M O N A S H

U N I V E R S I T Y

MUDET

AUSTRALIA

BAYESIAN STATISTICAL VARIABLE SELECTION

A REVIEW

Catherine M. Scipione

Working Paper No. 1/94

January 1994

DEPARTMENT OF ECONOMETRICS

ISSN 1032-3813

ISBN 0 7326 0390 0

BAYESIAN STATISTICAL VARIABLE SELECTION

A REVIEW

Catherine M. Scipione

Working Paper No. 1/94

January 1994

DEPARTMENT OF ECONOMETRICS

MONASH UNIVERSITY, CLAYTON, VICTORIA 3168, AUSTRALIA.

Bayesian Statistical Variable Selection A Review

Catherine M. Scipione
Monash University

Abstract

From a Bayesian viewpoint, the answer (in theory, at least) to the general model selection problem is known. However, the formalization of the selection problem does not realistically match the iterative process that occurs when selecting a model in practice. In addition, computational restrictions limit the applicability of the solution in general.

In the multiple linear regression variable selection setting, however, the Bayesian approach offers some practical procedures that can be used to at least reduce the possible number of models under consideration. 'Semi-automatic' methods for Bayesian variable selection have recently been developed by Mitchell and Beauchamp (1988) and George and McCulloch (1993) using relatively uniformative prior distributions for the unknown regression coefficients and variance parameter. In particular, their choices enable the computation of the general solution to be feasible.
Keywords: Model selection, multiple regression, Occam's razor, model validation, iterative learning, Markov chain Monte Carlo methods.

1 Introduction

The selection of the subset of predictor variables is central to the building of a multiple regression model. Typically the investigator is interested in a somewhat objective or automatic procedure to choose the set of predictor variables from amongst a set of p potential variables. Bayesian statistical methods, while commonly regarded as subjective, can be used to establish 'semi-automatic' procedures that may be quite useful for variable selection.

The main discussion of the paper is focused on Bayesian methods for statistical variable selection. Before investigating this question, however, it is of interest to review some general Bayesian concepts and their particular implementation in the multiple regression problem. Determining the subset of predictor variables to be used in a multiple regression

analysis is, of course, just one example of selecting one model from among a set of k models. Thus the parent problem of Bayesian model selection is discussed, so that the pros and cons of the general theory, which will carry through to the variable selection setting, can be highlighted. Finally, the variable selection problem is considered by looking at two particular approaches, put forward by Mitchell and Beauchamp (1988) and George and McCulloch (1993), as they appear to be the current most accessible approaches to so-called 'semi-automatic' Bayesian variable selection. The paper concludes with comments and some suggestions for future work.

2 Bayesian Multiple Regression

Consider the usual canonical multiple regression set up. Given an observed dependent variable Y and a set of known predictor variables $X = [X_1, \ldots, X_p]$, the linear relationship

$$Y = X\beta + \varepsilon$$

is assumed, where Y is $(n \times 1)$, X is an $(n \times p)$ known matrix of full rank, the vector of unknown regression coefficients, β, is $(p \times 1)$, and the vector of disturbances, ε, has an $n-$dimensional multivariate normal distribution with mean vector 0 and variance-covariance matrix $\sigma^2 I_n$, with σ^2 also unknown and where I_n denotes the $n-$dimensional identity matrix. A joint *prior distribution* is specified for the unknown parameters (β, σ^2), denoted $\pi(\beta, \sigma^2)$, and inference regarding these parameters is made using the joint *posterior distribution,* $\pi(\beta, \sigma^2 \mid y)$. For notational convenience, no distinction is made between a distribution and its density function with respect to Lebesgue measure. The posterior distribution can be found using Bayes' theorem

$$\pi(\beta, \sigma^2 \mid y) = \frac{f(y \mid \beta, \sigma^2)\pi(\beta, \sigma^2)}{f(y)}$$

where $f(y \mid \beta, \sigma^2)$ is the likelihood function given the observed data $Y = y$ and $f(y)$ is the marginal likelihood function of the observed data

$$f(y) = \int \int f(y \mid \beta, \sigma^2)\pi(\beta, \sigma^2) d\beta d\sigma^2;$$

$f(y)$ is also often called the integrated likelihood function. In practice, any prior distribution may be used and Bayes' theorem employed to calculate the joint posterior distribution $\pi(\beta, \sigma^2 \mid y)$. As this paper is concerned with 'semi-automatic' variable selection methods, two particular forms of prior distributions are considered here; one an improper prior and the other a hierarchical conjugate prior. These choices will be relevant to the variable selection discussion in Section 4.

The usual improper prior is specified according to

$$\pi(\beta, \sigma^2) \propto \frac{1}{\sigma^2},$$

with the resulting posterior density function

$$\pi(\beta, \sigma^2 \mid y) \propto \sigma^{-(n+2)} \exp[-\frac{-1}{2\sigma^2}(S^2 + (\beta - \hat{\beta})' X' X (\beta - \hat{\beta}))],$$

where $\hat{\beta} = (X'X)^{-1}X'y$, the least squares estimate of β, and $S^2 = (y - X\hat{\beta})'(y - X\hat{\beta})$. By integrating $\pi(\beta, \sigma^2 \mid y)$ with respect to σ^2, the resulting posterior distribution for β can be shown to be a $p-$dimensional Student t distribution with $n - p$ degrees of freedom, mean vector $\hat{\beta}$ and scale matrix $S^2(X'X)^{-1}/(n - p)$. Similarly, the marginal posterior distribution for σ^2 can be shown to be an inverse gamma distribution with mean $S^2/(n - p - 2)$ and variance $2S^4/[(n - p - 2)^2(n - p - 4)]$. See Zellner (1971) for explicit derivations and further details.

Notice that since $\pi(\beta, \sigma^2)$ is an improper prior and $\pi(\beta, \sigma^2 \mid y)$ is proper, it follows that $f(y)$ must also be improper. This may cause difficulties in assessing the fit of the model; see Gelfand, Dey and Chang (1992) and Section 3 for further comments. Of course, $f(y)$ can be made proper by restricting the joint prior distribution to be proper, for example, by restricting β to lie in $(-\beta_0, \beta_0)$ for some $\beta_0 > 0$ large and σ^2 to lie in $(\sigma_0^{-2}, \sigma_0^2)$ for some $\ln(\sigma_0^2)$ large. Mitchell and Beauchamp (1988) do this for their variable selection procedure; see Section 4 for further details.

The hierarchical conjugate prior considered here is specified by

$$\begin{aligned} \pi(\beta \mid \sigma^2) &= N_p(\mu_\beta, \tau^2 R) \\ \pi(\sigma^2) &= IG(\nu/2, \nu\lambda/2), \end{aligned}$$

where $N_p(\mu_\beta, \tau^2 R)$ denotes the $p-$dimensional normal density with mean vector μ_β and variance covariance matrix $\tau^2 R$, where R is the known prior correlation matrix, and $IG(\nu/2, \nu\lambda/2)$ denotes the inverse gamma density with shape parameter $\nu/2$ and scale parameter $\nu\lambda/2$, resulting in $\nu\lambda/\sigma^2 \sim \chi^2_\nu$. Here μ_β, τ^2, ν, and λ are considered known, typically with $\mu_\beta = 0$ and τ^2 taken very large (but finite). Choices of ν and λ can be selected by considering ν the number of observations from a (possibly imagined) prior experiment with $\nu\lambda/(\nu - 2)$ the prior estimate of σ^2. Given this prior specification, the posterior distribution can be written

$$\begin{aligned} \pi(\beta, \sigma^2 \mid y) \;\propto\; & \sigma^{-n} \exp\{-\frac{1}{2\sigma^2}(S^2 + (\beta - \hat{\beta})' X' X (\beta - \hat{\beta}))\} \\ & \cdot \exp\{-\frac{1}{2\tau^2}(\beta - \mu_\beta)' R^{-1} (\beta - \mu_\beta)\} \cdot \sigma^{-2(\nu/2+1)} \exp\{-\frac{\nu\lambda}{2\sigma^2}\}, \end{aligned}$$

although the marginal posterior distributions for β and σ^2 cannot be solved analytically. However, note that both full conditional distributions

$$\pi(\beta \mid \sigma^2, y) = N_p((\sigma^{-2}X'X + \tau^{-2}R^{-1})^{-1}(\sigma^{-2}X'y + \tau^{-2}R^{-1}\mu_\beta), (\sigma^{-2}X'X + \tau^{-2}R^{-1})^{-1})$$

$$\pi(\sigma^2 \mid \beta, y) = IG(\frac{n+\nu}{2}, \frac{(y - X\beta)'(y - X\beta) + \nu\lambda}{2})$$

are available, and thus Markov chain Monte Carlo methods could be employed to obtain estimates of the marginal posterior distributions. In this case, since the prior distribution $\pi(\beta, \sigma^2)$ is proper, $\pi(\beta, \sigma^2 \mid y)$ and $f(y)$ will both also be proper. Standard validation techniques are available; see Box (1980) and Geisser (1985).

Thus, given a particular model, ie likelihood function and joint prior distribution for the parameters of the likelihood function, a posterior distribution can be at least approximated and inferences regarding the process; ie, parameters, forecasts, etc; are available. However, how does a Bayesian choose the particular model upon which inferences are made? The answer in part lies in the solution to the Bayesian model selection problem.

3 Bayesian Model Selection

From a Bayesian viewpoint, the answer (in theory, at least) to the general model selection problem is known: Given a set of models, $M_1, M_2, \ldots, M_k$, and prior probabilities $\pi(M_j) = Pr(\text{model } j \text{ is the 'true' model})$, all relevant Bayesian model choices, or inferences, are based on the posterior probabilities

$$\pi(M_i \mid y) = \frac{f(y \mid M_i)\pi(M_i)}{\sum_{j=1}^{k} f(y \mid M_j)\pi(M_j)},$$

for $i = 1, \ldots, k$. Here y denotes the collection of observed data values and $f(y \mid M_i)$ the evaluation of the joint density function of the data, conditional on model M_i holding true. Typically, the model with the largest posterior probability is selected. Note that no requirement is made for the models under consideration to be nested. However, as other authors have noted, in particular Box (1980) and Gelfand et al. (1992), the simplicity of this solution can be misleading for several reasons.

Firstly, proceeding under the assumption that all uncertainty regarding the set of possible models is quantified in the prior distribution $\{\pi(M_j), j = 1, \ldots, k\}$, actually computing the desired posterior probabilities $\pi(M_j \mid y)$ in many interesting problems can be a formidable task. Typically, model M_j will consist of a likelihood, $f_j(y \mid \theta_j, M_j)$, and a prior density with respect to, say, Lebesgue measure for, $\pi_j(\theta_j, M_j)$. To obtain the desired posterior probabilities $\pi(M_j \mid y)$, the quantities

$$\pi(y \mid M_j) = \int f_j(y \mid \theta_j, M_j)\pi_j(\theta_j \mid M_j)d\theta_j,$$

for $j = 1, \ldots, k$, are required. As the parameter θ_j can be high dimensional the evaluation of the above integral may require cumbersome numerical integration techniques. Although some recent advances in the area of Bayesian computational methods have begun to reduce the computational difficulties for some particular model selection problems, in general these integrals can be quite difficult to evaluate. In addition, the number of models that are under consideration may be quite high, in which case the evaluation of many of these integrals can become prohibitive.

Secondly, it is assumed that each of the models under consideration are all 'valid' in some sense. That is, it is assumed that any one of the models could describe the behavior of the observed data without large and/or systematic departures from the modelling assumptions. This means that a considerable amount of checking of model assumptions must take place before the set of k possible models is determined. This can be done, as Box (1980) clearly points out, by using the marginal distributions $f(Y \mid M_i)$ and comparing them to the observed data. However, while checking models assumptions is (hopefully) standard practice, the formal selection procedure only considers model validity through the numerical value of $f(y \mid M_j)$, and relative to the prior probabilities, $\pi(M_j)$, for $j = 1, 2, \ldots, k$. While this may be precisely what is desired for consideration of a collection of models all of which are judged in some sense to be 'valid', hidden in the presentation is the potentially iterative process of discerning which models will be included for consideration.

Along these lines, Gelfand et al. (1992) discuss some of the problems of validating a given model, and point out the fact that $f(Y \mid M_i)$ need not be proper in the sense that it need not integrate to unity. As this makes $f(Y \mid M_i)$ difficult to use for model validation, and they suggest some alternative cross validatory techniques for checking model assumptions based on univariate predictive distributions $f(Y_r \mid Y_{(r)} = y_{(r)}, M_j)$, where $Y_{(r)}$ denotes the vector of observables, exluding the r^{th} component. They develop procedures using the univariate predictive distributions based on cross validatory ideas. However, this process is data analytic in nature, and again occurs before the final collection of models is assembled. See Geisser (1985) and Geisser and Eddy (1979) for related work on predictive approaches to model selection, and Stone (1974) for an early reference to work on cross validation techniques.

Finally, there nearly always exists some amount of uncertainty in the specification of the models under consideration. This may be in part due to the fact that the models are intended to be merely approximations to complex relationships. The Bayesian approach has an advantage over classical methods in that modelling the uncertainty in the parameters through prior distributions should account for at least some of the uncertainty in the likelihood.

The selection of a single model, though not specifically required from the Bayesian set up, is what is commonly done in practice. However, as Madigan and Raftery (1993) point out, conditioning inference on the 'truth' of a single selected model ignores model

uncertainty. They suggest averaging quantities of interest over the set of potential models, weighted by their respective posterior distributions. Specifically, suppose Δ is a particular quantity of interest. Then, given the set of posterior model probabilities, and the posterior density functions $\pi(\Delta \mid y, M_j)$ with respect to some dominating measure, then the so-called predictive density of Δ with respect to the same dominating measure can be calculated using

$$\pi(\Delta \mid y) = \sum_{j=1}^{k} \pi(\Delta \mid y, M_j)\pi(M_j \mid y).$$

Madigan and Raftery (1993) suggest basing inference for Δ on $\pi(\Delta \mid y)$, and not on $f(\Delta \mid y, M^*)$, where M^* is a single model chosen, for example, to have the highest posterior model probability. While this approach is certainly in concert with the Bayesian view of quantifying uncertainty through the laws of probability, it may not appear as desirable from the point of view of model description. Notice also that this approach, while accounting for some uncertainty, does not deal with the validation problem.

Thus, it appears that the problem of model selection is, in some sense, not well formulated. The author's general opinion is in line with Box (1980) and Gelfand et al. (1992), who suggest that the model selection problem is really a fusion of two distinct ideas, namely model criticism and model estimation. Model estimation, that is the usual procedure of calculating posterior model probabilities and related predictive distributions for a given collection of valid models, is well understood within the Bayesian context. Conditional on the validity of the set of models under consideration, the standard Bayesian approach has little difficulty. However, the notion of model criticism, or checking model validity, which must go together with model estimation in any practical setting, is typically not well posed. This is due to the problem of uncertainty in specifying the models under consideration, and the necessity to utilize observed data to validate the models.

In the multiple regression variables selection setting, typically the investigator is simply searching for linear relationships between dependent and predictor variables. In this setting, while the problems of model validation are still present, models outside the $k = 2^p$ possible models, where p is the number of predictor variables, are generally not of interest. The problem of validation still exists, however it is simply noted that the same criticisms will apply to any classical procedure that does not address the issue. In contrast, the Bayesian approach can deal with model uncertainty, in the sense of Madigan and Raftery (1993), whereas classical approaches generally do not. In addition, as will be shown, Bayesian methods can provide useful tools for at least reducing the number of possible models under consideration.

4 Variable Selection in Multiple Regression

The Bayesian model selection procedure described in Section 3 will now be applied to the multiple regression variable selection problem. Two recent papers, Mitchell and Beauchamp (1988), and George and McCulloch (1993), discuss Bayesian model selection in this important special case. In particular, both sets of authors are interested in developing a 'semiautomatic' approach to the Bayesian variable selection problem. Both begin with the canonical multiple regression set up described in Section 2, where now the collection of p known predictor variables, $X_1, \ldots, X_p$, are viewed only as *potential* predictors. The object of the selection procedure is to select a subset of the p predictors for the regression relation

$$Y = X^*\beta^* + \varepsilon$$

where the columns of X^* are the vectors of predictor variables selected and the β^* corresponds to the vector of nonzero regression coefficients corresponding to the columns of X^*.

The difference between the two approaches lies in the particular forms of the prior distribution placed on the components of the β vector and σ^2. Mitchell and Beauchamp explore a 'spike and slab' prior on each of the components of β, along with the standard noninformative prior on σ^2, while George and McCulloch place a mixture of normal distributions on the components of β, and an inverse gamma distribution on σ^2. Both formulations have the desired property that the priors involved can be defined to be relatively diffuse, while still having the ability to incorporate prior information regarding practically important variables where relevant. Both formulations rely on 'tuning parameters' that are used successfully to explore potential sets of predictor variables.

To fix notation, let M_j denote the set of assumptions associated with the j^{th} of the potential $k = 2^p$ possible regression models, including the subset of predictor variables and the form of the joint prior distribution placed on the regression coefficients, β, and the scale parameter, σ^2. (Note: Only first order terms are explicitly included, however higher ordered 'interaction' terms could be included into the set of p regressors if desired.) Let T_j denote the set of subscripts corresponding to the predictor variables included in the j^{th} model, and let k_j be the number of predictor variables included in M_j.

The 'spike and slab' prior used by Mitchell and Beauchamp on each of the nonzero $\beta_i, i = 1, 2, \ldots, p$, coefficients of the predictor variables included in any particular model is given by

$$\begin{aligned} Pr(\beta_i = 0) &= h_{0i} \\ Pr(|\beta_i| < b, \beta_i \neq 0) &= (b + \beta_{0i})h_{1i} \text{ for } -\beta_{0i} < b < \beta_{0i} \\ Pr(|\beta_i| > \beta_{0i}) &= 0 \end{aligned}$$

where $h_{0i} > 0, h_{1i} > 0$, and $h_{0i} + 2h_{1i}\beta_{0i} = 1$. The 'spike' naturally refers to the prior

probability mass at $\beta_i = 0$, and the 'slab' to a uniform density on $(-\beta_{0i}, \beta_{0i})$. The height of the 'spike,' h_{0i}, can be prespecified by the user prior as a prior belief (or perhaps the concensus of expert opinion) that β_i is 0 and hence the prior probability that the corresponding predictor variable, X_i, should not be included in the model. Hence, if a predictor variable is considered to have high practical significance, then h_{0i} should be set relatively low. The prior distribution for σ^2 is the standard noninformative prior

$$\pi(\sigma^2) \propto \frac{1}{\sigma^2},$$

restricted to $(\sigma_0^{-2}, \sigma_0^2)$, resulting in

$$\pi(\sigma^2) = \frac{1}{4\ln(\sigma_0^2)\sigma^2}$$

for some $\sigma_0^2 > 0$ large. Notice both $\pi(\beta \mid \sigma^2)$ and $\pi(\sigma^2)$ are proper density functions, unlike the similar prior discussed in Section 2.

A useful parameterization is to define $\gamma_i = h_{0i}/h_{1i}$, the ratio of the height of the spike to the height of the slab, so that $h_{0i} = 0$ if and only if $\gamma_i = 0$. Using this parameterization the marginal probability can be shown to be

$$Pr(M_j) = \prod_{i \in \bar{T}_j} \gamma_i \prod_{i \in T_j} 2\beta_{0i} \prod_{i=1}^{2^p} (\gamma_i + 2\beta_{0i})^{-1},$$

where $\bar{T}_j$ is the complement of the set T_j, and the posterior distribution

$$Pr(M_j \mid y) = g(\prod_{i \in T_j} \gamma_i)\pi^{k_j/2}\Gamma(\frac{n-k_j}{2}) \mid X_j'X_j \mid^{-1/2} (S_j^2)^{-[(n-k_j)/2]},$$

where g is a normalizing constant that does not dependent on j. Here $S_j^2 = (y - X_j\hat{\beta}_j)'(y - X_j\hat{\beta}_j)$, is the residual sum of squares for the least squares fit to the regression model with the predictor variables in the columns of X_j corresponding to the nonzero β coefficients in model M_j.

To obtain the posterior probabilities above, β_{0i} and σ_0^2 are taken to be 'sufficiently large' so that all integrals from $-\beta_{0i}$ to β_{0i} and σ_0^{-2} to σ_0^2 are well approximated by the same integrals from $-\infty$ to ∞ and 0 to ∞, respectively. From this point, the models can be ranked according to the magnitude of their posterior probabilities. For this, however, all of the γ_i values need to be specified *a priori*. An alternative approach is to restrict the β_i parameters to be given identical priors, that is $\gamma_i \equiv \gamma$. In this case γ can be treated as a tuning parameter, and used to consider the range of possible rankings of models as a function of γ. Notice also that the probabilities

$$Pr(\beta_j = 0 \mid y) = \sum_{\{i \mid j \in T_i\}} Pr(M_i \mid y)$$

are directly available. It is useful to plot $Pr(\beta_j = 0 \mid y)$ versus γ as a tool for exploring the variables that are important as well as those variables whose importance is difficult to separate from other variables, as is the case when collinearity is present.

Other graphical displays that may be useful include plotting the posterior entropy of the submodels

$$H = -\sum_{j=1}^{2^p} Pr(M_j \mid y) \ln(Pr(M_j \mid y)),$$

and the posterior expected number of terms in a submodel

$$E(k \mid y) = \sum_{j=1}^{2^p} k_j Pr(M_j \mid y),$$

both against γ. Finally, a cross validatory approach can also be used to calculate various types of 'checking functions' using univariate predictive distributions. Again, these 'checking functions' can be plotted against γ and can be useful in comparing the predictive properties of the submodels.

In contrast to the 'spike and slab' prior, George and McCulloch use a hierarchical prior for the regression coefficients. In this setup, a latent variable, δ_i, is introduced for each regression coefficient, β_i, for $i = 1, \ldots, p$. A normal mixture prior distribution is specified by

$$\pi(\beta_i \mid \delta_i) \sim (1 - \delta_i)N(0, \tau_i^2) + \delta_i N(0, c_i^2 \tau_i^2)$$

and

$$Pr(\delta_i = 1) = 1 - Pr(\delta_i = 0) = p_i,$$

where τ_i^2 is small so that if $\delta_i = 0$, then β_i will near 0 with high prior probability. τ_i^2 is chosen to specify a practical significance level, so that if $\delta = 0$, β_i can probably 'safely' be estimated by 0. The $c_i^2 > 1$ are set large so that if $\delta_i = 1$, β_i will have a high prior probability of being nonzero, or equivalently, the predictor variable associated with β_i will have a high prior probability of being included in the set of selected variables. Some suggestions for how to choose τ_i^2 and c_i^2 are given in their paper.

In addition to these marginal priors for the β_i parameters, a prior correlation structure can be introduced among the regression coefficients, if desired. While this may be difficult to specify in practice, George and McCulloch suggest using the prior correlation matrix of the regression coefficients, R, as a tuning parameter, with the range of possible values containing $R \propto (X'X)^{-1}$ to $R = I$. The prior covariance for β, given the vector δ, is $D_\delta R D_\delta$ where $D_\delta \equiv diag[d_1\tau_1, \ldots, d_p\tau_p]$, with $d_i = 1$ if $\delta_i = 0$ and $d_i = c_i$ if $\delta_i = 1$.

A prior distribution is also required for σ^2, the scale parameter on the noise component ε. George and McCulloch use an inverse gamma conjugate prior with shape parameter

$\nu_\delta/2$ and scale parameter $\nu_\delta\lambda_\delta/2$, which corresponds to

$$\frac{\nu_\delta\lambda_\delta}{\sigma^2} \sim \chi^2_{\nu_\delta}.$$

The dependence of ν_δ and λ_δ on δ can be used to induce dependence between β and σ^2. In particular, if more predictor variables are included in a model, the prior distribution $\pi(\sigma^2 \mid \delta)$ could be concentrated on smaller values of σ^2.

Finally, the prior distribution for $\delta, \pi(\delta)$, must be chosen in order to calculate the posterior distribution of $\delta, \pi(\delta \mid y)$. Here $\pi(\delta)$ corresponds directly to $\Pr(M_j)$ for $j = 1, 2, \ldots, 2^p$, as particular values of δ yield various combinations of predictor variables. For example, if $p = 3$, then there are $2^3 = 8$ various possible submodels. One of the eight possible models includes predictor variables 1 and 3 only. This corresponds, with high probability, to $\delta_1 = 1, \delta_2 = 0$, and $\delta_3 = 1$. Thus, interest would focus on the posterior probabilities of the various submodels, which for this example is $\pi(\delta_1 = 1, \delta_2 = 0, \delta_3 = 1 \mid y)$. While any prior distribution for δ can easily be used, George and McCulloch suggest that $\pi(\delta) = \prod_{i=1}^{p} p_i^{\delta_i}(1-p_i)^{(1-\delta_i)}$, which although it implies that the inclusion of predictor i is independent of the inclusion of predictor j, for all $i \neq j$, seems to perform well in their examples. A special case of this prior is $\pi(\delta) = 2^{-p}$, where each predictor has an equal prior probability of being included in the final model, with each $p_i = 1/2$. Parsimonious models can be favored if $\pi(\delta)$ is chosen to place higher probabilities on models with few number of parameters.

The real advantage to George and McCulloch's prior distributional structure is that the calculation of $\pi(\delta \mid y)$ can be well approximated using the Gibbs sampler algorithm. Recall the normal-inverse gamma prior distribution introduced in the multiple regression setting in Section 2. Noting the dependence on δ, the full conditional distributions from Section 2, now modified in accordance with the assumptions listed above, become

$$\begin{aligned}
\pi(\beta \mid \sigma^2, \delta, y) &= N_p((\sigma^{-2}X'X + D_\delta^{-1}R^{-1}D_\delta^{-1})^{-1}\sigma^{-2}X'y, (\sigma{-}2X'X + D_\delta^{-1}R^{-1}D_\delta^{-1})^{-1}) \\
\pi(\sigma^2 \mid \beta, \delta, y) &= IG(\frac{n+\nu_\delta}{2}, \frac{(y-X\beta)'(y-X\beta)+\nu_\delta\lambda_\delta}{2}).
\end{aligned}$$

The full conditional distributions for each δ_i are Bernoulli distributions with

$$\pi(\delta_i = 1 \mid \delta_{(i)}, \beta, \sigma^2, y) = \frac{a}{a+b},$$

where $\delta_{(i)}$ are the given values of all the components of δ *except* the i^{th} component and

$$\begin{aligned}
a &= \pi(\beta \mid \delta_{(i)}, \delta_i = 1) \cdot \pi(\sigma^2 \mid \delta_{(i)}, \delta_i = 1) \cdot \pi(\delta_{(i)}, \delta_i = 1) \\
b &= \pi(\beta \mid \delta_{(i)}, \delta_i = 0) \cdot \pi(\sigma^2 \mid \delta_{(i)}, \delta_i = 0) \cdot \pi(\delta_{(i)}, \delta_i = 0).
\end{aligned}$$

Iterative sampling between these three well known distributions via the Gibbs sampler is straightforward. See Gelfand and Smith (1990) for a general description of the Gibbs sampler algorithm.

Due to their use of this Markov chain Monte Carlo technique, they call their procedure SSVS, Stochastic Search Variable Selection. The Gibbs sampler induces a Markov chain whose unique equilibrium distribution is $\pi(\delta \mid y)$. As a result, realizations of the Markov chain 'visit' those models with the largest $\pi(\delta \mid y)$ probabilities most frequently, and thus pick out those submodels of interest. In cases where a large number of possible predictor variables are being considered, the number of submodels can be huge. Hence this algorithm can be quite useful in reducing the number of models under consideration.

5 Comments and Directions

Bayesian statistical methods combine relevant sample and prior information in a coherent fashion, using the laws of probability. It has been argued here and by other authors, that the model selection problem as currently formulated, does not match the iterative learning process that actually occurs in practical settings. Until a better formulation of the problem is determined, the model selection problem will continue to be a very difficult one in practical settings.

Despite this criticism, Bayesian approaches to model selection, and in particular to the variable selection problem in multiple regression, can lead to useful tools for the practicioner. The 'semi-automatic' procedures of Mitchell & Beauchamp (1988) and George & McCulloch (1993) offer graphical diagnostic tools and a computationally feasible means for at least reducing the number of models under consideration. These tools have been made possible primarily due to recent developments in Bayesian computing and the increasing availability of high speed computers. As these resources and methods improve, it is expected that Bayesian and non-Bayesian statistical methods will improve as well, particularly for more realistic, complex models.

References

[1] Berger, J. O. (1985), *Statistical Decision Theory and Bayesian Analysis,* New York: Springer-Verlag.

[2] Box, G. E. P. (1980), 'Sampling and Bayes' Inference in Scientific Modelling and Robustness' (with discussion), *Journal of the Royal Statistical Society: Series A,* **143** Part 4, 383-430.

[3] Cook, R. D. and Weisberg, S. (1982), *Residuals and Influence in Regression.* London: Chapman and Hall.

[4] Geisser, S. (1985), 'On the Prediction of Observables: a selective update,' in *Bayesian Statistics 2* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley, and A. F. M. Smith, eds.), Amsterdam: North-Holland, 203-230.

[5] Geisser, S. and Eddy, W. F. (1979), 'A Predictive Approach to Model Selection,' *Journal of the American Statistical Association,* **74** 153-160.

[6] Gelfand, A. E., Dey, D. K. and Chang, H. (1992), 'Model Determination using Predictive Distributions with Implementation via Sampling-Based Methods' (with discussion), in *Bayesian Statistics 4,* J. M. Bernardo, J. O. Berger, AA. P. Dawid and A. F. M. Smith, Eds., Oxford University Press, 147-167.

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

[8] George, E. I. and McCulloch, R. E. (1993), 'Variable Selection Via Gibbs Sampling,' *Journal of the American Statistical Association,* **88** 881-889.

[9] Madigan, D. and Raftery, A. E. (1993) 'Model Selection and Accounting for Model Uncertainty in Graphical Models using Occam's Window,' University of Washington Department of Statistics Technical Report #213 (revised).

[10] Mitchell, T. J. and Beauchamp, J. J. (1988), 'Bayesian Variable Selection in Linear Regression' (with discussion), *Journal of the American Statistical Association,* **83** 1023-1036.

[11] Smith, A. F. M. (1086), 'Some Bayesian Thoughts on Modelling and Model Choice,' *The Statistician,* **35** 97-102.

[12] Stone, M. (1974), 'Cross-validatory choice and assessment of statistical predictions,' *Journal of the Royal Statistical Society: Series B,* **36** 111-147.

[13] Zellner, A. (1971), *An Introduction to Bayesian Inference in Econometrics,* New York: John Wiley & Sons.