Journal of the Korean Statistical Society (1998), 27: 4, pp 407-420

Variable Selection in Linear Random Effects Models for Normal Data

Hea-Jung Kim[1]

ABSTRACT

This paper is concerned with selecting covariates to be included in building linear random effects models designed to analyze clustered response normal data. It is based on a Bayesian approach, intended to propose and develop a procedure that uses probabilistic considerations for selecting promising subsets of covariates. The approach reformulates the linear random effects model in a hierarchical normal and point mass mixture model by introducing a set of latent variables that will be used to identify subset choices. The hierarchical model is flexible to easily accommodate sign constraints in the number of regression coefficients. Utilizing Gibbs sampler, the appropriate posterior probability of each subset of covariates is obtained. Thus, in this procedure, the most promising subset of covariates can be identified as that with highest posterior probability. The procedure is illustrated through a simulation study.

Keywords: Linear Random Effects Models; Clustered Response Normal Data; Variable Selection; Hierarchical Normal and Point Mass Mixture Model; Gibbs Sampler; High Posterior Probability Model

1. INTRODUCTION

For clustered response data, linear random effects models provide an efficient tool for analyzing cluster-specific intercept and/or covariates effects. Followings are some examples of the clustered response data. In longitudinal research, repeated observations on each subject are unlikely to be independent; In genetic epidemiology observations on members of one family will be correlated; In sample surveys, responses from members of the same village are likely to be correlated. This dependence must be taken into account to correctly assess the relationship of the clustered response variable with covariates.

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

Linear random effects models as considered by Laird and Ware (1982) and Lindstrom and Bates (1988) extend the classical linear model for Gaussian response variables and have been widely used to account for the dependence within clusters. The linear models are generally defined as

$$y_{ij} = x'_{ij}\beta + u'_{ij}b_i + \varepsilon_{ij}, \quad j = 1, \ldots, n_i, \quad i = 1, \ldots, I, \tag{1.1}$$

where y_{ij} is the response for the jth observation in cluster i; x_{ij} and u_{ij} are $p \times 1$ and $q_i \times 1$ design vectors, respectively. u_{ij} often being a subvector of x_{ij}. β is the vector of regression coefficients representing population-specific effects and b_i is a $q_i \times 1$ vector of cluster-specific random effects that varies independently from one cluster to another according to a normal distribution with mean 0 and unknown covariance matrix $\tau_i^{-1} I_{q_i}$; and the disturbances ε_{ij} being uncorrelated normal random variables with mean 0 and variance ρ^{-1}, and the sequences $\{b_i\}$ and $\{\varepsilon_{ij}\}$ are assumed to be mutually uncorrelated. The primary objective in analyzing the linear random effects models is to estimate the parameters β, ρ, $\tau_i^{-1} I_{q_i}$, and the random effects b_i. Several authors have investigated estimation techniques of the model. Detailed expositions of the estimation techniques based on frequentist approach may be found in Jones (1993). As an alternative, Bayesian estimation methods can be found in Broemeling (1985) and Hobert and Casella (1996).

In many situations where β is of scientific interest, the practitioner may wish to use fewer than the full set of available covariates in x_{ij} for the model (1.1). There are various reasons why this might be the case, the two most obvious being the cost of processing the variables or the desire to identify the most relevant ones for analyzing the clustered data. However, except for the procedure using the usual information criteria such as AIC, SBC and backward elimination methods for selecting random effects (cf. Morrell et al., 1997), a formal variable selection method for the covariates having fixed effect has not been seen yet.

The purpose of this paper is to develop a Bayesian variable selection procedure, which takes into account the difficulty in using the information criteria (complexity of calculating the maximum likelihood and overwhelming burden of comparing all possible 2^p models) and makes use of the Bayesian approach of Geweke (1996). Prior and posterior distributions, and the computational algorithm are outlined in the next section. Construction of prior distributions and several aspects of the posterior are illustrated in Section 3 through a numerical example. The last section summarizes and discusses some possible extensions of this work.

2. BAYESIAN VARIABLE SELECTION

This section considers a Bayes procedure for identifying promising subsets of p covariates in (1.1). This procedure entails the specification of a hierarchical Bayes mixture prior which uses the data to assign larger posterior probability to the more promising model. To avoid the overwhelming burden of calculating the posterior probabilities of all 2^p models, the Gibbs sampler is used to search for promising models rather than compute the entire posterior.

To recapitulate, consider the matrix form of the linear random effects models (1.1),

$$Y = X\beta + Ub + \varepsilon, \quad b \sim N_q(0, A), \text{ and } \varepsilon \sim N_n(0, \rho^{-1} I), \tag{2.1}$$

where Y is a $n \times 1$ vector of clustered observations, X an $n \times p$ known full-rank design matrix of n corresponding observations on p covariates; $n = \sum_{i=1}^{I} n_i$. $U = (u_1, \ldots, u_I)$ a $n \times q, q = \sum_1^I q_i$, design matrix with $u_i = (0, \ldots, 0, u_{i1}, \ldots, u_{in_i}, 0, \ldots, 0)'$, $b = (b_1', \ldots, b_I')'$ a $q \times 1$ random real parameter vector, and A is the $q \times q$ block diagonal matrix with i-th diagonal matrix $\tau_i^{-1} I_{q_i}$, $i = 1, \ldots, I$.

2.1. HIERARCHICAL MIXTURE MODEL

We begin by describing a hierarchical mixture model which forms the basis for the covariate selection method considered in this paper. The model is constructed with proper prior distributions for all parameters of interest.

First stage of the hierarchical model is used to describe the relationship between the observed dependent variable and the set of all potential covariates $X_1, \ldots, X_p$, namely

$$Y|\beta, b, \rho \sim N(X\beta + Ub, \rho^{-1} I_n), \tag{2.2}$$

where $X = (X_1, \ldots, X_p)$ is an $n \times p$ matrix. The variable selection problem arises when there is some unknown subset of the covariates where corresponding regression coefficients are zero in (2.1). The key feature of the hierarchical model is that each component of β is modeled as having coming from a mixture of a point mass at zero and a normal distribution, possibly truncated to an interval. This is done by introducing a set of latent variables based on the data augmentation idea of Tanner and Wang (1987).

Under the assumption that distributions for each of the coefficients β_k, $k = 1, \ldots, p$, the random effects vector b_i and the parameter ρ are priori independent,

the second stage models can be simply expressed via the introduction of a set of distinct latent variables $\{\alpha_k = 0 \text{ or } 1,\ k = 1, \ldots, p\}$ so that β_k's are independent and random samples from the mixtures represented by

$$\beta_k|\alpha_k \sim (1-\alpha_k)I_0 + \alpha_k TN_{\{a_k \le \beta_k \le c_k\}}(\delta_k, \sigma_k^2),\ k = 1, \ldots, p, \tag{2.3}$$

and the prior of distributions $b = (b_1', \ldots, b_I')'$, ρ given $\tau = (\tau_1, \ldots, \tau_I)$ are

$$P(b, \rho|\tau) = P(\rho)\prod_{i=1}^{I} P(b_i|\tau_i) \tag{2.4}$$

with

$$\rho \sim Gamma(\gamma, \theta^{-1}) \text{ and } b_i|\tau_i \overset{ind}{\sim} N(0, \tau_i^{-1} I_{q_i}),\ i = 1, \ldots, I,$$

where I_0 is a point mass at 0, σ_k, δ_k, a_k, c_k, γ and θ are hyperparameters to be assessed, and $TN_{\{a_k \le \beta_k \le c_k\}}(\delta_k, \sigma_k^2)$ denotes the normal distribution $N(\delta_k, \sigma_k^2)$ truncated to interval $a_k \le \beta_k \le c_k$. In case β_k is not truncated in priori, we may simply set $a_k = -\infty$ and $c_k = \infty$.

The third, and final, stage specifies beliefs about $\alpha_k's$ and $\tau_i's$. This can be done by a reasonable choices of prior densities for $\alpha = (\alpha_1, \ldots, \alpha_p)'$ and $\tau = (\tau_1, \ldots, \tau_I)'$;

$$p(\alpha) = \prod_{k=1}^{p} \phi_k^{\alpha_k}(1-\phi_k)^{(1-\alpha_k)}, \tag{2.5}$$

$$p(\tau) \propto \prod_{i=1}^{I} \tau_i^{\gamma_i - 1} \exp\{-\tau_i \theta_i\}, \tag{2.6}$$

where ϕ_k's, γ_i's and θ_i's are hyperparameters.

The prior distribution is therefore proper, informative and coherent but non-conjugate. We choose this form because it is relatively easy to elicit one's subjective prior distribution about the coefficients of the covariates in this form, yet the computational problem remains fairly simple. Moreover, the informative prior specifications do not incur the problem of the improper posterior explained by Hobert and Casella (1996). If one believes these priors are not flexible enough to express one's prior opinion about β, ρ, and τ, one may use mixtures of distributions which will allow more flexibility.

The posterior distribution may be expressed up to a constant by combining the densities defined from (2.2) through (2.6):

$$
\begin{aligned}
& P(\beta, b, \rho, \alpha, \tau | Y, X, U) \\
& \propto \prod_{k=1}^{p} p_0(\beta_k) \rho^{\frac{n+2\gamma}{2}-1} \exp\left\{-\frac{\rho}{2}[2\theta + (Y - X\beta - Ub)'(Y - X\beta - Ub)]\right\} \\
& \times \prod_{i=1}^{I} \tau_i^{\frac{q_i+2\gamma_i}{2}-1} \exp\left\{-\frac{\tau_i}{2}(2\theta_i + b_i' b_i)\right\} p(\alpha) p(\tau), \qquad (2.7)
\end{aligned}
$$

where $p_0(\beta_k)$ is the prior density for β_k defined in (2.3).

Our main reason for embedding the linear random effects model (1.1) in the above hierarchical mixture model is to obtain the marginal posterior distribution $h(\alpha|Y) \propto f(Y|\alpha)P(\alpha)$, which contains the information related to variable selection. However, it is easily seen that the problem of analytically calculating the marginal posterior from the posterior distribution is a challenging one. Fortunatelly, recent developments of a MCMC method, say the Gibbs sampler, provides a method that directly addresses simulation based calculation of the marginal posterior.

2.2. THE GIBBS SAMPLER

The computational procedure employed here is a Gibbs sampler. The conditional distributions involved in the algorithm are simple.

Given $\beta_\ell (\ell \neq k)$ and b and ρ, from (1), we can define

$$
W_{ij}(k) = y_{ij} - \sum_{\ell \neq k} \beta_\ell x_{ij}(\ell) - u_{ij}' b_i, \; j = 1, \ldots, n_i, \; i = 1, \ldots, I, \qquad (2.8)
$$

where $x_{ij} = (x_{ij}(1), \ldots, x_{ij}(p))'$. Then the conditional distribution of β_k follows from the usual posterior density for the regression parameter in the normal linear model

$$
W_{ij}(k) = \beta_k x_{ij}(k) + \varepsilon_{ij}, \quad \text{where } \varepsilon_{ij} \overset{iid}{\sim} N(0, \rho^{-1}), \; k = 1, \ldots, p,
$$

obtained from proper I_0 and $TN_{\{a_k \leq \beta_k \leq c_k\}}(\delta_k, \sigma_k^2)$ priors of β_k according to $\alpha_k = 0$ and $\alpha_k = 1$, respectively (see, equation (2.3) for the notations).

Specifically, for $\beta_k = 0$ (i.e. $\alpha_k = 0$), the conditional posterior density kernel is proportional to

$$
\exp\{-\rho/2 \sum_{i=1}^{I} \sum_{j=1}^{n_i} W_{ij}(k)^2\}. \qquad (2.9)
$$

Conditional on $\beta_k \neq 0$, from (2.9) and the prior (2.3), the corresponding kernel density for β_k is

$$(2\pi\sigma_k^2)^{-1/2}\left[\Phi\left\{\frac{c_k-\delta_k}{\sigma_k}\right\}-\Phi\left\{\frac{a_k-\delta_k}{\sigma_k}\right\}\right]^{-1} \tag{2.10}$$
$$\times \quad exp\left\{-\frac{\rho}{2}\sum_{i=1}^{I}\sum_{j=1}^{n_i}(W_{ij}(k)-\beta_k x_{ij}(k))^2-\frac{1}{2\sigma_k^2}(\beta_k-\delta_k)^2\right\} I_{\{a_k,c_k\}}(\beta_k).$$

If we set $\hat{\beta}_k$ as the least square estimator of β_k, then

$$\rho\sum_{i=1}^{I}\sum_{j=1}^{n_i}(W_{ij}(k)-\beta_k x_{ij}(k))^2=\rho\sum_{i=1}^{I}\sum_{j=1}^{n_i}(W_{ij}(k)-\hat{\beta}_k x_{ij}(k))^2+\frac{(\beta_k-\hat{\beta}_k)^2}{2\omega_k^2}, \tag{2.11}$$

where

$$\hat{\beta}_k=\sum_{i=1}^{I}\sum_{j=1}^{n_i}x_{ij}(k)W_{ij}(k)/\sum_{i=1}^{I}\sum_{j=1}^{n_i}x_{ij}(k)^2 \quad \omega_k^2=(\rho\sum_{i=1}^{I}\sum_{j=1}^{n_i}x_{ij}(k)^2)^{-1}.$$

Substituting (2.11) into (2.10) and completing the square on β_k, we have

$$(2\pi)^{-1/2}\sigma_k^{-1}\left[\Phi\left\{\frac{c_k-\delta_k}{\sigma_k}\right\}-\Phi\left\{\frac{a_k-\delta_k}{\sigma_k}\right\}\right]^{-1} exp\left\{-\frac{M(k)}{2}\right\} I_{\{a_k,c_k\}}(\beta_k), \tag{2.12}$$

where

$$M(k)=\rho\sum_{i=1}^{I}\sum_{j=1}^{n_i}(W_{ij}(k)-\hat{\beta}_k x_{ij}(k))^2+\frac{(\beta_k-\Delta_k)^2}{\sigma_*(k)^2}+\frac{\hat{\beta}_k^2}{\omega_k^2}+\frac{\delta_k^2}{\sigma_k^2}-\frac{\Delta_k^2}{\sigma_*(k)^2},$$

$$\sigma_*(k)^2=(1/\omega_k^2+1/\sigma_k^2)^{-1} \text{ and } \Delta_k=\sigma_*(k)^2(\hat{\beta}_k/\omega_k^2+\delta_k/\sigma_k^2).$$

This leads to the following full conditional distribution of β_k for $\beta_k \neq 0$ (i.e. for $\alpha_k = 1$)

$$\begin{aligned}
\beta_k|\beta_{(k)},\rho,b,\tau,\alpha_{(k)},\alpha_k=1 \quad &\sim \quad TN_{\{a_k\leq\beta_k\leq c_k\}}(\Delta_k,\sigma_*(k)^2), \\
&\quad \text{for } \beta_k \text{ is truncated to an interval } [a_k\leq\beta_k\leq c_k], \\
\beta_k|\beta_{(k)},\rho,b,\tau,\alpha_{(k)},\alpha_k=1 \quad &\sim \quad N(\Delta_k,\sigma_*(k)^2), \\
&\quad \text{for } \beta_k \text{ is not truncated, } k=1,\ldots,p,
\end{aligned} \tag{2.13}$$

where $\beta_{(k)} = (\beta_1, \ldots, \beta_{k-1}, \beta_{k+1}, \ldots, \beta_p)$ and $\alpha_{(k)} = (\alpha_1, \ldots, \alpha_{k-1}, \alpha_{k+1}, \ldots, \alpha_p)$.

Now express the term $(Y - X\beta - Ub)'(Y - X\beta - Ub)$ in the jont posterior (2.7) as

$$Y'RY - \hat{b}'U'RU\hat{b} - (b - \hat{b})'U'RU(b - \hat{b}) + (\beta - \hat{\beta})'X'X(\beta - \hat{\beta}),$$

where $\hat{b} = (U'RU)^{-}U'RY$, $R = I_n - X(X'X)^{-1}X'$, and $\hat{\beta} = (X'X)^{-1}X'(Y - Ub)$, where B^{-} is the unique Moore-Penrose generalized inverse of the matrix B (cf. Broemeling, 1985). Then we have following full conditional posterior distributions from the joint posterior.

The full conditional posterior distribution of ρ given β, α, b and τ is

$$\rho \sim Gamma\left(\frac{n + 2\gamma}{2}, \frac{2}{2\theta + (Y - X\beta - Ub)'(Y - X\beta - Ub)}\right). \tag{2.14}$$

The full conditional posterior distribution of b given β, α, ρ, and τ is

$$b \sim N\left([\rho U'U + A(\rho)]^{-1}\rho U'(Y - X\beta), [\rho U'U + A(\tau)]^{-1}\right), \tag{2.15}$$

where $A(\tau)$ is the $q \times q$ block diagonal matrix with i-th diagonal matrix $\tau_i I_{q_i}$, $i = 1, \ldots, I$.

From (2.7), we see that the conditional probability of $\alpha_k = 1(p_k)$ is propotional to $\phi_k p(a_k \leq \beta_k \leq c_k | \beta_{(k)}, \rho, b, \tau)$ and that of $\alpha_k = 0$ is proportional to $(1 - \phi_k)p(-\epsilon \leq \beta_k \leq \epsilon | \beta_{(k)}, \rho, b, \tau)$, $\epsilon \to 0$. Thus p_k can be expressed in terms of the Bayes factor for $H_0 : \beta_k = 0$ against $H_1 : a_k \leq \beta_k \leq c_k$ obtained from the full conditional distribution of β_k.

To calculate the conditional Bayes factor in favor of H_0, versus H_1, it is necessary to integrate (2.13) over β_k and then take ratio of (2.9) to the integration result.

The ratio gives the conditional Bayes factor:

$$B_{01} = \exp\left\{\Delta_k^2/2\sigma_*(k)^2 - \delta_k^2/2\sigma_k^2\right\}\left[\Phi\left\{\frac{c_k - \Delta_k}{\sigma_*(k)}\right\} - \Phi\left\{\frac{a_k - \Delta_k}{\sigma_*(k)}\right\}\right] \tag{2.16}$$

$$\times \frac{\sigma_*(k)}{\sigma_k}\left[\Phi\left\{\frac{c_k - \delta_k}{\sigma_k}\right\} - \Phi\left\{\frac{a_k - \delta_k}{\sigma_k}\right\}\right]^{-1},$$

and then the conditional posterior probaility that $\beta_k = 0$ is computed from the conditional Bayes factor (2.15):

$$p_k = \frac{\phi_k}{\phi_k + (1 - \phi_k)B_{01}}. \tag{2.17}$$

This gives the full conditional distribution of the additional variable α_k as

$$\alpha_k|\beta_{(k)}, \rho, b, \tau, \alpha_{(k)} \sim Be(p_k), \quad k = 1, \ldots, p, \tag{2.18}$$

where $\alpha_{(k)} = (\alpha_1, \ldots, \alpha_{k-1}, \alpha_{k+1}, \ldots, \alpha_p)$ and $Be(p_k)$ denotes a Bernoulli distribution with parameter p_k. Based on a comprison of this probability p_k with a drawing from the uniform distribution on [0,1], the choice $\alpha_k = 0$ or $\alpha_k = 1$ is made.

Finally, the full conditional distribution of $\tau = (\tau_1, \ldots, \tau_I)$ given β, α, b, and ρ is

$$\tau_i \overset{ind}{\sim} Gamma\left(\frac{q_i + 2\gamma_i}{2}, \frac{2}{2\theta_i + b_i' b_i}\right) \quad i = 1, \ldots, I. \tag{2.19}$$

2.3. ITERATION SCHEME AND SUBSET SELECTION

The Gibbs sampling algorithm proceeds in the usual way. After an initial value for $(\beta, \rho, b, \alpha, \tau)$ is drawn from the prior distribution, the parameters β, , ρ, b, α, τ are drawn in succession from their respective conditional posterior distributions. A possible complication could be the simulation of $\beta_k \neq 0$ from the truncated normal distribution. This can be easily conducted by the algorithm of Devroye (1986). By repeated successive Gibbs sampling from (2.13) through (2.19), we would get the Gibbs sequence

$$\beta^{(0)}, \rho^{(0)}, b^{(0)}, \alpha^{(0)}, \tau^{(0)}, \beta^{(1)}, \rho^{(1)}, b^{(1)}, \alpha^{(1)}, \tau^{(1)}, \ldots, \beta^{(t)}, \rho^{(t)}, b^{(t)}, \alpha^{(t)}, \tau^{(t)}$$

that is an ergodic Markov chain. Therefore, as t approach infinity, the joint distribution of $\alpha^{(t)}$ can be shown to approach the marginal posterior 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 suggested by Cowles and Carlin (1996).

Once we determined the value of t^*, as practiced by Geman and Geman (1984) and 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 of every Tth value in a single long run of length $N = mT + t^*$, where the number of t^* is the 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 $\alpha^{(T)}(\ell)$'s for the Gibbs sample. Once we gets the Gibbs sample, we can use it to compute the empirical distribution of the

α which converges to the actual marginal posterior $h(\alpha|Y)$ (cf. Tierney, 1994). 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 a simple tabulation of the high-frequency values of $\alpha = (\alpha_1, \ldots, \alpha_p)$ can be used to identify the corresponding subsets of covariates as potentially promising.

3. NUMERICAL EXAMPLE

The Gibbs sampler for a linear model with quite general crossed or nested random effects structure has been implemented in SAS language on an 586 PC. Uniform, Gaussian, and Gamma random numbers are generated using SAS/IML. We now report result of a simulation study to illustrate the methodology.

In the simulation, we considered the following analysis of covariance model with random effects b_{1i} and b_{2i}

$$y_{ij} = \beta_1 Z1_{ij} + \beta_2 Z2_{ij} + \beta_3 Z3_{ij} + \beta_4 Z4_{ij} + \beta_5 Z5_{ij} + b_{1i} X_i + b_{2i}(Z1_{ij} \cdot X_i) + \varepsilon_{ij}, \tag{3.1}$$

where $X_i = 0$ for $i = 1$ and 1 for $i = 2$ and $j = 1, \ldots, 21$. Thus each data set was comprised of $I = 2$ clusters of size $n_i = 21$ $(i = 1, 2)$. The fixed effects coefficients were set at following values:

Model 1: $\beta_1 = -2$, $\beta_2 = 1$, $\beta_3 = 2$, $\beta_4 = 0$, and $\beta_5 = 0$,

Model 2: $\beta_1 = -4$, $\beta_2 = 2$, $\beta_3 = 4$, $\beta_4 = 0$, and $\beta_5 = 0$.

So that the best subset choice for the covariates will be $\{Z1, Z2, Z3\}$ for both models. $n = 42$ observations on $p = 5$ potential covariates were simulated from independent $U(0, 1)$ and corresponding ε_{ij}'s were simulated from $N(0, 1)$. Two random effects distributions were simulated from $\text{var}(b_1) = I_2$ and $\text{var}(b_2) = .5I_2$.

For the Gibbs sampler constructed for each model with a given prior, convergence diagnostic checking were done by use of 8 parallel chains with starting points drawn from what we believe is a distribution overdispersed with respect to

the stationary distribution. The convergence checking advocated by Cowles and Carlin (1996) showed that, regardless of the random effects models, 1000 iterations of the Gibbs sampling algorithm seemed to achieve the convergence. That is Gelman and Rubin shrinkage factors and the eight parallel traces of -2ln(joint posterior), obtained from differing starting points, approach to 1 and converges to a true stationary distribution, respectively. For the brevity of this paper, we selected a prior case for Model 1 and graphical results of the convergence checking in Figure 3.1 and Figure 3.2. We considered the prior case as follows: The first hierarchy of the model (2.2) was constructed according to model 1. The hyperparameters for the second hierarchy of the models in (2.3) and (2.4) were set to $\delta_k = 0$, $\gamma = \theta = \tau_i = 1$ for $i = 1, 2$, and $a_3 = c_1 = 0$, $c_3 = c_2 = \infty$, $a_2 = a_1 = -\infty$, reflecting non-positive and non-negative constraints on the corresponding coefficients, i.e. $\beta_1 \leq 0$ and $0 \leq \beta_3$; using the idea of "substantially significant determinant" of Geweke (1996), the parameters σ_k's were assessed by $\sigma_k = \sigma_k^* = \Delta y_{ij}/\Delta Zk_{ij},$, $k = 1, \ldots, 5$, where Δy_{ij} is the size of the maximum feasible change in y_{ij} and ΔZk_{ij} is one-half of range of Zk_{ij} in the generated data set. Finally, for the third hierarchy of the models in (2.5) and (2.6), we set $\gamma_i = \theta_i = 1$, $i = 1, 2$, and $\phi_k = 1/2$, for $k = 1, \ldots, 5$, because we favored no particular α_k. The figures were obtained by use of CODA software version 0.30 (cf. Best, Cowles, and Vines 1996).

For each data set generated from the above models having the same prior (except for varying σ_k's and ϕ_k's)) as in the convergence checking, the Gibbs sampler was run to take Gibbs sample of size $m = 1000$. After discarding the initial 1000 iterations to allow for "burn-in", every 10th output from 1001 through 11001 iterations was collected to construct each Gibbs sample of size $m = 1000$ (autocorrelation function of Gibbs sequence showed that it cut off after lag 10). Then we obtained the empirical distribution of α based upon each Gibbs sample obtained. One hundred repetitions of the estimation procedure were made and the results are given in Table 3.1. Table 3.1 lists four high-probable random effects models of each size, and the mean and standard deviation of the posterior probability. Table 3.1 shows that, regardless of the particular prior distribution the models with covariates $Z1, Z2$ and $Z3$ have posterior at least .023 and often much more than the alternative models. In the simulation study, we notice that a couple of systematic effects of the prior distributions on the posterior probabilities of the alternative models are evident. First, increases in ϕ_k, the prior probability that $\beta_k = 0$, favor smaller models. Second, increases in σ_k also favor small models. Table 3.1 also notes these evidences. From (2.14) and (2.15), we see that, as all

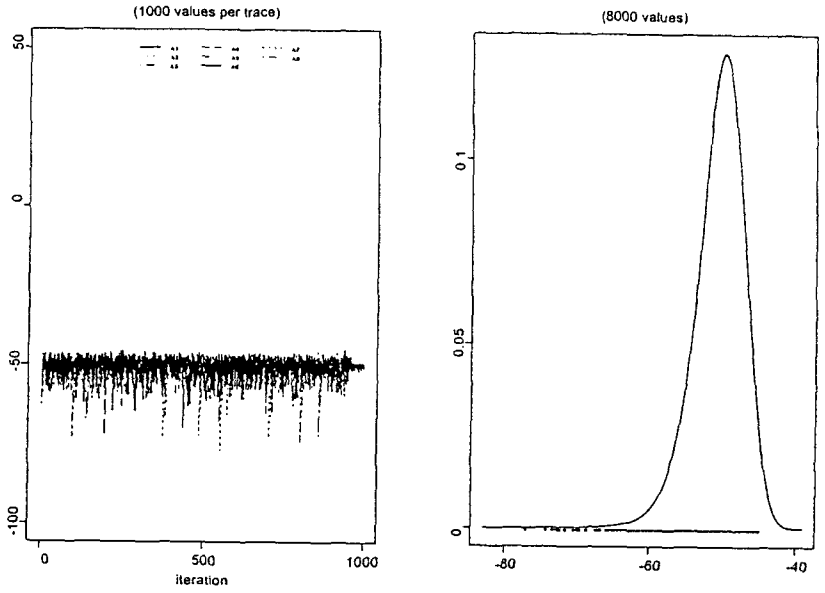

Figure 3.1: Traces and Normal Kernel Density Estimate of the Eight Parallel Chains of -2ln(joint posterior).

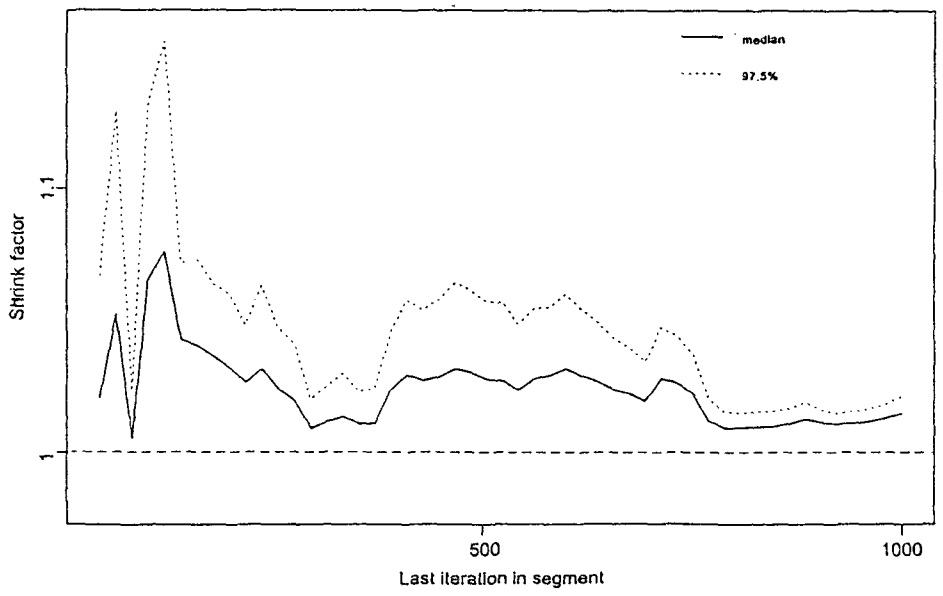

Figure 3.2: Gelman and Rubin Shrinkage Factors of the Eight Parallel Chains of -2ln(joint posterior).

Table 3.1: Mean and Standard Deviation (written in parenthesis) of Posterior Probabilities of Four High Frequency Subset of Covariates: "-" indicates unselective case

ϕ_k		0.2			0.5			0.8	
σ_k	$.5\sigma_k^*$	σ_k^*	$2\sigma_k^*$	$.5\sigma_k^*$	σ_k^*	$2\sigma_k^*$	$.5\sigma_k^*$	σ_k^*	$2\sigma_k^*$
				Model 1					
Covariates									
Z3	-	-	-	-	-	-	-	.190	.291
								(.012)	(.043)
Z1 Z3	-	-	-	-	.041	.110	.048	.065	.074
					(.009)	(.053)	(.008)	(.009)	(.012)
Z2 Z3	-	-	-	-	-	-	.172	.244	.226
							(.039)	(.054)	(.046)
Z1 Z2 Z3	.352	.663	.585	.620	.706	.732	.602	.474	.314
	(.074)	(.106)	(.097)	(.133)	(.138)	(.112)	(.141)	(.093)	(.061)
Z1 Z2 Z3 Z4	.275	.267	.197	.181	.132	.083	.034	-	-
	(.038)	(.041)	(.056)	(.034)	(.151)	(.007)	(.008)		
Z1 Z2 Z3 Z5	.201	.173	.119	.131	.065	.041	-	-	-
	(.041)	(.032)	(.018)	(.024)	(.011)	(.009)			
Z1 Z2 Z3 Z4 Z5	.164	.101	.045	.035	-	-	-	-	-
	(.034)	(.027)	(.007)	(.003)					
				Model 2					
Covariates									
Z3	-	-	-	-	-	-	-	.079	.213
								(.006)	(.012)
Z1 Z3	-	-	-	-	-	-	.027	.066	.077
							(.003)	(.008)	(.005)
Z2 Z3	-	-	.055	.056	.046	.314	.172	.344	.275
			(.013)	(.008)	(.051)	(.074)	(.008)	(.063)	(.071)
Z1 Z2 Z3	.516	.679	.768	.784	.831	.594	.701	.453	.349
	(.113)	(.095)	(.103)	(.132)	(.163)	(.103)	(.135)	(.097)	(.086)
Z1 Z2 Z3 Z4	.221	.153	.075	.063	.058	.015	.030	-	-
	(.043)	(.031)	(.014)	(.026)	(.009)	(.003)	(.004)		
Z1 Z2 Z3 Z5	.179	.124	.073	.071	.035	.016	-	-	-
	(.043)	(.021)	(.010)	(.016)	(.085)	(.003)			
Z1 Z2 Z3 Z4 Z5	.075	.029	-	-	-	-	-	-	-
	(.015)	(.004)							

$\sigma_k \to \infty$, all posterior probability becomes concentrated on the model with no covariates, consistent with Lindley's paradox (cf. Lindley, 1957).

4. CONCLUSION

We have proposed a family of nonconjugate priors for a Bayesian treatment of variable selection and model comparison in linear random effects models. We have focused on the clustered normal data case because it is the most common example and poses numerical difficulties (cf. Broemeling, 1985). The suggested method relies on the output of the Gibbs sampling algorithm. The algorithm is applied to reformulated linear random effects model setup constructed in a hierarchical truncated normal and point mass mixture model by introducing latent variables α_k that will be used to identify subset choices.

The illustrated example shows that the suggested method demonstrates good performance and is robust against the choice of hyperparameters. However, to avoid the subjective choice of hyperparameters, 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 hyperparameters will not be of closed forms. The Metropolis-Hastings algorithm may be used to construct a Markov chains for the hyperparameters.

The methodology has broad application to other statistical problems concerned with variable selection. Specific examples include variable selection in analysis of covariance model and generalized linear models with random effects. The study pertaining to applying the variable selection methodology to the other models is left as a future study of interest.

REFERENCES

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).

Broemeling, L. D. (1985). *Bayesian Analysis of Linear Models,* Marcel Dekker, Inc., New York.

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

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

Geman, S. and Geman, D. (1984). "Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images," *IEEE Transaction Pattern Analysis and Machine Intelligence,* **6**, 721-741.

Geweke, J. (1996). "Variable selection and model comparison in regression," *Bayesian Statistics 5*, Ed. by Bernardo, J. M., Berger, J. O., Dawid, A. P. and Smith, A. F. M., Oxford University Press, 609-620.

Hobert, J. P. and Casella, G. (1996). "The effect of improper priors on Gibbs sampling in hierarchical linear mixed models," *Journal of the American Statistical Association,* **91**, 1461-1473.

Jones, R. H. (1993). *Longitudinal Data with Serial Correlation: A State Space Approach,* Chapman and Hall, London.

Laird, N. M. and Ware, J. H. (1982). "Random effects models for longitudinal data," *Biometrics,* **38**, 963-974.

Lindley, D. V. (1957). "A statistical paradox," *Biometrika,* **44**, 187-192.

Lindstrom, M. and Bates, D. (1988). "Newton-Raphson and EM algorithms for linear mixed-effects models for repeated measure data," *Journal of the American Statistical Association,* **83**, 1014-1022.

Morrell, C. H., Pearson, J. D. and Brant, L. J. (1997). "Linear transformations of linear mixed-effects models," *The American Statistician,* **51**, 338-343.

Rao, C. and Kleffe, J. (1988). *Estimation of Variance Components and Applications,* North Holland, Amsterdam.

Tanner, M. and Wang, W. (1987). "The calculation of posterior distributions by data augmentation" (with discussion), *Journal of the American Statistical Association,* **82**, 528-550.

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