

# Variable Selection for Multivariate Logistic Regression Models

Ming-Hui Chen and Dipak K. Dey\*

*Journal of Statistical Planning and Inference*, 111, 37-55

## Abstract

In this paper, we use multivariate logistic regression models to incorporate correlation among binary response data. Our objective is to develop a variable subset selection procedure to identify important covariates in predicting correlated binary responses using a Bayesian approach. In order to incorporate available prior information, we propose a class of informative prior distributions on the model parameters and on the model space. The propriety of the proposed informative prior is investigated in detail. Novel computational algorithms are also developed for sampling from the posterior distribution as well as for computing posterior model probabilities. Finally, a simulated data example and a real data example from a prostate cancer study are used to illustrate the proposed methodology.

**AMS 1991 Subject Classification:** Primary 62F15; Secondary 62J12.

**Keywords and Phrases:** Bayesian computation; Hierarchical model; Markov chain Monte Carlo; Prior elicitation; Propriety.

---

\*Ming-Hui Chen is Associate Professor of Statistics and Dipak K. Dey is Professor of Statistics, Department of Statistics, University of Connecticut, 215 Glenbrook Road, U-4120, Storrs, CT 06269-4120, U.S.A.

# 1 Introduction

Many methods have been proposed in regression models for variable selection. Classical methods for variable selection include forward selection, backward elimination, and stepwise regression. They sequentially delete or add predictors by examining the mean squared error or a modified version of it. Various Bayesian methods have also been proposed. They include model determination using the Bayesian information criterion (BIC, Schwarz, 1978), and various types of Bayes factors. Mitchell and Beauchamp (1988) proposed 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. George and McCulloch (1993) proposed a stochastic search variable selection procedure where the subset selection is derived from a hierarchical normal mixture model. Recently Kuo and Mallick (1998) explored a simpler method of subset selection by embedding indicator variables in the regression equations that incorporate all submodels. They also extended their results to univariate generalized linear models. However there are no results available for the multivariate logistic regression models.

Logistic regression is widely used to model independent binary response data in medical and epidemiologic studies. However, in many applications, this independence is not a reasonable assumption. This is particularly obvious in longitudinal studies, where multiple measurements made on the same individual are likely to be correlated. Similarly, the observations taken within a subject or a cluster are also possibly related. It is important to incorporate the correlation in modeling such correlated data. Prentice (1988) provided a comprehensive review of various modeling strategies using generalized linear regression analysis of correlated binary data with covariates associated at each binary response. Following Liang and Zeger (1986) and Zeger and Liang (1986), Prentice used the generalized estimating equation (GEE) approach to obtain consistent and asymptotically normal estimators of regression coefficients.

In this paper we first generalize univariate logistic regression to multivariate logistic regression using a scale mixture of normals proposed by Chen and Dey (1998). The main objective of this paper is to identify important covariates in predicting binary outcomes. From a Bayesian perspective, historical data can be very helpful in interpreting the results of the current study. However, very few methods exist for the formal incorporation of historical data to construct the prior distribution. There is some literature addressing this issue for the linear model and generalized linear models. See for example, Ibrahim and Laud (1994), Laud and Ibrahim (1995), Ibrahim, Ryan, and Chen

(1998), Chen, Ibrahim, and Yiannoutsos (1999), and Bedrick, Christensen, and Johnson (1996). In all of these papers, the authors assume a univariate independent response variable. The literature on Bayesian variable subset selection using informative prior elicitation for models with correlated binary responses is essentially nonexistent.

In this paper, we propose classes of informative prior distributions for correlated binary response data. The prior specification is based on the notion of the existence of a similar previous study that measures the same covariates and the same responses as the current study. We call the data from the similar previous study the *historical data*. The priors considered here are very attractive for model selection problems. In particular, the proposed priors are attractive for variable subset selection, and this application serves as the primary motivation for the priors. This is so since their construction is based on observable quantities which do not change meaning from model to model. Thus, given the historical data, the elicitation scheme becomes automatic with very few prior hyperparameters needing to be specified. This is attractive in variable selection contexts since specifying meaningful prior distributions for the parameters in each subset model is a difficult task requiring contextual interpretations of a large number of parameters. In this paper, the Monte Carlo methods we propose will facilitate a very fast and efficient way of computing the posterior model probabilities using only a *single* posterior sample from a *single* model, that being the full model. Such a procedure has been proved to be quite feasible and powerful in model selection contexts (see, for example, Chen, Ibrahim, and Yiannoutsos, 1999). In addition, our proposed informative prior elicitation scheme allows us to incorporate historical data in a natural way. Since our priors are based on historical data, as shown in Section 4, the proposed priors greatly ease the massive prior elicitation task required for variable selection problems. Thus our methods will enable us to do inference that could not otherwise be feasibly conducted using frequentist or other Bayesian methods.

The rest of the paper is organized as follows. In Section 2, we present multivariate logistic regression models and derive the resulting likelihood functions. Section 3 is devoted to the development of the prior distributions and an examination of their propriety. In Section 4, we develop efficient computational algorithms to sample from the posterior distributions as well as to compute posterior model probabilities. In Section 5, a small scale simulation study is conducted and a real data set from a prostate cancer study is used to illustrate the proposed methodology. Section 6 gives brief concluding remarks.

## 2 The Multivariate Logistic Regression Models

We first introduce some notation which will be used throughout the paper. Suppose that we observe a binary (0-1) response  $Y_{ij}$  on the  $j^{\text{th}}$  variable for the  $i^{\text{th}}$  subject. Let  $x_{ij} = (x_{ij1}, x_{ij2}, \dots, x_{ijp_j})$  be the corresponding  $p_j$ -dimensional row vector of covariates for  $i = 1, 2, \dots, n$  and  $j = 1, 2, \dots, J$ , where  $x_{ij1} = 1$ , which corresponds to an intercept, and  $x_i = \text{bdiag}(x_{i1}, x_{i2}, \dots, x_{iJ})$  denotes the block diagonal matrix with the  $j^{\text{th}}$  diagonal block equal to  $x_{ij}$ . Let  $Y_i = (Y_{i1}, Y_{i2}, \dots, Y_{iJ})'$  and assume that  $Y_{i1}, Y_{i2}, \dots, Y_{iJ}$  are dependent and  $Y_1, Y_2, \dots, Y_n$  are independent. Let  $y_i = (y_{i1}, y_{i2}, \dots, y_{iJ})'$  and  $y = (y_1, y_2, \dots, y_n)$  be the observed data. Also let  $\beta_j = (\beta_{j1}, \beta_{j2}, \dots, \beta_{jp_j})'$  be a  $p_j$ -dimensional column vector of regression coefficients.

For the purpose of variable subset selection, we need the following additional notation. Let  $\mathcal{M}$  denote the model space. We enumerate the models in  $\mathcal{M}$  by  $m = 1, 2, \dots, \mathcal{K}$ , where  $\mathcal{K}$  is the dimension of  $\mathcal{M}$  and model  $\mathcal{K}$  denotes the full model. The full model is defined here as the model containing all of the available covariates in the study. Then, the dimension of the regression coefficients for the full model is  $k = \sum_{j=1}^J p_j$ . Also, let  $\beta^{(\mathcal{K})} = (\beta'_1, \dots, \beta'_J)'$  denote the regression coefficients for the full model including  $J$  intercepts, and let  $\beta^{(m)} = ((\beta_1^{(m)})', \dots, (\beta_J^{(m)})')'$  denote a  $k_m \times 1$  vector of regression coefficients for model  $m$  with  $J$  intercepts, and a specific choice of  $k_m - J$  covariates, where  $\beta^{(-m)}$  is  $\beta^{(\mathcal{K})}$  with  $\beta^{(m)}$  deleted. Corresponding to  $\beta^{(m)}$ , we write  $x_i^{(m)} = \text{bdiag}(x_{i1}^{(m)}, x_{i2}^{(m)}, \dots, x_{iJ}^{(m)})$ , which is the block diagonal matrix with the  $j^{\text{th}}$  diagonal block equal to  $x_{ij}^{(m)}$ .

In order to set up model  $m$  with a multivariate logistic link, we introduce a  $J$ -dimensional (latent) random vector  $w_i = (w_{i1}, w_{i2}, \dots, w_{iJ})'$  such that

$$Y_{ij} = \begin{cases} 1 & \text{if } w_{ij} > 0 \\ 0 & \text{if } w_{ij} \leq 0 \end{cases} \quad (2.1)$$

and assume that

$$w_i \sim N(x_i^{(m)}\beta^{(m)}, \kappa(\lambda)\Sigma), \quad (2.2)$$

and the mixing variable

$$\lambda \sim \pi_K(\lambda), \quad (2.3)$$

where  $\kappa(\lambda) = 4\lambda^2$ , and  $\pi_K(\lambda)$  is the density of an asymptotic Kolmogorov distribution, which takes the form

$$\pi_K(\lambda) = 8 \sum_{k=1}^{\infty} (-1)^{k+1} k^2 \lambda \exp\{-2k^2 \lambda^2\}. \quad (2.4)$$

In (2.2), we take  $\Sigma = (\rho_{jj^*})_{J \times J}$  to be a correlation matrix such that  $\rho_{jj} = 1$  to ensure identifiability of the parameters. See Chen and Dey (1998) for a detailed discussion. Such a  $w_i$  is sometimes called a tolerance variable since in a bioassay setting  $w_i$  can be a lethal dose of a drug.

Although the expression of the asymptotic Kolmogorov distribution  $\pi_K(\lambda)$  appears complicated,  $\pi_K(\lambda)$  has some nice computational properties. Chen and Dey (1998) presented a comprehensive study of this distribution. Using an appropriate  $t$  approximation to the logistic distribution, Chen and Dey (1998) showed that a good proposal density for  $\pi_K(\lambda)$  given by (2.4) is

$$g_L(\lambda|\nu, b) = \frac{\left(\frac{\nu}{8b^2}\right)^{\nu/2}}{\Gamma\left(\frac{\nu}{2}\right)(\lambda^2)^{\nu/2+1}} \exp\left\{-\left(\frac{\nu}{8b^2}\right)\frac{1}{\lambda^2}\right\} 2\lambda. \quad (2.5)$$

They further showed that the best choices of  $\nu$  and  $b$  are  $\nu = 5$  and  $b = .712$  and they also provided an efficient way to evaluate the infinite series of  $\pi_K(\lambda)$ . It is interesting to mention that when we take

$$\lambda^2 \sim \mathcal{IG}\left(\frac{\nu}{2}, \frac{\nu}{8b^2}\right),$$

where  $\mathcal{IG}(u, v)$  is an inverse gamma distribution with a pdf  $\pi_{\mathcal{IG}}(\lambda|u, v) = \frac{v^u}{\Gamma(u)\lambda^{u+1}} e^{-v/\lambda}$ ,  $\lambda > 0$ , then

$$\lambda \sim g_L(\lambda|\nu, b).$$

This property is useful in the implementation of Markov chain Monte Carlo sampling from the posterior distribution. We will elaborate on this further in Section 4.

Finally, we mention that the distribution of  $w_i$  determines the joint distribution of  $Y_i$  through (2.1) and the correlation matrix  $\Sigma$  captures the correlations among the  $Y_{ij}$ 's. More specifically, the joint distribution of the responses is given by

$$\begin{aligned} p(y_i|\beta^{(m)}, \Sigma, \lambda_i, x_i^{(m)}) &= P(Y_{i1} = y_{i1}, Y_{i2} = y_{i2}, \dots, Y_{iJ} = y_{iJ}|\beta^{(m)}, \Sigma, \lambda_i, x_i^{(m)}) \\ &= \int_{A_{i1}} \int_{A_{i2}} \dots \int_{A_{iJ}} \frac{1}{(2\pi\kappa(\lambda_i))^{J/2} |\Sigma|^{\frac{1}{2}}} \exp\left\{-\frac{\kappa^{-1}(\lambda_i)}{2}(w_i - x_i^{(m)}\beta^{(m)})'\Sigma^{-1}(w_i - x_i^{(m)}\beta^{(m)})\right\} dw_i, \end{aligned} \quad (2.6)$$

where

$$A_{ij} = \begin{cases} (-\infty, 0] & \text{if } y_{ij} = 0 \\ (0, \infty) & \text{if } y_{ij} = 1 \end{cases}. \quad (2.7)$$

Let  $D^{(m)} = (n, y, x^{(m)})$  denote the current data. From (2.6), the likelihood function based on the observed data  $D^{(m)}$  is given by

$$L(\beta^{(m)}, \Sigma|D^{(m)}) = \prod_{i=1}^n \int_0^\infty p(y_i|\beta^{(m)}, \Sigma, \lambda_i, x_i^{(m)})\pi_K(\lambda_i)d\lambda_i. \quad (2.8)$$

### 3 The Prior Distributions

In this section, we develop a class of informative prior distributions for the model parameters  $(\beta^{(m)}, \Sigma)$  as well as a prior distribution on the model space.

#### 3.1 Prior Distribution on Model Parameters $(\beta^{(m)}, \Sigma)$

Informative prior elicitation is an important part of a Bayesian analysis, especially in variable subset selection since proper prior distributions are required to compute posterior model probabilities. We propose a class of informative priors for the regression coefficients  $\beta^{(m)}$ , since these parameters are of primary inferential interest in the variable selection problem. Our prior construction for  $\beta^{(m)}$  is based on the availability of historical data. For ease of exposition, we assume only one previous study, as the extension to multiple previous studies is straightforward. To this end, let  $D_0^{(m)} = (n_0, y_0, x_0^{(m)})$  be the data from the historical study and let  $w_{0i} = (w_{0i1}, \dots, w_{0iJ})'$  be the latent variable vector associated with the historical study.

Let  $\pi_0(\beta^{(m)}, \Sigma | c_0)$  be an *initial prior* distribution for  $(\beta^{(m)}, \Sigma)$ , which takes the form

$$\pi_0(\beta^{(m)}, \Sigma | c_0) \propto \exp \left\{ -\frac{1}{2c_0} (\beta^{(m)})' B_0^{(m)} \beta^{(m)} \right\}, \quad (3.1)$$

where  $B_0^{(m)}$  is a precision matrix,  $c_0$  is a scalar parameter, and both  $B_0^{(m)}$  and  $c_0$  are prespecified. Typically, we choose  $B_0^{(m)} = I_{k_m}$ , where  $I_{k_m}$  is the  $k_m \times k_m$  identity matrix. Also in (3.1),  $vec^*(\Sigma) = (\rho_{12}, \rho_{13}, \dots, \rho_{J-1, J})' \in V$ , where the region  $V$  is a subset of the region  $[-1, 1]^{J(J-1)/2}$  that leads to a proper correlation matrix. As mentioned by Chib and Greenberg (1998) and also shown by Rousseeuw and Molenberghs (1994), the region  $V$  forms a convex solid body in the hypercube  $[-1, 1]^{J(J-1)/2}$ . Note that in (3.1), we use a uniform prior on the region  $V$  for  $\Sigma$ . Also note that when  $c_0 < \infty$ ,  $\pi_0(\beta^{(m)}, \Sigma | c_0)$  is proper; but when  $c_0 = \infty$ ,  $\pi_0(\beta^{(m)}, \Sigma | c_0)$  is an improper uniform prior.

The informative prior based on historical data takes the form

$$\pi(\beta^{(m)}, \Sigma | a_0, D_0^{(m)}) \propto \pi^*(\beta^{(m)}, \Sigma | a_0, D_0^{(m)}) \pi_0(\beta^{(m)}, \Sigma | c_0), \quad (3.2)$$

where

$$\begin{aligned} \pi^*(\beta^{(m)}, \Sigma | a_0, D_0^{(m)}) &= \prod_{i=1}^{n_0} \int_0^\infty \int_{A_{0i1}} \dots \int_{A_{0iJ}} \frac{a_0^{J/2} |\Sigma|^{-1/2}}{(2\pi\kappa(\lambda_{0i}))^{J/2}} \exp \left\{ -\frac{a_0\kappa^{-1}(\lambda_{0i})}{2} \right. \\ &\quad \left. (w_{0i} - x_{0i}^{(m)} \beta^{(m)})' \Sigma^{-1} (w_{0i} - x_{0i}^{(m)} \beta^{(m)}) \right\} \pi_K(\lambda_{0i}) dw_{0i} d\lambda_{0i}, \quad (3.3) \end{aligned}$$

and

$$A_{0ij} = \begin{cases} (-\infty, 0] & \text{if } y_{0ij} = 0 \\ (0, \infty) & \text{if } y_{0ij} = 1 \end{cases}. \quad (3.4)$$

From (3.2), the *initial prior*  $\pi_0(\beta^{(m)}, \Sigma | c_0)$  may be viewed as the prior for  $(\beta^{(m)}, \Sigma)$  before observing the historical data. The prior parameter  $c_0$  controls the impact of  $\pi_0(\beta^{(m)}, \Sigma | c_0)$  on the entire prior. In (3.2),  $0 \leq a_0 \leq 1$  is a scalar prior precision parameter that weights the historical data relative to the likelihood of the current study. Small values of  $a_0$  give little prior weight to the historical data relative to the likelihood of the current study, whereas values of  $a_0$  close to 1 give roughly equal weight to the prior and the likelihood of the current study. The case  $a_0 = 1$  corresponds to the formal Bayesian update of  $\pi_0(\beta^{(m)}, \Sigma | c_0)$  using Bayes theorem. Thus, with  $a_0 = 1$ , the prior and likelihood of the current study are equally weighted, and (3.2) corresponds to the posterior distribution of  $(\beta^{(m)}, \Sigma)$  based on the data  $D_0^{(m)}$ . The case  $a_0 = 0$  results in no incorporation of historical data, and in this case, the prior reduces to the initial prior. The parameter  $a_0$  allows the investigator to control the influence of the historical data on the current study. Such control is important in cases where there is heterogeneity between the historical data and the current study, or when the sample sizes of the two studies are quite different.

The prior specification is completed by specifying a prior distribution for  $a_0$ . We take a beta prior for  $a_0$ , and thus we propose a joint prior distribution for  $(\beta^{(m)}, \Sigma, a_0)$  of the form

$$\pi(\beta^{(m)}, \Sigma, a_0 | D_0^{(m)}) \propto \pi^*(\beta^{(m)}, \Sigma | a_0, D_0^{(m)}) \pi_0(\beta^{(m)}, \Sigma | c_0) a_0^{\delta_0 - 1} (1 - a_0)^{\lambda_0 - 1}, \quad (3.5)$$

where  $(\delta_0, \lambda_0)$  are specified prior hyperparameters.

In the context of Bayesian variable subset selection, computing posterior model probabilities requires *proper* priors. Therefore, it is important to show that (3.5) is proper when  $\pi_0(\beta, \Sigma | c_0)$  is an improper uniform prior, i.e.,  $c_0 = \infty$ . The following theorem characterizes the propriety of (3.5) with an improper uniform initial prior.

Before presenting the theorem, we introduce the following notation. Let  $z_{0ij} = 1$  if  $y_{0ij} = 0$  and  $z_{0ij} = -1$  if  $y_{0ij} = 1$ . Also let  $x_{0ij}^* = z_{0ij} x_{0ij}^{(m)}$ ,  $x_{0i}^* = \text{bdiag}(x_{0i1}^*, x_{0i2}^*, \dots, x_{0iJ}^*)$ , and

$$X_0^* = \begin{pmatrix} x_{01}^* \\ \vdots \\ x_{0n_0}^* \end{pmatrix}.$$

**Theorem 3.1** *Assume that the following conditions are satisfied:*

(C1)  $X_0^*$  is of full rank,

(C2) There exists an  $n_0 \times 1$  positive vector  $a$  such that

$$a'X_0^* = 0,$$

(C3)  $\delta_0 > k_m/2$ , where  $k_m$  is the dimension of  $\beta^{(m)}$ ,

(C4)  $\pi_0(\beta, \Sigma | c_0) \propto 1$ , i.e.,  $c_0 = \infty$ .

Then, the joint prior  $\pi(\beta^{(m)}, \Sigma, a_0 | D_0^{(m)})$  given in (3.5) is proper, that is,

$$\int_0^1 \int_V \int_{R^{k_m}} \pi^*(\beta^{(m)}, \Sigma | a_0, D_0^{(m)}) a_0^{\delta_0-1} (1-a_0)^{\lambda_0-1} d\beta^{(m)} d\text{vec}^*(\Sigma) da_0 < \infty, \quad (3.6)$$

where  $R^{k_m}$  denotes  $k_m$  dimensional Euclidean space.

The proof of the theorem is technical, and thus left to the appendix.

### 3.2 Prior Distribution on the Model Space

Let

$$p_0^*(\beta^{(m)} | D_0^{(m)}) = \int_0^1 \int_V \pi^*(\beta^{(m)}, \Sigma | a_0, D_0^{(m)}) \pi_0(\beta^{(m)}, \Sigma | c_0) a_0^{\delta_0-1} (1-a_0)^{\lambda_0-1} d\text{vec}^*(\Sigma) da_0. \quad (3.7)$$

We see that  $p_0^*(\beta^{(m)} | D_0^{(m)})$  is proportional to the marginal prior of  $\beta^{(m)}$ . We propose to take the prior probability of model  $m$  as

$$p(m) = \frac{\int_{R^{(m)}} p_0^*(\beta^{(m)} | D_0^{(m)}) d\beta^{(m)}}{\sum_{j=1}^{\mathcal{K}} \int_{R^{(m)}} p_0^*(\beta^{(j)} | D_0^{(j)}) d\beta^{(j)}}. \quad (3.8)$$

This choice for  $p(m)$  in (3.8) is a natural one since the numerator is just the normalizing constant of the joint prior of  $(\beta^{(m)}, a_0, \sigma^2)$  under model  $m$ . The prior model probabilities in (3.8) are based on coherent Bayesian updating and this results in several attractive interpretations. Firstly,  $p(m)$  in (3.8) corresponds to the posterior probability of model  $m$  based on the data  $D_0^{(m)}$  using a uniform prior for the previous study,  $p_0(m) = 2^{-k}$  for  $m \in \mathcal{M}$  as  $\delta_0 \rightarrow \infty$ . That is,  $p(m) \propto p(m | D_0^{(m)})$ , and thus  $p(m)$  corresponds to the usual Bayesian update of  $p_0(m)$  using  $D_0^{(m)}$  as the data. Secondly, as  $\lambda_0 \rightarrow \infty$ ,  $p(m)$  reduces to a uniform prior on the model space. Therefore, as  $\lambda_0 \rightarrow \infty$ , the historical data  $D_0^{(m)}$  have a minimal impact in determining  $p(m)$ . On the other hand,  $p(m)$  in (3.8) has a nice theoretical property, which greatly eases the computational burden for calculating posterior model probabilities using the Markov chain Monte Carlo (MCMC) output. These properties are discussed in more detail in the next section.

## 4 Computational Development

In this section, we develop necessary tools for sampling from the posterior distribution and for computing the posterior model probabilities.

### 4.1 Sampling from the Posterior Distribution

We only need to consider how to sample from the posterior distribution under the full model. For ease of exposition, we drop the model index  $\mathcal{K}$  to present our MCMC sampling algorithm. From (3.5), the joint posterior distribution of  $(\beta, \Sigma)$  and  $a_0$  can be written as

$$\begin{aligned} & \pi(\beta, \Sigma, a_0 | D, D_0) \\ & \propto L(\beta, \Sigma | D) \pi^*(\beta, \Sigma | a_0, D_0^{(m)}) \pi_0(\beta, \Sigma | c_0) a_0^{\delta_0 - 1} (1 - a_0)^{\lambda_0 - 1}, \end{aligned} \quad (4.1)$$

where  $L(\beta, \Sigma | D)$  is given by (2.8). To sample from the posterior distribution  $\pi(\beta, \Sigma, a_0 | D, D_0)$ , we introduce the vectors of latent variables and the mixing variables  $w = (w_1, w_2, \dots, w_n)$  and  $\lambda = (\lambda_1, \lambda_2, \dots, \lambda_n)$  for the current study, and  $w_0 = (w_{01}, w_{02}, \dots, w_{0n})$  and  $\lambda_0 = (\lambda_{01}, \lambda_{02}, \dots, \lambda_{0n})$  for the historical study. Then, the joint posterior distribution of  $(\beta, \Sigma, a_0, w, \lambda, w_0, \lambda_0)$  is given by

$$\begin{aligned} & \pi(\beta, \Sigma, a_0, w, \lambda, w_0, \lambda_0 | D, D_0) \\ & \propto \prod_{i=1}^n \left[ \frac{1}{(\kappa(\lambda_i))^{J/2} |\Sigma|^{\frac{1}{2}}} \exp \left\{ -\frac{\kappa^{-1}(\lambda_i)}{2} (w_i - x_i \beta)' \Sigma^{-1} (w_i - x_i \beta) \right\} \pi_K(\lambda_i) \right] \\ & \quad \times \prod_{i=1}^{n_0} \left[ \frac{a_0^{J/2} |\Sigma|^{-1/2}}{(\kappa(\lambda_{0i}))^{J/2}} \exp \left\{ -\frac{a_0 \kappa^{-1}(\lambda_{0i})}{2} (w_{0i} - x_{0i} \beta)' \Sigma^{-1} (w_{0i} - x_{0i} \beta) \right\} \pi_K(\lambda_{0i}) \right] \\ & \quad \times \pi_0(\beta, \Sigma | c_0) a_0^{\delta_0 - 1} (1 - a_0)^{\lambda_0 - 1}, \end{aligned} \quad (4.2)$$

where  $\pi_0(\beta, \Sigma | c_0) \propto \exp \left\{ -\frac{1}{2c_0} \beta' B_0 \beta \right\}$ . To run the Gibbs sampler (see, for example, Gelfand and Smith 1990; Geman and Geman 1984), we need to sample from the following full conditional distributions:

- (i)  $[\beta | \Sigma, a_0, w, \lambda, w_0, \lambda_0, D, D_0]$ ;
- (ii)  $[\Sigma | \beta, a_0, w, \lambda, w_0, \lambda_0, D, D_0]$ ;
- (iii)  $[a_0 | \beta, \Sigma, w_0, \lambda_0, D, D_0]$ ;
- (iv)  $[w, w_0 | \beta, \Sigma, a_0, \lambda, \lambda_0, D, D_0]$ ; and

(v)  $[\lambda, \lambda_0 | \beta, \Sigma, a_0, w, w_0, D, D_0]$ .

We briefly describe how to sample from the above conditional distributions. For (i), it is easy to show that  $[\beta | \Sigma, a_0, w, \lambda, w_0, \lambda_0, D, D_0]$  is a multivariate normal distribution  $N(\hat{\beta}, B^{-1})$ , where

$$B = (1/c_0)B_0 + a_0 \sum_{i=1}^{n_0} \kappa^{-1}(\lambda_{0i})x'_{0i}\Sigma^{-1}x_{0i} + \sum_{i=1}^n \kappa^{-1}(\lambda_i)x'_i\Sigma^{-1}x_i$$

and

$$\hat{\beta} = B^{-1} \left( a_0 \sum_{i=1}^{n_0} \kappa^{-1}(\lambda_{0i})x'_{0i}\Sigma^{-1}w_{0i} + \sum_{i=1}^n \kappa^{-1}(\lambda_i)x'_i\Sigma^{-1}w_i \right).$$

Thus, sampling  $\beta$  from its full conditional distribution is straightforward. To sample the correlation matrix  $\Sigma$  from  $[\Sigma | \beta, a_0, w, \lambda, w_0, \lambda_0, D, D_0]$ , we use a Metropolized hit-and-run algorithm of Chen and Dey (1998), which is a generalization of the Metropolis algorithm of Chib and Greenberg (1998). The details for generating  $\Sigma$  can be found in Chen and Dey (1998). For (iii) it can be shown that the conditional posterior density of  $[a_0 | \beta, \Sigma, w_0, \lambda_0, D, D_0]$  is log-concave when  $\delta_0 > 0$  and  $\lambda_0 \geq 1$ . Thus, we use the adaptive rejection algorithm of Gilks and Wild (1992) to sample  $a_0$ . For (iv), we use a cycle of  $J$  Gibbs steps to generate  $w_{ij}$ , and  $w_{0ij}$  from their respective conditional distributions for  $j = 1, 2, \dots, J$  in turn. We use the algorithm of Geweke (1991) to generate  $w_{ij}$  and  $w_{0ij}$ , since their respective conditional posterior distributions are truncated multivariate normals over intervals defined by, for example, (2.7) or (3.4). Finally, we briefly discuss how to generate the mixing variables  $\lambda_i$  and  $\lambda_{0i}$ . Using a proposal density  $g_L(\lambda_i | \nu, b)$  or  $g_L(\lambda_{0i} | \nu, b)$  given by (2.5), Chen and Dey (1998) developed an efficient Metropolis algorithm (Metropolis *et al.* 1953). For illustrative purposes, we consider only how to draw  $\lambda_i$ , since drawing  $\lambda_{0i}$  proceeds in a similar fashion. Let  $\lambda_i$  be the current value. Generate

$$\lambda_i^{*2} \sim \mathcal{IG} \left( \frac{J + \nu}{2}, \frac{1}{8} \left[ (w_i - x_i\beta)' \Sigma^{-1} (w_i - x_i\beta) + \frac{\nu}{b^2} \right] \right). \quad (4.3)$$

Then, a move to the proposal point  $\lambda_i^*$  is made with probability

$$\min \left\{ \frac{\pi_K(\lambda_i^*)/g_L(\lambda_i^* | \nu, b)}{\pi_K(\lambda_i)/g_L(\lambda_i | \nu, b)}, 1 \right\}, \quad (4.4)$$

where  $\pi_K(\lambda_i)$  and  $g_L(\lambda_i | \nu, b)$  are given in (2.4) and (2.5), respectively.

## 4.2 Computation of Model Probabilities

Using Bayes theorem, the posterior probability of model  $m$  is given by

$$p(m | D^{(m)}) = \frac{p(D^{(m)} | m) p(m)}{\sum_{j=1}^{\mathcal{K}} p(D^{(j)} | j) p(j)}, \quad (4.5)$$

where  $p(D^{(m)}|m)$  denotes the marginal distribution of the data  $D^{(m)}$  for the current study under model  $m$ , and  $p(m)$  denotes the prior probability of model  $m$  in (3.8).

Now, we consider how to compute  $p(m|D^{(m)})$ . Letting  $c_m$  and  $c_{0m}$  be the respective normalizing constants for the joint posterior and prior distributions under model  $m$ , we have

$$c_m = \int_0^1 \int_V \int_{R^{(m)}} L(\beta^{(m)}, \Sigma | D^{(m)}) \pi^*(\beta^{(m)}, \Sigma | a_0, D_0^{(m)}) \\ \times \pi_0(\beta^{(m)}, \Sigma | c_0) a_0^{\delta_0-1} (1-a_0)^{\lambda_0-1} d\beta^{(m)} d\text{vec}^*(\Sigma) da_0,$$

and

$$c_{0m} = \int_0^1 \int_V \int_{R^{(m)}} \pi^*(\beta^{(m)}, \Sigma | a_0, D_0^{(m)}) \pi_0(\beta, \Sigma | c_0) a_0^{\delta_0-1} (1-a_0)^{\lambda_0-1} d\beta^{(m)} d\text{vec}^*(\Sigma) da_0,$$

where  $\pi^*(\beta^{(m)}, \Sigma | a_0, D_0^{(m)})$  is given in (3.3). Then, the marginal distribution of the data  $D^{(m)}$  is

$$p(D^{(m)}|m) = \int_0^1 \int_V \int_{R^{(m)}} L(\beta^{(m)}, \Sigma | D) \pi(\beta^{(m)}, \Sigma, a_0 | D_0^{(m)}) d\beta^{(m)} d\text{vec}^*(\Sigma) da_0 = \frac{c_m}{c_{0m}}. \quad (4.6)$$

It immediately follows from (4.6) that (4.5) can be rewritten as

$$p(m|D^{(m)}) = \frac{(c_m/c_{0m}) p(m)}{\sum_{j=1}^{\mathcal{K}} (c_j/c_{0j}) p(j)}. \quad (4.7)$$

From (4.7), it can be seen that the calculation of posterior probabilities requires evaluating  $\mathcal{K}$  ratios of two normalizing constants and  $\mathcal{K}$  prior model probabilities. Since each quantity in (4.7) involves an analytically intractable high dimensional integral, the computation of the posterior model probabilities is a difficult task. However, the following key theoretical result will greatly ease such a challenging computational problem, which makes the implementation of Bayesian variable selection feasible for this model.

Recall that  $\beta^{(\mathcal{K})} = (\beta^{(m)'}, \beta^{(-m)'})'$  where  $\beta^{(-m)}$  is  $\beta^{(\mathcal{K})}$  with  $\beta^{(m)}$  deleted. Then we have the following result.

**Theorem 4.1** *Assume that  $\pi_0(\beta^{(m)}, \Sigma | c_0) = \pi_0(\beta^{(m)}, \Sigma | \beta^{(-m)} = 0, c_0)$ , and conditions (C1)–(C3) given in Theorem 3.1 hold. Then, the posterior probability  $p(m|D^{(m)})$  of model  $m$  is given by*

$$p(m|D^{(m)}) = \frac{\pi(\beta^{(-m)} = 0 | D^{(\mathcal{K})}, D_0^{(\mathcal{K})})}{\sum_{j=1}^{\mathcal{K}} \pi(\beta^{(-j)} = 0 | D^{(\mathcal{K})}, D_0^{(\mathcal{K})})}, \quad (4.8)$$

$m = 1, \dots, \mathcal{K}$ , where  $\beta^{(-m)}$  is  $\pi(\beta^{(-m)} = 0 | D^{(\mathcal{K})}, D_0^{(\mathcal{K})})$  is the marginal posterior density of  $\beta^{(-m)}$  evaluated at  $\beta^{(-m)} = 0$ , and we write  $\pi(\beta^{(-\mathcal{K})} = 0 | D^{(\mathcal{K})}, D_0^{(\mathcal{K})}) = 1$ .

**Proof:** Under conditions (C1)–(C3) given in Theorem 3.1, using Lemmas 1 and 2 of Chen, Ibrahim, and Yiannoutsos (1999) and the Savage-Dicky density ratio (see Verdinelli and Wasserman, 1995), it can be shown that

$$\frac{p(D^{(m)}|m)}{p(D^{(\mathcal{K})}|\mathcal{K})} = \frac{\pi(\beta^{(-m)} = 0|D^{(\mathcal{K})}, D_0^{(\mathcal{K})})}{\pi(\beta^{(-m)} = 0|D_0^{(\mathcal{K})})}, \quad m = 1, \dots, \mathcal{K}, \quad (4.9)$$

where  $\pi(\beta^{(-m)} = 0|D_0^{(\mathcal{K})})$  is the marginal prior density of  $\beta^{(-m)}$  evaluated at  $\beta^{(-m)} = 0$ . Then, after some algebra, we get

$$p(m) = c_0^* \pi(\beta^{(-m)} = 0|D_0^{(\mathcal{K})}), \quad (4.10)$$

where  $c_0^*$  is a constant that does not depend on the model index  $m$ . Combining (4.9) and (4.10) together gives (4.8).  $\square$

We first emphasize that Theorem 4.1 is valid only if the prior distribution  $\pi(\beta, \Sigma, a_0|D_0^{(\mathcal{K})})$  is proper. Therefore, a proper prior distribution plays an important role in our Bayesian variable selection procedure. The result in (4.8) is very attractive, since it shows that the posterior model probability  $p(m|D^{(m)})$  is simply a function of the marginal posterior density functions of  $\beta^{(-m)}$  for the full model evaluated at  $\beta^{(-m)} = 0$ . This formula does not algebraically depend on the prior model probability  $p(m)$ , since it cancels out in the derivation due to the structure of  $p(m)$ . This is an important feature since it allows us to compute the posterior model probabilities directly *without* numerically computing the prior model probabilities. This has a clear computational advantage and as a result, allows us to compute posterior model probabilities very efficiently. We note that this computational device works best if all of the covariates are standardized to have mean 0 and variance 1. This is not restrictive since this is a typical transformation taken quite often in practice to numerically stabilize the MCMC sampling algorithms.

Due to the complexity of our model, the analytical evaluation of  $\pi(\beta^{(-m)} = 0|D^{(\mathcal{K})}, D_0^{(\mathcal{K})})$  does not appear possible. However,  $\pi(\beta^{(-m)} = 0|D^{(\mathcal{K})}, D_0^{(\mathcal{K})})$  can be estimated by the conditional marginal density estimation (CMDE) method using the MCMC output  $\{(\beta_{(l)}^{(\mathcal{K})}, \Sigma_{(l)}, a_0^{(l)}, w^{(l)}, \lambda^{(l)}, w_0^{(l)}, \lambda_0^{(l)}), l = 1, 2, \dots, L\}$  from the full model joint posterior distribution (4.2). Gelfand, Smith, and Lee (1992), Chen (1994), and Chen and Shao (1997) have shown that the CMDE is the most efficient Monte Carlo method for estimating marginal posterior densities when a joint posterior density is known up to a normalizing constant. It directly follows from Chen and Shao (1997) that a simulation consistent estimator of  $\pi(\beta^{(-m)} = 0|D^{(\mathcal{K})}, D_0^{(\mathcal{K})})$  is given by

$$\hat{\pi}(\beta^{(-m)} = 0|D^{(\mathcal{K})}, D_0^{(\mathcal{K})}) = \frac{1}{L} \sum_{l=1}^L N_{k-k_m}(\beta^{(-m)} = 0|\beta_{(l)}^{(-m)}, \hat{\beta}_{(l)}, (B_{(l)})^{-1}), \quad (4.11)$$

where  $N_{k-k_m}(\beta^{(-m)} = 0 | \beta_{(l)}^{(-m)}, \hat{\beta}_{(l)}^{(\mathcal{K})}, (B_{(l)})^{-1})$  is the  $(k - k_m)$ -dimensional conditional normal density function of  $N_k(\hat{\beta}_{(l)}^{(\mathcal{K})}, (B_{(l)})^{-1})$  given  $\beta_{(l)}^{(m)}$  evaluated at  $\beta^{(-m)} = 0$ ,

$$B_{(l)} = (1/c_0)B_0 + a_0^{(l)} \sum_{i=1}^{n_0} \kappa^{-1}(\lambda_{0i}^{(l)}) \left(x_{0i}^{(\mathcal{K})}\right)' \Sigma_{(l)}^{-1} x_{0i}^{(\mathcal{K})} + \sum_{i=1}^n \kappa^{-1}(\lambda_i^{(l)}) \left(x_i^{(\mathcal{K})}\right)' \Sigma_{(l)}^{-1} x_i^{(\mathcal{K})}$$

and

$$\hat{\beta}^{(\mathcal{K})} = B^{-1} \left( a_0^{(l)} \sum_{i=1}^{n_0} \kappa^{-1}(\lambda_{0i}^{(l)}) \left(x_{0i}^{(\mathcal{K})}\right)' \Sigma_{(l)}^{-1} w_{0i}^{(l)} + \sum_{i=1}^n \kappa^{-1}(\lambda_i^{(l)}) \left(x_i^{(\mathcal{K})}\right)' \Sigma_{(l)}^{-1} w_i^{(l)} \right).$$

## 5 Examples

### Example 1: Simulation Study

In this example, we first demonstrate variable subset selection with our proposed methodology and the computational feasibility of our methods using simulated data sets. We also study the impact of the dependence among the binary responses on the posterior model probabilities.

The data for the current study is generated as follows. We generate  $n = 400$  independent observations from a bivariate logit model defined by (2.1)–(2.4) with the mean of the latent random vector given by

$$x_i \beta = \begin{pmatrix} \beta_{10} + \beta_{11}x_{i11} + \beta_{12}x_{i12} + \beta_{13}x_{i13} + \beta_{14}x_{i14} \\ \beta_{20} + \beta_{21}x_{i21} + \beta_{22}x_{i22} + \beta_{23}x_{i23} + \beta_{24}x_{i24} \end{pmatrix},$$

$\beta'_1 = (\beta_{10}, \beta_{11}, \beta_{12}, \beta_{13}, \beta_{14}) = (1, 1.3, 1, 0, 0)$ , and  $\beta'_2 = (\beta_{20}, \beta_{21}, \beta_{22}, \beta_{23}, \beta_{24}) = (1, 1.5, 1, 0, 0)$ .

Also,  $(x_{i11}, x_{i13})'$  are generated as *i.i.d.* bivariate normal random vectors with mean  $(0, 0)'$  and covariance matrix  $\begin{pmatrix} 1 & 0.3 \\ 0.3 & 1 \end{pmatrix}$ , and we generate two other covariates  $(x_{i12}, x_{i14})$  which are *i.i.d.* normal random variables each with mean 0 and variance 1, and independent of  $(x_{i11}, x_{i13})$ . Then,

we take  $(x_{i21}, x_{i22}, x_{i23}, x_{i24}) = (x_{i11}, x_{i12}, x_{i13}, x_{i14})$ . Thus, the true model contains the two covariates  $[(x_1, x_2); (x_1, x_2)]$  for the bivariate response vector  $(y_1, y_2)$ , and the “full” model contains the eight covariates  $[(x_1, \dots, x_4); (x_1, \dots, x_4)]$ . Therefore, our model space  $\mathcal{M}$  contains 256 models, with an intercept included in each. The historical data were generated in a similar fashion.

We take  $n_0 = 300$  with all other parameters the same as for the current data. In addition,  $[(x_{0i11}, \dots, x_{0i14}); (x_{0i21}, \dots, x_{0i24})]$  are generated in exactly the same way as the current data. In addition, we take the correlation,  $\rho$ , between latent variables  $(w_{i1}, w_{i2})$  for the current data or  $(w_{0i1}, w_{0i2})$  for the historical data, to be one of the following values: 0, 0.5, 0.7, 0.95, which reflects independent, moderately correlated, and strongly correlated logit models. In total, we generate four pairs of current and historical data sets.

We implement the MCMC sampling algorithm described in Section 4.1. For each of the current and historical data sets, we generate 20,000 iterations after a burn-in of 1,000 iterations for computing posterior model probabilities using the computational procedure proposed in Section 4.2. Table 1 shows posterior probabilities for the top five models with  $c_0 = 100$ , and  $(\delta_0, \lambda_0) = (10, 10)$ , which gives a prior mean of 0.5 for  $a_0$ . From Table 1, we see that the true model obtains the largest posterior probability for all of the four simulated current and historical data sets. This result implies that our proposed variable selection method is able to identify the true model. It can also be observed that the posterior model probability of the true model is increasing in  $\rho$ , which is used to generate dependent bivariate binary responses. This finding is interesting, since the strong correlation among the correlated binary responses helps in selecting the true model. In addition, for all of the four simulated data sets, we obtain the same five top models. In particular,  $[(x_1, x_2); (x_1, x_2, x_3)]$  consistently obtains the second largest posterior model probability.

**Table 1: Posterior Model Probabilities For Simulated Data**

$\rho$	Model	$p(m D^{(m)})$	$\rho$	Model	$p(m D^{(m)})$
0.0	$[(x_1, x_2); (x_1, x_2)]$	0.407	0.7	$[(x_1, x_2); (x_1, x_2)]$	0.431
	$[(x_1, x_2); (x_1, x_2, x_3)]$	0.118		$[(x_1, x_2); (x_1, x_2, x_3)]$	0.114
	$[(x_1, x_2, x_3); (x_1, x_2)]$	0.101		$[(x_1, x_2); (x_1, x_2, x_4)]$	0.109
	$[(x_1, x_2); (x_1, x_2, x_4)]$	0.096		$[(x_1, x_2, x_3); (x_1, x_2)]$	0.088
	$[(x_1, x_2, x_4); (x_1, x_2)]$	0.095		$[(x_1, x_2, x_4); (x_1, x_2)]$	0.087
0.5	$[(x_1, x_2); (x_1, x_2)]$	0.427	0.95	$[(x_1, x_2); (x_1, x_2)]$	0.572
	$[(x_1, x_2); (x_1, x_2, x_3)]$	0.111		$[(x_1, x_2); (x_1, x_2, x_3)]$	0.091
	$[(x_1, x_2); (x_1, x_2, x_4)]$	0.103		$[(x_1, x_2, x_3); (x_1, x_2)]$	0.083
	$[(x_1, x_2, x_3); (x_1, x_2)]$	0.095		$[(x_1, x_2); (x_1, x_2, x_4)]$	0.079
	$[(x_1, x_2, x_4); (x_1, x_2)]$	0.093		$[(x_1, x_2, x_4); (x_1, x_2)]$	0.075

We also fit the simulated data using two independence logit models for binary responses. For example, for the simulated data with  $\rho = 0.95$ , the independence logit models select the true model as the top model with a posterior model probability of 0.405. Although the independence logit models select the same top model as the multivariate logit model, the posterior model probability for the independence logit models is 0.708 times smaller than the one for the multivariate logit model.

Finally, we note that other choices of  $c_0$  and  $(\delta_0, \lambda_0)$  and other values of regression coefficients were also tried, and similar results were obtained. We also note that the computation of all posterior model probabilities took 1.5 hours on a digital alpha machine. The computer code was written in FORTRAN 77 using double precision accuracy.

## Example 2: Prostate Cancer Study

To further illustrate the proposed methodology, we consider data from a prostate cancer study. Adenocarcinoma of the prostate is the second-leading cause of cancer mortality in men. In order to examine the relation between several prostate cancer response variables and the important preoperative staging system predictors, two similar studies were conducted at the University of Pennsylvania in Philadelphia and Brigham and Women’s Hospital in Boston. All patients involved in these two studies had undergone surgery. Two data sets, called the PENN data and the MASS data, were collected from these two studies. The PENN data were collected from 1989 to 1995, and the MASS data were collected between August of 1995 and April of 1996. In the analyses, we take the PENN data as the historical data and the MASS data as the current data. The sample size of the MASS data is  $n = 103$ , and the sample size of the PENN data is  $n_0 = 713$ . We consider Pathological Extracapsular Extension (PECE) and Pathological Positive Surgical Margins (PPSM) as two binary response variables ( $y = (y_1, y_2)$ ). PECE takes the values of “0” and “1”, indicating whether or not cancer has penetrated the prostatic capsule. A value of 0 indicates that there is no cancer present in or near the capsule at all, and a “1” indicates that the disease extends into or penetrates through the capsule. PPSM= 1 denotes that the cancer has penetrated the prostate wall to a point that is not removable by surgery, and PPSM= 0 indicates otherwise. Clearly, these binary responses are correlated. For illustrative purposes, we consider only four important clinical variables, which are Prostate Specific Antigen (PSA,  $x_1$ ), Clinical Gleason Score (GLEAS,  $x_2$ ), Clinical Stage (CLINS,  $x_3$ ), and Calculated Volume of Cancer Volume (CALVCA,  $x_4$ ). Summary statistics of these two data sets and detailed descriptions of the clinical variables can be found in Desjardin (1997). We standardize all covariates in the posterior computations. Using the PENN data as historical data and the MASS data as current data, we apply the proposed variable selection methods to identify the best subset of four clinical variables in predicting two correlated binary responses.

As in Example 1, we generate 20,000 iterations after a burn-in of 1,000 iterations for computing posterior model probabilities. Table 2 shows posterior probabilities for the top five models with  $c_0 = 100$ ,  $(\delta_0, \lambda_0) = (10, 50)$ ,  $(10, 10)$ ,  $(100, 10)$ , and  $a_0 = 1$  with probability 1 ( $\delta_0 \rightarrow \infty$ ). These choices reflect little, moderate, and high prior weight to the historical data. In Table 2,  $E(a_0|D)$  denotes the posterior mean of  $a_0$ . From Table 2, we see that model  $[(x_1, x_2, x_3); (x_1, x_2)]$ , i.e.,  $[(\text{PSA}, \text{GLEAS}, \text{CLINS}), (\text{PSA}, \text{GLEAS})]$ , obtains the largest posterior model probability for all four

choices of  $(\delta_0, \lambda_0)$ . Also, the posterior model probability of the top model is not very sensitive to the choice of the prior weight to the historical data. However, the set of the top five models using a high prior weight is different than the one using a low prior weight. In addition, we compute the posterior estimates of  $\rho$  under the above four choices of  $(\delta_0, \lambda_0)$ . The posterior means and the 95% highest posterior density intervals for  $\rho$  are 0.792 (0.732, 0.852), 0.792 (0.729, 0.850), 0.794 (0.731, 0.855), 0.795 (0.734, 0.854) for  $(\delta_0, \lambda_0) = (10, 50), (10, 10), (100, 10)$ , and  $a_0 = 1$  with probability 1, respectively. Therefore, the posterior estimate of  $\rho$  is not sensitive to the choice of  $(\delta_0, \lambda_0)$ . Finally, we also use the independence logit models to fit the prostate cancer data. For all four choices of  $(\delta_0, \lambda_0)$ , [(PSA, GLEAS, CLINS), (PSA, GLEAS)] obtains the largest posterior model probability, and the corresponding values are 0.247, 0.282, 0.263, and 0.263 for  $(\delta_0, \lambda_0) = (10, 50), (10, 10), (100, 10)$ , and  $a_0 = 1$  with probability 1, respectively. Compared to Table 2, the largest posterior model probabilities for the independence logit models are at least 0.89 times smaller than those for the multivariate logit model when moderate to high prior weights for the historical data are taken.

**Table 2: Posterior Model Probabilities For Prostate Cancer Data**

$E(a_0 D)$	Model	$p(m D^{(m)})$	$E(a_0 D)$	Model	$p(m D^{(m)})$
0.27	$[(x_1, x_2, x_3); (x_1, x_2)]$	0.278	0.91	$[(x_1, x_2, x_3); (x_1, x_2)]$	0.354
	$[(x_1, x_2, x_3); (x_1, x_2, x_4)]$	0.185		$[(x_1, x_2, x_3); (x_1, x_2, x_3)]$	0.126
	$[(x_1, x_2, x_3, x_4); (x_1, x_2, x_4)]$	0.130		$[(x_1, x_2, x_3); (x_1)]$	0.117
	$[(x_1, x_2, x_3); (x_1, x_2, x_3)]$	0.110		$[(x_1, x_2, x_3); (x_1, x_2, x_3)]$	0.083
	$[(x_1, x_2, x_3, x_4); (x_1, x_2)]$	0.103		$[(x_1, x_2, x_3); (x_1, x_2, x_4)]$	0.080
0.54	$[(x_1, x_2, x_3); (x_1, x_2)]$	0.388	1.0	$[(x_1, x_2, x_3); (x_1, x_2)]$	0.367
	$[(x_1, x_2, x_3); (x_1, x_2, x_3)]$	0.135		$[(x_1, x_2, x_3); (x_1, x_2, x_3)]$	0.127
	$[(x_1, x_2, x_3); (x_1, x_2, x_4)]$	0.125		$[(x_1, x_2, x_3); (x_1)]$	0.102
	$[(x_1, x_2, x_3, x_4); (x_1, x_2)]$	0.099		$[(x_1, x_2, x_3); (x_1, x_2, x_4)]$	0.079
	$[(x_1, x_2, x_3, x_4); (x_1, x_2, x_4)]$	0.069		$[(x_1, x_2, x_3, x_4); (x_1, x_2)]$	0.073

## 6 Discussion

In this paper, we have developed a Bayesian variable subset selection procedure for correlated binary responses for multivariate logit models. Our proposed methods are quite natural and useful when historical data are available. The proposed priors have some very attractive properties and are proper under some very general conditions. Our methodology can be easily extended to correlated ordinal or general polychotomous response data. We have also developed novel computational methods for sampling from the posterior distribution and for computing posterior model probabilities for variable subset selection. The expressions obtained for the posterior model probabilities

facilitate a very quick and efficient method of calculation. In addition, the algorithms developed for sampling from the posterior distribution are quite efficient and feasible even for large data sets with a large number of covariates. The examples presented in Section 5 demonstrate the feasibility and the power of our methods. The Bayesian approach proposed here for this class of models appears to have a clear advantage over frequentist based procedures or other Bayesian procedures. In Section 5, we also investigated the impact of correlation on the posterior model probabilities, and empirically found that a strong correlation helps in identifying the best subset model.

### Appendix: Proof of Theorem 3.1

To prove Theorem 3.1, we need the following lemma.

**Lemma A.1** *Let  $M_0$  be an  $n_0 \times k_m$  matrix. Assume that  $n_0 > k_m$ ,  $M_0$  is of full rank, and there exists a positive vector  $a$  such that*

$$a' M_0 = 0. \tag{A.1}$$

*Then there exists a constant  $K_0$  depending only on  $M_0$  such that*

$$\|\beta^{(m)}\| \leq K_0 \|u\| \tag{A.2}$$

*whenever*

$$M_0 \beta^{(m)} \leq u, \tag{A.3}$$

*where  $\|\cdot\|$  denotes the Euclidean norm.*

**Proof:** Let  $\mathcal{E} = \{\varepsilon = (\varepsilon_1, \dots, \varepsilon_{k_m})' \in R^{k_m} : \varepsilon_i = \pm 1\}$ . Since  $M_0$  is of full rank, for every  $\varepsilon \in \mathcal{E}$ , there is a  $b_\varepsilon \in R^{n_0}$  such that

$$b'_\varepsilon M_0 = \varepsilon'. \tag{A.4}$$

Let  $a = (a_1, \dots, a_{n_0})' \in R^{n_0}$  be the positive vector satisfying (A.1). Put

$$\delta = \frac{\min_{1 \leq i \leq n_0} (a_i)}{2 \max_{\varepsilon \in \mathcal{E}} \|b_\varepsilon\|}.$$

For  $\varepsilon = \varepsilon_{\beta^{(m)}} = \text{sign}(\beta^{(m)'}) = (\text{sign}(\beta_1^{(m)}), \dots, \text{sign}(\beta_{k_m}^{(m)}))'$ , we have  $\delta > 0$  and  $a + \delta b_\varepsilon > 0$ . Hence, it follows from (A.3) that

$$\begin{aligned} (a + \delta b_\varepsilon)' u &\geq (a + \delta b_\varepsilon)' M_0 \beta^{(m)} \\ &= a' M_0 \beta^{(m)} + \delta b'_\varepsilon M_0 \beta^{(m)} \\ &\geq \delta b'_\varepsilon M_0 \beta^{(m)} = \delta \text{sign}(\beta^{(m)'}) \beta^{(m)} \\ &\geq (\delta/k_m) \|\beta^{(m)}\|, \end{aligned}$$

as desired.  $\square$

**Proof of Theorem 3.1:** Let  $\lambda_{01}, \dots, \lambda_{0n_0}$  be independent random variables with the common asymptotic Kolmogorov probability density function  $\pi_K(\lambda_0)$ . Let  $\tilde{w}_{0i} = (\tilde{w}_{0i1}, \dots, \tilde{w}_{0iJ})'$  be independent random variables such that

$$\tilde{w}_{0i} | \lambda_{0i}, a_0 \sim N(0, (1/a_0)\kappa(\lambda_{0i})\Sigma),$$

that is, given  $\lambda_{0i}$ ,  $\tilde{w}_{0i}$  is normally distributed with mean zero and covariance matrix  $(1/a_0)\kappa(\lambda_{0i})\Sigma$ .

Put

$$A_{0i} = A_{0i1} \times A_{0i2} \times \dots \times A_{0iJ}.$$

Thus, we can rewrite  $\pi^*(\beta^{(m)}, \Sigma | a_0, D_0^{(m)})$  as

$$\begin{aligned} \pi^*(\beta^{(m)}, \Sigma | a_0, D_0^{(m)}) &= E1\{(\tilde{w}_{0i} + x_{0i}^{(m)}\beta^{(m)}) \in A_{0i}, 1 \leq i \leq n_0\} \\ &= E\left(1\{\tilde{w}_{0ij} + x_{0ij}^{(m)}\beta_j^{(m)} \in A_{0ij}, 1 \leq j \leq J, 1 \leq i \leq n_0\}\right), \end{aligned} \quad (\text{A.5})$$

where  $1\{B\}$  denotes the indicator function so that  $1\{B\} = 1$  if  $B$  is true and 0 otherwise. It is easy to see that

$$\begin{aligned} &\{\tilde{w}_{0ij} + x_{0ij}^{(m)}\beta_j^{(m)} \in A_{0ij}, 1 \leq j \leq J, 1 \leq i \leq n_0\} \\ &= \{\tilde{w}_{0ij} + x_{0ij}^{(m)}\beta_j < 0, y_{0ij} = 0, 1 \leq i \leq n_0\} \cap \{\tilde{w}_{0ij} + x_{0ij}^{(m)}\beta_j^{(m)} \geq 0, y_{0ij} = 1, 1 \leq i \leq n_0\} \\ &\subset \{z_{0ij}x_{0ij}^{(m)}\beta_j^{(m)} \leq -z_{0ij}\tilde{w}_{0ij}, 1 \leq j \leq J, 1 \leq i \leq n_0\} \\ &= \{X_0^*\beta^{(m)} \leq w^*\}, \end{aligned} \quad (\text{A.6})$$

where

$$w^* = \begin{pmatrix} w_1^* \\ \vdots \\ w_{n_0}^* \end{pmatrix}, \quad w_i^* = \begin{pmatrix} w_{i1}^* \\ \vdots \\ w_{iJ} \end{pmatrix}, \quad \text{and} \quad w_{ij}^* = -z_{0ij}\tilde{w}_{0ij}.$$

Therefore,

$$\pi^*(\beta^{(m)}, \Sigma | a_0, D_0^{(m)}) \leq E\left(1\{X_0^*\beta^{(m)} \leq w^*\}\right). \quad (\text{A.7})$$

Conditions (C1) and (C2) and Lemma A.1 yield

$$\begin{aligned} &\int_0^1 \int_V \int_{R^{k_m}} \pi^*(\beta^{(m)}, \Sigma | a_0, D_0^{(m)}) a_0^{\delta_0-1} (1-a_0)^{\lambda_0-1} d\beta^{(m)} dvec^*(\Sigma) da_0 \\ &= \int_0^1 \int_V \int_{R^{k_m}} E\left(1\{X_0^*\beta^{(m)} \leq w^*\}\right) a_0^{\delta_0-1} (1-a_0)^{\lambda_0-1} d\beta^{(m)} dvec^*(\Sigma) da_0 \end{aligned} \quad (\text{A.8})$$

$$\begin{aligned}
&\leq \int_0^1 \int_V \int_{R^{k_m}} E\left(1_{\{\|\beta^{(m)}\| \leq K_0 \|w^*\|\}}\right) a_0^{\delta_0-1} (1-a_0)^{\lambda_0-1} d\beta^{(m)} dvec^*(\Sigma) da_0 \\
&\leq K \int_0^1 \int_V E\left\{\left(\max_{1 \leq i \leq n_0} \|\tilde{w}_{0i}\|\right)^{k_m}\right\} a_0^{\delta_0-1} (1-a_0)^{\lambda_0-1} dvec^*(\Sigma) da_0 \\
&\leq K \int_0^1 \int_V \left[\sum_{i=1}^{n_0} \sum_{j=1}^J E\left(|\tilde{w}_{0ij}|^{k_m}\right)\right] a_0^{\delta_0-1} (1-a_0)^{\lambda_0-1} dvec^*(\Sigma) da_0 \\
&\leq K \int_0^1 \int_V \left[\sum_{j=1}^J E\left(|\kappa(\lambda_0)/a_0|^{k_m/2}\right)\right] a_0^{\delta_0-1} (1-a_0)^{\lambda_0-1} dvec^*(\Sigma) da_0 \\
&\leq K \left[\sum_{j=1}^J E\left(|\kappa(\lambda_0)|^{k_m/2}\right)\right] \int_0^1 a_0^{\delta_0 - \frac{k_m}{2} - 1} (1-a_0)^{\lambda_0-1} da_0 \\
&< \infty
\end{aligned}$$

by (C3), the existence of the  $k_m$ th moment for the asymptotic Kolmogorov distribution, and the fact that  $V$  is a finite region.  $\square$

## Acknowledgement

The authors wish thank the Guest Editor and the two referees for their helpful comments and suggestions, which have led to an improvement in this article. They also thank Dr. Anthony V. D'Amico of the Joint Center for Radiation Therapy at Harvard Medical School for providing the prostate cancer data sets. Dr. Chen's research was partially supported by NSF grant No. DMS-9702172, and NIH grants #CA 70101-01 and #CA 74015-01.

## References

- Albert, J.H. and Chib, S. (1993), Bayesian Analysis of Binary and Polychotomous Response Data, *Journal of the American Statistical Association*, 88, 669-679.
- Bedrick, E.J., Christensen, R., and Johnson, W. (1996), A New Perspective on Priors for Generalized Linear Models, *Journal of the American Statistical Association*, 91, 1450-1460.
- Chen, M.-H. (1994), Importance-weighted Marginal Bayesian Posterior Density Estimation, *Journal of the American Statistical Association*, 89, 818-824.
- Chen, M.-H. and Dey, D.K. (1998), Bayesian Modeling of Correlated Binary Responses via Scale Mixture of Multivariate Normal Link Functions. *Sankhyā, Series A*, 60, 322-343.

- Chen, M.-H., Ibrahim, J.G., and Yiannoutsos, C. (1999), Prior Elicitation, Variable Selection, and Bayesian Computation for Logistic Regression Models, *Journal of the Royal Statistical Society, Series B*, 61, 223-242.
- Chen, M.-H. and Shao, Q.-M. (1997), Performance Study of Marginal Posterior Density Estimation via Kullback-Leibler Divergence, *Test, A Journal of the Spanish Society of Statistics and O.R.*, 6, 321-350.
- Chib, S. and Greenberg, E. (1998), Bayesian Analysis of Multivariate Probit Models, *Biometrika*, 85, 347-361.
- Desjardin, A.M. (1997), Statistical Inference on Studies of Adenocarcinoma of the Prostate, *Unpublished M.S. Thesis*, Department of Mathematical Sciences, Worcester Polytechnic Institute, Worcester, MA.
- 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.
- Gelfand, A.E., Smith, A.F.M. and Lee, T.M. (1992), Bayesian Analysis of Constrained Parameter and Truncated Data Problems Using Gibbs Sampling, *Journal of the American Statistical Association*, 87, 523-532.
- Geman, S. and Geman, D. (1984), Stochastic Relaxation, Gibbs Distributions and the Bayesian Restoration of Images, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6, 721-741.
- George, E.I. and McCulloch, R.E. (1993), Variable Selection via Gibbs Sampling, *Journal of the American Statistical Association*, 88, 881-889.
- Geweke, J. (1991), Efficient Simulation from the Multivariate Normal and Student-t Distributions Subject to Linear Constraints, *Computing Science and Statistics: Proceedings of the Twenty-Third Symposium on the Interface*, 571-578.
- Gilks, W.R. and Wild, P. (1992), Adaptive Rejection Sampling for Gibbs Sampling, *Applied Statistics*, 41, 337-348.
- Ibrahim, J.G., and Laud, P.W. (1994), A Predictive Approach to the Analysis of Designed Experiments, *Journal of the American Statistical Association*, 89, 309-319.

- Ibrahim, J.G., Ryan, L.M., and Chen, M.-H. (1998), Use of Historical Controls to Adjust for Covariates in Trend Tests for Binary Data, *Journal of the American Statistical Association*, *93*, 1282-1293.
- Kuo, L. and Mallick, B.K. (1998), Variable Selection for Regression Models, *Sankhyā, Series B*, *60*, 65-81.
- Laud, P.W., and Ibrahim, J.G. (1995), Predictive Model Selection, *Journal of the Royal Statistical Society, Series B*, *57*, 247-262.
- Liang, K.-Y. and Zeger, S.L. (1986), Longitudinal Data Analysis Using Generalized Linear Models, *Biometrika*, *73*, 13-22.
- Marsaglia, G. and Olkin, I. (1984), Generating Correlation Matrices, *SIAM Journal on Scientific and Statistical Computations*, *5*, 470-475.
- Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H. and Teller, E. (1953), Equations of State Calculations by Fast Computing Machines, *Journal of Chemical Physics*, *21*, 1087-1092.
- Mitchell, T.J. and Beauchamp, J.J. (1988), Bayesian Variable Selection in Linear Regression (with discussion), *Journal of the American Statistical Association*, *83*, 1023-1036.
- Prentice, R.L. (1988), Correlated Binary Regression with Covariate Specific to Each Binary Observation, *Biometrics*, *44*, 1033-1048.
- Rousseeuw, P. and Molenberghs, G. (1994), The Shape of Correlation Matrices, *American Statistician*, *48*, 276-279.
- Schwarz, G. (1978), Estimating the Dimension of a Model, *The Annals of Statistics*, *6*, 461-464.
- Verdinelli, I. and Wasserman, L. (1995), Computing Bayes Factors Using a Generalization of the Savage-Dickey Density Ratio, *Journal of the American Statistical Association*, *90*, 614-618.
- Zeger, S.L. and Liang, K.-Y. (1986), Longitudinal Data Analysis for Discrete and Continuous Outcomes, *Biometrics*, *42*, 121-130.