Sankhyā : The Indian Journal of Statistics
1998, Volume spl, Series B, Pt. , pp.

VARIABLE SELECTION FOR REGRESSION MODELS

By LYNN KUO
University of Connecticut, Storrs
and
BANI MALLICK
Imperical College, London

SUMMARY. A simple method for subset selection of independent variables in regression models is proposed. We expand the usual regression equation to an equation that incorporates all possible subsets of predictors by adding indicator variables as parameters. The vector of indicator variables dictates which predictors to include. Several choices of priors can be employed for the unknown regression coefficients and the unknown indicator parameters. The posterior distribution of the indicator vector is approximated by means of the Markov chain Monte Carlo algorithm. We select subsets with high posterior probabilities. In addition to linear models, we consider generalized linear models.

1. Introduction

Many methods have been proposed for selecting suitable predictors in multiple regression. Classical methods for variable selection include backward elimination, forward selection, and stepwise regression. They sequentially delete or add predictors by means of mean squared error or modified mean squared error criteria. Various Bayesian methods have also been proposed. They include model determination by means of the following criteria: Bayesian information criterion (BIC, Schwarz, 1978), asymptotic information criterion (AIC, Akaike, 1974), Bayes factor, and pseudo-Bayes factor. But the power explosion of the number of possible submodels (2^p) being considered for p predictors often handicaps the computation. A more automatic data driven tool is needed for the data analyst to identify a parsimonious model.

AMS (1991) *subject classification.* 62J05, 62J02.
Key words and phrases. Bayesian inference, F-tests, generalized linear model, Gibbs sampling, linear model, subset selection.

Mitchell and Beauchamp (1988) propose a Bayesian variable selection method assuming the prior distribution of each regression coefficient is a mixture of a point mass at 0 and a diffuse uniform distribution elsewhere. They also review other methods. Recently, George and McCulloch (1993) propose a stochastic search variable selection procedure where the subset selection is derived from a hierarchical normal mixture model. Gibbs sampling was developed for computing the posterior distribution of subset selection. The promising predictors are then identified by their more frequent appearance in the sequence of Gibbs sample. Their methods, however, require sophisticated choices of the tuning factors that specify the two variances in the normal mixture models in the first stage of the hierarchical prior.

Motivated by the work of George and McCulloch, we explore a simpler method of subset selection. Instead of building a hierarchical model, we embed indicator variables in the regression equation that incorporates all 2^p submodels. Let γ_j be an indicator variable supported at two points 1 and 0. We write the regression model for the ith subject, $i = 1, \ldots, n$, by

$$y_i = \sum_{j=1}^{p} \beta_j \gamma_j x_{ij} + \epsilon_i, \qquad \ldots (1.1)$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ is the usual unknown column vector of regression coefficients, and x_{ij} is the known j^{th} covariate for the ith subject. When $\gamma_j = 1$, we include the j^{th} predictor in the regression model. When $\gamma_j = 0$, we omit the j^{th} predictor when building the model. As usual, we assume ϵ_i are i.i.d. with a normal $N(0, \sigma^2)$ distribution.

The idea of adding indicator function in the regression equation has potential to be applied to more complex modeling in data analysis. For example, it is often natural to retain the main effect if we find the interaction effect is significant. Then an easy extension of model (1.1) for two predictors with an interaction term could be

$$\begin{aligned} y_i &= \beta_0 + \beta_1 \max(\gamma_1, \gamma_3) x_{i1} + \beta_2 \max(\gamma_2, \gamma_3) x_{i2} + \beta_3 \gamma_3 x_{i1} x_{i2} + \epsilon_i, \\ &= \beta_0 + \beta_1 [1 - (1-\gamma_1)(1-\gamma_3)] x_{i1} + \beta_2 [1 - (1-\gamma_2)(1-\gamma_3)] x_{i2} \\ &\quad + \beta_3 \gamma_3 x_{i1} x_{i2} + \epsilon_i. \end{aligned} \qquad \ldots (1.2)$$

The basic idea can be extended to more than two predictors.

We can specify quite general classes of priors for $\boldsymbol{\beta}$ and $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p)^T$ and σ. For simplicity, we assume independent priors for $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, and σ; $\boldsymbol{\beta} \sim N_p(\boldsymbol{\beta}_0, \mathbf{D}_0)$, a multivariate normal; γ_j, $j = 1, \ldots, p$, are chosen independently, each with Bernoulli distribution $B(1, p_j)$; and σ has an inverse gamma distribution for conjugacy. The choice of $\boldsymbol{\beta}_0$ and $\mathbf{D}_0$ reflects the statistician's prior belief about the mean and covariance matrix of $\boldsymbol{\beta}$ in the full model with all $\gamma_j = 1$. In the absence of such prior information, we can consider the following choice.

Assume the intercept term is always included in building the model. After centering and scaling, for each of the $\mathbf{y} = (y_1, \ldots, y_n)^T$ and $\mathbf{x_j} = (x_{1j}, \ldots, x_{nj})^T$, $j = 1, \ldots, p$, it is reasonable to choose $\boldsymbol{\beta}_0 \equiv (0, \ldots, 0)^T$ in the prior. We also choose $\mathbf{D}_0$ with moderately large diagonal terms, so the analysis is focused on the likelihood. Nevertheless, we cannot let the diagonal terms go to ∞ because that would support no predictors being selected. Let I be a $p \times p$ dimensional identity matrix. We think $\mathbf{D}_0 = k^2 I$ for any $\frac{1}{2} \leq k \leq 4$ are reasonable. Geweke (1996) recommends a procedure by looking at the changes of y_i versus the changes of x_{ij} for each j to determine the diagonal terms σ_j^2 of $\mathbf{D}_0$. Geweke's suggestion is akin to the centering, scaling, and choosing $\boldsymbol{\beta}_0$ and $\mathbf{D}_0$ procedure we just discussed. The strategies for eliciting prior distributions proposed by Kadane *et al.* (1980), and Garthwaite and Dickey (1992) could also be used to determine the hyperparameters $\boldsymbol{\beta}_0$ and $\mathbf{D}_0$. The probability p_j reflects the statistician's preference for including the j^{th} predictor in model building. Often, we choose $p_j = 1/2$ for all j to reflect the equally likely likelihood prior for all possible 2^p submodels in the absence of any prior preference for the predictors. Alternative choices of prior are given in the next section.

The posterior distribution $\boldsymbol{\gamma}$ is supported on each of the 2^p models. It measures the likelihood of each submodel. Therefore, we select submodels with high posterior probabilities. In fact, selecting the model with the highest posterior probability is the Bayes decision rule with respect to a 0-1 loss function with the loss value of 0 for the correct model and 1 for all other models. The posterior probabilities can be evaluated by means of the Markov chain Monte Carlo (MCMC) method. We can also relate the steps in determining whether or not to include the jth predictor in the MCMC to the classical F-tests in subset selection with the following difference. In MCMC, the decision on whether or not retaining the j^{th} predictor can be related to hypothesis testing ($\beta_j = 0$ versus $\neq 0$) using the Bayes factor rule; the classical F-test is based on statistical significance. In a vague sense, we can think of our method as an automated stochastic F-test for subset selection. Although we set up our problem as if we need to compute the posterior probabilities for each of the 2^p submodels (a problem we intended to avoid), we never have to do this computation. In the MCMC, we see we can zoom in to the submodels with promising predictors very quickly. They are the submodels that occur with high frequencies in the Gibbs sample. Many of the uninteresting submodels never or rarely appear in the Gibbs sample.

Why do we consider variable selection? One may argue that the Bayesian paradigm already offers an alternative to variable selection, namely the use of prior information. For example, consider a problem with a large number of predictors that can be classified into three groups (variables known to be important, variables thought to be important, variables included as pure speculation). One could include all variables by incorporating the prior information in the form of different prior variances for the three sets of regression coefficients. However, this does not serve the purpose of subset selection. Mitchell and Beauchamp

(1988) discuss succintly on the reasons for undertaking variable selection: (a) to express the relationship between the response and predictors as simply as possible; (b) to identify important and negligible predictors; (c) to reduce future cost of prediction; and (d) to increase the precision of statistical estimates and predictions. In short, in order to have a parsimonious and interpretable model, we need to consider subset selection.

Let ϑ_j denote $\beta_j\gamma_j$ and let $\boldsymbol{\vartheta}$ denote $(\vartheta_1,\ldots,\vartheta_p)^T = (\beta_1\gamma_1,\ldots,\beta_p\gamma_p)^T$, the vector of coefficients of the regression equation. Let us consider the simplest case with $\mathbf{D}_0$ to be diagonal with σ_j^2 in the j^{th} entry of the diagonal. Then our prior on ϑ_j is the spike and bell prior, that is, it is a mixture of a point mass at 0 with probability $1-p_j$ and a normal density $N(\beta_{0,j},\sigma_j^2)$ with probability p_j. This is equivalent to a prior considered by Geweke (1996). Geweke also considers a mixture of a spike and a truncated normal prior. Although we do not discuss the truncated normal component here, we can easily modify our argument as in Geweke to accommodate for the truncated normal component.

Although our formulation is similar to that of George and McCulloch (1993), it differs in the following sense. George and McCulloch essentially propose a shrinkage method with mixture of two normal priors. The variables retained in their models are the ones with coefficients significantly different from zero. Because they don't exactly allow the coefficients to be zero, they need to decide on how small they are to be essentially treated as zero. Therefore, they need to deal with the complex issue of choosing tuning factors for the hyperparameters in their hierarchical setup. Our priors, mixtures of point mass and continuous distributions, put mass on regression coefficients ϑ_j being exactly 0, for each $j=1,\ldots,p$. Our method assigns positive proabability to submodels a priori by allowing the coefficients to be exactly zero with positive probability. Therefore, our Bayesian variable selection method is basically a discrete process where the predictors are either retained or dropped from the model. Our method has both advantages of selecting the important variables and further shrinking the coefficients to reduce predictive variability. Therefore, we arrive at the more parsimonious and interpretable model more efficiently. To use our methods, users only need to specify a prior on $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, and σ, a relatively simple task. Our methods retain the desirable features of George and McCulloch, such as avoiding the forbidding problem of evaluating posterior probabilities of all 2^p models and identifying the promising submodels from the data and prior. We test our programs (Sections 4 and 5) on several simulated data sets. Our results reveal that we make good decisions on the correct models.

We can compute the posterior covariance matrix of $\boldsymbol{\vartheta}$ from the Gibbs sampler. This matrix measures the variation of $\boldsymbol{\vartheta}$ that incorporates model uncertainty. In our opinion, reporting this measure is more realistic than reporting the usual covariance matrix for a fixed model, presently in practice. Having determined the model, one can compute the posterior conditional covariance matrix of $\boldsymbol{\beta}$ (with only the coefficients specified by the model included) by rerunning the

Gibbs sampler. This conditional posterior covariance matrix corresponds to the measure commonly used. Recently, several papers discuss measurement for model uncertainty, for example, Draper (1995), Raftery, Madigan and Hoeting (1997). Our paper provides an alternative perspective for measuring model uncertainty.

In addition to the usual linear model (Section 2), we also consider generalized linear models (Section 3) as in McCullagh and Nelder (1989).

Bayeisan variable selection is an active research area recently. In addition to the papers we have cited, the numerous related articles include Clyde, Desimone, and Parmigiani (1996), Clyde and Parmigiani (1996), Green (1995), Hoeting, Raftery, and Madigan (1996, 1997), Madigan and Raftery (1994), and Raftery, Madigan, and Hoeting (1997).

We assume the readers are familiar with the MCMC method for Bayesian computation. Geman and Geman (1984), Tanner and Wong (1987), Gelfand and Smith (1990), Casella and George (1992) and Tierney (1994) provide general tutorials on the MCMC algorithm.

Section 2 develops two MCMC algorithms for the usual regression model. One algorithm updates $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ in a cycle. It has the advantages of being easy to implement and to extend to nonconjugacy cases. Moreover, it can be easily modified for the nested model in (1.2). The other algorithm is similar to the one in Geweke (1996). It updates $\boldsymbol{\vartheta}$ directly. We implemented both algorithms and compared the efficiency of these two algorithms. We found they are quite comparable. Section 3 develops the general methodology for the generalized linear model. Section 4 provides numerical examples on the linear model: Subsection 4.1 describes our results on three simulated data sets and Subsection 4.2 applies our method to the Hald data set in Draper and Smith (1981) and to an aerobic fitness data set given in SAS (SAS Institute, 1985). Subsection 5.1 gives simulated data examples for the generalized linear model and subsection 5.2 applies a generalized linear model to a real data set given in Feigl and Zelen (1965).

2. The Expanded Linear Regression Model

In this section, we describe our regression model that incorporates all possible subsets. We describe two Markov chain Monte Carlo algorithms that identify the subsets of promising predictors.

Let us consider the following expanded linear regression model

$$\mathbf{y}|\boldsymbol{\beta}, \boldsymbol{\gamma}, \sigma^2 \sim N_n(\mathbf{X}\boldsymbol{\vartheta}, \sigma^2 I), \qquad \ldots (2.1)$$

where $\mathbf{y}$ is the $n \times 1$ response variables, $\mathbf{X} = [\mathbf{x_1}, \ldots, \mathbf{x_p}]$ is the $n \times p$ matrix of covariates with $\mathbf{x}_j = (x_{1j}, \ldots, x_{nj})^T$, $\boldsymbol{\vartheta} = (\beta_1\gamma_1, \ldots, \beta_p\gamma_p)^T$, and σ is a scalar.

Let us first consider the following simple prior that chooses $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, and σ independently, where $\boldsymbol{\beta} \sim N_p(\boldsymbol{\beta}_0, \mathbf{D}_0)$, $\gamma_j \sim B(1, p_j)$ independently for $j =$

$1, \ldots, p$, and

$$\pi(\sigma) \propto \frac{1}{\sigma^{\alpha+1}} \exp\{-\frac{\eta}{2\sigma^2}\},\ \alpha > 0,\ \eta > 0.$$

We denote this prior density of σ by $IG'(\alpha, \eta)$, the modified inverse gamma density. It is equivalent to choosing σ^2 with the inverse gamma density $IG(\alpha/2, 2/\eta)$ as defined by Berger (1985 p. 561). We can also let $\alpha \to 0$ and $\eta \to 0$ to mimic the non-informative prior $\pi(\sigma) = 1/\sigma$.

In the following, we develop two algorithms. One updates $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ sequentially. The other updates the regression coefficients $\boldsymbol{\vartheta}$ directly as in Geweke (1996) and keeps track of the indicator variables $\boldsymbol{\gamma}$ for evaluating the posterior probabilities of different submodels.

Now, let us discuss the first algorithm. The potentially promising predictors can be identified from $\boldsymbol{\gamma}$'s that have high posterior probabilities. Therefore, we are interested in evaluating $\mathrm{P}(\boldsymbol{\gamma}|\mathbf{y})$. This can be done by the Gibbs sampling. Starting with an initial choice of $\boldsymbol{\beta}^0,\ \boldsymbol{\gamma}^0,\ \sigma^0$, we generate a Gibbs sample for $\boldsymbol{\beta}^1,\ \boldsymbol{\gamma}^1,\ \sigma^1,\ \boldsymbol{\beta}^2,\ \boldsymbol{\gamma}^2, \sigma^2$, etc., using the following conditional densities. The least squared estimates of $\boldsymbol{\beta}$ and σ for the full model with $\gamma_j = 1$ in (1.1) can be used for $\boldsymbol{\beta}^0$ and σ^0; $\boldsymbol{\gamma}^0$ can be set to $(1, \ldots, 1)^T$ initaially. Then $\mathrm{P}(\boldsymbol{\gamma}|\mathbf{y})$ is tabulated from the frequencies of $\boldsymbol{\gamma}$ in the Gibbs sample.

Now we describe the conditional densities needed in the Gibbs algorithm. Let $\mathbf{X}^* = [\gamma_1\mathbf{x_1}, \ldots, \gamma_p\mathbf{x_p}]$. Given the above prior, we can show the posterior distribution of $\boldsymbol{\beta}$ given $\boldsymbol{\gamma},\ \sigma,\ \mathbf{y}$ is $N_p(\tilde{\boldsymbol{\beta}}, \mathbf{D})$, where the posterior mean is $\tilde{\boldsymbol{\beta}} = (\mathbf{D}_0^{-1} + \sigma^{-2}\mathbf{X^*}^T\mathbf{X}^*)^{-1}(\mathbf{D}_0^{-1}\boldsymbol{\beta}_0 + \sigma^{-2}\mathbf{X^*}^T\mathbf{y})$, and the posterior covariance is $\mathbf{D} = (\mathbf{D}_0^{-1} + \sigma^{-2}\mathbf{X^*}^T\mathbf{X}^*)^{-1}$. To obtain the posterior density of $\boldsymbol{\gamma}$ given $\boldsymbol{\beta},\ \sigma,\ \mathbf{y}$, we sample variates γ_j with $j = 1, \ldots, p$, preferably in random order from the the posterior distribution of γ_j given $\boldsymbol{\gamma}_{-j},\ \boldsymbol{\beta},\ \sigma,\ \mathbf{y}$, where $\boldsymbol{\gamma}_{-j} = (\gamma_1, \ldots, \gamma_{j-1}, \gamma_{j+1}, \ldots, \gamma_p)$. The posterior distribution of γ_j given $\boldsymbol{\gamma}_{-j},\ \boldsymbol{\beta},\ \sigma,\ \mathbf{y}$, is Bernoulli $B(1, \tilde{p_j})$ with $\tilde{p_j} = c_j/(c_j + d_j)$, where

$$c_j = p_j \exp\{-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\vartheta}_j^*)^T(\mathbf{y} - \mathbf{X}\boldsymbol{\vartheta}_j^*)\}$$

and

$$d_j = (1 - p_j) \exp\{-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\vartheta}_j^{**})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\vartheta}_j^{**})\}.$$

The vector $\boldsymbol{\vartheta}_j^*$ is the column vector of $\boldsymbol{\vartheta}$ with the jth entry replaced by β_j, similarly, $\boldsymbol{\vartheta}_j^{**}$ is obtained from $\boldsymbol{\vartheta}$ with the jth entry replaced by 0. The posterior distribution of σ given $\boldsymbol{\beta},\ \boldsymbol{\gamma}$, and $\mathbf{y}$ is $IG'(\alpha + n, \eta + (\mathbf{y} - \mathbf{X}\boldsymbol{\vartheta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\vartheta}))$.

Frequentists may be concerned with the identifiability of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ in (1.1). This concern may be eased by noting that we are really interested in ϑ_j for all j. They are always identifiable. If we insist upon to identify β_j and γ_j separately, we can assume that β_j does not take on the 0 value. Bayesian analysis is not restricted by this assumption. Bayesian identifiability concerns the issue of whether or not the data and prior provide information about the parameters $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$. Let us examine the simpler case where $\mathbf{D}_0$ is diagonal with σ_j^2 in the j^{th} entry of the diagonal. Let $\boldsymbol{\beta}_{-j}$ denote $\boldsymbol{\beta}^T$ with β_j deleted. Let $\beta_{0,j}$ denote the j^{th} row of $\boldsymbol{\beta}_0$. Then the conditional density of $\boldsymbol{\beta}$ given $\boldsymbol{\gamma}$, σ and $\mathbf{y}$, $N_p(\tilde{\boldsymbol{\beta}}, \mathbf{D})$, can be obtained by sampling β_j, $j = 1, \ldots, p$, sequentially or in random order, from the distribution

$$\beta_j | \boldsymbol{\beta}_{-j}, \boldsymbol{\gamma}, \mathbf{y} \sim N\left(\frac{\beta_{0,j}\sigma^2 + b_j\sigma_j^2}{\sigma^2 + a_j\sigma_j^2}, \frac{\sigma^2\sigma_j^2}{\sigma^2 + \sigma_j^2 a_j}\right), \qquad \ldots (2.2)$$

where

$$a_j = \gamma_j^2 \sum_{i=1}^{n} x_{ij}^2,$$

and

$$b_j = \gamma_j \sum_{i=1}^{n} x_{ij}(y_i - \sum_{l \neq j} \beta_l \gamma_l x_{il}).$$

When $\gamma_j = 0$, then $a_j = b_j = 0$. Therefore it follows from (2.2) that data do not provide information about β_j while the prior does. This is expected because the expanded regression equation when $\gamma_j = 0$ does not include β_j in the model.

The sampling from prior part of the above procedure is also related to the MCMC developed by Carlin and Chib (1995). Our prior can also be thought as a pseudo-prior as discussed in Carlin and Chib when the predictors are not selected in the model.

Now let us consider the second algorithm that is given in Geweke (1996) where updating is done on the $\boldsymbol{\vartheta}$. Let $\boldsymbol{\vartheta}_{-j}$ denote $\boldsymbol{\vartheta}^T$ with ϑ_j deleted. Because our prior assigns positive probability to zero for $\boldsymbol{\vartheta}_j$, the posterior distribution of ϑ_j given $\boldsymbol{\vartheta}_{-j}$ is also a mixture of point mass at zero and a normal density. Let us define $z_i = y_i - \sum_{l \neq j} \vartheta_l x_{il}$, $\tilde{b}_j = \sum_{i=1}^n x_{ij} z_i / \sum_{i=1}^n x_{ij}^2$, $w_j^2 = \sigma^2 / \sum_{i=1}^n x_{ij}^2$, $\tilde{\sigma}_j^2 = 1/(w_j^{-2} + \sigma_j^{-2})$, and $\tilde{\mu}_j = \tilde{\sigma}_j^2(\tilde{b}_j/w_j^2 + \beta_{0j}/\sigma_j^2)$. A straightforward computation shows that

$$\Pr(\vartheta_j = 0 | \boldsymbol{\vartheta}_{-j}, \sigma, \mathbf{y}) = \frac{1 - p_j}{1 - p_j + p_j \mathrm{BF}}, \qquad \ldots (2.3)$$

where

$$\mathrm{BF} = \frac{\tilde{\sigma}_j}{\sigma_j} \exp\left\{-\frac{(\beta_{0j} - \tilde{b}_j)^2}{2(\sigma_j^2 + w_j^2)} + \frac{\tilde{b}_j^2}{2w_j^2}\right\}.$$

Therefore, the Gibbs sampler consists of updating ϑ_j given $\boldsymbol{\vartheta}_{-j}$, σ, and data in random order for $j = 1, \ldots, p$, and updating σ given $\boldsymbol{\vartheta}$ and $\mathbf{y}$. The first updating is done by setting ϑ_j to zero with probability given in (2.3) or generating ϑ_j from the $N(\tilde{\mu}_j, \tilde{\sigma}_j^2)$ with the remaining probability, for $j = 1, \ldots, p$, in random order. The second updating is done similarly as before, that is generate σ from the $IG'(\alpha + n, \eta + (\mathbf{y} - \mathbf{X}\boldsymbol{\vartheta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\vartheta}))$. We set γ_j to zero, when ϑ_j equals zero in each iteration; and γ_j to 1, otherwise. Each submodel is uniquely represented by the $\boldsymbol{\gamma}$ vector. The probabilities for submodels are obtained exactly as before by counting the frequency of each submodel.

We can also consider alternative priors. For the prior of $\boldsymbol{\gamma}$, we may wish to assign weight according to model size as suggested by George and McCulloch (1993) where $\pi(\boldsymbol{\gamma}) = w_{|\boldsymbol{\gamma}|}\binom{p}{|\boldsymbol{\gamma}|}^{-1}$ and $|\boldsymbol{\gamma}|$ denotes the number of ones (size) of $\boldsymbol{\gamma}$. We can assign more weight to parsimonious models by setting $w_{|\boldsymbol{\gamma}|}$ large for small $|\boldsymbol{\gamma}|$. Instead of assuming independent priors for $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, and σ, we can also formulate dependent priors as in George and McCulloch: $\beta_j|\gamma_j \sim (1-\gamma_j)N(0,\tau_j^2)+\gamma_j N(0,c_j^2\tau_j^2)$ independently with known c_j and τ_j, where $\gamma_j = 0$ or 1; $\sigma^2|\boldsymbol{\gamma}$ is an inverse gamma where the parameter can depend on $\boldsymbol{\gamma}$; and the prior on $\boldsymbol{\gamma}$ can be either the above choice or the independent Bernoulli distributions with known parameters p_j discussed much earlier. This mixture-of-normal-prior formulation has the desirable feature of modeling β_j close to 0 with very small variance when γ_j is 0 (j^{th} predictor not in the model). Having discussed earlier how the prior drives the generation of β_j in the Gibbs sampler, when γ_j is 0, we see this prior can reduce the variance of β_j in the Gibbs sampler. Instead of independent prior for $\boldsymbol{\beta}$ and σ, we can also consider prior for σ and conditional prior for $\boldsymbol{\beta}$ given σ. We could adopt the g-prior proposed by Zellner (1986). In addition to the normal prior for $\boldsymbol{\beta}$, we can also consider other non-conjugate priors that include the double exponential distribution (Tibshirani, 1996). The MCMC algorithms can be developed for these priors.

3. The Generalized Linear Model

In this section, we extend our treatment of the usual regression model to the generalized linear model (GLM) of McCullagh and Nelder (1989). We describe the MCMC algorithm that identifies the promising subsets of predictors.

In GLM, the distribution of y_i is assumed to belong to an exponential family

$$f(y_i|\boldsymbol{\beta},\boldsymbol{\gamma},\phi) = \exp\{(y_i\theta_i - b(\theta_i))/a_i(\phi) + c(y_i,\phi)\} \qquad \ldots (3.1)$$

where $\mu_i = E(y_i) = b'(\theta_i)$ and $\mathrm{var}(y_i) = b''(\theta_i)/a_i(\phi)$. The functions a_i, b, and c are known. Furthermore, the linear predictor $\boldsymbol{\eta}$ is related to the mean $\boldsymbol{\mu}$ by a link function g such that $g(\boldsymbol{\mu}) = \boldsymbol{\eta} = \sum_{j=1}^{p} \mathbf{x}_j\beta_j\gamma_j$, where $\boldsymbol{\eta} = (\eta_1,\ldots,\eta_n)^T$ and $\boldsymbol{\mu} = (\mu_1,\ldots,\mu_n)^T$. The link function g can be any monotonic differentiable function. It is usually taken to be the canonical link $g(\boldsymbol{\mu}) = (b')^{-1}(\boldsymbol{\mu})$. In the binomial and Poisson models, $a_i(\phi)$ is a known constant. In other models, $a_i(\phi)$ is commonly $a_i(\phi) = \phi/w_i$, where the weights w_i's are known prior weights ("sample size") and the dispersion parameter ϕ can also be denoted by σ^2. Therefore, the likelihood function can be written as

$$L(\boldsymbol{\beta},\boldsymbol{\gamma},\phi|\mathbf{y}) = \prod_{i=1}^{n} \exp\{w_i(\theta_i y_i - b(\theta_i))/\phi + c(y_i,\phi)\}, \qquad \ldots (3.2)$$

where $\theta_i = (b')^{-1}(\mu_i)$ and $\mu_i = g^{-1}(\mathbf{x}_{(i)}\boldsymbol{\vartheta})$. Recall $\boldsymbol{\vartheta} = (\beta_1\gamma_1, \ldots, \beta_p\gamma_p)^T$. Note previously we have used $\mathbf{x}_j$ to denote the j^{th} column vector of the matrix $\mathbf{X}$, whereas we use $\mathbf{x}_{(i)}$ to denote the i^{th} row vector of the matrix $\mathbf{X}$, i.e., the p dimensional covariates of the i^{th} subject. The likelihood in (3.2) reduces to $L(\boldsymbol{\beta}, \boldsymbol{\gamma}|\mathbf{y})$ for the Poisson and binomial models.

For simplicity, we assume independent priors for $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, and ϕ. This independence assumption can be relaxed. We can assume quite general classes of priors for $\boldsymbol{\beta}$. The Metropolis algorithm (1953) can be used to generate $\boldsymbol{\beta}$ given $\boldsymbol{\gamma}$, ϕ, and $\mathbf{y}$ when this conditional density is not easily identified. The Metropolis-within-Gibbs method is used to generate the sample $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, ϕ. Müller (1994), Tierney (1994) and Chib and Greenberg (1995) provide further details on the method. We can also take advantage of the adaptive rejection sampling method of Gilks and Wild (1992) by considering a log-concave prior for $\boldsymbol{\beta}$. This includes the multivariate normal with mean $\boldsymbol{\beta}_0$ and covariance $\mathbf{D}_0$. On sampling for the $\boldsymbol{\beta}$ given $\boldsymbol{\gamma}$, ϕ, $\mathbf{y}$, the adaptive rejection sampling method within Gibbs can be used to sample $\beta_1, \ldots, \beta_p$ sequentially or in random order. The adaptive rejection method constructs piecewise linear upper and lower bounds for the concave function to be used in the rejection step. It is adaptive in the sense the density function is closer to the upper and lower functions constructed to squeeze it when more random variates are sampled from the envelop function. As pointed out by Dellaportas and Smith (1993), the adaptive rejection method can be used for all the canonical link functions in GLM. They also list several non-canonical link functions for which Gilks and Wild methods can be applied where the log-concavity of the likelihood function is essential. The prior on the $\boldsymbol{\gamma}$ is the same as the one in the linear model. The prior on ϕ can be either inverse gamma or arbitrary.

As in linear models, our objective is to find the posterior distribution of $\boldsymbol{\gamma}$ from the MCMC algorithm. Then we identify the subsets of predictors with high posterior probabilities. We summarize the MCMC algorithm briefly. Starting with an initial choice of $\boldsymbol{\beta}^0$, γ^0, σ^0, we generate the Gibbs sample of $\boldsymbol{\beta}^1$, γ^1, ϕ^1, $\boldsymbol{\beta}^2$, γ^2, ϕ^2, etc., using the following conditional densities. Then $\mathrm{P}(\boldsymbol{\gamma}|\mathbf{y})$ is tabulated from the frequencies of $\boldsymbol{\gamma}$ in the Gibbs sample. We first describe how to sample the variate $\boldsymbol{\beta}$ given $\boldsymbol{\gamma}$, ϕ, $\mathbf{y}$. It can be done by either the Metropolis algorithm or the Gilks and Wild method. For the Metropolis algorithm, let us assume the current $\boldsymbol{\beta}$ is $\boldsymbol{\beta}^{(i)}, i = 0$, where the superscript i is the number of iterations in the Metropolis step. Then we generate a multivariate normal variate $\mathbf{z}$ from $N_p(0, \Sigma)$, where Σ is the current estimate of the posterior covariance matrix of $\boldsymbol{\beta}$. Let $\boldsymbol{\beta}^* = \boldsymbol{\beta}^{(i)} + c\mathbf{z}$. Then,

$$\boldsymbol{\beta}^{(i+1)} = \begin{cases} \boldsymbol{\beta}^* & \text{with probability } p^{(i)} \\ \boldsymbol{\beta}^{(i)} & \text{with probability } 1 - p^{(i)} \end{cases}$$

where

$$p^{(i)} = \min\{1, L(\boldsymbol{\beta}^*, \boldsymbol{\gamma}, \phi|\mathbf{y})\pi(\boldsymbol{\beta}^*)/(L(\boldsymbol{\beta}^{(i)}, \boldsymbol{\gamma}, \phi|\mathbf{y})\pi(\boldsymbol{\beta}^{(i)}))\}.$$

We continue this iteration until reaching equilibrium, say at $i = I$. Then $\boldsymbol{\beta} = \boldsymbol{\beta}^{(I)}$ is the desired vector for the Metropolis algorithm. In our experience, I varying from 20 to 50 steps suffices. The scaler c is adjusted to achieve a desirable staying rate. If the log-concavity conditions are satisfied for prior and likelihood, then we can replace the above Metropolis algorithm by the Gilks and Wild method, i.e., we update β_j given $\boldsymbol{\beta}_{-j}$, $\boldsymbol{\gamma}$, $\mathbf{y}$, one at a time in random order for $j = 1, \ldots, p$ using the adaptive rejection method. Now we describe how to update $\boldsymbol{\gamma}$. We update the variate γ_j given $\boldsymbol{\gamma}_{-j}$, $\boldsymbol{\beta}$, ϕ, $\mathbf{y}$ by a Bernoulli distribution $B(1, \tilde{c}_j/(\tilde{c}_j + \tilde{d}_j)$ We compute $\tilde{c}_j$ and $\tilde{d}_j$ by

$$\tilde{c}_j = p_j L(\boldsymbol{\beta}, \boldsymbol{\gamma}_j^*, \phi|\mathbf{y})$$

and

$$\tilde{d}_j = (1 - p_j) L(\boldsymbol{\beta}, \boldsymbol{\gamma}_j^{**}, \phi|\mathbf{y}),$$

where the likelihood functions L are given in (3.2), the vector $\boldsymbol{\gamma}_j^*$ ($\boldsymbol{\gamma}_j^{**}$) is obtained from $\boldsymbol{\gamma}$ with the j^{th} entry replaced by 1 (0). In many cases, the posterior distribution of ϕ given $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, and $\mathbf{y}$ is the updated inverse gamma distribution when the prior is inverse gamma. The Metropolis algorithm can be used again for cases of non-conjugacy.

4. Numerical Examples for Linear Models

4.1 *Simulated examples.* In this subsection we illustrate the performance of our method on simulated examples. The two problems of example 4.1.1 treat small problems involving five potential predictors. Example 4.1.2 considers a large problem with 30 potential predictors, which is comparable to a feasible range of most practical problems. All of our simulation examples are designed to be similar to George and McCulloch (1993) because readers may be interested in the comparison.

EXAMPLE 4.1.1. This example considers two simple, variable selection problems with $p = 5$ predictors of length $n = 60$. In problem 1, the predictors were obtained as independent standard normal vectors, $\mathbf{x}_1, \ldots, \mathbf{x}_5$ i.i.d. $\sim N_{60}(0, \mathbf{I})$. The dependent variable was generated according to the model

$$\mathbf{y} = \mathbf{x}_4 + 1.2\mathbf{x}_5 + \boldsymbol{\epsilon}$$

where $\boldsymbol{\epsilon} \sim N_{60}(0, \sigma^2 I)$ with $\sigma = 2.5$. Thus $\boldsymbol{\beta} = (0, 0, 0, 1, 1.2)^T$. The least squares estimates for these data were $\hat{\boldsymbol{\beta}} = (.01, .44, .48, .95, 1.31)^T$, with standard errors $\hat{\sigma}_{\boldsymbol{\beta}}$=(.42,.39,.41,.36,.43) and $\hat{\sigma}$=2.738.

Problem 1. We applied our method to this problem with priors chosen as $p_j = .5$, for $j = 1, \ldots, 5$; $\boldsymbol{\beta}_0 = (0, \ldots, 0)^T$; $D_0 = 16\mathbf{I}$; and $\sigma \sim IG'(.01, .01)$.

The same priors with dimension modifications are used in the next problem and the next example. Note that we chose the prior standard deviation of β_j to be 4, for all j, much larger than $\hat{\sigma}_{\boldsymbol{\beta}}$ reported above to represent a relatively diffuse prior for $\boldsymbol{\beta}$. The prior on σ is chosen to be moderately non-informative. The choices are much easier to specify than they are for the method of George and McCulloch. A sample of 8,000 iterations in a single Markov chain was then simulated and tabulated. Table 1 displays the frequencies of the four highest frequency models.

Table 1. HIGH FREQUENCY MODELS, EXAMPLE 4.1.1, PROBLEM 1

Model variables	proportions
4 5	.67
2 4 5	.16
3 4 5	.11
5	.02

It is clear our analysis predicts the right model which is $\hat{\mathbf{y}} = f(\mathbf{x}_4, \mathbf{x}_5)$. Also it allows other promising models with $\mathbf{x}_2$ and $\mathbf{x}_3$ but always keeping the right covariates $\mathbf{x}_4, \mathbf{x}_5$. As the t-value of $\mathbf{x}_5$ (2.99) is slightly larger than that of $\mathbf{x}_4$ (2.64), so we see $\mathbf{x}_5$ singly in the model but for a very low proportion of the sample. From the Gibbs sample, we can also compute the posterior mean and standard deviation of $\boldsymbol{\vartheta}$. The posterior means are 1.28 and 1.55 for ϑ_4 and ϑ_5 with standard deviations .58 and .84. Each standard deviation incorporates our measure of model uncertainty. Each is larger than that used by the frequentists (.36, .43) as expected.

Problem 2. Problem 2 is identical to problem 1 except that $\mathbf{x}_3$ is replaced by $\mathbf{x}_3^* = \mathbf{x}_5 + .15\mathbf{z}$ where $\mathbf{z} \sim N_{60}(0, \mathbf{I})$, yielding corr$(\mathbf{x}_3, \mathbf{x}_5)$=.989. This $\mathbf{x}_3^*$ is a substantial proxy for $\mathbf{x}_5$. This problem is meant to illustrate how our method performs in the presence of extreme collinearity. Now the least squares estimates of the $\boldsymbol{\beta}$ were $\hat{\boldsymbol{\beta}}$=(.22,.13,-2.11,1.34,3.13) and $\hat{\sigma}_{\boldsymbol{\beta}}$=(.37,.38,2.4,.31,2.4). The result is different from problem 1 due to collinearity. The classical analysis based on p-values tells us to throw out everything except the 4th covariate.

We did the same analysis as for problem 1. The result is in Table 2. This is a problem where we have a strong proxy (almost identical) variables, so either one of $\mathbf{x}_3$ or $\mathbf{x}_5$ will do and our method still identifies the most promising models among them. The posterior mean and standard deviation of $(\vartheta_3, \vartheta_4, \vartheta_5)$ are (4.93,1.29,2.67) and (3.83,.59,3.06) computed from the Gibbs sampler. The larger standard deviation than that of the frequentist's seems to be more desirable due to model uncertainty.

Table 2. HIGH FREQUENCY MODELS, EXAMPLE 4.1.1, PROBLEM 2

Model variables	proportions
3 4	.7
4 5	.23
3 4 5	.04
3	.02

Example 4.1.2. This example is created to demonstrate the practical potential of our approach for data sets with a relatively large number of covariates. We constructed $p = 30$ predictors, $\mathbf{x}_1, \ldots, \mathbf{x}_{30}$, for a sample size of $n = 60$. They were obtained as $\mathbf{x}_j = \mathbf{x}^*{}_j + \mathbf{z}$, where $\mathbf{x}^*{}_j$ i.i.d. $\sim N_{60}(0, \mathbf{I})$, $j = 1, \ldots, 30$, independently of $\mathbf{z} \sim N_{60}(0, \mathbf{I})$. This induced pairwise correlations of about .5. The dependent variables were generated according to the model $\mathbf{y} = [\mathbf{x}_1, \cdots, \mathbf{x}_{30}]\boldsymbol{\beta} + \boldsymbol{\epsilon}$ where $\boldsymbol{\epsilon} \sim N_{60}(0, \sigma^2\mathbf{I})$ with $\sigma = 2$. The coefficients $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{30})$ were set at $(\beta_1, \ldots, \beta_{10}) = (0, \ldots, 0)$, $(\beta_{11}, \ldots, \beta_{20}) = (1, \ldots, 1)$, and $(\beta_{21}, \ldots, \beta_{30}) = (2, \ldots, 2)$. A sample of $20,000$ observations of the Gibbs sequence was then simulated and tabulated.

Table 3 lists the highest frequencies of false identified choices. False choice is defined to be at least one of the predictors in 1 to 10 is included or at least one of the predictors in 11 to 30 is deleted. So 68% of the times the Gibbs sample contain the right variables that are the variables numbered from 11 to 30; 9% of the Gibbs sample contain false choice with $\mathbf{x}_2$ included; 6% with $\mathbf{x}_2$ included and $\mathbf{x}_{14}$ excluded, etc. So here our majority result is correct, but some variation is allowed for lower probability models too.

Table 3. HIGH FREQUENCY MODELS FOR FALSE CHOICES, EXAMPLE 4.1.2

Relative frequency	false choice
.68	None
.09	2
.06	2, 14
.03	8
.02	16

4.2 *Real data examples.* In this subsection we apply our method to two real data sets.

Example 4.2.1. The data for our first real example is the familiar Hald data (Draper and Smith 1981), which have been used by various authors to illustrate variable selection procedures. The data consist of n=13 observations on a dependent variable $\mathbf{y}$ (heat evolved during a chemical reaction) and p=4 independent variables $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4$ (inputs to the reaction). Thus 2^4=16 possible models are under consideration. George and McCulloch (1993) state: "As described by Draper and Smith (1981), three models were favored by conventional selection procedures. The model $\hat{y} = f(\mathbf{x}_1, \mathbf{x}_2)$ yielding $R^2 = 97.9\%$, was favored by all subsets regression, backward elimination, and stepwise regression; the model $\hat{y} = f(\mathbf{x}_1, \mathbf{x}_4)$ was also favored by all subset regression; and the model $\hat{y} = f(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_4)$ was favored by forward selection."

We applied our procedure: $30,000$ iterations of the Gibbs sequence were then simulated, and the higher frequency models were tabulated in Table 4. Our results capture $\hat{\mathbf{y}} = f(\mathbf{x}_1, \mathbf{x}_2)$ as the best model.

Table 4. HIGH FREQUENCY MODELS, EXAMPLE 4.2.1

model variable	proportions
1,2	.290
1,4	.006
1,2,3	.190
1,2,4	.210
1,3,4	.120
2,3,4	.005
1,2,3,4	.179

Example 4.2.2. Now we take an example from *SAS User's Guide: Statistics* (1985, p. 696) with the aerobic fitness data. Aerobic fitness (measured by the ability to consume oxygen) is fit to the results of some simple exercise tests. The goal is to develop an equation to predict fitness based on the exercise tests rather than on the expensive and cumbersome oxygen consumption measurement. The variables are age (years), weight (KG), oxygen uptake rate (ML per KG body weight per minute, the response variable), time to run 1.5 miles in minutes (runtime), heart rate while resting (rstpulse), heart rate while running (runpulse) (same time oxygen rate measured), and maximum heart rate records while running (maxpulse). The SAS analysis shows all variables except weights and heart rate while resting are significant. The p-values for runtime, age, maxpulse, runpulse, weight, and rstpulse are .0001, .0051, .0322, .0360, .1869, .7473 respectively.

We used our methods generating a Gibbs sample of size 8000 and tabulated the higher frequency models in Table 5. Our results agree with SAS on the variables to be deleted.

Table 5. HIGH FREQUENCY MODELS, EXAMPLE 4.2.2

Model variables	proportions
runtime, runpulse, maxpulse	.58
age, runtime, runpulse	.23
age, runtime, runpulse, maxpulse	.15
runtime	.03

5. Numerical Examples for Generalized Linear Models

5.1 *Simulated examples.* In this subsection we illustrate the performance of our method for the generalized linear model with simulated data.

EXAMPLE 5.1. This example considers two simple, variable selection problems with $p = 5$ predictors of length $n = 60$. The predictors were obtained as independent standard normal vectors, $\mathbf{x}_1, \ldots, \mathbf{x}_5$ i.i.d. $\sim N_{60}(0, \mathbf{I})$. The dependent variable was generated according to the model

$$\boldsymbol{\eta} = \log(E(\mathbf{y})) = .8\mathbf{x}_4 + \mathbf{x}_5$$

where $\mathbf{y} \sim$ Poisson($\exp\{\boldsymbol{\eta}\}$). Thus $\boldsymbol{\beta} = (0, 0, 0, .8, 1.0)^T$. The GLIM analysis for these data were $\hat{\boldsymbol{\beta}}$=(-.14,-.02,-.05.,.65,1.27), with standard errors $\hat{\sigma}_{\boldsymbol{\beta}} = (.1, .12, .16, .14, .16)$.

Problem 1. We applied our method to this problem with prior probability of choice as $p_j = .5, j = 1, \ldots, 5$ and independent normal priors on β's with mean 0 and variance 16. A sample of 8,000 observations of the Gibbs sequence was then simulated and tabulated. We got the right model, with only 4th and 5th covariates, an extremely high percentage of the time (more than 90%). Our posterior estimates for ϑ_4 and ϑ_5 are respectively .89 and 1.07 with standard deviations .24 and .32.

Problem 2 is identical to problem 1 except that $\mathbf{x}_3$ is replaced by $\mathbf{x}_3^* = \mathbf{x}_5 + .15\mathbf{z}$ where $\mathbf{z} \sim N_{60}(0, \mathbf{I})$, yielding corr$(\mathbf{x}_3, \mathbf{x}_5)$ =.989. This $\mathbf{x}_3^*$ is a substantial proxy for $\mathbf{x}_5$. This problem is meant to illustrate how our method performs in the presence of extreme collinearity. We did the same analysis as for problem 1. Table 6 shows the result. So again the presence of proxy variables can affect the result.

Table 6. FREQUENCY MODELS, EXAMPLE 5.1, PROBLEM 2

Model variables	proportions
3 4	.6
4 5	.25
3 4 5	.1
4	.01

5.2 *Real data example.* The data set presented by Feigl and Zelen (1965) involves 33 patients suffering from leukemia. The response is the patient's survival time in weeks. There are two covariates, a discrete covariate of AG-factor which has 2 levels, positive and negative and a continuous covariate of patient blood cell count (in log units). We consider a model with three predictors: AG-factor, log blood count and the interaction term between them. We assume an exponential distribution for the survival times and a log link function. Here the log-linear model for the mean also implies a log-linear model for the hazard function. We want to keep the hierarchical structure (don't want to get interaction without any of the main effect). We can either model it using (1.2) with independent priors or using the following dependent prior distribution over γ_js. If γ_1, γ_2 and γ_3 are the indicator variables corresponding to AG-factor, blood cell count and the interaction, then the conditional prior distribution for γ_3 given γ_1, γ_2 will be Bernoulli $B(1, p_3 = 0)$ if any one of γ_1, γ_2 is zero. So the distribution of γ_3 given γ_1, γ_2 is $B(1, p_3)$ where $p_3 \neq 0$ if $\gamma_1 = 1$ and $\gamma_2 = 1$, $p_3 = 0$ otherwise.

We applied our procedure with 20,000 iterations of the Gibbs sequence and presented the results in table 7. They show the presence of the main factors but not the interaction term. So our methodology could be easily extended to survival models like the proportional hazard models.

Table 7. FREQUENCY MODELS FOR EXAMPLE 5.2:

model variable	proportions
AG-Factor	.11
Blood count	.13
AG-factor, Blood Count	.63
AG-factor, Blood Count, and Interaction	.13

6. Conclusion

We propose a simple method for subset selection of independent variables in regression models. We expand the usual regression equation to an equation that incorporates all possible subsets of predictors by adding indicator variables as parameters. The vector of indicator variables dictates which predictors to include. The posterior distribution of the indicator vector is approximated by means of the Markov chain Monte Carlo algorithm. We select subsets with high posterior probabilities. This expansion method can be thought as a technique to implement the mixture prior (for the regression coefficient) with a point mass and a continuous part. The simplicity of this method allows us to easily apply it to more complex models for subset selection that include (1) non-conjugate cases where one cannot integrate out the other parameters to get the posterior distribution of the model $\boldsymbol{\gamma}$; (2) generalized linear models; and (3) model building where the main effects are retained when the interaction effects are also present.

Acknowledgement. This paper is revised from the Technical Report #94-26, University of Connecticut, Statistics Department. The authors wish to thank Michael P. Cohen, Alan Gelfand, Jennifer Hoeting, and a referee for their discussions.

References

BERGER, J. O. (1985), *Statistical Decision Theory and Bayesian Analysis* 2nd ed., New York: Springer-Verlag

AKAIKE, H. (1974), A New look at the statistical identification model, *IEEE Transactions on Automatic Control,* **19**, 716-723.

CARLIN, B. P., AND CHIB, S. (1995), Bayesian model choice via Markov Chain Monte Carlo, *Jour. Royal Statist. Soc.,* Ser. B, **57**, 473-484.

CASELLA, G. AND GEORGE, E.I. (1992), Explaining the Gibbs sampler, *The American Statistician*, **46**, 167-174.

CHIB, S. AND GREENBERG, E. (1995), Understanding the Metropolis-Hastings Algorithm, *the American Statistician*, **49**, 327-335.

DELLAPORTAS, P. AND SMITH, A.F.M. (1993), Bayesian inference for generalized linear and proportional hazards models via Gibbs sampling, *Applied Statistics*, **42**, 443-459.

CLYDE, M., DESIMONE, H., AND PARMIGIANI, G. (1996), Prediction via orthogonalized model mixing, *Jour. Amer. Statist. Assoc.,* **91**, 1197-1208.

CLYDE, AND PARMIGIANI (1996), Orthogonalizations and prior distributions for orthogonalized model mixing, in *Modelling and Prediction: Honoring Seymour Geisser,* edited by Jack C. Lee, Wesley O. Johnson and Arnold Zellner, Springer-Verlag, 206-227.

DRAPER, D. (1995), Assessment and propagation of model uncertainty (with discussion), *Jour. Royal Statist. Soc.,* **57**, 45-98.

DRAPER, N. AND SMITH, H. (1981), *Applied Regression Analysis* 2nd ed, New York: John Wiley.

FEIGL, P. AND ZELEN, M. (1965), Estimation of exponential probabilities with concomitant information. *Biometrics* **21**, 826-838.

GELFAND, A.E., AND SMITH, A.F.M. (1990), Sampling Based approaches to calculating marginal densities, *Jour. Amer. Statist. Assoc.,* **85**, 398-409.

GEMAN, S., AND GEMAN, D. (1984), Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images, *IEEE Transactions on Pattern Analysis and Machine Intelligence,* **6**, 721-741.

GEORGE, E.L., AND MCCULLOCH, R.E. (1993), Variable selection via Gibbs sampling, *Jour. Amer. Statist. Assoc.,* **88**, 881-889.

GEWEKE, J. (1996), Variable selection and model comparison in regression, *Bayesian Statistics 5,* eds: J. M. Bernard, J.O. Berger, A. P. Dawid, and A.F.M. Smith. Oxford University Press, 609-620.

GILKS, W.R. AND WILD, P. (1992), Adaptive rejection sampling for Gibbs sampling, *Applied Statistics,* **41**, 337-348.

GREEN, P. J. (1995), Reversible jump MCMC computation and Bayesian model determination, *Biometrika,* **82**, 711-732.

MCCULLAGH, P. AND NELDER, J.A. (1989), *Generalized Linear Models,* 2nd ed., London: Chapman and Hall.

HOETING, RAFTERY, AND MADIGAN (1996), A method for simultaneous variable selection and outlier identification in linear regression, *Computational Statistics and Data Analysis,* **22**, 251-270.

– – –– (1997), A method for simultaneous variable and transformation selection in linear regression, to appear in *Statistics and Computing.*

MADIGAN AND RAFTERY (1994), Model selection and accounting for model uncertainty in graphical models using Occam's window, *Jour. Amer. Statist. Assoc.,* **89**, 1335-1346.

METROPOLIS, N., ROSENBLUTH, A. W., ROSENBLUTH, M. N., TELLER, A. H., AND TELLER, E. (1953), Equation of state calculations by fast computing machines, *Journal of Chemical Physics,* **21**, 1087-1092.

MITCHELL, T. J. AND BEAUCHAMP, J. J. (1988), Bayesian variable selection in linear regression (with discussion), *Jour. Amer. Statist. Assoc.,* **83**, 1023-1036.

MÜLLER P. (1994), Metropolis posterior integration schemes, *ISDS Technical Report.*

RAFTERY, A. E., MADIGAN, D.M., AND HOETING, J. (1997), Model selection and accounting for model uncertainty in linear regression models, *Jour. Amer. Statist. Assoc.,* **92**, 179-191.

SCHWARZ, G. (1978), Estimating the dimension of a model, *Annals of Statistics,* **6**, 461-464.

SAS INSTITUTE (1985), *SAS User's Guide: Statistics,* ver.**5**, SAS Institute. Inc.

TANNER, M. AND WONG, W. (1987), The calculation of posterior distributions by data augmentation (with discussion), *Jour. Amer. Statist. Assoc.,* **82**, 528-550.

TIBSHIRANI, R. (1996), Regression shrinkage and selection via lasso, *Jour. Royal Statist. Soc.,* Ser. B, 267-288.

TIERNEY, L. (1994), Markov Chains for exploring posterior distributions, (with discussion), *Ann. Statist.,* **22**, 1701-1758.

ZELLNER, A. (1986), On assessing prior distributions and Bayesian regression analysis with g-prior distributions, *Bayesian Inference and Decision Techniques,* eds P.Goel and A. Zellner, Elsevier Science Publishers, 233-243.

LYNN KUO
DEPARTMENT OF STATISTICS U-120
UNIVERSITY OF CONNECTICUT
STORRS CT 06269-3120
U.S.A.
e-mail: lynn@stat.uconn.edu

BANI MALLICK
DEPARTMENT OF MATHEMATICS
IMPERIAL COLLEGE
LONDON SW7 2BZ
U.K.
b.mallick@ic.ac.uk