▨연구논문

A Bayesian Method for Narrowing the Scope of Variable Selection in Binary Response Logistic Regression

Hea-Jung Kim · Ae-Kyung Lee
Dept. of Statistics, Dongguk University

Abstract

This article is concerned with the selection of subsets of predictor variables to be included in building the binary response logistic regression model. It is based on a Bayesian approach, intended to propose and develop a procedure that uses probabilistic considerations for selecting promising subsets. This procedure reformulates the logistic regression setup in a hierarchical normal mixture model by introducing a set of hyperparameters that will be used to identify subset choices. It is done by use of the fact that cdf of logistic distribution is approximately equivalent to that of $t_{(8)}/.634$ distribution. The appropriate posterior probability of each subset of predictor variables is obtained by the Gibbs sampler, which samples indirectly from the multinomial posterior distribution on the set of possible subset choices. Thus, in this procedure, the most promising subset of predictors can be identified as that with highest posterior probability. To highlight the merit of this procedure a couple of illustrative numerical examples are given.

1. Introduction

A vast literature in quality management, statistics, and biometrics is concerned with the analysis of binary response data. When the dependent variable of a regression model is observed to be qualitative variable expressed as binary output, we may consider a model given by

$$Y_i = H(\boldsymbol{X}_i'\boldsymbol{\beta}) + \varepsilon_i, \qquad (1)$$

where Y_i is a binary output, $\boldsymbol{X}_i$ is a $p \times 1$ predictor vector, $\boldsymbol{\beta}$ is a vector of unknown coefficients and ε_i's are uncorrelated with $E(\varepsilon_i) = 0$, $i = 1, \cdots, n$, respectively. Here $H(\cdot)$ is a known cdf linking the probabilities $p_i = \Pr(Y_i = 1)$ with the linear structure $\boldsymbol{X}_i'\boldsymbol{\beta}$, so that $p_i = H(\boldsymbol{X}_i'\boldsymbol{\beta})$. In particular, when the link cdf $H(\cdot)$ (having linking function $H^{-1}(\cdot)$) is taken to be the cdf of the logistic distribution, the model is called linear logistic regression model. The model is discussed extensively in Nelder and McCullagh(1989) and Collett(1991).

At some point during the analysis with the logistic regression model, one may wish to delete some predictors from the model. The search for a best submodel is called variable selection or subset selection. Some reasons for the variable selection are (a) to express the relationship between the binary response and the predictors as simple as possible; (b) to identify important and negligible predictors; or (c) to increase the precision of statistical estimates and predictions. A wide variety of selection procedures based on a comparison of all 2^p possible submodels have been proposed, including AIC, BIC, and the marginal likelihood criterion by Chip(1995). It is well known that, in case p is large, the computational requirements for these procedures can be prohibitive. To mitigate the computational burden, one may use heuristic methods to restrict attention to a smaller number of potential subsets. Based on this idea, the stepwise procedures have been suggested, such as forward selection or backward elimination, which sequentially include or exclude variables based on the deviance considerations (cf. Collett 1991).

The purpose of this article is to develop and suggest a variable selection procedure that avoids the overwhelming comparison of all 2^p possible submodels for the logistic regression model. The procedure selects potentially promising subsets of the predictor variables, $x_1, \cdots, x_p$, so that it may narrow the scope of possible models for further considerations. This procedure, initiated by George and McCulloch(1993), is based on a synthesis of the hierarchical Bayes modeling (cf. Mitchell and Beauchamp 1988) and Gibbs sampling (cf. Casella and George 1992): The procedure reformulates the logistic regression setup in a hierarchical normal mixture model by introducing a set of hyperparameters that will be used to identify subset choices. Then the appropriate posterior probability of each subset of predictor variables is obtained by means of the Gibbs sampler, which samples indirectly from the multinomial posterior distribution on the set of possible subset choices. Consequently, the most promising subset of predictors with highest posterior probability can then be identified by its most frequent appearance in the Gibbs sample.

In Section 2 we define and motivate the hierarchical framework for the logistic regression model that serves as the basis for the stochastic search variable selection. In Section 3 we show how the hierarchical model can be used to identify highly promising logistic regression models via the Gibbs sampler. In Section 4 we illustrate the suggested procedure on both simulated and a real data sets.

2. A Hierarchical Model for Variable Selection

Suppose that we have n binary response observations Y_i, $i=1,\cdots,n$, where $E(Y_i)=p_i$ which is the success probability corresponding to the i-th observation. The binary response logistic regression model for the dependence of p_i on p explanatory variables vector, $X_i=(x_{1i},x_{2i},\cdots,x_{pi})'$ is

$$\text{logit}\,(p_i) = \log(p_i/(1-p_i)) = \boldsymbol{\beta}'\boldsymbol{X}_i, \quad i=1,\cdots,n, \tag{2}$$

where $\boldsymbol{\beta}=(\beta_1,\cdots,\beta_p)'$ is an unknown coefficient vector. As a result of some arrangement,

$$p_i = \frac{\exp(\boldsymbol{\beta}'\boldsymbol{X}_i)}{1+\exp(\boldsymbol{\beta}'\boldsymbol{X}_i)}. \tag{3}$$

Since Y_i is an observation from a Bernoulli distribution with mean p_i, corresponding model for the expected value of y_i is $E(Y_i)=\exp(\boldsymbol{\beta}'\boldsymbol{X}_i)/(1+\exp(\boldsymbol{\beta}'\boldsymbol{X}_i))$. For the model (2), selecting a subset of predictors is equivalent to setting to 0 those β_i's corresponding to the unselected predictors. Afterwards, we shall assume that $x_1,\cdots,x_p$ contains no variable that would be included in every possible model. If an intercept was to be included in the variable selection (as is usually the case), then one should set $x_{1i}=1$, $i=1,\cdots,n$.

The likelihood function of the model (2) is given by

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} p_i^{Y_i}(1-p_i)^{1-Y_i}, \tag{4}$$

where p_i is defined by (3). This likelihood depends on the unknown success

probabilities p_i, which in turn depends on the β through (3), so that the likelihood function may be regarded as a function of β.

To extract information relevant to variable selection, we consider the following hierarchical model structure (cf. Bernardo and Smith 1994). In conventional terminology, the first stage of the hierarchy relates data to parameters via (4). The key feature of this hierarchical model is that each component of β's modeled as having come from a mixture of two normal distribution with different variances.

Thus the second stage models can be simply expressed via the introduction of a set of distinct hyperparameters $\{\alpha_j = 0 \text{ or } 1;\ j = 1, \cdots, p\}$, so that our parameter β is a random sample from a normal mixture represented by

$$\beta_j \mid \alpha_j \sim (1-\alpha_j)N(0, \sigma_i^2) + \alpha_j N(0, c_j^2\sigma_j^2), \quad j = 1, \cdots, p, \tag{5}$$

where $p(\alpha_j = 1) = 1 - p(\alpha_j = 0) = q_j$ and hyperparameters σ_j, q_j and c_j are known. A similar setup in this context was considered by Mitchell and Beauchamp (1988) and George and McCulloch(1993).

If we set small σ_j and large c_j in the above formulation, we have the following interpretations: (a) If $\alpha_j = 0$, β_j would probably be so small that it could be safely estimated by 0; (b) If $\alpha_j = 1$, then non-zero estimate of β_j should probably be included in the final model. Therefore, q_j may be thought of as the prior probability that β_j will require a non-zero estimate, or equivalently that j-th predictor variable x_j should be included in the logistic regression model. The second stage of the hierarchy thus provides the joint prior for $\beta_j \mid \alpha_j$'s in (5) as a multivariate normal prior given by

$$\beta \mid \alpha \sim N_p(0, D_\alpha R D_\alpha), \tag{6}$$

where $\alpha = (\alpha_1, \cdots, \alpha_p)$, R is the prior correlation matrix, and with $D_\alpha \equiv diag\{a_1\sigma_1, \cdots, a_p\sigma_p\}$, with $a_j = 1$ if $\alpha_j = 0$ and $a_j = c_j$ if $\alpha_j = 1$. For choosing $c_j (>1)$ and σ_j in (6), a useful guide is the following. The density of $N(0, c_i^2\sigma_j^2)$ is larger than that of $N(0, \sigma_j^2)$ iff $|\beta| > \delta(c_j)\sigma_j$, where $\delta(c_j) = (2\ln(c_j)c_j^2/(c_j^2-1))^{1/2}$. It may be also useful to note that c_j is the ratio of the heights of $N(0, c_j^2\sigma_j^2)$ and $N(0, \sigma_j^2)$ at 0, indicating the prior odds of excluding x_j when β_j is very close to 0.

The third, and final, stage specifies beliefs about α_j's. This can be done via a reasonable choice of the prior density for $\boldsymbol{\alpha}$:

$$p(\boldsymbol{\alpha}) = \prod_{j=1}^{p} q_j^{\alpha_j}(1-q_j)^{(1-\alpha_j)}$$

Therefore, the complete model structure of the hierarchy has the form.

$$p(\boldsymbol{Y} \mid \boldsymbol{\beta}) = \prod_{i=1}^{n} p_i^{Y_i}(1-p_i)^{1-Y_i},$$
$$p(\boldsymbol{\beta} \mid \boldsymbol{\alpha}) = (2\pi)^{-p/2}|D_\alpha R D_\alpha|^{-1/2}\exp\{-\beta'(D_\alpha R D_\alpha)^{-1}\beta\},$$
$$p(\boldsymbol{\alpha}) = \prod_{j=1}^{p} q_j^{\alpha_j}(1-q_j)^{(1-\alpha_j)}.$$

In many applications, it may be of interest to make inferences both about the unit characteristics, the β_j's, and the population characteristics, the α_j's. In either case, straightforward probability manipulations involving Bayes' theorem provide the required joint posterior density of $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ from which one can make the inference of interest:

$$\begin{aligned} f(\boldsymbol{\beta}, \boldsymbol{\alpha} \mid \boldsymbol{Y}) = {} & C(2\pi)^{-p/2}|D_\alpha R D_\alpha|^{-1/2}\exp\{-\frac{1}{2}\boldsymbol{\beta}'(D_\alpha R D_\alpha)^{-1}\boldsymbol{\beta}\} \\ & \times \prod_{j=1}^{p} q_j^{\alpha_j}(1-q_j)^{(1-\alpha_j)} \prod_{i=1}^{n} p_i^{Y_i}(1-p_i)^{1-Y_i}, \end{aligned} \tag{7}$$

where C in the above equation is a generic proportionality constant. Our main reason for embedding the logistic model (2) in the above hierarchical mixture model is to obtain the marginal posterior distribution $h(\boldsymbol{\alpha} \mid \boldsymbol{Y}) \propto f(\boldsymbol{Y} \mid \boldsymbol{\alpha})\pi(\boldsymbol{\alpha})$, which contains the information relevant to variable selection. However, it is easily seen that the problem of analytically calculating the marginal from (7) is a challenging one. Fortunately, recent developments of a MCMC method, say the Gibbs sampler, provides a method that directly addresses simulation based calculation of the marginal posterior (cf. Gelfand and Dey 1994).

3. Variable Selection

As the cdf of logistic distribution is approximately equivalent to that of $t_{(8)}/.638$ (cf. Albert and Chip 1993), the variable selection for $t_{(8)}$ link regression model is the same as that for logistic regression model. Let $M(t)$ be a class of t-link functions, and let consider the model $M_\nu \in M(t)$, $\nu = 8$, wherein

$$f(Y \mid \boldsymbol{\beta}, M_\nu) = \prod_{j=1}^{n} T_\nu(\boldsymbol{X}_i'\boldsymbol{\beta})^{Y_i}(1 - T_\nu(\boldsymbol{X}_i'\boldsymbol{\beta}))^{1-Y_i}$$

is the sampling density(likelihood function). Here $T_\nu(\cdot)$ is the cdf of t distribution with ν degrees of freedom. To allow for the possibility that the posterior simulation requires data augumentation, so let n latent variables $Z_1, Z_2, \cdots, Z_n$ be independently distributed from t distributions with locations parameter $\boldsymbol{X}_i'\boldsymbol{\beta}$, scale parameter 1, and degrees of freedom ν such that

$$Z_i \sim t_\nu(\boldsymbol{X}_i'\boldsymbol{\beta}, 1) \text{ and } Y = I(Z_i > 0), \quad i = 1, \cdots, n, \tag{8}$$

where $I(A)$ is an indicator function of the event A. Since $\Pr(Z_i > 0) = T_\nu(X_i'\boldsymbol{\beta})$, (8) defines t_ν link linear regression model. Let us introduce additional independent random variables λ_i, $i = 1, \cdots, n$, then the distribution of Z_i can be written as the following scale mixture of normal distributions: $Z_i \mid \lambda_i$ is distributed $N(X_i'\boldsymbol{\beta}, \lambda_i^{-1})$ and λ_i is distributed Gamma $(\nu/2, \ 2/\nu)$, $i = 1, \cdots, n$, with pdf proposional to $\lambda_i^{\nu/2-1}\exp(-\nu\lambda_i/2)$. Thus, under this data agumentation approach, we can rewrite the likelihood in (4) as that of the unobserables $\boldsymbol{\beta}$, $\boldsymbol{\lambda}$, and Z_i's;

$$\begin{aligned} L(\boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{Z}) = \prod_{i=1}^{n} \{ &\{I(Z_i > 0)I(Y_i = 1) + I(Z_i \le 0)I(Y_i = 0)\} \\ &\times \phi(Z_i; \boldsymbol{X}_i'\boldsymbol{\beta}, \lambda_i^{-1}) c(\nu) \lambda_i^{\nu/2-1} e^{-\nu\lambda_i/2} \}, \end{aligned} \tag{9}$$

where $c(\nu) = [\Gamma(\nu/2)(2/\nu)^{\nu/2}]^{-1}$ and $\phi(; \boldsymbol{X}_i'\boldsymbol{\beta}, \lambda_i^{-1})$ is the $N(\boldsymbol{X}_i'\boldsymbol{\beta}, \lambda_i^{-1})$ pdf. Under the hierarchical model, the joint posterior density of the unobservables $\boldsymbol{\beta}$, $\boldsymbol{\alpha}$, $\boldsymbol{\lambda}$ and $\boldsymbol{Z}$, given the data $Y = (Y_1, \cdots, Y_n)'$, is thus obtained by

$$f(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\lambda}, \boldsymbol{Z} \mid \boldsymbol{Y}) = C(2\pi)^{-p/2}|D_\alpha RD_\alpha|^{-1/2}\exp\{-\frac{1}{2}\boldsymbol{\beta}'(D_\alpha RD_\alpha)^{-1}\boldsymbol{\beta}\}$$
$$\times \prod_{i=1}^{n}\{I(Z_i > 0)I(Y_i = 1) + I(Z_i \leq 0)I(Y_i = 0)\} \quad (10)$$
$$\times \prod_{i=1}^{n}\{\phi(Z_i \boldsymbol{X}_i'\boldsymbol{\beta}, \lambda_i^{-1})c(\nu)\lambda_i^{\nu/2-1}e^{-\nu\lambda_i/2}\}\prod_{j=1}^{p}\{q_j^{\alpha_j}(1-q_j)^{(1-\alpha_j)}\},$$

where C here is a generic proportionality constant and $c(\nu) = [\Gamma(\nu/2)(2/\nu)^{\nu/2}]^{-1}$. Computation of the marginal posterior distribution of $\boldsymbol{\alpha}$ using the Gibbs sampling algorithm requires only the posterior distribution of $\boldsymbol{\alpha}$ conditional on $\boldsymbol{\beta}$, $\boldsymbol{\lambda}$ and $\boldsymbol{Z}$, the posterior distribution of $\boldsymbol{\beta}$ conditional on $\boldsymbol{\alpha}$, $\boldsymbol{\lambda}$ and Z, the posterior distribution of $\boldsymbol{\lambda}$ conditional on $\boldsymbol{\beta}$, $\boldsymbol{\alpha}$ and Z and the posterior of Z conditional on $\boldsymbol{\beta}$, $\boldsymbol{\lambda}$ and $\boldsymbol{\alpha}$, and these fully conditional distributions are of standard forms. From (10), the posterior density of $\boldsymbol{\beta}$ given $\boldsymbol{\alpha}$, $\boldsymbol{\lambda}$ and Z is given by

$$\pi(\boldsymbol{\beta} \mid Y, \boldsymbol{Z}, \boldsymbol{\lambda}, \boldsymbol{\alpha}) \propto |D_\alpha RD_\alpha|^{-1/2}\exp\{-\frac{1}{2}\boldsymbol{\beta}'(D_\alpha RD_\alpha)^{-1}\boldsymbol{\beta}\}\prod_{i=1}^{n}\phi(Z_i; \boldsymbol{X}_i'\boldsymbol{\beta}, \lambda_i^{-1}). \quad (11)$$

It is noted that this fully conditional posterior density is the usual posterior density for the regression parameter in the normal linear model

$$\boldsymbol{Z} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}, \quad \text{where} \quad \boldsymbol{e} \sim N_n(0, D_\lambda^{-1}), \quad (12)$$

where $\boldsymbol{\beta}$ is assigned to the proper $N_p(0, D_\alpha RD_\alpha)$ prior, $D_\lambda = diag\{\lambda_1, \lambda_2, \cdots, \lambda_n\}$ and $\boldsymbol{X} = (X_1, \cdots, X_n)'$.

Thus, the result by Zellner(1971) gives the conditional posterior of $\boldsymbol{\beta}$ as

$$\boldsymbol{\beta} \mid Y, \boldsymbol{Z}, \boldsymbol{\lambda}, \boldsymbol{\alpha} \sim N_p(\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{B}}), \quad (13)$$

where $\tilde{\boldsymbol{\beta}} = \{(D_\alpha RD_\alpha)^{-1} + \boldsymbol{X}'D_\lambda\boldsymbol{X}\}^{-1}(\boldsymbol{X}'D_\lambda Z)$, $\tilde{\boldsymbol{B}} = \{(D_\alpha RD_\alpha)^{-1} + \boldsymbol{X}'D_\lambda\boldsymbol{X}\}^{-1}$ and $D_\lambda = diag\{\lambda_1, \lambda_2, \cdots, \lambda_n\}$.

The fully conditional distributions of $Z_1, \cdots, Z_n$ are independently distributed as truncated normal distributions :

$$Z_i \mid Y, \beta, \lambda, \alpha \sim N(X_i'\beta, 1)I(Z_i > 0), \quad \text{if } Y_i = 1,$$
$$Z_i \mid Y, \beta, \lambda, \alpha \sim N(X_i'\beta, 1)I(Z_i \leq 0), \quad \text{if } Y_i = 0. \tag{14}$$

Additional variables $\lambda_1, \lambda_2, \cdots, \lambda_n$ are independent with conditional distributions

$$\lambda_i \mid \beta, Z, \alpha \sim \text{Gamma}\left(\frac{\nu+1}{2}, \frac{2}{\nu+(Z_i - X_i'\beta)^2}\right). \tag{15}$$

Full conditional distributions of $\alpha_1, \cdots, \alpha_p$ are

$$\alpha_j \mid Y, Z, \beta, \lambda, \alpha_{(j)} \sim Be\left(\frac{b_j}{b_j + d_j}\right), \quad j = 1, \cdots, p, \tag{16}$$

where $\alpha_{(j)} = (\alpha_1, \cdots, \alpha_{j-1}, \alpha_{j+1}, \cdots, \alpha_p)$, $Be(\gamma)$ denotes a Bernoulli distribution with parameter γ,

$$b_j = \left\{|D_\alpha R D_\alpha|^{-1/2} \exp\{-\frac{1}{2}\beta'(D_\alpha R D_\alpha)^{-1}\beta\}\right\}_{\alpha_j = 1} \times q_j$$

and

$$d_j = \left\{|D_\alpha R D_\alpha|^{-1/2} \exp\{-\frac{1}{2}\beta'(D_\alpha R D_\alpha)^{-1}\beta\}\right\}_{\alpha_j = 0} \times (1 - q_j).$$

In case, if we choose the prior correlation $R = I_p$ in (7), the dependence through out (16) may be eliminated so that

$$\frac{b_j}{b_j + d_j} = \frac{\exp\{-\beta_j^2/(2c_j^2\sigma_j^2)\}q_j}{\exp\{-\beta_j^2/(2c_j^2\sigma_j^2)\}q_j + c_j \exp\{-\beta_j^2/(2\sigma_j^2)\}(1-q_j)}.$$

This simplifies the calculation required. By repeated successive Gibbs sampling from (13), (14), (15) and (16), we get the Gibbs sequence

$$\beta^{(0)}, Z^{(0)}, \lambda^{(0)}, \alpha^{(0)}, \beta^{(1)}, Z^{(1)}, \lambda^{(1)}, \alpha^{(1)}, \cdots, \beta^{(\tau)}, Z^{(\tau)}, \lambda^{(\tau)}, \alpha^{(\tau)}, \tag{17}$$

that is an ergodic Markov chain. Therefore, as τ approaches infinity, the joint distribution of $\alpha^{(\tau)}$ can be shown to approach the joint distribution of α. Thus,

for large τ, say τ^*, $\boldsymbol{\alpha}^{(\tau^*)}$ can be regarded as one simulated value from the marginal posterior of $\boldsymbol{\alpha}$. For the determination of τ^*, we may use variety of diagnostic tools (cf. Cowles and Carlin, 1996) :

(i) Run a several parallel chains with starting points drawn from what we believe is a distribution overdispersed with respect to the stationary distribution. Then we visually inspect these chains by overlaying their sampled values on a comman graph for -2ln(the joint posterior in (10)) (because we are dealing with high dimensional models).
(ii) Check the graph of Gelman and Rubin (1992) shrink factors and lag 1 autocorrelation of -2ln(the joint posterior).

By use of the above tools, we may check the convergence of the Gibbs sequence and determine appropriate value of τ^*. Once we determine the value of τ^*. as practiced by Geman and Geman (1984) and Besag, York and Mollie (1991), a single long chain of the Gibbs sampler is used to get the Gibbs sample of size m, $\{\alpha^{(T)}(1), \cdots, \alpha^{(T)}(m)\}$. This method consists of picking off every Tth value in a single long run of lengh $N = mT + \tau^*$, where the number of τ^* is initial iterations that should be discarded to allow for "burn-in". The autocorrelation function of the long run chain gives the value of T that secures the indepence of $\alpha^{(T)}$'s for the Gibbs sample. The Gibbs sample can be used to compute the empirical distribution of the $\boldsymbol{\alpha}$ which converges to the actual marginal posterior $h(\boldsymbol{\alpha} \mid \boldsymbol{Y})$ (cf. Casella and George 1992; Tierney 1994). In particular, the empirical distribution of the $\boldsymbol{\alpha}$ would have following implications:

(i) the distribution corresponding to the most promising subsets of $x_1, \cdots, x_p$ will appear with the highest frequency, because it is just those values which have largest probability under $h(\boldsymbol{\alpha} \mid \boldsymbol{Y})$.
(ii) The low-frequency or zero-frequency values of $\boldsymbol{\alpha}$ may simply be ignored, because these correspond to the least promising models.
(iii) If no high-frequency values of $\boldsymbol{\alpha}$ appeared in the empirical distribution, then we would conclude that the data contain little information for discriminating between models.

Thus a simple tabulation of the high-frequency values of $\boldsymbol{\alpha}$ can be used to identify the corresponding subsets of predictors as potentially promising. The

starting value of β, $\beta^{(0)}$, may be taken to be the maximum likelihood estimate, $\lambda^{(0)} \equiv (1, \cdots, 1)'$ and $\alpha^{(0)} \equiv (1, \cdots, 1)'$. Note that it is computationally easy to simulate from the multivariate normal distribution (13) and the truncated normal distributions in (14) (see Devroye 1986 for simulation algorithm).

4. Numerical Example

In this section we illustrate the performance of the variable selection approach on both simulated and a real data examples.

4.1 Simulated Example

This subsection illustrates the performance of the suggested variable selection approach on a simulated example. This example treats problems involving $p = 5$ potential predictors of size $n = 50$. The predictors were obtained as independent standard normal variables $x_1, \cdots, x_5$, iid $\sim N(0, 1)$ so that they were practically uncorrelated. The dependent variable was generated according to the logistic model (2):

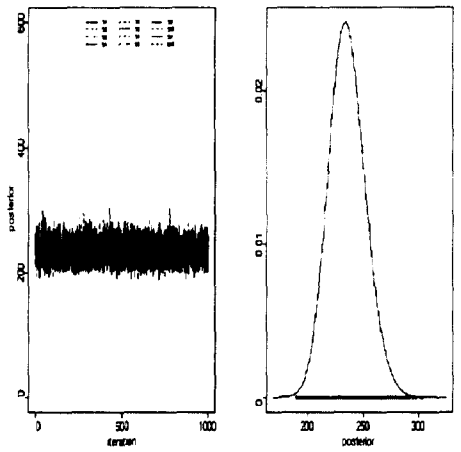
$$p_i = P(Y_i = 1) = \frac{\exp(\beta_4 x_4 + \beta_5 x_5)}{1 + \exp(\beta_4 x_4 + \beta_5 x_5)} . \tag{18}$$

Thus $\beta = (0, 0, 0, \beta_4, \beta_5)'$. We applied the suggested variable selection method with the indifference prior

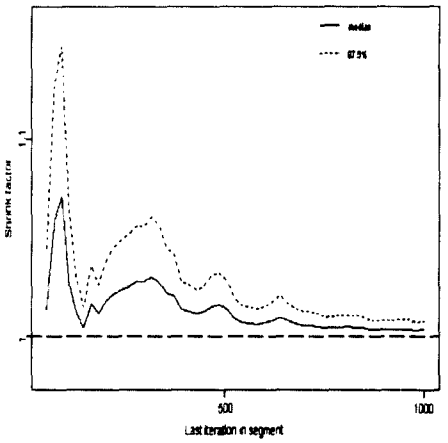
$$p(\alpha) \equiv (1/2)^5 , \ \sigma_1 = \cdots = \sigma_5 = .5, \ c_1 = \cdots = c_5 = c, \ \text{and} \ R = I_5 .$$

Using the simulated data of size $n = 50$, we ran twelve parallel chains for the Gibbs sampler(formulated by use of each prior in <Table 1>). The parallel chains were obtained by differing starting points overdispersed to provide good coverage of the posterior. The figures a1 and a2 show the traces from each chain as separate series on the same graph and a kenel density estimate calculated by combining the values of -2ln(the point posterior) from all chains. The figures b1 and b2 indicate that the estimated shrink factor for -2 ln(the joint posterior) appears to have stabilized around 1.0 within 1000 iterations. Using the same

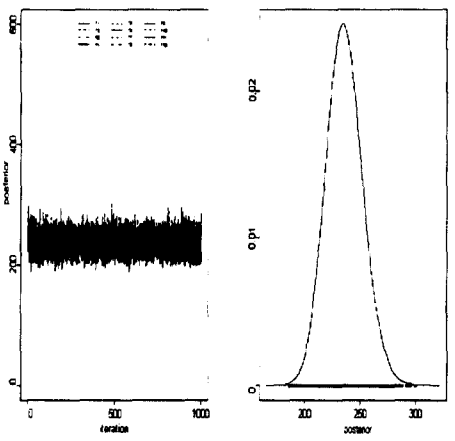
artificial data set of size $n = 50$ a Gibbs sample of $m = 1000$ observations from the Gibbs sequence was obtained from each Gibbs sampler having different prior.



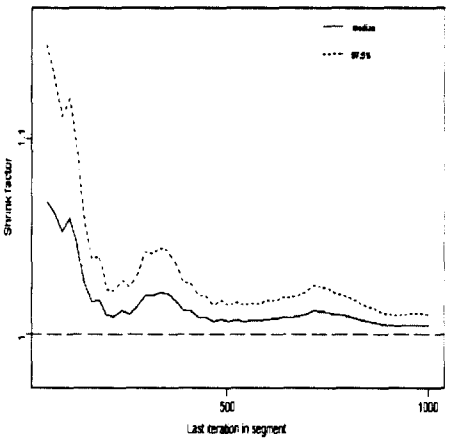
a1 ($\beta_4 = \beta_5 = 2$, $c = 5$)



b1 ($\beta_4 = \beta_5 = 2$, $c = 5$)



a2 ($\beta_4 = \beta_5 = 4$, $c = 5$)



b2 ($\beta_4 = \beta_5 = 4$, $c = 5$)

< Figure 1 > a1 and a2 indicate Trace plot and Kernel density for -2ln(the joint posrerior); b1 and b2 indicate Shrink Factor plot for -2ln(the joint posrerior).

< Table 1 > High Frequency Models

True Model	x_4x_5 ($\beta_4=2, \beta_5=2$)		x_4x_5 ($\beta_4=4, \beta_5=4$)	
	selected variables	proportion	selected variables	proportion
prior 1 ($c=5$)	x_4x_5	38.3	x_4x_5	41.1
	$x_2x_4x_5$	10.3	$x_3x_4x_5$	14.5
	$x_3x_4x_5$	9.7	$x_2x_4x_5$	10.8
prior 2 ($c=10$)	x_4x_5	47.2	x_4x_5	55.6
	x_4	14.2	$x_3x_4x_5$	10.8
	$x_1x_4x_5$	6.7	$x_2x_4x_5$	7.5
prior 3 ($c=15$)	x_4x_5	42.8	x_4x_5	58.6
	x_4	18.3	$x_3x_4x_5$	8.5
	x_5	7.0	x_5	7.1

The sampling scheme adopted here was to allow initial 1000 iterations for "burn-in" and then to pick up every 30th observation until Gibbs sample of size $m=1000$ was collected. <Table 1> displays the three high-frequency models corresponding to the frequencies of $\boldsymbol{\alpha}=(\alpha_1, \cdots, \alpha_5)'$ that appeares for each combination of c, β_4, and β_5. In each case of the priors, the true model is included in the three high-frequency value of $\boldsymbol{\alpha}$, suggesting reasonable robustness with respect to prior specifications. Aside from the robustness, the table notes the following implications: (i) It shows how the suggested variable selection method successful in identifying several promising models rather than the single best model. This feature is similar to the way in which stepwise methods are used to narrow the scope of model selection. (ii) For every prior, the true model is included in three most probable models selected. Prior 1, which had smallest c_i, seemed to favor more saturated models. The prior 3, which used the largest c_i, seemed to favor more parsimonious models.

4.2. Real Data Example

The real data in <Table 2> (reported in Collett 1991) are those of the presence of prostatic nodal involvement collected on 53 patients with cancer of the prostate. The data include a binary response variable Y that takes the value 1 if cancer to spreaded to the surrounding lymph nodes and value 0 otherwise. The objective is to explain the binary response with a constant term and five variables: age of the patient in years at diagnosis (x_1); logarithmic level of serum acid phosphate ($\ln(x_2)$); the result of an X-ray examination, coded 0 if negative and 1 if positive (x_3); the size of the tumor, coded 0 if small and 1 if large (x_4); and the

pathological grade of the tumor, coded 0 if less serious and 1 if more serious (x_5). The probability of positive response can be explained through a logit link function. If interactions and powers of predictor variables are excluded, then there are 2^5 possible models that can be fit. Instead of conventional variable selection method that searches the best fitted model among 32 possible models, we have applied the suggested variable selection approach to select promising subsets of $x_1, \cdots, x_5$. For the purpose of robustness, we have considered various choices of hyperparameters σ_j, c_j and R for the second hierarchy of the model (6). For each σ_j we have considered the low and high settings, $\sigma_j = .3$ and $\sigma_j = .5$. For each c_j we have considered the low and high settings, $c_j = 3$ and $c_j = 6$. These

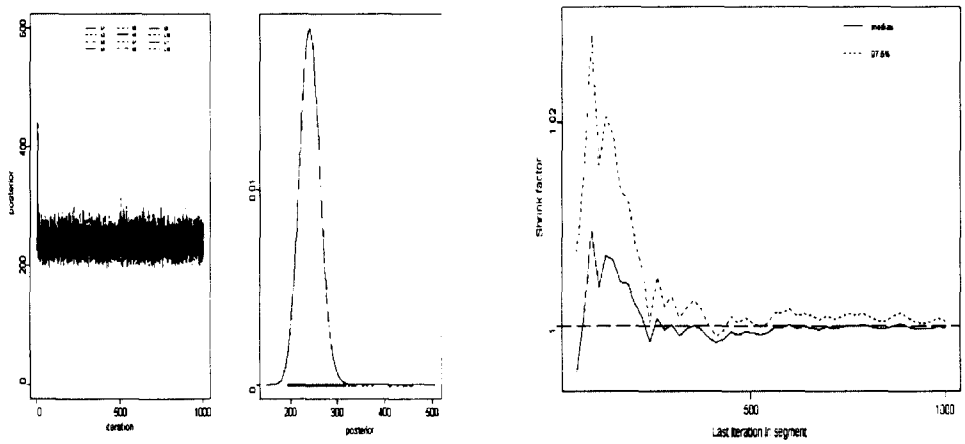
< Table 2 > Nodal Involvement Data

case	y	x_1	x_2	x_3	x_4	x_5	case	y	x_1	x_2	x_3	x_4	x_5
1	0	66	.48	0	0	0	28	0	61	.50	0	1	0
2	0	68	.56	0	0	0	29	0	64	.50	0	1	1
3	0	66	.50	0	0	0	30	0	63	.40	0	1	0
4	0	56	.52	0	0	0	31	0	52	.55	0	1	1
5	0	58	.50	0	0	0	32	0	66	.59	0	1	1
6	0	60	.49	0	0	0	33	1	58	.48	1	1	0
7	0	65	.46	1	0	0	34	1	57	.51	1	1	1
8	0	60	.62	1	0	0	35	1	65	.49	0	1	0
9	1	50	.56	0	0	1	36	0	65	.48	0	1	1
10	0	49	.55	1	0	0	37	0	59	.63	1	1	1
11	0	61	.62	0	0	0	38	0	61	1.01	0	1	0
12	0	58	.71	0	0	0	39	0	53	.76	0	1	0
13	0	51	.65	0	0	0	40	0	67	.95	0	1	0
14	1	67	.67	1	0	1	41	0	53	.66	0	1	1
15	0	67	.47	0	0	1	42	1	65	.84	1	1	1
16	0	51	.49	0	0	0	43	1	50	.81	1	1	1
17	0	56	.50	0	0	1	44	1	60	.76	1	1	1
18	0	60	.78	0	0	0	45	1	45	.70	0	1	1
19	0	52	.83	0	0	0	46	1	56	.78	1	1	1
20	0	56	.98	0	0	0	47	1	46	.70	0	1	0
21	0	67	.52	0	0	0	48	1	67	.67	0	1	0
22	0	63	.75	0	0	0	49	1	63	.82	0	1	0
23	1	59	.99	0	0	1	50	1	57	.67	0	1	1
24	0	64	1.87	0	0	0	51	1	51	.72	1	1	0
25	1	61	1.36	1	0	0	52	1	64	.89	1	1	0
26	1	56	.82	0	0	0	53	1	68	1.26	1	1	1
27	0	64	.40	0	1	1		-	-	-	-	-	-

< Table 3 > The Four Priors

prior	1	2	3	4
σ_j	0.3	0.3	0.5	0.5
c_j	3	6	3	6
R	I_5	I_5	I_5	I_5

choices provid substantial separation between the two mixture components in (5) while still allowing for plausible values of β_j when $\alpha_j = 1$. Moreover, for the prior correlation, $R = I_5$ is considered. When $R = I_5$, the components of $\boldsymbol{\beta}$ are independent under (6). The priors (6) under the above combinations of hyper-parameters are noted in <Table 3>. Finally, for the four hierarchy of the model, we used the indifference prior $p(\boldsymbol{\alpha}) \equiv (1/2)^5$, because we favored no particular $\boldsymbol{\alpha}$. For each prior, convergence diagnostic checking was done by the same way as in the previous artificial data example. <Figure 2> shows that 1000 iterations of the Gibbs sampling algorithm seem to achieve the convergence. After the initial 1000 iterations every 10th output from 1001 through 11001 iterations was collected to construct Gibbs sample of size $m = 1000$ for a given prior.



(a) (b)

< Figure 2 > (a) Trace plot and Kernel density for -2ln (the joint posrerior) ; (b) Shrink Factor plot for -2ln (the joint posrerior) from Nodal Involevment Data.

< Table 4 > Four High Frequency Models and Marginal Likelihood

prior	selected models	prop.	log(marginal likelihood)
prior 1	$\ln(x_2)+x_3+x_4$	15.4	-25.0745
	$\ln(x_2)+x_3+x_4+x_5$	14.4	-25.7683
	x_3+x_4	7.0	-37.1907
	x_2+x_3	6.4	-36.0531
prior 2	$\ln(x_2)+x_3+x_4$	20.5	-25.0745
	x_3+x_4	11.4	-37.1907
	$\ln(x_2)+x_3+x_4+x_5$	9.8	-25.7683
prior 3	x_2+x_3	8.9	-36.0531
	$\ln(x_2)+x_3+x_4$	12.0	-25.0745
	x_2+x_3	10.4	-36.0531
	x_3+x_4	7.4	-37.1907
prior 4	$\ln(x_2)+x_3+x_4+x_5$	6.6	-25.7683
	x_2+x_3	11.6	-36.0531
	$\ln(x_2)+x_3+x_4$	10.3	-25.0745
	x_3	10.2	-35.9913
	x_2	9.8	-39.9913

<Table 4> displays the four high-frequency models and corresponding logarithmic values of marginal likelihood obtained from each of the four priors. See, Kim(1997) for the algorithm of calculating the marginal likelihood. The table notes that, as in the simulation example, the suggested variable selection method is robust against the the choice of the prior specification. This is shown from the fact for each prior, the selection method yields similar set of high-frequency models including the best fitting model(see <Table 5>).

<Table 5> is the summary output (obtained from SAS PROGRAM) for the binary response logistic regression on all 6 predictor variables (including an intercept) and the best fitting model. In this full model, the weakest variables x_1 and x_5 obtained p values larger than .25 when Ward test(cf. McCullagh and Nelder 1989) for the effect of a predictor variable, given that the other variables are already in the model, is applied. Using the deviance criterion (cf. Collett 1991), we can see that the best fitting model among 2^5 possible logistic regression models has the predictors $\ln(x_2)+x_3+x_4$ (difference in the deviances between the full model and the best fitting model is 2.43 on chi-square distribution with 2 d.f. which is not significant). This result is consistent with that obtained by Chip(1995). Upon comparison between <Table 4> and <Table 5>, we see that this example once again illustrates how the suggested variable selection method narrows the scope of possible models for further consideration.

< Table 5 > Parameter Estimates of the Logistic Regression Model to the Data on Nodal Involvement

		full model				best fitting model			
predictor	d.f	Coef.	Std.	Chi-Sq.	p-val.	Coef.	Std.	Chi-Sq.	p-val.
constant	1	2.4598	3.5222	0.4877	0.4849	-1.1994	0.7162	2.8046	0.094
x_1	1	-0.0637	0.0587	1.1763	0.278	-	-	-	-
$\ln(x_2)$	1	2.5725	1.1970	4.6188	0.0316	2.2922	1.1387	4.0520	0.0441
x_3	1	2.0401	0.8288	6.0583	0.0138	2.0550	0.7976	6.6380	0.0100
x_4	1	1.5466	0.7811	3.9205	0.0477	1.7638	0.7483	5.5562	0.0184
x_5	1	0.8345	0.7889	1.1188	0.2902	-	-	-	-

The choice of a single best model at this point could proceed by applying standard model selection criteria, such as AIC, the deviance criterion, and the marginal likelihood criterion (cf. Chip 1995), to the more manageable selected subsets, i.e. selected high-frequency models.

5. Concluding Remarks

This article has developed and illustrated a Bayesian approach to narrow the scope of possible models in the variable selection for the binary response logistic regression model. Though the suggested approach would not directly lead to a single best fitting model, it is demonstrated as a way to save the overwhelming job of comparing all the 2^p possible submodels for the logistic regression model with p predictor variables. Thus, as an alternative to usual optimal subset selection procedure (involving the overwhelming comparisons of all 2^p possible subset models), a two-stage variable selection procedure can be constructed: First, select $m \ll 2^p$ promising subset models via the suggested approach. In the second stage, choose a best fitting model by means of usual variable selection criteria such as AIC, BIC, the deviance criterion (cf. Collett 1991) and the marginal likelihood by Chip(1995). For the full Bayesian two-stage procedure, we may adopt the marginal likelihood criterion in the second stage.

The suggested approach relies on the output of the Gibbs sampling algorithm and demonstrates good performances in a couple of examples. The algorithm is applied to a reformulated logistic regression setup constructed in a hierarchical normal mixture model by introducing hyperparameters that will be used to identify

subset choices. Among the hyperparameters, c_j and σ_j, $j=1,\cdots,p$, are assumed to be known even though values of them are not readily available. We have given some useful guidelines to select them. The illustrated examples showed that the approach is robust against the choice of the parameters. However, to avoid the subjective choice of the parameters, we may assume vague priors for the parameters in the hierarchical model setting. This will lead to the algorithm more complicated, because the full conditional distributions of c_j and σ_j will not be of closed forms. The metropolis-Hastings algorithm (cf. Smith and Roberts 1993) may be used to construct a Markov chains for c_j and σ_j. The study pertaining to the performance of the suggested approach obtained by the vague priors is no less important and left as a future study of interest.

References

[1] Bernardo, J.M. and Smith, A.F.M.(1994), *Bayesian theory*, Wiley, New York.

[2] Besag, G.E., York, J., and Mollie.(1991), "Bayesian image restoration, with two applications in spatial statistics," *Annals of Institute of Statistical Mathematics,* Vol. 43, pp. 1-59.

[3] Casella, G. and George, E.I.(1992), "Explaning the Gibbs sampler," *American Statistician*, Vol. 46, pp. 167-174.

[4] Chip, S.(1995), "Marginal likelihood from the Gibbs output," *Journal of American Statistics Association*, Vol. 90, pp. 1313-1321.

[5] Cowles, M.K. and Carlin, B.P.(1996), "Markov Chain Mote Carlo Convergence Diagnostics: A Comparitive Review," *Journal of the American Statistical Association,* Vol. 91, pp. 883-904.

[6] Collett, D.(1991), *Modelling binary data*, Chapman and Hall, New York.

[7] Devorye, L.(1986), *Non-uniform random generation*, Springer Verlag, New York.

[8] George, E.I. and McCulloch, R.E.(1993), "Variable selection via Gibbs sampling," *Journal of the American Statistical Association*, Vol. 88, pp. 881- 889.

[9] Gelfand, A.E. and Dey, D.K.(1994), "Bayesian model choice: asymptotics and exact calculations," *Journal of the Royal Statistical Society, Ser.* B, Vol. 56, pp. 501-514.

[10] Geman, S. and Geman, D.(1984), "Stochastic relaxation, Gibbs distribution and the bayesian restoration of image," *IEEE Transaction Pattern Analysis Machine Intell,* 6, pp. 721-741.

[11] Gelman, A.E. and Roubin, D.B.(1992a), "Inference from iterative simulation using multiple sequence(with disscessed)," *Statistical Science*, 7, pp. 457-511.

[12] Kim, H.J.(1997), "On a Bayes Criterion for the Goodness-of-Link Test for Binary Response Regression Models: Probit Link versus Logit Link" *Journal of the Korean Statistical Society,* Vol. 26, No. 2, pp. 261-276.

[13] Mitchell, T.J. and Beauchamp, J.J.(1988), "Bayesian variable selection in linear regression (with discussion)," *Journal of the American Statistical Association,* Vol. 83, pp. 1023-1036.

[14] Nelder, J.A. and McCullagh, P.(1989), *Generalized linear models,* Capman and Hall, New York.

[15] Smith, A.F.M. and Roberts, G.O.(1993), "Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods," *Journal of the Royal Statistical Society, Ser* B, Vol. 55, pp. 3-23.

[16] Tierney, L.(1994), "Markov chains for exploring posterior distributions," *Annals of Statistics*, Vol. 22, pp. 1701-1762.

[17] Zellner, A.(1971), *An introduction to Bayesian inference in Econometrics*, Wiley, New York.