Journal of the Korean Statistical Society (1999), 28: 2, pp 167–182

A Bayesian Variable Selection Method for Binary Response Probit Regression†

Hea-Jung Kim[1]

ABSTRACT

This article is concerned with the selection of subsets of predictor variables to be included in building the binary response probit 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 probit regression setup in a hierarchical normal mixture model by introducing a set of hyperparameters that will be used to identify subset choices. The appropriate posterior probability of each subset of predictor variables is obtained through 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 the one with highest posterior probability. To highlight the merit of this procedure a couple of illustrative numerical examples are given.

Keywords: Binary response probit regression; Variable selection; Hierarchical normal mixture model; Data augmentation; Gibbs sampler; High frequency model

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

$$Y_i = H(X_i'\beta) + \varepsilon_i, \tag{1.1}$$

where Y_i is a binary output, X_i is a $p \times 1$ predictor vector, β is a vector of unknown coefficients and ε_i's are uncorrelated with $E(\varepsilon_i) = 0$, $i = 1, \ldots, n$,

†The author wishes to acknowledge the financial support of the Korea Research Foundation made in the program year of 1998

[1]Department of Statistics, Dongguk University, Seoul, 100-715, Korea

respectively. Here $H(\cdot)$ is a known cdf linking the probabilities $p_i = Pr(Y_i = 1)$ with the linear structure $X_i'\beta$, so that $p_i = H(X_i'\beta)$. In particular, when the link cdf $H(\cdot)$(having linking function $H^{-1}(\cdot)$) is taken to be the cdf of the standard normal distribution, $\Phi(\cdot)$, the model is called probit regression model. The model is discussed extensively in Nelder and McCullagh(1987) and Collett(1991).

At some point during the analysis with the probit 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. 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 Chib (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). However, It has long been known that one needs extreme care to use the stepwise procedures in the probit regression. When one performs many significance tests in the course of the stepwise procedures, each at a level of α, the overall probability of rejecting at least one true null hypothesis is much larger than α. Furthermore, none of the p-values for the parameter estimates have the conventional meaning because none of the test statistics has a normal or chi-square distribution.

The purpose of this article is to develop and suggest a variable selection procedure that not only avoids the overwhelming comparison of all 2^p possible submodels for the probit regression model, but eliminates the problems of the stepwise procedures. The procedure selects potentially promising subsets of the predictor variables, $x_1, \ldots, x_p$, 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).

2. HIERARCHICAL MODEL FOR VARIABLE SELECTION

Suppose that we have n binary response observations Y_i, $i = 1, \ldots, n$, where $E(Y_i) = p_i$ which is the success probability corresponding to the i-th observation. The binary response probit regression model for the dependence of p_i on p

explanatory variables vector, $X_i = (x_{1i}, x_{2i}, \ldots, x_{pi})'$ is

$$\text{probit}(p_i) \;=\; \Phi^{-1}(p_i) \;=\; \beta' X_i, \;\; i = 1, \ldots, n. \tag{2.1}$$

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

$$p_i \;=\; \Phi(\beta' X_i) \;=\; (2\pi)^{-1/2} \int_{-\infty}^{\beta' X_i} \exp(-\frac{1}{2}u^2) du. \tag{2.2}$$

Since Y_i is an observation from a Bernoulli distribution with mean p_i, corresponding model for the expected value of Y_i is $E(Y_i) = \Phi(\beta' X_i)$. For the model (2.1), 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 $x_1, \ldots, x_p$ 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 $x_{1i} = 1$, $i = 1, \ldots, n$.

The likelihood function of the model (2.1) is given by

$$L(\beta) \;=\; \prod_{i=1}^{n} p_i^{Y_i} (1 - p_i)^{1 - Y_i}, \tag{2.3}$$

where p_i is defined by (2.2). This likelihood depends on the unknown success probabilities p_i, which in turn depends on β through (2.2), 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 (2.3). A key feature of this hierarchical model is that each component of β is 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 $\{\alpha_j = 0 \text{ or } 1, \; j = 1, \ldots, p\}$, so that our parameter β is a random sample from a normal mixture represented by

$$\beta_j | \alpha_j \;\sim\; (1 - \alpha_j) N(0, \sigma_j^2) + \alpha_j N(0, c_j^2 \sigma_j^2), \; j = 1, \ldots, p, \tag{2.4}$$

where $Pr(\alpha_j = 1) = 1 - Pr(\alpha_j = 0) = q_j$ 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 probit regression model.

The second stage of the hierarchy thus provides the joint prior for $\beta_j|\alpha_j$'s in (2.4) as a multivariate normal prior given by

$$\beta|\alpha \sim N_p(0, D_\alpha R D_\alpha), \tag{2.5}$$

where $\alpha = (\alpha_1, \ldots, \alpha_p)$, R is the prior correlation matrix, and

$$D_\alpha \equiv \text{diag}\{a_1\sigma_1, \ldots, a_p\sigma_p\},$$

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 (2.5), a useful guide is the following. The density of $N(0, c_j^2\sigma_j^2)$ is larger than that of $N(0, \sigma_j^2)$ iff $|\beta_j| > \delta(c_j)\sigma_j$, 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 α:

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

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

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

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 β and α from which one can make the inference of interest:

$$f(\beta, \alpha|Y) = C(2\pi)^{-p/2}|D_\alpha R D_\alpha|^{-1/2}\exp\{-\frac{1}{2}\beta'(D_\alpha R D_\alpha)^{-1}\beta\} \tag{2.6}$$

$$\times \prod_{j=1}^{p} q_j^{\alpha_j}(1-q_j)^{(1-\alpha_j)} \prod_{i=1}^{n} p_i^{Y_i}(1-p_i)^{1-Y_i},$$

where C in the above equation is a generic proportionality constant.

Our main reason for embedding the probit model (2.1) in the above hierarchical mixture model is to obtain the marginal posterior distribution $h(\alpha|Y) \propto f(Y|\alpha)\pi(\alpha)$, which contains the information relevant to variable selection. However, it is easily seen that the problem of analytically calculating the marginal from (2.6) is a challenging one. Fortunately, recent developments of a MCMC method, say the Gibbs sampler, provide a method that directly addresses simulation based calculation of the marginal posterior (cf. Chib, 1995).

3. GIBBS SAMPLING SCHEME

As in Albert and Chib (1993), introduce a set of latent variables $\{Z_i,\ i = 1,\ldots,n\}$, where the Z_i are independent $N(X_i'\beta, 1)$, and $Y_i = 1$ if $Z_i > 0$ and $Y_i = 0$ otherwise. It can be easily shown that Y_i are independent Bernoulli random variables with $p_i = Pr(Y_i = 1) = \Phi(X_i'\beta)$. Thus, under this data augmentation approach, we can rewrite the likelihood in (2.3) as that of the unobservables β and Z_i's:

$$L(\beta, Z) = \prod_{i=1}^{n}\{I(Z_i > 0)I(Y_i = 1) + I(Z_i < 0)I(Y_i = 0)\}\phi(Z_i; X_i'\beta, 1), \quad (3.1)$$

where $Z = (Z_1,\ldots,Z_n)'$, $\phi(\ ; X_i'\beta, 1)$ is the $N(X_i'\beta, 1)$ pdf, and $I(W \in A)$ is the indicator function that is equal to 1 if the random variable W is contained in the set A.

Under the hierarchical model, the joint posterior density of the unobservables β, α and Z, given the data $Y = (Y_1,\ldots,Y_n)'$, is thus obtained by

$$f(\beta, \alpha, Z|Y) = C(2\pi)^{-p/2}|D_\alpha R D_\alpha|^{-1/2}\exp\{-\frac{1}{2}\beta'(D_\alpha R D_\alpha)^{-1}\beta\} \quad (3.2)$$

$$\times \prod_{j=1}^{p} q_j^{\alpha_j}(1-q_j)^{(1-\alpha_j)} \prod_{i=1}^{n}\{I(Z_i > 0)I(Y_i = 1) + I(Z_i < 0)I(Y_i = 0)\}\phi(Z_i; X_i'\beta, 1),$$

where C here is a generic proportionality constant.

3.1. The Gibbs Sampler

Computation of the marginal posterior distribution of α using the Gibbs sampling algorithm requires only the posterior distribution of α conditional on β and

Z, the posterior distribution of β conditional on α and Z and the posterior of Z conditional on β and α, and these fully conditional distributions are of standard forms.

From (3.2), the posterior density of β given α and Z is given by

$$\pi(\beta|Y,Z,\alpha) \propto |D_\alpha R D_\alpha|^{-1/2} \exp\{-\frac{1}{2}\beta'(D_\alpha R D_\alpha)^{-1}\beta\} \prod_{i=1}^{n} \phi(Z_i; X_i'\beta, 1). \quad (3.3)$$

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

$$Z = \mathbf{X}\beta + \epsilon, \quad \text{where } \epsilon \sim N_n(0, I_n), \quad (3.4)$$

where β is assigned to the proper $N_p(0, D_\alpha R D_\alpha)$ prior and $\mathbf{X} = (X_1, \ldots, X_n)'$. Thus, the result by Zellner (1971) gives the conditional posterior of β as

$$\beta|Y,Z,\alpha \sim N_p(\tilde{\beta}, \tilde{B}), \quad (3.5)$$

where $\tilde{\beta} = \{(D_\alpha R D_\alpha)^{-1} + \mathbf{X}'\mathbf{X}\}^{-1}(\mathbf{X}'Z)$ and $\tilde{B} = \{(D_\alpha R D_\alpha)^{-1} + \mathbf{X}'\mathbf{X}\}^{-1}$.

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

$$\begin{aligned} Z_i|Y,\beta,\alpha &\sim N(X_i'\beta, 1)I(Z_i > 0), \text{ if } Y_i = 1, \\ Z_i|Y,\beta,\alpha &\sim N(X_i'\beta, 1)I(Z_i \le 0), \text{ if } Y_i = 0. \end{aligned} \quad (3.6)$$

Additional variables $\alpha_1, \ldots, \alpha_p$ are conditionally distributed as

$$\alpha_j|Y,Z,\beta,\alpha_{(j)} \sim Be\left(\frac{b_j}{b_j + d_j}\right), \quad (3.7)$$

where $\alpha_{(j)} = (\alpha_1, \ldots, \alpha_{j-1}, \alpha_{j+1}, \ldots, \alpha_p)$, $Be(\gamma)$ 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

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

Remark 3.1. If we choose the prior correlation $R = I_p$ in (2.6), then the dependence through out (3.7) 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.

Remark 3.2. For large p, it is inefficient to invert the $p \times p$ matrix $\{(D_\alpha R D_\alpha)^{-1} + \mathbf{X}'\mathbf{X}\}$ in each pass of the Gibbs sampling algorithm. Instead, let L_α be a factor of $D_\alpha R D_\alpha$ such that $L_\alpha L_\alpha' = D_\alpha R D_\alpha$. Let $L_\alpha' \mathbf{X}'\mathbf{X} L_\alpha$ have orthogonal factorization $P_\alpha \Lambda_\alpha P_\alpha'$, i.e. P_α is the corresponding ordered eigenvectors so that $P_\alpha P_\alpha' = P_\alpha' P_\alpha = I_p$. Finally, let $H_\alpha = L_\alpha P_\alpha$. Then

$$\left\{(D_\alpha R D_\alpha)^{-1} + \mathbf{X}'\mathbf{X}\right\}^{-1} = H_\alpha (I_p + \Lambda_\alpha)^{-1} H_\alpha'. \tag{3.8}$$

This leads to an efficient computation for the inverse matrix for large p.

3.2. Gibbs Sampling Scheme and Subset Selection

The hierarchical nature of the model gives relatively straightforward implementation of the Gibbs sampling scheme. A possible complication could be the simulation from truncated normal distribution. This can be easily conducted by the algorithm of Devroye (1986). By repeated successive Gibbs sampling from (3.5) through (3.7), we would get the Gibbs sequence

$$\beta^{(0)}, Z^{(0)}, \alpha^{(0)}, \beta^{(1)}, Z^{(1)}, \alpha^{(1)}, \beta^{(2)}, Z^{(2)}, \alpha^{(2)}, \ldots, \beta^{(t)}, Z^{(t)}, \alpha^{(t)} \tag{3.9}$$

that is an ergodic Markov chain. Therefore, as t approaches infinity, the joint distribution of $\alpha^{(t)}$ can be shown to approach the joint distribution of α. Thus, for large t, say t^*, $\alpha^{(t^*)}$ can be regarded as one simulated value from the marginal posterior of α.

For the determination of t^*, we may use a 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 common graph for -2*ln*(the joint posterior in (3.2) whether they converge to a true distribution.

(ii) Check the convergence (converging to 1) of Gelman and Rubin (1992) shrinkage factor of the -2*ln*(the joint posterior) values.

By use of the above tools, we may check the convergence of the Gibbs sequence and determine appropriate value of t^*. Once we check the convergence of the Gibbs sampler and determine appropriate value of t^*, as practiced by Besag, York and Mollie (1991), a single long run chain of the Gibbs sampler is used to get the Gibbs sample of size m, $\{\alpha^{(T)}(1), \ldots, \alpha^{(T)}(m)\}$. This method consists of picking off every Tth value in a single long run of length $N = mT + t^*$, where

the number of t^* 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 independence of $\alpha^{(T)}$'s for the Gibbs sample.

The Gibbs sample can be used to compute the empirical distribution of the α which converges to the actual marginal posterior $h(\alpha|Y)$ (cf. Casella and George, 1992). In particular, the empirical distribution of the α would have following implications:

(i) the distribution corresponding to the most promising subsets of $x_1, \ldots, x_p$ will appear with the highest frequency, because it is just those values which have largest probability under $h(\alpha|Y)$.

(ii) The low-frequency or zero-frequency values of α may simply be ignored, because these correspond to the least promising models.

(iii) If no high-frequency values of α appeared in the empirical distribution, then we would conclude that the data contain little information for discriminating between models.

Thus, instead of estimating $h(\alpha|Y)$, one can simply identify potentially promising subsets of predictors from a tabulation of high-frequency values of α.

4. NUMERICAL EXAMPLE

In this section we illustrate the performance of the variable selection approach on both artificial and a real data examples. The real data are presented in Collett(1991) and are often used to illustrate techniques for selecting predictors. The objectives in these examples are to demonstrate a convenient method for the formulation of subjective priors, illustrate favorable performance of the procedure, and study the relation between prior and posterior distributions for the coefficients of some predictor variables.

4.1. Artificial Data Example

This example treats problems involving p=5 potential predictors of size n=50. The predictors were obtained as independent standard normal variables $x_1, \ldots, x_5$ iid $\sim N(0,1)$, so that they were practically uncorrelated. The dependent variable was generated according to the probit model (2.1):

$$p_i = Pr(Y_i = 1) = \Phi(\beta_4 x_4 + \beta_5 x_5). \tag{4.1}$$

Thus $\beta = (0, 0, 0, \beta_4, \beta_5)'$.

We applied the suggested variable selection method with the indifference prior

$$P(\alpha_j = 1) = q_j = q,\ \sigma_j = \sigma,\ c_j = c,\ \text{for}\ j = 1, \ldots, 5 \text{ and } R = I_5.$$

Different prior beliefs will, of course, lead to other choices for q_j, σ_j and c_j. For instance, it is thought that a certain predictor may not be enter the model at all, the corresponding σ_j and p_j would be smaller, while c_j would be larger and their values may be set employing the same kind of reasoning about marginal effects. We considered various choices of the hyperparameters for the indifference priors. For each σ_j, we considered the low and high settings, $\sigma_j = .3$ and $\sigma_j = .5$. For each c_j we considered the low and high settings, $c_j = 4$ and $c_j = 9$. These choices provided substantial separation between the two mixture components in (2.4) while still allowing for plausible values of β_j when $\alpha_j = 1$. As a base probability that each predictor is included in the model, we took $q_j = .5$. To study the relation between the prior and posterior distributions, we also considered $q_j = .2$ and $q_j = .8$. Thus we set up following twelve priors for the example.

Table 4.1: Twelve Indifference Priors

prior	1	2	3	4	5	6	7	8	9	10	11	12
q	0.2	0.2	0.2	0.2	0.5	0.5	0.5	0.5	0.8	0.8	0.8	0.8
σ	0.3	0.3	0.5	0.5	0.3	0.3	0.5	0.5	0.3	0.3	0.5	0.5
c	4	9	4	9	4	9	4	9	4	9	4	9

For the probit model given in (4.1), corresponding Gibbs sampler was formulated for each prior specified in Table 1, and then it was checked for convergence.

Using SAS/IML we generated an artificial data set of size 50 from the model (4.1) with given values of β_j's, and ran twelve parallel chains for the Gibbs sampler (formulated by using each prior of Table 1). The parallel chains were obtained by differing starting points overdispersed to provide good coverage of the posterior. Twelve sets of starting points considered for each model were combinations of following parameter values:

(i) β_j, $j = 1, \ldots, 5$: mle of β_j, mle $\pm$ (s.d. of mle);

(ii) $(\alpha_1, \ldots, \alpha_5)$: (0, ..., 0), (0, 1, 0, 1, 0), (1, 0, 1, 0, 1), (1, ..., 1).

We considered 48 ((4 different (β_4, β_5)) $\times$ 12 (priors)) sets of the twelve parallel chains for the model (4.1). As convergence diagnostic tools, we used the

trace and Gelman and Rubin shrink factor of -2 ln(the joint posterior in (3.2)) obtained from each set of the parallel chains. Since they revealed almost the same convergence diagnostic results, we present the result of only one set of the chains of each models in Figure 4.1 and Figure 4.2. They were produced by selecting plots option from the "CODA Output Analysis Menu" by Best et al. (1996). The figures show that all twelve chains appear to settle to the same (or similar) distribution within 1000 iterations. This is a clear indication that convergence has been achieved within 1000 iterations for the Gibbs sampler.

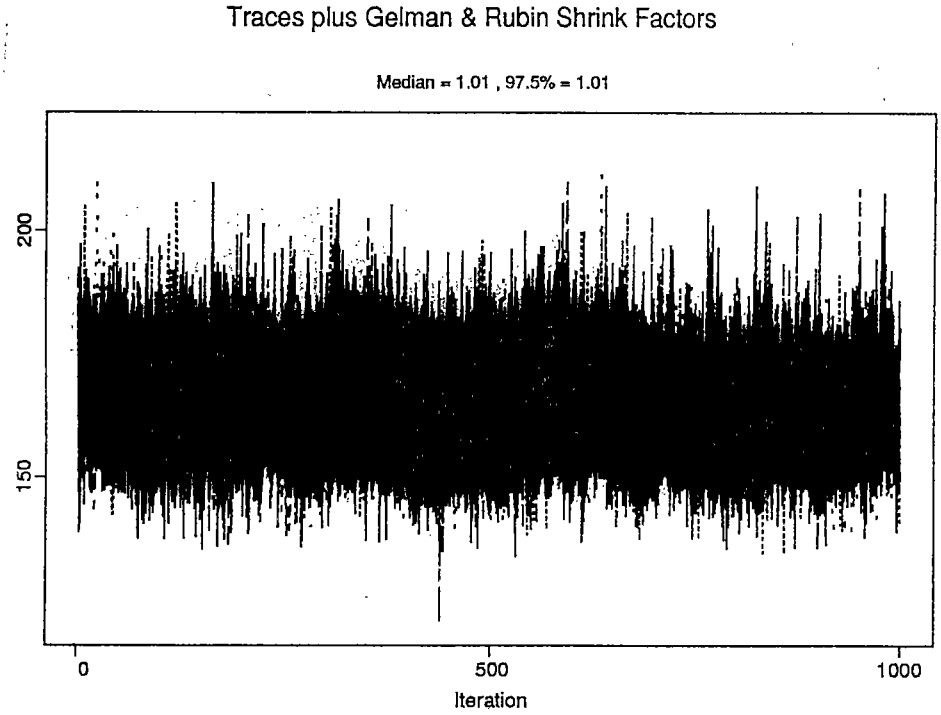

Figure 4.1: Graphical Summary of the Twelve Parallel Chains of the Gibbs Sampler: Probit Model ($\beta_1 = 4$, $\beta_2 = 4$) with Prior 4; Trace of the twelve chains (B1 to B12) and Gelman and Rubin shrinkage factor.

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.

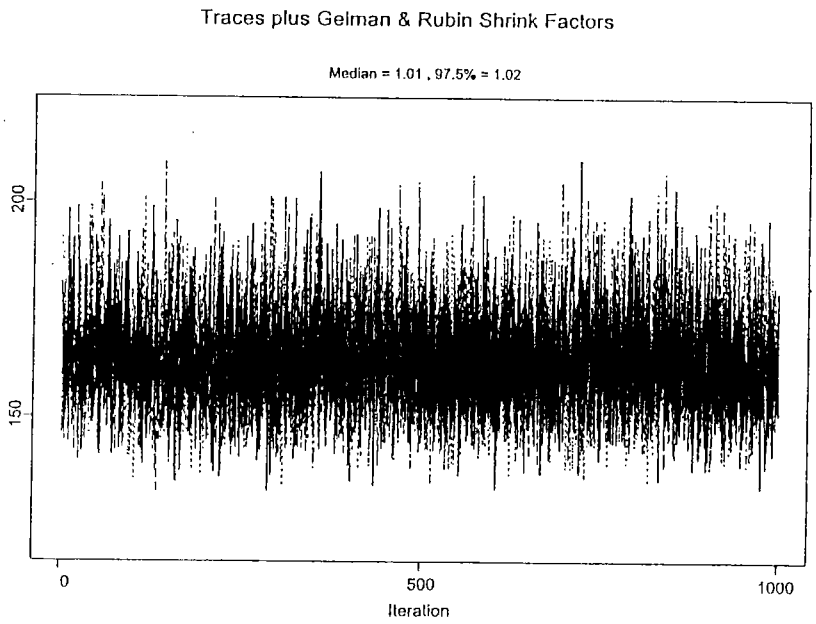

Figure 4.2: Graphical Summary of the Twelve Parallel Chains of the Gibbs Sampler: Probit Model ($\beta_1 = -2$, $\beta_2 = -2$) with Prior 4; Trace of the twelve chains (B1 to B12) and Gelman and Rubin shrinkage factor.

The sampling scheme adopted here was to allow initial 1000 iterations for "burn-in" and then to pick up every 10th observation until Gibbs sample of size m=1000 was collected. For each Gibbs sampling, we used the mle for $\beta^{(0)}$, $\alpha_j^{(0)} = 1$ and $\lambda_j^{(0)} = 1$, $j = 1, \ldots, 5$, as starting values. Table 2 displays respective three high-frequency probit models corresponding to the frequencies of $\alpha = (\alpha_1, \ldots, \alpha_5)'$ that appeared for each prior with given (β_4, β_5). In each case of the priors, the true model is included in the first three high-frequency values among 2^5 different frequency values of α, suggesting reasonable robustness with respect to prior specifications. Aside from the robustness, the table notes 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.

4.2. Real Data Example

These data are presented in Collett (1991) and are often used to illustrate techniques for selecting predictors of binary response regression models (see Collett (1991) for the logistic regression and Chib(1995) for the probit regression).

The data are the presence of prostatic nodal involvement collected on 53 patients with cancer of the prostate. They include a binary response variable Y that takes the value 1 if cancer spread 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 probit and a logistic link functions. If interactions and powers of predictor variables are excluded, then there are 2^6 possible models that can be fitted (including the constant term).

Table 4.2: High Frequency Probit Models and Relative Frequencies (%)

Prior	$(\beta_4 = \beta_5 = -2)$		$(\beta_4 = \beta_5 = 4)$		Prior	$(\beta_4 = \beta_5 = -2)$		$(\beta_4 = \beta_5 = 4)$	
	Model	Freq.	Model	Freq.		Model	Freq.	Model	Freq.
1	x_4x_5	22.5	x_4x_5	37.1	7	x_4	24.1	x_4x_5	15.9
	x_4	18.5	$x_2x_4x_5$	12.5		x_4x_5	10.7	x_5	13.1
	$x_3x_4x_5$	10.3	$x_3x_4x_5$	10.0		x_1x_4	6.3	x_4	13.0
2	x_4	36.2	x_4x_5	52.6	8	x_4	30.2	x_4	15.5
	x_4x_5	27.5	x_4	12.9		x_4x_5	9.0	x_5	14.3
	x_3x_4	6.3	$x_3x_4x_5$	6.7		x_4	5.1	x_4x_5	10.5
3	x_4	22.0	x_4x_5	16.7	9	x_4x_5	25.2	x_4x_5	30.1
	x_4x_5	12.4	x_5	12.3		x_4	17.0	x_4	10.9
	x_3x_4	6.5	x_4	12.1		$x_3x_4x_5$	10.4	$x_2x_4x_5$	8.9
4	x_4	32.0	x_4	17.7	10	x_4	23.5	x_4x_5	56.4
	x_4x_5	8.8	x_5	17.4		x_4x_5	10.6	x_4	10.2
	x_1x_4	4.8	x_4x_5	9.9		x_3x_4	5.7	$x_1x_4x_5$	8.1
5	x_4x_5	22.2	x_4x_5	37.5	11	x_4	23.5	x_4x_5	15.1
	x_4	17.9	$x_2x_4x_5$	11.3		x_4x_5	10.6	x_5	12.4
	$x_3x_4x_5$	9.9	$x_3x_4x_5$	10.3		x_3x_4	6.6	x_4	10.4
6	x_4	38.8	x_4x_5	48.9	12	x_4	32.8	x_4	16.7
	x_4x_5	26.6	x_5	11.0		x_4x_5	9.9	x_5	16.1
	x_3x_4	7.8	x_4	7.2		x_5	5.9	x_4x_5	12.1

Instead of conventional variable selection method that searches the best fitted model among 64 possible models, we have applied the suggested variable

selection approach to select promising subsets of constant term with predictor x_1 and $\ln(x_2), \ldots, x_6$. For this example, we used the probit model to illustrate the suggested stochastic variable search method. 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 (2.5). 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 choices provided substantial separation between the two mixture components in (2.4) while still allowing for plausible values of β_j when $\alpha_j = 1$. Moreover, for the prior correlation choice, $R = I_6$ and $R \propto Q$, $Q = diag((\mathbf{X}'\mathbf{X})^{-1})^{-1/2}(\mathbf{X}'\mathbf{X})^{-1}diag((\mathbf{X}'\mathbf{X})^{-1})^{-1/2}$ are considered. The eight priors considered are tabulated in Table 3. Here $diag((\mathbf{X}'\mathbf{X})^{-1})$ denotes a diagonal matrix whose diagonal elements are those of $(\mathbf{X}'\mathbf{X})^{-1}$. When $R = I_6$, the components of β are independent under (2.5). When $R \propto Q$, the prior correlation is identical to the design correlation, a generalization of the g prior of Zellner (1986). These choices provided substantial separation between the two mixture components in (2.4) while still allowing for plausible values of β_j when $\alpha_j = 1$. Finally, for the third hierarchy of the model, we set the prior $\pi(\alpha) = (1/2)^5$, so that we made the intercept C is always included in the final model by setting α_1 =1 and $q_j = .5, j = 2, \ldots, 5$, because we favored no particular α_j except for α_1.

For the Gibbs sampler constructed for the probit model with each of the priors in Table 3, convergence diagnostic checking was done by the same way as in the previous artificial data example. The checking showed that 1000 iterations of the Gibbs sampling algorithm seemed to achieved the convergence.

Table 3 displays four high-frequency probit models of each size obtained from each of eight combinations of the hyperparameters. They were obtained from Gibbs samples of each size $m = 1000$. After the initial 1000 iterations, every 10th output from 2001 through 12001 iterations was collected to construct each Gibbs sample of size $m = 1000$. The table notes that, as in the previous example, the suggested variable selection method safely includes best fitting model (cf. Chib 1995) regardless of the choice of the prior specification. As expected the high setting of σ_j tends to select smaller model than the low setting of σ_j does. Certain variables, such as x_5, are included more often under the low setting than under the high setting. It is also interesting that although there are some overlappings in models selected using $R = I_6$ and using $R \propto Q$, there is a pronounced difference. $R = I_6$, lessening posterior correlations, tends to select smaller model than $R \propto Q$ does. Certain variables, such as x_5, are included more often under $R \propto Q$ than under $R = I_6$.

Table 4.3: Four High Frequency Probit Regression Models(C= Constant term)

	Case: $R = I_6$		Case: $R \propto Q$	
(σ_j, c_j)	Selected Probit Models	prop.(%)	Selected Probit Models	prop.(%)
(0.3, 3)	$C + \ln(x_2) + x_3 + x_4$	15.3	$C + \ln(x_2) + x_3 + x_4$	17.5
	$C + \ln(x_2) + x_3 + x_4 + x_5$	10.9	$C + \ln(x_2) + x_3 + x_4 + x_5$	11.5
	$C + x_3 + x_4$	8.7	$C + \ln(x_2) + x_3$	8.5
	$C + \ln(x_2) + x_3$	7.6	$C + x_3 + x_4$	8.0
(0.3, 6)	$C + \ln(x_2) + x_3 + x_4$	18.1	$C + \ln(x_2) + x_3 + x_4$	19.6
	$C + \ln(x_2) + x_3$	12.7	$C + \ln(x_2) + x_3 + x_4 + x_5$	10.5
	$C + x_3 + x_4$	9.3	$C + \ln(x_2) + x_3$	10.0
	$C + \ln(x_2) + x_3 + x_4 + x_5$	7.4	$C + x_3 + x_4$	8.1
(0.5, 3)	$C + \ln(x_2) + x_3$	8.7	$C + \ln(x_2) + x_3 + x_4$	12.0
	$C + \ln(x_2) + x_3 + x_4$	8.3	$C + \ln(x_2) + x_3$	9.1
	$C + x_3$	7.0	$C + x_3 + x_4$	7.3
	$C + x_3 + x_4$	6.4	$C + \ln(x_2) + x_3 + x_4 + x_5$	6.8
(0.5, 6)	$C + x_3$	11.1	$C + x_3$	14.1
	$C + \ln(x_2) + x_3$	10.9	$C + \ln(x_2) + x_3$	10.2
	$C + \ln(x_2)$	8.9	$C + x_3 + x_4$	9.9
	$C + \ln(x_2) + x_3 + x_4$	7.3	$C + \ln(x_2) + x_3 + x_4$	7.9

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 (see Chib 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 a probit 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 probit 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 $m << 2^p$ 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 (Collett 1991) and the marginal likelihood by Chib (1995). For the full Bayesian two-stage procedure, we may adopt the marginal likelihood criterion at 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 was applied to reformulated probit regression setup constructed in a hierarchical truncated normal mixture model by introducing hyperparameters that will be used to identify subset choices. Among the hyperparameters, c_j and σ_j, $j = 1, \ldots, p$ were assumed to be known even though values of them were not readily available. We gave some useful guidelines to select them. The illustrated examples showed that the approach was 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 a more complicated algorithm, 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 will be left as a future study of interest.

REFERENCES

Albert, J. H. and Chib, S. (1993). "Bayesian analysis of binary and polychotomous response data," *Journal of the American Statististical Association,* **88**, 669-679.

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

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

Best, N., Cowles, M. K., and Vines, K. (1996). *CODA; Convergence diagnosis and output analysis software for Gibbs sampling output version 0.30,* MRC Biostatistics Unit, Cambridge.

Casella, G. and George, E. I. (1992). "Explaining the Gibbs sampler," *American Statistician,* **46**, 167-174.

Chib, S. (1995). "Marginal likelihood from the Gibbs output," *Journal of the American Statististical Association,* **90**, 1313-1321.

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

Cowles, M. K. and Carlin, B. P. (1996). "Markov chain Monte Carlo convergence diagnostics: a comparative review," *Journal of the American Statististical Association,* **91**, 883-904.

Devroye, L. (1986). *Non-uniform random generation,* Springer Verlag, New York.

Gelman, A. and Rubin, D. B. (1992). "Inference from iterative simulation using multiple sequences (with discussion)," *Statistical Science,* **7**, 457-511.

George, E. I. and McCulloch, R. E. (1993). "Variable selection via Gibbs sampling," *Journal of the American Statististical Association,* **88**, 881-889.

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

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

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,* B, **55**, 3-23.

Zellner, A. (1971). *An Introduction to Bayesian Inference in Econometrics,* Wiley, New York.

Zellner, A. (1986). "On assessing prior distributions and Bayesian Regression Analysis with g prior distributions," in *Bayesian Inference and Decision Techniques,* eds. P. Goel and A. Zellner, New York: Elsevier, 233-243.