beta, \gamma, Y) = \pi(\sigma|\beta, Y) \\ \pi(\gamma_i|\beta, \sigma, \gamma_{(i)}, Y) = \pi(\gamma_i|\beta, \gamma_{(i)}),\ i = 1, \ldots, p, \end{gathered} \tag{13}$$

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. The most costly step in simulating (13) is the generation of β from

$$\pi(\beta|\sigma, \gamma, Y) = N_p((X'X + \sigma^2(D_\gamma R_\gamma D_\gamma)^{-1})^{-1}X'Y, \sigma^2(X'X + (D_\gamma R_\gamma D_\gamma)^{-1})^{-1}), \tag{14}$$

which requires recomputing $(X'X+\sigma^2(D_\gamma R_\gamma D_\gamma)^{-1})^{-1}$ on the basis of new values of σ^2 and γ. This can be done quickly and efficiently by using the Cholesky decomposition (see Thisted (1988)). We note that as a result of this step, $O(p^3)$ operations are required to generate each value of γ.

When $v_{0\gamma_{(i)}} = 0$, the Gibbs sampler implementation (13) cannot be used to simulate (11) because simulation schemes such as (13) generate reducible, and hence nonconvergent, Markov chains. Effectively, the Gibbs sampler gets stuck when it generates a value $\beta_i = 0$. To avoid this problem, Geweke (1996) proposed an alternative implementation which jointly draws (γ_i, β_i) one at a time given σ and the other (γ_j, β_j) pairs. This implementation might also be preferable for the case $v_{0\gamma_{(i)}} > 0$ with $v_{1\gamma_{(i)}}/v_{0\gamma_{(i)}}$ chosen extremely large, which leads to very slow convergence of the Markov chain generated by (13). In this case, the Geweke (1996) alternative seems to offer improved convergence rates at the expense of computational speed. Carlin and Chib (1995), Green (1995) and Phillips and Smith (1996) have proposed alternative, computationally intensive MCMC methods which may also be used to simulate (11) for the case $v_{0\gamma_{(i)}} = 0$.

4. Conjugate Hierarchical Setups

In this section, we consider the special case of the Section 2 model when $\Upsilon_{(\sigma,\gamma)}$ in (2) is of the form

$$\pi(\beta|\sigma,\gamma) = N_p(0, \sigma^2 D^*_\gamma R_\gamma D^*_\gamma), \tag{15}$$

where analogously to (6), D^*_γ is diagonal and R_γ is a correlation matrix. We denote the ith diagonal element of D^{*2}_γ by

$$(D^{*2}_\gamma)_{ii} = \begin{cases} v^*_{0\gamma_{(i)}} & \text{when} \quad \gamma_i = 0 \\ v^*_{1\gamma_{(i)}} & \text{when} \quad \gamma_i = 1. \end{cases} \tag{16}$$

Because the conditional distribution of β and σ given γ is conjugate for (1), we refer to the resulting hierarchical mixture prior as "the conjugate prior". As opposed to the nonconjugate prior of the last section, β and σ can here be eliminated by routine integration from the full posterior $\pi(\beta, \sigma, \gamma|Y)$. As will be seen in subsequent sections, this feature yields attractive computational methods for posterior evaluation and exploration.

4.1. Conjugate hyperparameter settings

Under the prior (15), each component of β is again modeled as having come from a scale mixture of normals

$$\pi(\beta_i|\sigma,\gamma) = (1-\gamma_i)N(0, \sigma^2 v^*_{0\gamma_{(i)}}) + \gamma_i N(0, \sigma^2 v^*_{1\gamma_{(i)}}). \tag{17}$$

As with the nonconjugate prior, the idea is that $v^*_{0\gamma_{(i)}}$ and $v^*_{1\gamma_{(i)}}$ are to be set "small and large" respectively, so that when the data supports $\gamma_i = 0$ over $\gamma_i = 1$, then β_i is probably small enough so that X_i will not be needed in the model. However, the way in which $v^*_{0\gamma_{(i)}}$ and $v^*_{1\gamma_{(i)}}$ determine "small and large" is affected by the unknown value of σ, thereby making specification more difficult than in the nonconjugate case.

If σ were known, the nonconjugate and conjugate priors would be simple reparametrizations of each other according to $v_{0\gamma_{(i)}} = \sigma^2 v^*_{0\gamma_{(i)}}$ and $v_{1\gamma_{(i)}} = \sigma^2 v^*_{1\gamma_{(i)}}$. Thus, if a reasonable estimate $\hat{\sigma}^2$ of σ^2 were available, perhaps a least squares estimate based on the data, the practical significance strategy from Section 3.1 could be used to first select $v_{0\gamma_{(i)}}$ and $v_{1\gamma_{(i)}}$, and then $v^*_{0\gamma_{(i)}}$ and $v^*_{1\gamma_{(i)}}$ could be obtained from

$$v^*_{0\gamma_{(i)}} = v_{0\gamma_{(i)}}/\hat{\sigma}^2 \text{ and } v^*_{1\gamma_{(i)}} = v_{1\gamma_{(i)}}/\hat{\sigma}^2. \tag{18}$$

Of course, no matter how $v^*_{0\gamma_{(i)}}$ and $v^*_{1\gamma_{(i)}}$ are chosen, the conjugate and nonconjugate priors are different. Indeed, in the conjugate case, the marginal distribution of β_i given γ is

$$\pi(\beta_i|\gamma) = (1-\gamma_i)T(\nu, 0, \lambda_\gamma v^*_{0\gamma_{(i)}}) + \gamma_i T(\nu, 0, \lambda_\gamma v^*_{1\gamma_{(i)}}), \tag{19}$$

where $T(\nu, 0, \lambda_\gamma v^*_{j\gamma_{(i)}})$ is the t distribution with ν degrees of freedom and scale parameter $\lambda_\gamma v^*_{j\gamma_{(i)}}$. For a chosen threshold of practical significance $\delta_{i\gamma}$, the pdf $\pi(\beta_i|\gamma_{(i)}, \gamma_i = 0) = T(\nu, 0, \lambda_\gamma v^*_{0\gamma_{(i)}})$ is larger than the pdf $\pi(\beta_i|\gamma_{(i)}, \gamma_i = 1) = T(\nu, 0, \lambda_\gamma v^*_{1\gamma_{(i)}})$ precisely on the interval $(-\delta_{i\gamma}, \delta_{i\gamma})$, when $v_{0\gamma_{(i)}}$ and $v_{1\gamma_{(i)}}$ satisfy

$$(v^*_{0\gamma_{(i)}}/v^*_{1\gamma_{(i)}})^{\nu/(\nu+1)} = [v^*_{0\gamma_{(i)}} + \delta^2_{i\gamma}/(\nu\lambda_\gamma)]/[v^*_{1\gamma_{(i)}} + \delta^2_{i\gamma}/(\nu\lambda_\gamma)]. \tag{20}$$

Although one could choose $v^*_{0\gamma_{(i)}}$ and $v^*_{1\gamma_{(i)}}$ to satisfy (20), the strategy described in the previous paragraph provides a simple, good approximation. Presumably λ_γ would be set equal to the prior estimate $\hat{\sigma}^2$, since as ν gets large, the priors (17) and (19) become similar when $\lambda_\gamma \equiv \sigma^2$.

A special case which has received substantial attention in the literature (see, for example, Clyde, DeSimone and Parmigiani (1996), Raftery, Madigan and Hoeting (1993) and Smith and Kohn (1996)), is the conjugate formulation (15) with $v^*_{0\gamma_{(i)}} \equiv 0$ and $v^*_{1\gamma_{(i)}}$ chosen to be consistent with reasonable values of β_i. Under this choice (19) becomes

$$\pi(\beta_i|\sigma, \gamma_{(i)}) = (1-\gamma_i)I_0 + \gamma_i T(\nu, 0, \lambda_\gamma v^*_{1\gamma_{(i)}}), \tag{21}$$

where I_0 is a point mass at 0. Just as for the nonconjugate case (9), the threshold of practical significance is $\delta_{i\gamma} \equiv 0$, corresponding to the preference that any $\beta_i \neq 0$

be included in the model. Again, this criterion will select β_i on the basis of how well they can be distinguished from 0 rather than their absolute size. As in (10) for the nonconjugate case, it may also be convenient to represent the conditional prior (15) as $\pi(\beta|\gamma) = \pi(\beta_\gamma|\gamma)\pi(\beta_{\bar\gamma}|\gamma)$ where β_γ and $\beta_{\bar\gamma}$ are subvectors of β such that

$$\pi(\beta_\gamma|\gamma) = N_{q_\gamma}(0, \sigma^2 D^*_{1\gamma} R_{1\gamma} D^*_{1\gamma}) \text{ and } \pi(\beta_{\bar\gamma} = 0|\gamma) = 1. \tag{22}$$

Another potentially valuable specification of the conjugate 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 $v^*_{0\gamma_{(i)}} = 1$ and $v^*_{1\gamma_{(i)}} = K > 0$ 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 (1995).

Finally, the choice of a prior correlation matrix R_γ for the conjugate prior entails the same considerations as in the nonconjugate case. Again, appealing choices are $R_\gamma \equiv I$, $R_\gamma \propto (X'X)^{-1}$ for $v^*_{0\gamma_{(i)}} > 0$ and $R_{1\gamma} \propto (X'_\gamma X_\gamma)^{-1}$ for $v^*_{0\gamma_{(i)}} \equiv 0$.

4.2. Eliminating β and σ from $\pi(\beta, \sigma, \gamma|Y)$

The principal advantage of using the conjugate hierarchical prior is that it enables analytical margining out of β and σ from $\pi(\beta, \sigma, \gamma|Y) = f(Y|\beta, \sigma)\pi(\beta|\sigma, \gamma)$ $\pi(\sigma)\pi(\gamma)$. We consider the two cases $v^*_{0\gamma_{(i)}} > 0$ and $v^*_{0\gamma_{(i)}} \equiv 0$ separately.

When $v^*_{0\gamma_{(i)}} > 0$, combining the likelihood from (1) with the priors (3) and (15) yields the joint posterior

$$\begin{aligned}\pi(\beta, \sigma, \gamma|Y) \propto \quad & \sigma^{-(n+p+\nu+1)}|D^*_\gamma R_\gamma D^*_\gamma|^{-1/2}\exp\{-\frac{1}{2\sigma^2}|\tilde Y - \tilde X\beta|^2\} \\ & \exp\{-\frac{\nu\lambda}{2\sigma^2}\}\pi(\gamma),\end{aligned}$$

where

$$\tilde Y = \begin{bmatrix} Y \\ 0 \end{bmatrix} \qquad \text{and} \qquad \tilde X = \begin{bmatrix} X \\ (D^*_\gamma R_\gamma D^*_\gamma)^{-1/2} \end{bmatrix}. \tag{23}$$

Integrating out β and σ yields

$$\pi(\gamma|Y) \propto g(\gamma) \equiv |\tilde X'\tilde X|^{-1/2}|D^*_\gamma R_\gamma D^*_\gamma|^{-1/2}(\nu\lambda + S^2_\gamma)^{-(n+\nu)/2}\pi(\gamma), \tag{24}$$

where

$$S_\gamma^2 = \tilde{Y}'\tilde{Y} - \tilde{Y}'\tilde{X}(\tilde{X}'\tilde{X})^{-1}\tilde{X}'\tilde{Y} = Y'Y - Y'X(X'X + (D_\gamma^* R_\gamma D_\gamma^*)^{-1})^{-1}X'Y. \quad (25)$$

When $v_{0\gamma_{(i)}}^* \equiv 0$, combining the likelihood from (1) with the priors (3) and (22) yields the joint posterior

$$\pi(\beta, \sigma, \gamma|Y) \propto \quad \sigma^{-(n+q_\gamma+\nu+1)}|D_{1\gamma}^* R_{1\gamma} D_{1\gamma}^*|^{-1/2}\exp\{-\frac{1}{2\sigma^2}|\tilde{Y} - \tilde{X}_\gamma\beta_\gamma|^2\}$$
$$\exp\{-\frac{\nu\lambda}{2\sigma^2}\}\pi(\gamma),$$

where

$$\tilde{Y} = \begin{bmatrix} Y \\ 0 \end{bmatrix} \qquad \text{and} \qquad \tilde{X}_\gamma = \begin{bmatrix} X_\gamma \\ (D_{1\gamma}^* R_{1\gamma} D_{1\gamma}^*)^{-1/2} \end{bmatrix}, \quad (26)$$

and β_γ and X_γ are defined at the end of Section 3.1. Integrating out β_γ and σ yields

$$\pi(\gamma|Y) \propto g(\gamma) \equiv |\tilde{X}_\gamma'\tilde{X}_\gamma|^{-1/2}|D_{1\gamma}^* R_{1\gamma} D_{1\gamma}^*|^{-1/2}(\nu\lambda + S_\gamma^2)^{-(n+\nu)/2}\pi(\gamma), \quad (27)$$

where

$$S_\gamma^2 = \tilde{Y}'\tilde{Y} - \tilde{Y}'\tilde{X}_\gamma(\tilde{X}_\gamma'\tilde{X}_\gamma)^{-1}\tilde{X}_\gamma'\tilde{Y} = Y'Y - Y'X_\gamma(X_\gamma'X_\gamma + (D_{1\gamma}^* R_{1\gamma} D_{1\gamma}^*)^{-1})^{-1}X_\gamma'Y. \quad (28)$$

Note that $g(\gamma)$ in (24) and (27) only gives $\pi(\gamma|Y)$ up to a normalization constant. To obtain this constant exactly would require evaluation of the sum of $g(\gamma)$ over all possible γ values. However, as will be seen in Section 4.5 below, this normalization constant can be estimated by sampling from the posterior.

4.3. Some fast updating schemes

When either

$$v_{0\gamma_{(i)}}^* > 0 \text{ and } R_\gamma \equiv I, \quad (29)$$

or

$$v_{0\gamma_{(i)}}^* \equiv 0 \text{ and } D_{1\gamma}^* R_{1\gamma} D_{1\gamma}^* = c(X_\gamma'X_\gamma)^{-1}, \quad (30)$$

the value of $g(\gamma)$ can be rapidly updated as γ is changed one component at a time. As will be seen in Sections 4.4 and 4.5, these rapid updating schemes can be used to speed up algorithms for evaluating and exploring the posterior $\pi(\gamma|Y)$.

When condition (29) holds, fast updating of $g(\gamma)$ can be obtained from the Chambers (1971) regression updating algorithm. This is based on the representation of $g(\gamma)$ in (24) and (25) as

$$g(\gamma) = \Big(\prod_{i=1}^{p} T_{ii}^2[(1-\gamma_i)v_{0\gamma_{(i)}}^* + \gamma_i v_{1\gamma_{(i)}}^*]\Big)^{-1/2}(\nu\lambda + Y'Y - W'W)^{-(n+\nu)/2}\pi(\gamma), \quad (31)$$

where $T'T = \tilde{X}'\tilde{X}$ for T upper triangular and $W = T'^{-1}\tilde{X}'\tilde{Y}$. T may be obtained by the Cholesky decomposition. The representation (31) follows by noting that $|\tilde{X}'\tilde{X}|^{-1/2} = |T'T|^{-1/2} = |T|^{-1} = \prod_{i=1}^{p} T_{ii}^{-1}$, $|D_\gamma^* R D_\gamma^*|^{-1/2} = |D_\gamma^*|^{-1} = (\prod_{i=1}^{p}[(1-\gamma_i)v_{0\gamma(i)}^* + \gamma_i v_{1\gamma(i)}^*])^{-1/2}$ when $R_\gamma \equiv I$, and $\tilde{Y}'\tilde{X}(\tilde{X}'\tilde{X})^{-1}\tilde{X}'\tilde{Y} = \tilde{Y}'\tilde{X}(T'T)^{-1}\tilde{X}'\tilde{Y} = W'W$. Chambers (1971) provides an algorithm for fast updating of T and W when a row is added or deleted from $(\tilde{X}, \tilde{Y})$. However, it is straightforward to see that updating $(\tilde{X}'\tilde{X}, \tilde{X}'\tilde{Y})$, and hence (T, W), based only on changing one component of γ, corresponds to having a row added to or deleted from $(\tilde{X}, \tilde{Y})$ in (23). Thus, the Chambers (1971) algorithm may be used to update (31). This algorithm requires $O(p^2)$ operations per update.

When condition (30) holds, Smith and Kohn (1996) observed that fast updating of $g(\gamma)$ can be obtained as follows. This is based on the representation of $g(\gamma)$ in (27) and (28) as

$$g(\gamma) = (1+c)^{-q_\gamma/2}(\nu\lambda + Y'Y - (1+1/c)^{-1}W'W)^{-(n+\nu)/2}\pi(\gamma), \qquad (32)$$

where $T'T = X_\gamma'X_\gamma$ for T upper triangular and $W = T'^{-1}X_\gamma'Y$. T may be obtained by the Cholesky decomposition. The representation (32) follows by noting that when $D_{1\gamma}^* R_{1\gamma} D_{1\gamma}^* = c(X_\gamma'X_\gamma)^{-1}$, $|\tilde{X}_\gamma'\tilde{X}_\gamma|^{-1/2}|D_{1\gamma}^* R_{1\gamma} D_{1\gamma}^*|^{-1/2} = (1+c)^{-q_\gamma/2}$ and $Y'X_\gamma(X_\gamma'X_\gamma + (D_{1\gamma}^* R_{1\gamma} D_{1\gamma}^*)^{-1})^{-1}X_\gamma'Y = (1+1/c)^{-1}W'W$. Dongarra, Moler, Bunch and Stewart (1979), Ch. 10 provide an algorithm for fast updating of T, and hence W, whenever $X_\gamma'X_\gamma$ is changed by only one row and column. But this is precisely what happens when only one component of γ is changed. Thus, the algorithm of Dongarra et al. (1979) may be used to update (32). This algorithm requires $O(q_\gamma^2)$ operations per update, where γ is the changed value.

4.4 Exhaustive calculation of $\pi(\gamma|Y)$

Under the conjugate hierarchical prior, exhaustive calculation of $\pi(\gamma|Y)$ is feasible 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, under conditions (29) or (30), the calculation of $g(\gamma)$ for every γ value can be substantially speeded up by using the updating schemes described in Section 4.3.

This can be done by exploiting an ordering of the 2^p γ values where consecutive γ's differ by just one component. Such an ordering is provided by the *Gray Code* (see Press, Teukolsky, Vetterling and Flannery (1994)). After computing $g(\gamma)$, T and W for an initial γ value, subsequent values of T and W can be obtaining with the appropriate fast updating scheme by proceeding in the Gray Code order. As each new value of T and W is obtained, $g(\gamma)$ in (31) or (32) can be quickly recomputed. Related applications of the Gray Code for exhaustive computation are described in Diaconis and Holmes (1994).

By using sequential updating as γ is varied according to the Gray Code, this exhaustive calculation is feasible for p less than about 25. For example, using the fast updating algorithm with condition (29), performing the computations on a Sun Sparcstation 10 took about 7 seconds for $p = 15$, and about 5 minutes for $p = 20$. The entire calculation under (29) requires $O(2^p p^2)$ operations since each update requires $O(p^2)$ operations. The entire calculation under (30) is apt to be even faster since each update requires fewer operations.

A referee has pointed out to us that a brute force computation going through the models in any order requires $O(2^p p^3)$ operations. Thus, exhaustive calculation by brute force should be feasible in any problem with p less than about 20. In terms of feasibility, the Gray code approach is only buying us models with about 20-25 regressors. Of course, when possible, the Gray code approach will always be substantially faster.

Finally, note that there is a possibility for round-off error buildup with these sequential computations. In particular, round-off error can occur with the fast updating algorithms when $v^*_{1\gamma_{(i)}}/v^*_{0\gamma_{(i)}}$ is set extremely large under (29), or when $X'X$ is extremely ill-conditioned under (30). In any case, we recommend that the final value of $g(\gamma)$ be fully recomputed using the Cholesky decomposition to check that no round-off error has occurred. In our experience, we have always found complete agreement using double precision.

4.5. Conjugate MCMC exploration of the posterior

For large p, exhaustive calculation of $\pi(\gamma|Y)$ is not feasible. However, it is still possible to use MCMC methods to search for high probability γ values. The availability of $g(\gamma)$ in (24), (27), (31) and (32) allows for the easy construction of MCMC algorithms (described below) for simulating a Markov chain

$$\gamma^{(1)}, \ldots, \gamma^{(K)} \tag{33}$$

which is converging in distribution to $\pi(\gamma|Y)$. Although similar in spirit to the simulation search in Section 3.2, the sequence (33) is a Markov chain generated directly from the conditional form $g(\gamma)$, whereas the sequence (12) is a subsequence of the auxiliary Markov chain (11) obtained by applying the Gibbs sampler to an expression for the full posterior.

Just as for the sequence (12), the sequence (33) will, in many cases, have the property that high probability γ values will appear more quickly than low probability values, thereby facilitating the exploration for the more "promising" models. Indeed, the empirical frequencies of the γ values will be consistent estimates of their probability under $\pi(\gamma|Y)$. As with (12), the length of the sequence (33) can be much smaller than 2^p and still serve to identify at least

some of the high probability values. Even for moderate p, where exhaustive calculation is feasible, it may be cost effective (in terms of time) to instead use MCMC search when the goal is simply the identification of high probability γ values.

Instead of using empirical frequency estimates as in the nonconjugate setup, an attractive feature of the conjugate prior is the availability of the exact $g(\gamma)$ values which provides useful information about $\pi(\gamma|Y)$. First of all, the exact relative probability of two values γ_0 and γ_1 is obtained as $g(\gamma_0)/g(\gamma_1)$. This allows for the more accurate identification of the high probability models among those selected. Furthermore, only minimal additional effort is required to obtain these relative probabilities since $g(\gamma)$ must be calculated for each of the visited γ values in the execution of the MCMC algorithms described in Sections 4.5.1 and 4.5.2.

The availability of $g(\gamma)$ also makes it possible to estimate the normalizing constant C,

$$\pi(\gamma|Y) = Cg(\gamma), \tag{34}$$

as follows. Let A be a preselected subset of γ values and let $g(A) = \sum_{\gamma\in A} g(\gamma)$ so that $\pi(A|Y) = Cg(A)$. For a simulation of (33), a consistent estimate of C is obtained by

$$\hat{C} = \frac{1}{g(A)K}\sum_{k=1}^{K} I_A(\gamma^{(k)}), \tag{35}$$

where $I_A(\,)$ is the indicator of the set A. Note that if (33) were an uncorrelated sequence, then $\mathrm{Var}(\hat{C}) = (C^2/K)(1-\pi(A|Y))/\pi(A|Y)$ suggesting that (35) will be a better estimator of C when $\pi(A|Y)$ is large. It is also desirable to choose A such that $I_A(\gamma^{(k)})$ will be inexpensive to evaluate. Because the selection of A cannot depend on the simulation sequence used to evaluate (35), we choose A to be a set of γ values visited by a preliminary simulation of (33). As can be seen in Section 5, these estimates can be remarkably accurate.

Inserting $\hat{C}$ into (34) yields improved estimates of the probability of individual γ values,

$$\hat{\pi}(\gamma|Y) = \hat{C}g(\gamma), \tag{36}$$

as well as an estimate of the total visited probability,

$$\hat{\pi}(B|Y) = \hat{C}g(B), \tag{37}$$

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

$$|(\hat{C}/C) - 1|. \tag{38}$$

Since $\hat{\pi}(\gamma|Y)/\pi(\gamma|Y) \equiv \hat{C}/C$, (38) measures of the uniform accuracy of the probability estimates. It also measures the total probability discrepancy since $\sum_\gamma |\hat{\pi}(\gamma|Y) - \pi(\gamma|Y)| = |\hat{C} - C| \sum_\gamma g(\gamma) = |(\hat{C}/C) - 1|$.

In the next two subsections, we describe various MCMC algorithms which should be useful for generating a Markov chain (33). These algorithms are obtained as variants of the Gibbs sampler and Metropolis-Hastings algorithms applied to $g(\gamma)$. In Section 5 we compare the relative merits of the various algorithms on simulated sample problems.

4.5.1. Gibbs sampling algorithms

A variety of MCMC algorithms for generating the Markov chain (33) can be conveniently obtained by applying the Gibbs sampler to $g(\gamma)$. As opposed to the nonconjugate case, obtaining a convergent Markov chain does not require $\upsilon_{0\gamma_{(i)}} > 0$. Perhaps the simplest Gibbs sampler implementation is obtained by generating each γ value componentwise from the full conditionals,

$$\gamma_i|\gamma_{(i)}, Y \quad i = 1, \ldots, p, \tag{39}$$

(recall $\gamma_{(i)} = (\gamma_1, \gamma_2, \ldots, \gamma_{i-1}, \gamma_{i+1}, \ldots, \gamma_p)$) where the γ_i may be drawn in any fixed or random order. The results of Liu, Wong and Kong (1994) suggest that, because of the margining out of β and σ, the sequence (33) obtained by this algorithm should converge faster than the sequence (12), making it more effective on a per iteration basis for learning about $\pi(\gamma|Y)$.

The generation of the components in (39) in conjunction with $g(\gamma)$ in (24) or (27), can be obtained trivially as simulations of Bernoulli draws. Furthermore, under conditions (29) or (30), the required sequence of Bernoulli probabilities can be computed fast and efficiently by exploiting the appropriate updating scheme for $g(\gamma)$ from Section 4.3. To see this, note that the Bernoulli probabilities are simple functions of the ratio

$$\frac{\pi(\gamma_i = 1, \gamma_{(i)}|Y)}{\pi(\gamma_i = 0, \gamma_{(i)}|Y)} = \frac{g(\gamma_i = 1, \gamma_{(i)})}{g(\gamma_i = 0, \gamma_{(i)})}. \tag{40}$$

At each step of the iterative simulation from (39), one of the values of $g(\gamma)$ in (40) will be available from the previous component simulation. The other value of $g(\gamma)$ can then be obtained by using the appropriate updating scheme from Section 4.3. Since γ is varied by exactly one component. As a valuable byproduct, this sequence of $g(\gamma)$ values can be stored to obtain exact relative probabilities or probability estimates via (35), (36) and (37) of the visited γ values.

As noted by Smith and Kohn (1995), this Gibbs sampler can be substantially faster under condition (30) where fast updating to γ requires $O(q_\gamma^2)$ operations,

than under (29) where fast updating requires $O(p^2)$ operations. This is likely to happen when $\pi(\gamma|Y)$ is concentrated on those γ for which q_γ is small, namely the parsimonious models. This advantage could be especially pronounced in large problems with many useless predictors.

Simple variants of the componentwise Gibbs sampler 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 Gibbs sampler can be obtained by drawing the components of γ in groups, rather than one at a time. Let $\{I_k\}$, $k = 1, \ldots, m$ be a partition of $\{1, \ldots, p\}$ so that, $I_k \subseteq \{1, \ldots, p\}$, $\cup I_k = \{1, \ldots, p\}$ 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 $\gamma_{(I_k)} = \{\gamma_i | i \notin I_k\}$. The grouped Gibbs sampler generates the Markov chain (33) by iterative simulation from

$$\gamma_{I_k} | \gamma_{(I_k)}, Y \quad k = 1, \ldots, m. \tag{41}$$

As before, when condition (29) or (30) holds, the conditional distribution in (41) can be computed fast and efficiently by exploiting the appropriate updating scheme for $g(\gamma)$ from Section 4.3. This can be done by computing the conditional probabilities of each γ_{I_k} in the Gray Code order.

The potential advantage of using the grouped Gibbs sampler is improved convergence of the Markov chain (33). 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 Gibbs sampler can be obtained by generating the γ_{I_k} in a different fixed or random order.

4.5.2. Metropolis-Hastings algorithms

Another class of MCMC algorithms for generating the Markov chain (33) from $g(\gamma)$ is the Metropolis-Hastings (MH) algorithms. To construct an MH algorithm, one begins with a Markov transition kernel, say $q(\gamma^0, \gamma^1)$, called a proposal. For each γ^0 value, $q(\gamma^0, \gamma^1)$ is a probability distribution over γ^1 values. For a given proposal $q(\gamma^0, \gamma^1)$, the corresponding MH algorithm generates each transition from $\gamma^{(j)}$ to $\gamma^{(j+1)}$ in (33) as follows.

1. Generate a candidate value γ^* with probability distribution $q(\gamma^{(j)}, \gamma^*)$.
2. Set $\gamma^{(j+1)} = \gamma^*$ with probability

$$\alpha^{MH}(\gamma^{(j)}, \gamma^*) = \min \left\{ \frac{q(\gamma^*, \gamma^{(j)})}{q(\gamma^{(j)}, \gamma^*)} \frac{g(\gamma^*)}{g(\gamma^{(j)})}, 1 \right\}. \tag{42}$$

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

Under weak conditions on $q(\gamma^0, \gamma^1)$, the sequence (33) obtained by this algorithm will be a Markov chain which is converging to $\pi(\gamma|Y)$ (see Chib and Greenberg (1995)) or Tierney (1994).

When condition (29) or (30) holds, the acceptance probability α^{MH} in (42) can be computed fast and efficiently by exploiting the appropriate updating scheme for $g(\gamma)$ described in Section 4.3. Just as for the Gibbs sampler described in Section 4.5.1, when $\pi(\gamma|Y)$ is concentrated on those γ for which q_γ is small, such an MH algorithm can be substantially faster under condition (30) than under (29).

A special class of MH algorithms, the Metropolis algorithms, are obtained from the class of proposals $q(\gamma^0, \gamma^1)$ which are symmetric in (γ^0, γ^1). For this class, the form of (42) simplifies to

$$\alpha^M(\gamma^{(j)}, \gamma^*) = \min\left\{\frac{g(\gamma^*)}{g(\gamma^{(j)})}, 1\right\}. \tag{43}$$

Perhaps the simplest symmetric proposal is

$$q(\gamma^0, \gamma^1) = 1/p \quad \text{if} \quad \sum_1^p |\gamma_i^0 - \gamma_i^1| = 1. \tag{44}$$

This yields the Metropolis algorithm

1. Generate a candidate γ^* by randomly changing one component of $\gamma^{(j)}$.
2. Set $\gamma^{(j+1)} = \gamma^*$ with probability $\alpha^M(\gamma^{(j)}, \gamma^*)$. 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 (1993) for model averaging, and was suggested for the SSVS context by Clyde and Parmigiani (1994) based on an earlier version of this paper.

It is interesting to observe that the algorithm obtained by replacing $\alpha^M(\gamma^{(j)}, \gamma^*)$ in (43) with

$$\alpha^G(\gamma^{(j)}, \gamma^*) = \frac{g(\gamma^*)}{g(\gamma^{(j)}) + g(\gamma^*)}, \tag{45}$$

is a componentwise Gibbs sampler (39) which randomly chooses the next component to generate. Because it will always be the case that $\alpha^M(\gamma^{(j)}, \gamma^*) > \alpha^G(\gamma^{(j)}, \gamma^*)$, the Metropolis algorithm is more likely to move at each step. Liu (1995) has shown that this Gibbs sampler is inferior to the Metropolis algorithm under the asymptotic variance criterion of Peskun (1995), and has proposed the Metropolized Gibbs sampler, an improved alternative which may also be interest of here.

The proposal (44) is a special case of the class of symmetric proposals of the form

$$q(\gamma^0, \gamma^1) = q_d \quad \text{if} \quad \sum_1^p |\gamma_i^0 - \gamma_i^1| = d. \tag{46}$$

Such proposals yield Metropolis algorithms of the form

1. Generate a candidate γ^* by randomly changing d components of $\gamma^{(j)}$ with probability q_d.
2. Set $\gamma^{(j+1)} = \gamma^*$ with probability $\alpha^M(\gamma^{(j)}, \gamma^*)$. 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 (44) which only moves locally, such algorithms require more computation per iteration.

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

$$q(\gamma^0, \gamma^1) = q_d \text{ if } \sum_1^p (\gamma_i^0 - \gamma_i^1) = d. \tag{47}$$

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.

5. Nonconjugate and Conjugate MCMC Performance

In this section, we illustrate and compare the performance of three MCMC methods for posterior exploration under the nonconjugate and the conjugate setups: the SSVS Gibbs algorithm (13) for the nonconjugate setup, denoted as NG (for nonconjugate Gibbs); the fixed order componentwise Gibbs algorithm (39) for the conjugate setup, denoted as CG (for conjugate Gibbs); and the simple Metropolis algorithm based on (44) for the conjugate setup, denoted as CM (for conjugate Metropolis). Although many other methods are available for comparison, these three algorithms capture some of the basic differences among the different MCMC choices for the nonconjugate and conjugate setups.

The performance of each of these methods is compared on three different, simulated variable selection problems. To facilitate performance comparisons, we used the same hyperparameter settings for all three methods: prior correlation matrix $R_\gamma \equiv I$, inverse Gamma parameters $\nu = 10$ and $\lambda = \hat{\sigma}^2$ (the least squares estimate based on the full model), and γ prior (4) with $w_i \equiv 0.5$ which yields

$\pi(\gamma) \equiv 1/2^p$. We used identical values of $\delta_{i\gamma} \equiv \delta = 1$, $v_{0\gamma_{(i)}} \equiv v_0 > 0$, $v_{1\gamma_{(i)}} \equiv v_1$, $v^*_{0\gamma_{(i)}} \equiv v^*_0 > 0$ and $v^*_{1\gamma_{(i)}} \equiv v^*_1$ across all models and coefficients. To match the nonconjugate and conjugate priors, we set $v^*_0 = v_0/\hat{\sigma}^2$ and $v^*_1 = v_1/\hat{\sigma}^2$ as suggested in (18).

5.1. Computational speed

We begin with a comparison of the computational speed of the NG, CG and CM algorithms. For each value of p, we simulated sample regression data $Y \sim N_{6p}(X\beta, \sigma^2 I)$ as in (1), using $X_1, \ldots, X_p$ i.i.d. $\sim N_{6p}(0, I)$, $\beta_i = 1.5i/p$ for $i = 1, \ldots, p$, and $\sigma = 2.5$. The algorithms were applied with the previously described settings and $v_1/v_0 = v^*_1/v^*_0 = 100$. For each such data set, we measured the time it took each algorithm to generate p components of γ, which we call one iteration. For the NG and CG algorithms, one iteration entails simulating all p γ_i's once, whereas for the CM algorithm, one iteration entails simulating p γ values, each of which differs from the preceding value by at most one component. We note that all three algorithms here require $O(p^3)$ operations per iteration. Nonetheless, their speeds appear to be different.

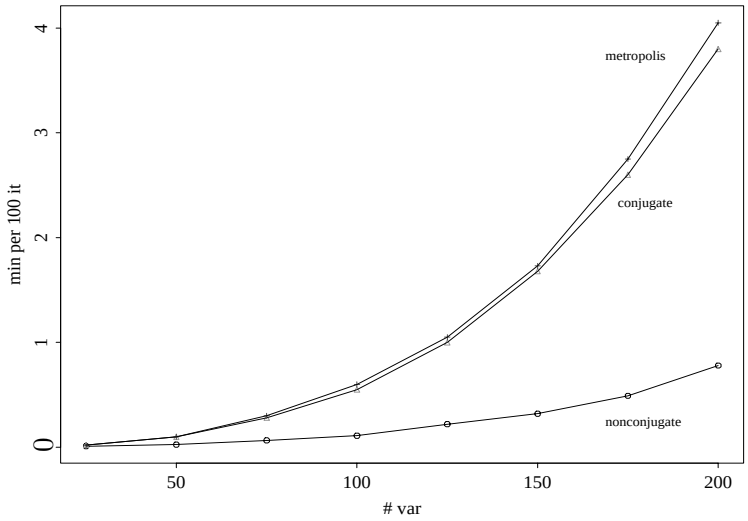

Figure 2. Time in minutes for 100 iterations versus p = number of X_i: ∘=NG, △=CG, +=CM.

Figure 2 plots minutes to compute 100 iterations versus p for each algorithm on a Sun SPARCstation 10 using a Fortran program compiled with the fast option. We use ∘ to denote NG values, △ to denote CG values, and + to denote CM values. Clearly, NG generates iterations much faster than CG and CM,

which appear to be roughly equivalent. Furthermore, the speed advantage of NG is more pronounced for large p. Nonetheless, it is practical to generate thousands of iterations for large problems with all three algorithms. For example, generating 1000 iterations of γ when $p = 100$, takes about 30 seconds using NG and about 5 minutes using CG and CM. When $p = 200$, it takes about 8 minutes using NG and about 40 minutes using CG and CM. It is interesting that analytical simplification has actually led to increased computational requirements per iteration for CG and CM.

Of course, the amount of information provided by these algorithms will depend on the dependence structure of the Markov chain output. To investigate this issue in the following sample problems, we compare the performance of NG, CG and CM for a fixed amount of execution time rather than a fixed number of iterations. To make the comparisons fair, we use one value of γ per iteration from the output of each of the three algorithms. Note that for CM this only uses every pth value of the output sequence. Alternatively, we could have used p values of γ per iteration for all three algorithms by taking each implicit new value of γ created each time a component γ_i is generated. Although this is a more efficient use of the output for CG, it is somewhat less satisfactory for NG which generates a new value of β for each iteration.

5.2. Three sample problems

We proceeded to compare the performance of NG, CG and CM on three simulated sample problems to illustrate the relative strengths of the MCMC approaches. In all three problems, we constructed $n = 180$ observations on $p = 15$ potential regressors. We used p small enough so that we could easily obtain the actual posterior characteristics for comparisons. In particular, we computed exact posterior distributions $\pi(\gamma|Y)$ under the conjugate prior using the exhaustive calculation based on the Gray Code sequencing described in Section 4.4. We also used very long runs of the NG algorithm to verify that the nonconjugate and conjugate posterior probabilities were close enough for our comparisons.

The first sample problem is a simple one in which the posterior is concentrated on a relatively small number of models. In this situation we find that all three algorithms perform quite well. The second problem is constructed to have severe multicollinearity. Here we see that all the algorithms perform surprisingly well, although the conjugate algorithms have an advantage. The third problem is designed to be very difficult by constructing data which provide very little coefficient information and by setting the variance ratio $\upsilon_1/\upsilon_0 = \upsilon_1^*/\upsilon_0^*$ large. In this situation, the conjugate algorithms are seen to have a substantial advantage.

5.2.1. A straightforward problem

For our first sample problem, we constructed $n = 180$ observations on $p = 15$ potential regressors by generating $Z_1, \ldots, Z_{15}, Z$ i.i.d. $\sim N_{180}(0, I)$ and setting $X_i = Z_i + 2Z$ for $i = 1, \ldots, 15$. This induced a pairwise correlation of about 0.8 among the X_i. We set evenly spaced values $\beta_i = 2i/15$ for $i = 1, \ldots, 15$, and set $\sigma = 2.5$. Our draw of $Y \sim N_{180}(X\beta, \sigma^2 I)$ resulted in least squares estimates of the β_i with classical standard errors ranging from .18 to .22 and t-statistics ranging from 1.1 to 11.7. For this problem, we ran NG, CG, and CM using the previously described settings with $v_1/v_0 = v_1^*/v_0^* = 100$.

We begin by comparing NG, CG and CM on the basis of how well each marginal probability $\pi(\gamma_i = 1|Y)$ is estimated by the empirical frequency of $\gamma_i = 1$ in a short simulation run. To account for computational speed differences, we ran each algorithm for the same amount of *time* rather than for the same number of iterations. We set the time to be that required to generate 500 iterations of NG (about one second). This entailed 290 iterations of both CG and CM.

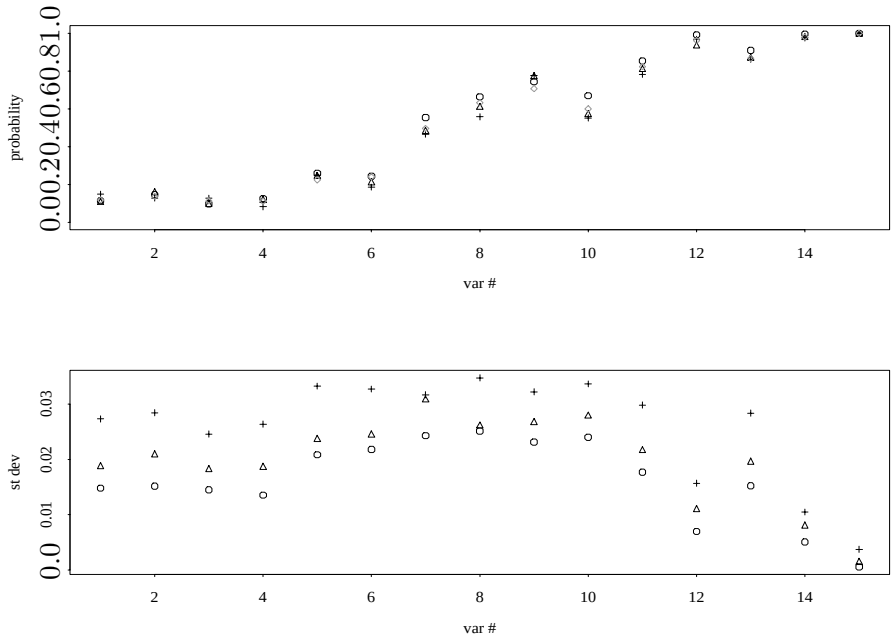

Figure 3. Straightforward problem, top panel: estimates of $\pi(\gamma_i = 1|Y)$ vs i for NG ($\circ$), CG ($\triangle$), and CM ($+$) and exact values from Gray Code ($\diamond$). Bottom panel: Monte Carlo standard errors.

The top panel of Figure 3 plots the estimates obtained by NG ($\circ$), CG ($\triangle$) and CM ($+$) on a single short run together with the actual $\pi(\gamma_i = 1|Y)$ values

($\diamond$) (under the conjugate prior). All three estimates are very close to the true values. As an interesting aside, we note that $\pi(\gamma_i = 1|Y)$ tends to increase with i, starting off close to 0 when β_i is small, increasing to about 0.5 when β_i is close to 1, and ending up at 1 when β_i is close to 2. This is exactly the desired effect of setting $\delta = 1$ in this problem.

As suggested by Geweke (1992) and Geyer (1992), we proceeded to measure the precision of these estimates by their Monte Carlo standard errors, $\mathrm{SE}(\bar{\gamma}_i) = ((1/K)\sum_{|h|<K}(1 - |h|/K)\phi_i(h))^{1/2}$, where ϕ_i is the autocovariance function of the γ_i sequence (see Brockwell and Davis (1991)) and $\bar{\gamma}_i$ is the empirical frequency of $\gamma_i = 1$ in a sequence of length K. To obtain ϕ_i, we simulated additional independent long runs of 25,000 iterations for NG, CG, and CM. In all cases ϕ_i dies off after 45 lags so we set $\phi_i(h) = 0$ for $h > 45$.

The bottom panel of Figure 3 plots Monte Carlo standard errors, $\mathrm{SE}(\bar{\gamma}_i)$, for one second runs of NG, CG and CM. From the plot we see that the NG estimates ($\circ$) have the smallest standard errors, followed by CG ($\triangle$), and lastly CM (+), although the difference between NG and CG is slight. Note that all the Monte Carlo standard errors are small with the largest being about 0.035. In this problem, the posterior is sufficiently well behaved that computational speed is the dominant performance factor. This suggests that in problems where p is large and the posterior is still reasonably informative, the computational efficiency of NG may render it superior to the other algorithms.

We next compare NG, CG and CM on the basis of what can be learned about the high posterior probability γ values in a moderately long simulation run. We set the time to be that required to generate 5000 iterations of NG (about ten seconds). This entailed 2900 iterations of CG and CM. Out of the $2^{15} = 32,768$ possible γ values, on such a run, NG, CG, and CM visited 984, 840, and 818 different values, respectively. These visited values accounted for 86.3%, 83.5% and 82.5% of the total (conjugate) posterior probability, respectively, and included virtually all of the high probability values. To see this, observe the plots in Figure 4 of the actual posterior probability for all 2^{15} γ values on the vertical axis. Points to the left of the vertical line correspond to γ values which were not drawn in our ten second runs and points to the right of the vertical line correspond to γ values which were drawn. Within each subset the posterior probabilities are plotted in ascending order. The top, middle, and bottom plots correspond to NG, CG, and CM respectively. It is clear from the plots that,in this problem, all three algorithms found virtually all the high posterior probability γ values.

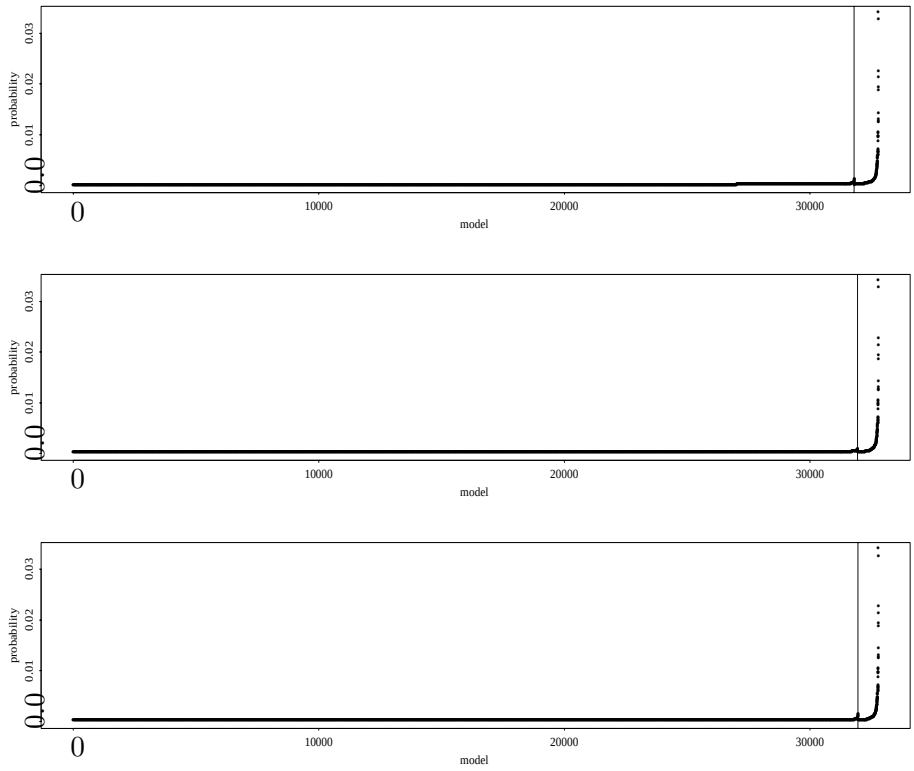

Figure 4. Straightforward problem, non-visited and visited model probabilities. NG, CG, and CM in top, middle and bottom panels respectively.

As described in Section 4.5, a valuable feature of using the conjugate prior is the availability of $g(\gamma)$ in (24) and (27). In addition to providing the exact relative posterior probabilities of the visited γ values, this allows for the estimation of the norming constant C for which $\pi(\gamma|Y) \equiv Cg(\gamma)$ in (34). Applying (35) to the CG output, with A determined by a very short preliminary run of 100 iterations, we obtained an estimate $\hat{C}$ for which $\hat{C}/C = .976$. This $\hat{C}$ provides estimates (36) of the probability of individual γ values, and an estimate (37) of the total visited probability with a uniform accuracy (38) of .024. For example, the estimate of the total visited probability is 81.5% which is remarkably close to the actual value of 83.5%. Such estimates are unavailable with the nonconjugate formulation.

5.2.2. A multicollinear problem

Our second sample problem is constructed to have severe and complicated multicollinearity. We start by constructing some of the X_i in the same manner

as in problem 5.2.1. Again generating $Z_1, \ldots, Z_{15}, Z$ i.i.d. $\sim N_{180}(0, I)$, we set $X_i = Z_i + 2Z$ for $i = 1, 3, 5, 8, 9, 10, 12, 13, 14, 15$ only. To induce strong multicollinearity, we then set $X_2 = X_1 + .15Z_2$, $X_4 = X_3 + .15Z_4$, $X_6 = X_5 + .15Z_6$, $X_7 = X_8 + X_9 - X_{10} + .15Z_7$, and $X_{11} = X_{14} + X_{15} - X_{12} - X_{13} + .15Z_{11}$. This construction resulted in a correlation of about .998 between X_i and X_{i+1} for $i = 1, 3, 5$. A similarly strong linear relationship was present within the sets (X_7, X_8, X_9, X_{10}) and $(X_{11}, X_{12}, X_{13}, X_{14}, X_{15})$.

To make the problem difficult we set $\beta = (1.5, 0, 1.5, 0, 1.5, 0, 1.5, -1.5, 0, 0, 1.5, 1.5, 1.5, 0, 0)$. Although we have not put X_2 directly in the model, it is so highly correlated with X_1 that we can only expect to conclude (since $\delta = 1$) that at least one of X_1 and X_2 is needed. Similarly, $X_3, X_5, (X_7 - X_8)$ and $(X_{11} + X_{12} + X_{13})$ are in the model but are highly correlated with $X_4, X_6, (X_9 - X_{10})$ and $(X_{14} + X_{15})$ respectively. Finally, we set $\sigma = 2.5$ as in problem 5.2.1. Our draw of $Y \sim N_{180}(X\beta, \sigma^2 I)$ resulted in least squares estimates of the β_i with classical standard errors ranging from 1.1 to 1.4, which were much larger than those of problem 5.2.1. Their t-statistics ranged from -2.7 to 3.5. We ran NG, CG and CM as before using $v_1/v_0 = v_1^*/v_0^* = 100$.

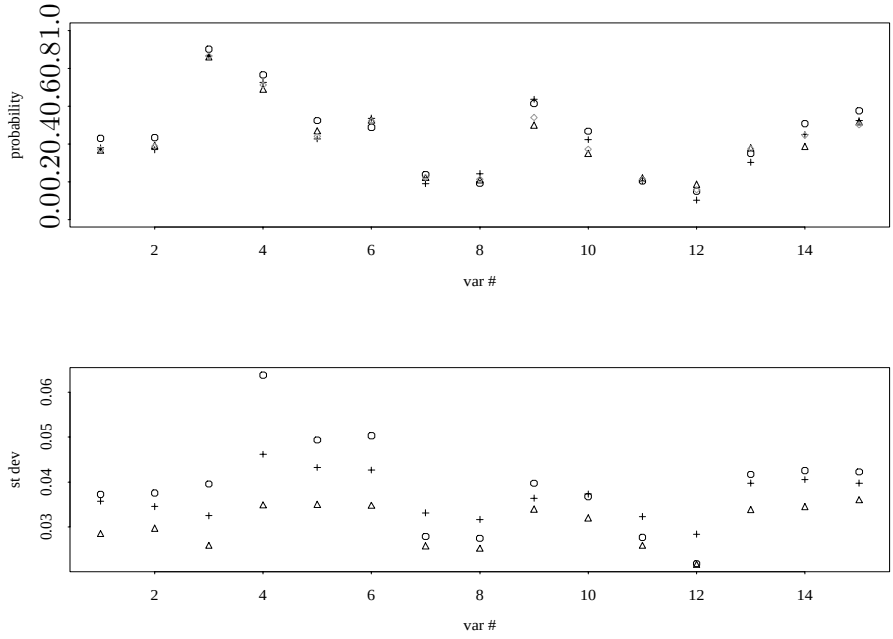

Figure 5. Multicollinear problem, top panel: estimates of $\pi(\gamma_i = 1|Y)$ vs i for NG ($\circ$), CG ($\triangle$), and CM ($+$) and exact values from Gray Code ($\diamond$). Bottom panel: Monte Carlo standard errors.

Proceeding as in problem 5.2.1, we compared the performance of NG, CG and CM on the basis of a short and a moderate simulation run. For the short run, we again set the time to be that required to generate 500 iterations of NG (about one second) which entailed 300 iterations of CG and CM. For the moderate run, we again set the time to be that required to generate 5000 iterations of NG (about ten seconds) which entailed 3000 iterations of CG and CM.

For the short run comparisons, the top panel of Figure 5 shows that, even in this problem, NG, CG and CM all yield very good estimates of the marginals $\pi(\gamma_i = 1|Y)$. Many of the probabilities are quite close to 0.5. Although one might be tempted to conclude from the marginals that there is little evidence for the inclusion of X_5 or X_6, we also computed the NG estimate of $\pi(\gamma_5 = 1$ and/or $\gamma_6 = 1|Y)$ which turned out to be 0.88. The Monte Carlo standard errors in the bottom panel of Figure 5 reveal that here, the conjugate methods, CG and CM, are superior to NG. Generally, CG appears to perform best.

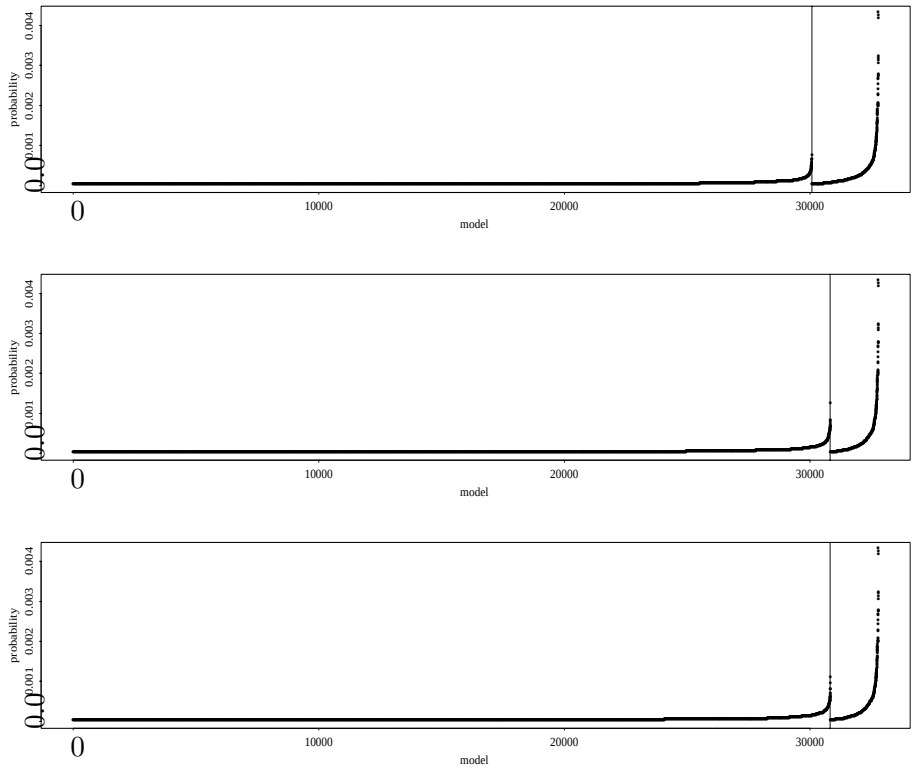

Figure 6. Multicollinear problem, non-visited and visited model probabilities. NG, CG, and CM in top, middle and bottom panels.

For the moderate run comparisons, Figure 6 illustrates the ability of all three algorithms to find the high probability models in this multicollinear problem. Again, NG, CG, and CM visited essentially all of the high probability models. Out of the 32,768 possibilities, NG, CG, and CM visited 2701, 1954 and 1952 different γ values accounting for 59.7%, 51.1% and 51.7% of the posterior probability . All three algorithms visited more γ values and less total probability than in problem 5.2.1. This is not surprising because the posterior probability $\pi(\gamma|Y)$ is less concentrated here.

Applying (35) to the CG output here, with A again determined by a very short preliminary run of 100 iterations, we obtained an estimate of the norming constant $\hat{C}$ for which $\hat{C}/C = 1.059$. This $\hat{C}$ provides estimates (36) of the probability of individual γ values, and an estimate (37) of the total visited probability with a uniform accuracy (38) of .059. The total visited probability estimate here is 54.1% as opposed to the actual value of 51.1%.

5.2.3. A weak information problem

Our third sample problem is identical to problem 5.2.1 with two important exceptions: $\sigma = 200$ rather than 2.5, and $\upsilon_1/\upsilon_0 = \upsilon_1^*/\upsilon_0^* = 2500$ rather than 100. The effect of such a large σ is to vastly diminish the information provided by the data for determining if $|\beta_i| > \delta = 1$. Indeed, our draw of $Y \sim N_{180}(X\beta, \sigma^2 I)$ here resulted in least squares estimates of the β_i with classical standard errors ranging from 13 to 16, which were much larger than in problems 5.2.1 and 5.2.2. Their t-statistics ranged from -1.25 to 1.48. In this case, the likelihood is relatively flat so that the posterior and prior are very similar. The effect of increasing υ_1/υ_0 and $\upsilon_1^*/\upsilon_0^*$ is to increase the separation between the components of the mixture prior for each β_i.

With υ_1/υ_0 large, we expect NG to perform poorly in this example. To see this, consider the extreme case where the likelihood in β is flat so that each conditional draw of β by NG is a draw from the prior (7). If γ_i starts at 0, then the next draw of β_i will be from the $N(0, \upsilon_0)$ distribution. With υ_1/υ_0 large, such a draw is, with very high probability unlikely to look like a draw from the $N(0, \upsilon_1)$ distribution so that the next draw of γ_i will very likely be a 0. Eventually we will get a draw from $N(0, \upsilon_0)$ so far out in the tail that the subsequent γ_i draw is 1. But then we start drawing from the β_i from the $N(0, \upsilon_1)$ and it becomes highly unlikely that a draw will look like it comes from the $N(0, \upsilon_0)$ distribution and γ_i is stuck at 1. As a result, the NG transition probabilities for γ_i from 0 to 1 and from 1 to 0 will be very small. This will cause NG to generate long sequences of 0's and long sequences of 1's, yielding slowly converging estimates.

Proceeding as in problems 5.2.1 and 5.2.2, we compared the performance of NG, CG and CM on the basis of a short and a moderate simulation run. For the short run, we again set the time to be that required to generate 500 iterations of NG (about one second) which entailed 320 iterations of CG and CM. For the moderate run, we again set the time to be that required to generate 5000 iterations of NG (about ten seconds) which entailed 3200 iterations of CG and CM.

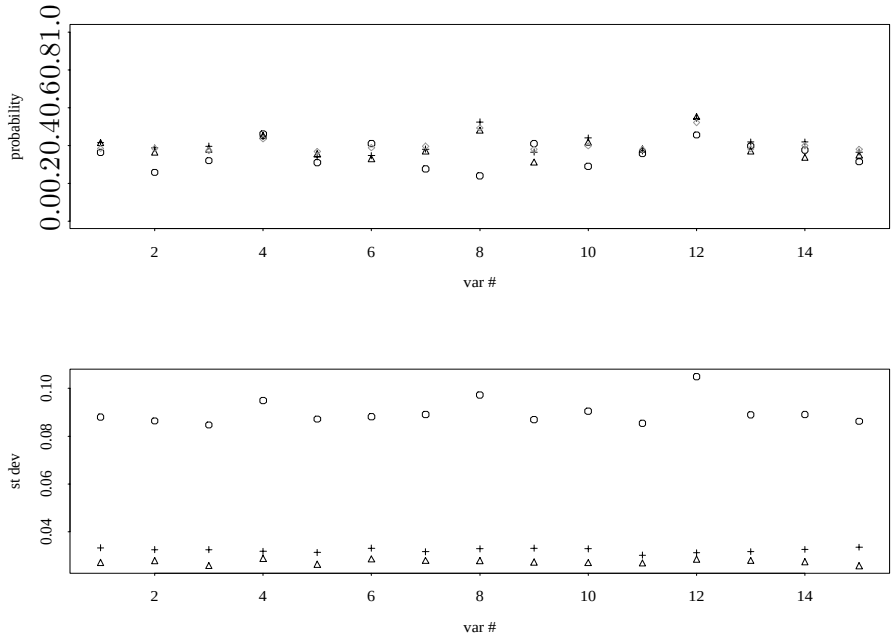

Figure 7. Noninformative problem, top panel: estimates of $\pi(\gamma_i = 1|Y)$ vs i for NG ($\circ$), CG ($\triangle$), and CM ($+$) and exact values from Gray Code ($\diamond$). Bottom panel: Monte Carlo standard errors.

The top panel of Figure 7 shows that with a short run, NG, CG and CM still produced extremely accurate estimates of the marginal probabilities $\pi(\gamma_i = 1|Y)$. Note that for this problem, the actual posterior probabilities are slightly less than the prior probabilities $w_i \equiv 0.5$, a consequence of setting $\upsilon_1/\upsilon_0 = \upsilon_1^*/\upsilon_0^*$ to be so large. The bottom panel of Figure 7 clearly shows the superiority of CG and CM over NG in this problem. The Monte Carlo standard errors for NG are around 0.1 while those of CG and CM are around 0.02. Given the noninformative nature of the problem, the methods perform remarkably well at estimating the marginal probabilities $\pi(\gamma_i = 1|Y)$.

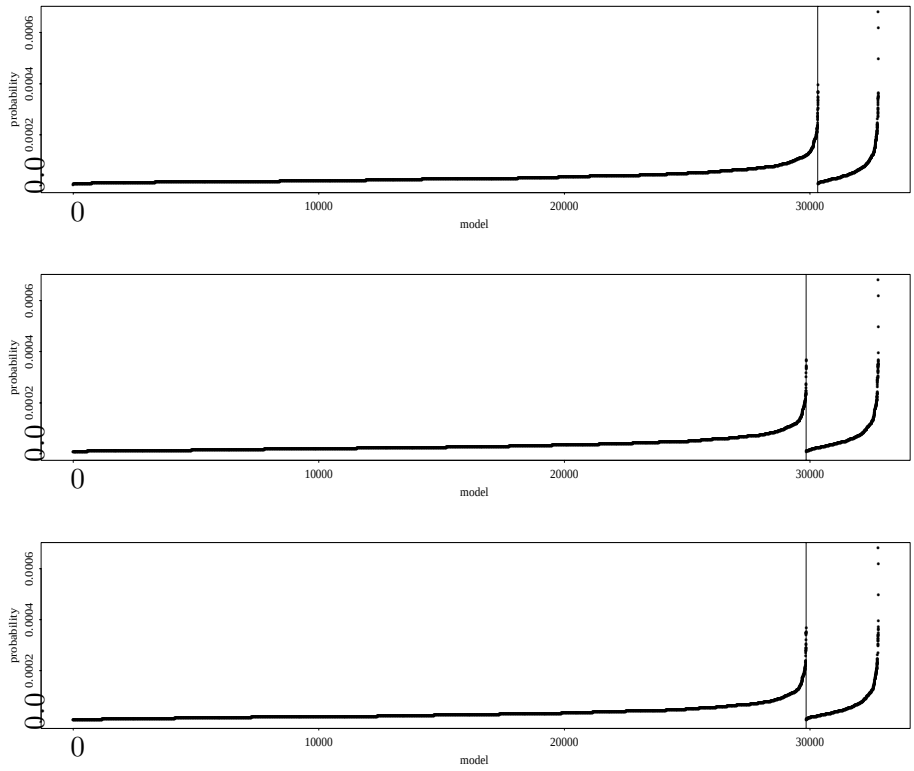

Figure 8. Noninformative problem, non-visited and visited model probabilities. NG, CG, and CM in top, middle and bottom panels.

Figure 8 illustrates how the methods search the model space in this problem. Here is a clear indication that the methods can fail. All three methods miss models which have relatively high probability. Indeed, NG, CG, and CM visit 2453, 2932, and 2937 distinct γ values, accounting for 13.3%, 16.5% and 16.3% of the total probability, respectively. This weak performance appears to be a consequence of a relatively flat posterior which provides much less information than the posteriors in problems 5.2.1 and 5.2.2. Of course, one might argue that the performance here is not so bad since many of the models with relatively high posterior probability were actually identified.

Applying (35) to the CG output here, with A again determined by a very short preliminary run of 100 iterations, we obtained an estimate of the norming constant $\hat{C}$ for which $\hat{C}/C = 1.250$. This $\hat{C}$ provides estimates (36) of the probability of individual γ values, and an estimate (37) of the total visited probability with a uniform accuracy (38) of .250. The total visited probability estimate here is 20.7% as opposed to the actual value of 16.5%.

6. Constructing Financial Index Tracking Portfolios Via Bayesian Variable Selection

In this section, we illustrate the application of Bayesian variable selection to a real problem involving $p = 200$ potential regressors. In addition to showing how the prior may be reasonably constructed for this problem, we show how estimates of the marginal probabilities $\pi(\gamma_i = 1|Y)$ and calculation of exact relative probabilities using $g(\gamma)$ can be useful on very large problems.

The problem addressed here is that of constructing, from a large pool of stocks, a portfolio (linear combination) of a small number of stocks such that the portfolio returns are highly correlated with the returns on a stock market index such as the Standard and Poor's stock market index (SP500), itself a large portfolio of 500 stocks. This problem is of interest because investment portfolios which track broad market indices seem to manifest desirable risk-to-return tradeoffs. By using only a small number of stocks, the transaction costs of constructing and maintaining such a tracking portfolio can be considerably reduced. It is also of some academic interest to identify the nature of a small number of stocks which explain a substantial portion of the behavior of the market.

By considering the regression of the index returns on the individual stock returns, this portfolio construction problem can be treated as a variable selection problem. To illustrate this application, we considered the specific problem of constructing a portfolio to track the SP500 from a pool of $p = 200$ candidate stocks. The data we used consisted of 362 observations of weekly returns on Y, the SP500 index, and on $X_1, \ldots, X_{200}$, 200 randomly selected stocks for the period January 1985 to December 1991. The data were obtained from the Center for Research in Security Prices database (CRSP) at the University of Chicago Graduate School of Business.

An important consideration in constructing a tracking portfolio is to keep the size of the portfolio weights above a certain level so that transaction costs don't outweigh the benefits of diversification. Because the regression coefficients here correspond to the portfolio weights, this requirement can be satisfied by using a threshold of practical significance. In this particular problem, we chose such a threshold as follows. Roughly speaking, suppose a tracking portfolio of 50 stocks was the largest which would be considered. An acceptable equally weighted portfolio of that size would then have $\beta_i \approx .02$. Not wanting to stray too far from such a weight, we chose $\delta_{i\gamma} \equiv .008$ as our threshold. In practice, other considerations, such as transaction costs and the value of the portfolio, would be used to choose such a threshold.

Because of the vast number of possible models in this problem, 2^{200}, we began by using one of the Bayesian approaches to search for a more manageable

number of candidate models. To do this, we used the nonconjugate hierarchical prior with $R_\gamma \equiv I$ in order to take advantage of the computational speed of the nonconjugate Gibbs algorithm (NG) from Section 5. The values of $\upsilon_{0\gamma_{(i)}}$ and $\upsilon_{1\gamma_{(i)}}$ were chosen to satisfy (8) with $\delta_{i\gamma} \equiv .008$ and $\upsilon_{1\gamma_{(i)}}/\upsilon_{0\gamma_{(i)}} \equiv 625$. The values $\nu = 5$ and $\lambda_\gamma \equiv .007^2$ for the inverse gamma prior (3) were chosen so that the full model least squares estimate s^2_{LS} was in the left tail of $\pi(\sigma^2|\gamma)$ and the sample variance s^2_Y was in the right tail. Finally, we used the independence prior $\pi(\gamma) = \prod w_i^{\gamma_i}(1-w_i)^{(1-\gamma_i)}$ in (4) with $w_i \equiv 0.05$. This choice of w_i small put increased prior weight on parsimonious models, and provided a more peaked posterior.

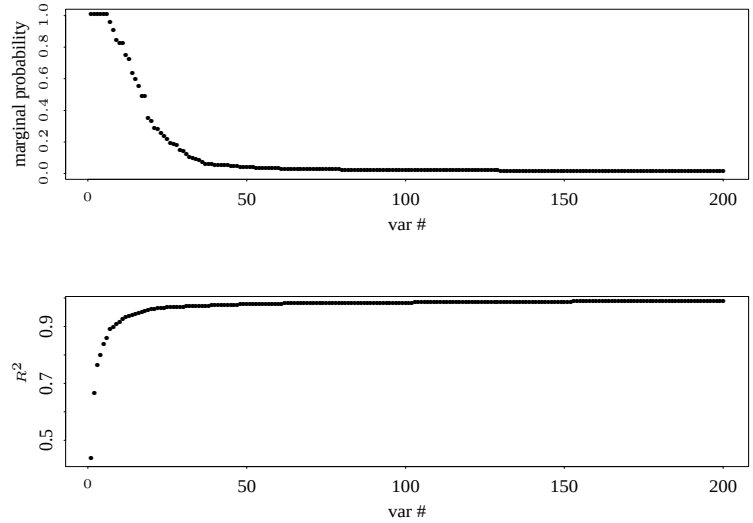

Figure 9. Top panel, $\pi(\gamma_i|Y)$ vs i for 200 stocks. Bottom panel, R^2 values.

Using the NG algorithm with this prior and all $p = 200$ candidate stocks, we simulated 11,000 γ values and kept the last 10,000. The top panel of Figure 9 displays the sorted values of the $\pi(\gamma_i = 1|Y)$ estimates in descending order. Initially, there are 6 stocks that have estimates close to 1. After that the $\pi(\gamma_i = 1|Y)$ estimates decline, getting close to zero at about the 50th stock. As a check on the prior inputs, we fit 200 nested regressions where stocks were added one at a time in order of the decreasing $\pi(\gamma_i = 1|Y)$ estimates. The R^2 values from these regressions, displayed in the bottom panel of Figure 9, increase most rapidly for the variables with large $\pi(\gamma_i = 1|Y)$ estimates. This suggests that these marginal probability estimates correspond to the explanatory power of the variables. The rapid increase of the R^2 values also suggests that the SP500 can be fit reasonably well with a relatively small subset of the 200 stocks.

Based on the above, we felt it would be informative to run the conjugate Gibbs algorithm (CG) from Section 6 with the 50 stocks which obtained the largest $\pi(\gamma_i = 1|Y)$ estimates above. This would allow us to compute the exact relative probabilities using $g(\gamma)$, and would allow for many more iterations in a reasonable amount of time. Using the matching prior obtained by (18) with $\hat{\sigma}^2 = .005^2$, and setting $w_i \equiv .01$ to obtain an even more peaked posterior, we simulated 450,000 γ values with the CG algorithm. We note that the $\pi(\gamma_i = 1|Y)$ estimates converged quickly and remained stable throughout the simulation. By recording $g(\gamma)$ as the simulation progressed, we were able to quickly identify the most probable models which were visited.

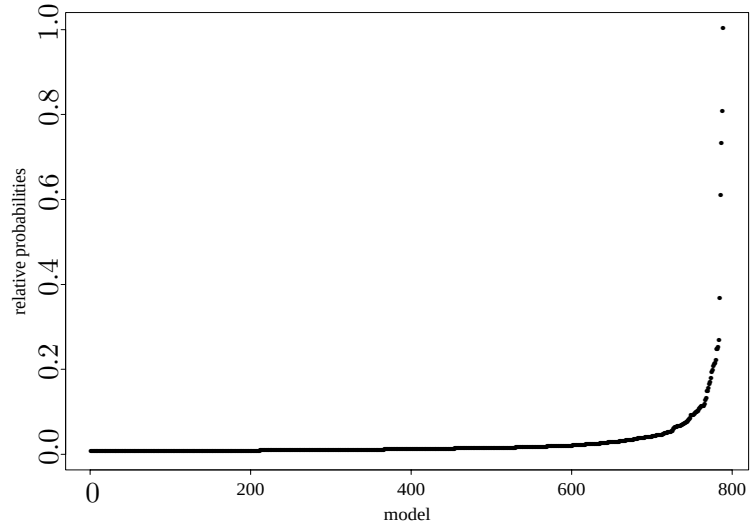

Figure 10. Exact relative posterior probabilities of the 789 most probable models.

The relative probabilities of the 789 most probable models are displayed in order in Figure 10. These were all the models visited whose relative probability was within a factor of .00674 (= −5 on the log posterior scale) of the best model. This relative probability distribution is very peaked suggesting that a small subset of models are far more promising than the rest. It was surprising that many of these 789 models were visited only once during the simulation, highlighting the value of being able to use $g(\gamma)$ to identify promising models. Of course, in such a vast problem it is unlikely that we have found the very highest probability model. Nevertheless, it does appear that at least some of these models are good. For example, the model with the highest relative probability included 21 variables and yielded $R^2 = 95.6\%$ which was virtually identical to the 21 variable model

obtained using stepwise regression. Interestingly, the posterior probability of this stepwise model was only .00176 of our maximum probability model.

In this example, we have illustrated how one might use Bayesian variable selection with various heuristics on a large problem to select a small set of models for further consideration. Having done this, we anticipate that the practitioner would then carefully examine the features of these models to select the one which best met the goals of the particular application. For the construction of tracking portfolios, this would entail looking at various measures of cost, risk and return of the implied portfolios as well as standard regression diagnostics. Although reductions to a smaller set of models can also be accomplished with frequentist methods such as stepwise selection, such methods are apt to lead to a different set of models because they are based on different criteria (see George and McCulloch (1995)). At the very least, Bayesian variable selection can provide a promising set of alternatives.

7. Discussion

In this paper, we have described and compared a variety of approaches for prior specification and posterior computation for Bayesian variable selection. The main differences between these approaches can be summed up as follows.

To begin with, prior specification corresponding to the removal of a predictor x_i can be obtained by either a continuous distribution on β_i which is concentrated at 0, or by assigning an atom of probability to the event $\beta_i = 0$. The continuous prior is useful for removing predictors whose coefficients are too small to be practically significant. The potential of this approach is illustrated by Wakefield and Bennett (1996) who use it in a pharmacokinetic context to select dosage predictors on the basis of "clinical" significance rather than statistical significance. The atom at $\beta_i = 0$ prior is useful for eliminating only those predictors whose coefficients cannot be distinguished from 0. The potential of this approach is illustrated by Smith and Kohn (1996) who use it to select knots for spline smoothing.

Another distinguishing characteristic of prior specification is the difference between nonconjugate and conjugate forms for the coefficient priors. Nonconjugate forms offer the advantage of precise specification of a nonzero threshold of practical significance, and appear to allow for more efficient MCMC exploration with approximately uncorrelated predictors. Conjugate forms offer the advantage of analytical simplification which allows for exhaustive posterior evaluation in moderately sized problems (p less than about 25). In larger problems where posterior evaluation is not feasible, conjugate forms allow for exact calculation of relative posterior probabilities and estimates of total visited probability by

MCMC posterior exploration. Furthermore, conjugate forms appear to allow for more efficient MCMC exploration with more correlated designs.

For the purpose of posterior exploration, a large variety of MCMC algorithms can be constructed based on the Gibbs sampler and Metropolis-Hastings algorithms. The simplest of these are the basic Gibbs sampler and the Metropolis algorithm which succesively update models by changing a single coordinate at a time. In terms of identifying high probability models, both of these algorithms performed comparably in our simulation examples. This comparability was also observed by Clyde, Desimone and Parmigiani (1996) who labeled Gibbs as SSVS and the Metropolis as MC^3. Interestingly, Clyde et al found that a mixture of SSVS and MC^3 performed better than either one. As discussed in Section 4.5, many other extensions of these algorithms are straightforward to construct.

Finally, we should point out that when the goal is prediction, it will usually be more appropriate to average predictions over the posterior distribution rather than using predictions from any single model (see Geisser (1993)). The potential of prediction averaging in the context of variable selection uncertainty has been nicely illustrated by Clyde et al. (1996) and Raftery et al. (1993). Of course, in situations where a single model is needed, such as in the dosage problem of Wakefield and Bennett (1996) and in the portfolio problem of Section 6, then the Bayesian approach will entail selection from among the high posterior probability models.

Acknowledgements

This paper is a major revision of our previous paper "Fast Bayes Variable Selection". We would like to thank Merlise Clyde, George Easton, Giovani Parmigiani and Luke Tierney for helpful comments and suggestions. This research was supported by NSF grant DMS-9404408, Texas ARP grant 003658130 and by the Graduate Schools of Business at the University of Chicago and the University of Texas at Austin.

References

Brockwell, P. J. and Davis, R. A. (1991). *Time Series: Theory and Methods*, Second edition. Springer-Verlag, New York.

Carlin, B. P. and Chib, S. (1995). Bayesian model choice via Markov Chain Monte Carlo Methods. *J. Roy. Statist. Soc. Ser. B* **57**, 473-484.

Chambers, J. M. (1971). Regression updating. *J. Amer. Statist. Assoc.* **66**, 744-748.

Chib, S. and Greenberg, E. (1995). Understanding the Metropolis-Hastings algorithm. *Amer. Statist.* **49**, 327-335.

Chipman, H. (1996), Bayesian variable selection with related predictors. *Canad. J. Statist.* **24**, 17-36.

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

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

Diaconis, P. and Holmes, S. (1994). Gray codes for randomization procedures. *Statist. Comput.* **4**, 287-302.

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

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

George, E. I. and McCulloch, R. E. (1995). Stochastic search variable selection. In *Practical Markov Chain Monte Carlo in Practice* (Edited by W. R. Gilks, S. Richardson and D. J. Spiegelhalter), 203-214. Chapman & Hall, London.

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* (Edited by D. Berry, K. Chaloner and J. Geweke), 339-348. Wiley, New York.

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

Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In *Bayesian Statistics* 4 (Edited by J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith), 169-194. Oxford Press.

Geweke, J. (1996). Variable selection and model comparison in regression. In *Bayesian Statistics* 5 (Edited by J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith), 609-620. Oxford Press.

Geyer, C. J. (1992). Practical Markov chain Monte Carlo. *Statist. Sci.* **7**, 473-511.

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

Hoeting, J., 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.

Kuo, L. and Mallick, B. (1994). Variable selection for regression models. Department of Statistics, University of Connecticut and Department of Mathematics, Imperial College, London, England.

Liu J. S. (1996). Peskun's theorem and a modified discreete-state Gibbs sampler. *Biometrika* **82**, 681-682.

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

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

Meehan, P. M., Dempster, A. P. and Brown, E. N. (1994). A belief function approach to likelihood to updating in a gaussian linear model. Manuscript, Frontier Science and Technology Research Foundation, Harvard University and Massachusetts General Hospital.

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

Peskun, P. H. (1975). Optimum Monte Carlo sampling using Markov chains. *Biometrika* **60**, 607-612.

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

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

Press, W. H., Teukolsky, S. A., Vetterling, W. T. and Flannery, B. P. (1992). *Numerical Recipes in C, Second edition*. Cambridge University Press.

Raftery, A. E., Madigan, D. M. and Hoeting, J. (1993). Model selection and accounting for model uncertainty in linear regression models. Technical Report no. 262, Department of Statistics, University of Washington. (To appear in *J. Amer. Statist. Assoc.*)

Raftery, A. E., Madigan, D. M. and Volinsky, C. T. (1996). Accounting for model uncertainty in survival analysis improves predictive performance (with discussion). In *Bayesian Statistics* 5 (Edited by J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith), 323-350. Oxford Press.

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**, 3-23.

Smith, M. and Kohn, R. (1996). Nonparametric regression using Bayesian variable selection. *J. Econometrics* **75**, 317-343.

Thisted, R. A. (1988). *Elements of Statistical Computing*. Chapman and Hall.

Tierney, L. (1994). Markov chains for exploring posterior distributions (with discussion). *Ann. Statist.* **22**, 1701-1762.

Wakefield, J. C. and Bennett, J. E. (1996). The Bayesian modelling of covariates for population pharmacokinetic models. *J. Amer. Statist. Assoc.* **91**, 917-927.

Department of Mgmt. Science and Info. Systems, University of Texas at Austin, CBA 5.202, Austin, TX 78712-1175, U.S.A.

Graduate School of Business, University of Chicago, IL 60637, U.S.A.

(Received December 1994; accepted January 1997)