Variable Selection for Regression Models

LYNN KUO and BANI MALLICK [*]

Abstract

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.

KEY WORDS: Bayesian Inference; F-tests; Generalized linear model; Gibbs sampling; Linear model; Subset selection.

[*] Lynn Kuo is Associate Professor, Department of Statistics, U-120, University of Connecticut, Storrs, CT 06269-3120, and Bani Mallick is Lecturer, Department of Mathematics, Imperial College, London, England.

1. Introduction

Many methods have been proposed for selecting suitable predictors in multiple regression. Classical methods 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.

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 selections. The promising predictors are then identified by their more frequent appearance in the sequence of Gibbs samplers. 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

$$y_i = \sum_{j=1}^{p} \beta_j \gamma_j x_{ij} + \epsilon_i, \tag{1.1}$$

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.

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 large diagonal terms, so the analysis is focused on the likelihood. 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. The posterior probabilities can be evaluated by means of the Markov chain Monte Carlo (MCMC) method. In fact, 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 for 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 samplers. Many of the uninteresting submodels never or rarely appear in the Gibbs samplers.

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

Although our formulation is similar to that of George and McCulloch (1993), it differs in the following sense. We put all the regression coefficients and the indicator variables of whether or not to include a particular predictor directly in the model. This allows us to assume independent priors for the regression coefficients and indicator variables. This avoids the hierarchical setup of George and McCulloch where the prior of $\boldsymbol{\beta}$ is defined conditional on $\boldsymbol{\gamma}$. Our methods, therefore, avoid the complex issue of choosing tuning factors for the hyperparameters in the hierarchical setup. 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 better decisions on the correct models than George and McCulloch (1993).

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. We can compute the posterior covariance matrix of $\boldsymbol{\vartheta}$ from the Gibbs samplers. 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 samplers. This conditional posterior covariance matrix corresponds to the measure commonly used.

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 the MCMC algorithm for the usual regression model. 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). Section 5 gives simulated data examples for the generalized linear model.

2. The Expanded Linear Regression Model

In this section, we describe our regression model that incorporates all possible subsets. We describe the Markov Chain Monte Carlo algorithm that identifies the subsets of promising predictors.

Let us consider the following expanded linear regression model

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

where $\mathbf{y}$ is the $n \times 1$ response variables, $\mathbf{X} = [\mathbf{x_1}, \dots, \mathbf{x_p}]$ is the $n \times p$ matrix of covariates with $\mathbf{x}_j = (x_{1j}, \dots, x_{nj})^T$, $\boldsymbol{\vartheta} = (\beta_1\gamma_1, \dots, \beta_p\gamma_p)^T$, 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 $j = 1, \dots, p$, 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 noninformative prior $\pi(\sigma) = 1/\sigma$.

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 $\boldsymbol{\beta}^0$, $\boldsymbol{\gamma}^0$, σ^0, we generate Gibbs samplers for $\boldsymbol{\beta}^1$, $\boldsymbol{\gamma}^1$, σ^1, $\boldsymbol{\beta}^2$, $\boldsymbol{\gamma}^2$, σ^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 $(1, \dots, 1)^T$ initially. Then $\mathrm{P}(\boldsymbol{\gamma}|\mathbf{y})$ is tabulated from the frequencies of $\boldsymbol{\gamma}$ in the Gibbs samplers.

Now we describe the conditional densities needed in the Gibbs algorithm. Let $\mathbf{X}^* = [\gamma_1\mathbf{x_1}, \dots, \gamma_p\mathbf{x_p}]$. Given the above prior, we can show the posterior distribution of $\boldsymbol{\beta}$ given

$\boldsymbol{\gamma}$, σ, $\mathbf{y}$ 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 $\boldsymbol{\beta}$, σ, $\mathbf{y}$, we sample variates γ_j with $j = 1, \ldots, p$, preferably in random order from the the posterior distribution of γ_j given $\boldsymbol{\gamma}_{-j}$, $\boldsymbol{\beta}$, σ, $\mathbf{y}$, where $\boldsymbol{\gamma}_{-j} = (\gamma_1, \ldots, \gamma_{j-1}, \gamma_{j+1}, \ldots, \gamma_p)$. The posterior distribution of γ_j given $\boldsymbol{\gamma}_{-j}$, $\boldsymbol{\beta}$, σ, $\mathbf{y}$, 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 can be circumvented by assuming 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

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

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.

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 samplers, when γ_j is 0, we see this prior can reduce the variance of β_j in the Gibbs samplers. The MCMC algorithms can be developed for these priors. The algorithms can be obtained from the authors and are omitted here.

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

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

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 $g(\boldsymbol{\mu}) = \boldsymbol{\eta} = \sum_{j=1}^{p} \mathrm{x}_j\beta_j\gamma_j$, 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

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

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 samplers $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, ϕ. Müller (1994) and Chib and Greenberg (1994) 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 $\boldsymbol{\beta}^0$, $\boldsymbol{\gamma}^0$, σ^0, we generate Gibbs samplers of $\boldsymbol{\beta}^1$, $\boldsymbol{\gamma}^1$, ϕ^1, $\boldsymbol{\beta}^2$, $\boldsymbol{\gamma}^2$, 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 samplers. We first describe how to sample the variate $\boldsymbol{\beta}$ given $\boldsymbol{\gamma}$, ϕ, $\mathbf{y}$. It can be done by either the Metholopis 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 nonconjugacy.

4. Numerical Examples for Linear Models

4.1 Simulated Examples

In this subsection we illustrate the performance of our method on simulated examples. Example 4.1.1 treats small problems involving five potential predictors. Example 4.1.2 considers a large problem with 30 potential predictors, which is feasible range of most practical problems. Our methods should be applicable to problems with more predictors. 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.

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 noninformative. The choices are much easier to specify than they are for the method of George and McCulloch. A sample of 8,000 Gibbs samplers in a single Markov chain was then simulated and tabulated. Table 1 displays the frequencies of the four highest frequency models.

(Insert Table 1 around here)

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 samplers. From the Gibbs samplers, 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 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 samplers. The larger standard deviation than that of the frequentist's seems more desirable due to model uncertainty.

(Insert Table 2 around here)

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, $\mathbf{x}_1, \ldots, \mathbf{x}_{30}$, which is a plausible range for most of the practical problems for sample size $n = 60$. These 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 samplers contain the right variables that are variable numbers from 11 to 30; 9% of the Gibbs samplers 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.

(Insert Table 3 around here)

4.2 Real Data Examples

In this subsection we apply our method to two real data examples.

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: 5000 samplers of the Gibbs sequence were then simulated, and the higher frequency models were tabulated in Table 4. Our results capture the $\hat{\mathbf{y}} = f(\mathbf{x}_1, \mathbf{x}_2)$ model.

(Insert Table 4 around here)

Example 4.2.2. Now we took an example from *SAS User's Guide: Statistics* (1985, p. 696) which is known as 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 expensive and cumbersome oxygen consumption measurement. The variables are age (years), weight (KG), oxygen uptake rate (ML per KG body weight per minute), 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 collecting Gibbs samplers of size 8000 and tabulated the higher frequency models in Table 5. Our results agree with SAS on the variables to be deleted.

(Insert Table 5 around here)

5. Simulated Examples for Generalized Linear Models

In this section we illustrate the performance of our method in the case of generalized linear model, on simulated examples. Example 5.1 treats small problems involving five potential predictors.

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 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).

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 repaced 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.

(Insert Table 6 around here)

References

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

[2] Akaike, H. (1974), "A New Look at the Statistical Identification Model," *IEEE Transactions on Automatic Control,* 19, 716-723.

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

[4] Chib, S. and Greenberg, E. (1994), "Understanding the Metropolis-Hastings Algorithm," preprint.

[5] 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.

[6] Draper, N. and Smith, H. (1981), *Applied Regression Analysis* 2nd ed, New York: John Wiley.

[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] 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.

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

[10] Gilks, W.R. and Wild, P. (1992), "Adaptive Rejection Sampling for Gibbs Sampling," *Applied Statistics,* 41, 337-348.

[11] McCullagh, P. and Nelder, J.A. (1989), *Generalized Linear Models,* 2nd ed., London: Chapman and Hall.

[12] 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.

[13] 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.

[14] Müller P. (1994), "Metropolis Posterior Integration Schemes," *Statistics and Computing,* in press.

[15] Schwarz, G. (1978), "Estimating the Dimension of a Model," *Annals of Statistics,* 6, 461-464.

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

[17] Tanner, M. and Wong, W. (1987), "The Calculation of Posterior Distributions by Data Augmentation (with discussion)," *Journal of the American Statistical Association,* 82, 528-550.

[18] Tierney, L. (1991), "Markov Chains for Exploring Posterior Distributions," Technical Report No. 560, School of Statistics, University of Minnesota.

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

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

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

Table 4: High Frequency Models, Example 4.2.1

Model variables	proportions
1 2	.81
3 4	.14
2 3 4	.02
1 2 3 4	.03

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

Table 6: High Frequency Models, Example 5.1, Problem 2

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