Statistica Sinica **10**(2000), 971-987

BAYESIAN VARIABLE SELECTION FOR TIME SERIES COUNT DATA

Joseph G. Ibrahim*†, Ming-Hui Chen# and Louise M. Ryan*†

Harvard School of Public Health, †Dana-Farber Cancer Institute and #Worcester Polytechnic Institute

Abstract: We consider a parametric model for time series of counts by constructing a likelihood-based generalization of a model considered by Zeger (1988). We consider a Bayesian approach and propose a class of informative prior distributions for the model parameters that are useful for variable subset selection. The prior specification is motivated from the notion of the existence of data from similar previous studies, called historical data, which is then quantified in a prior distribution for the current study. We derive theoretical and computational properties of the proposed priors and develop novel methods for computing posterior model probabilities. To compute the posterior model probabilities, we show that only posterior samples from the full model are needed to estimate the posterior probabilities for all of the possible subset models. We demonstrate our methodology with a simulated and a real data set.

Key words and phrases: Correlated counts, Gibbs sampling, hierarchical centering, historical data, Poisson regression, posterior distribution.

1. Introduction

Data from similar previous studies, called historical data, is often available in applied research settings where the investigator has access to previous studies measuring the same response and covariates as the current study. 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 into a prior distribution. There is some literature addressing this issue for the linear model and generalized linear models. See for example, Ibrahim, Ryan and Chen (1998), Chen, Ibrahim and Yiannoutsos (1999), and Bedrick, Christensen and Johnson (1996). In these papers, the authors assume a univariate independent response variable. The literature for informative prior elicitation for models with correlated responses is essentially nonexistent.

In this paper, we propose classes of informative prior distributions for time series count data. The prior specification is based on the notion of specifying an

$n_0 \times 1$ vector y_0 of prior predictions for the response vector, y, of the current study, along with a covariate matrix X_0 corresponding to y_0. Then (n_0, y_0, X_0) are used to specify an automated parametric informative prior for the regression coefficients. The quantity y_0 can be taken as the raw response vector from the historical data, a vector of fitted values based on the historical data, a vector obtained from a theoretical prediction model, or a vector specified from expert opinion or case-specific information. Thus y_0 can be viewed as a prior "prediction" for y, the actual data in the current study. Similarly, X_0 can be taken as the raw covariate matrix based on the historical data or it can be specified in other ways. In any case, taking (n_0, y_0, X_0) to be the raw historical data results in a more natural, interpretable, and automated specification. The Monte Carlo methods we propose will facilitate a very fast and efficient way of computing posterior model probabilities using only a *single* posterior sample from a *single* model, that being the full model. Such a procedure has proved to be quite feasible and powerful in the model selection context (see for example, Chen, Ibrahim and Yiannoutsos, (1999)). In addition, our proposed informative prior elicitation schemes allow us to incorporate historical data in a natural way.

2. The Method

2.1. The likelihood function

Let $\mathcal{M}$ denote the model space. We enumerate the models in $\mathcal{M}$ by $m = 1, \ldots, \mathcal{K}$, 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. Further, let I denote a model indicator, so that $I = m$ means that model m is selected. If k is the number of covariates for the full model, our model space contains 2^k models. Let $\beta^{(\mathcal{K})} = (\beta_0, \ldots, \beta_k)'$ denote the regression coefficients for the full model including an intercept, and let $\beta^{(m)}$ denote a $k_m \times 1$ vector of regression coefficients for model m with an intercept, and a specific choice of $k_m - 1$ covariates. We write $\beta^{(\mathcal{K})} = (\beta^{(m)'}, \beta^{(-m)'})'$, where $\beta^{(-m)}$ is $\beta^{(\mathcal{K})}$ with $\beta^{(m)}$ deleted.

Consider a time series of counts y_t, $t = 1, \ldots, n$, where each y_t has a corresponding $k_m \times 1$ covariate vector $x_t^{(m)}$ under model m. Under model m, conditional on $\beta^{(m)}$ and a stationary unobserved process ϵ_t, the y_t's are assumed to be independent discrete random variables from a distribution in the exponential family. This leads to the conditional density

$$p(y \mid \beta^{(m)}, \epsilon, I = m) = \prod_{t=1}^{n} p(y_t \mid \beta^{(m)}, \epsilon_t)$$

$$= \prod_{t=1}^{n} \exp\left\{\tau_t^{-1}\left[y_t \theta(x_t^{(m)}, \beta^{(m)}, \epsilon_t) - q(\theta(x_t^{(m)}, \beta^{(m)}, \epsilon_t))\right] - c(y_t)\right\}, \quad (2.1)$$

indexed by the canonical parameter $\theta_t \equiv \theta(x_t^{(m)}, \beta^{(m)}, \epsilon_t)$ and the scale parameter τ_t, where $y = (y_1, \ldots, y_n)'$, and $\epsilon = (\epsilon_1, \ldots, \epsilon_n)'$. Further suppose $\theta(x_t^{(m)}, \beta^{(m)}, \epsilon_t)$ satisfies the equation

$$\theta(x_t^{(m)}, \beta^{(m)}, \epsilon_t) = h((x_t^{(m)})'\beta^{(m)} + \epsilon_t), \quad t = 1, \ldots, n, \tag{2.2}$$

where h is a monotonic differentiable function, often referred to as the link function. In (2.1), the functions q and c determine a particular family in the class, such as the binomial or Poisson distributions. For example, if we take y_t to have a Poisson distribution with conditional mean $\lambda_t = \exp((x_t^{(m)'}\beta^{(m)} + \epsilon_t)$, then $\tau_t = 1$, $h((x_t^{(m)})'\beta^{(m)} + \epsilon_t) = (x_t^{(m)})'\beta^{(m)} + \epsilon_t$, $q(\theta(x_t^{(m)}, \beta^{(m)}, \epsilon_t)) = \exp\left\{(x_t^{(m)})'\beta^{(m)} + \epsilon_t\right\}$, and $c(y_t) = \ln(y_t!)$. We emphasize here that the likelihood in (2.1) is a general exponential family model for discrete outcomes, with the Poisson model being a special case. For ease of exposition, we assume $\tau_t = 1$ throughout, since this is in fact the case for many models in the exponential family, including the binomial and the Poisson. In addition, it will be convenient to write (2.1) in vector notation as

$$p(y \mid \beta^{(m)}, \epsilon, I = m) = \exp\left\{y'\theta(X^{(m)}, \beta^{(m)}, \epsilon) - J_n'Q(X^{(m)}, \beta^{(m)}, \epsilon) - J_n'C(y)\right\}, \tag{2.3}$$

where $X^{(m)}$ is the $n \times k_m$ matrix of covariates with tth row equal to $(x_t^{(m)})'$, J_n is an $n \times 1$ vector of ones, $\theta(X^{(m)}, \beta^{(m)}, \epsilon)$, $Q(X^{(m)}, \beta^{(m)}, \epsilon)$, and $C(y)$ are $n \times 1$ vectors with the tth components equal to $\theta_t = h((x_t^{(m)})'\beta^{(m)} + \epsilon_t)$, $q_t = q(\theta_t)$, and $c_t = c(y_t)$, respectively.

The latent process ϵ_t is assumed to have a normal distribution with mean 0. We assume an AR(1) structure for the covariance matrix of ϵ. This structure is well motivated in the statistical literature and is one of the most widely used in the time series setting (see Zeger (1988)). It proves to be quite appropriate for our purposes, as demonstrated in Section 4. Thus we assume that ϵ has a multivariate normal distribution with mean 0 and covariance matrix $\sigma^2\Sigma$, where the (i, j)th element of Σ has the form $\sigma_{ij} = \rho^{|i-j|}$, $-1 \leq \rho \leq 1$. The unobserved process ϵ_t is analogous to a "random effect" in a random effects model except for the correlation. We note that the mean and variance of ϵ_t do not depend on t. Zeger (1988) constructs a similar model for Poisson count data through the mean and covariance of the latent process, which then define the estimating equations. He does not specify a parametric distribution for the latent process as is done here, and only considers Poisson count models.

Let $\phi_n(\epsilon|\mu, \sigma^2\Sigma)$ denote the n-dimensional normal density of the latent process ϵ with mean μ and covariance matrix $\sigma^2\Sigma$, i.e.,

$$\phi_n(\epsilon|\mu, \sigma^2\Sigma) = (2\pi\sigma^2)^{-n/2}|\Sigma|^{-1/2}\exp\left\{-\frac{1}{2\sigma^2}(\epsilon - \mu)'\Sigma^{-1}(\epsilon - \mu)\right\}. \tag{2.4}$$

We note that $|\Sigma| = (1-\rho^2)^{n-1}$. Then the joint density of (y, ϵ) can be written as

$$p(y, \epsilon \mid \beta^{(m)}, \sigma^2, \rho, I = m)$$
$$= \exp\left\{y'\theta(X^{(m)}, \beta^{(m)}, \epsilon) - J_n' Q(X^{(m)}, \beta^{(m)}, \epsilon) - J_n' C(y)\right\} \times \phi_n(\epsilon|0, \sigma^2\Sigma). \quad (2.5)$$

To induce the correlation structure on y we integrate out ϵ from (2.5), leading to the "marginal" likelihood of $\beta^{(m)}$ given by

$$p(y \mid \beta^{(m)}, \sigma^2, \rho, I = m) = \int p(y, \epsilon \mid \beta^{(m)}, \sigma^2, \rho, I = m)\, d\epsilon, \quad (2.6)$$

where $p(y, \epsilon \mid \beta^{(m)}, \sigma^2, \rho, I = m)$ is given by (2.5). The marginal likelihood of $\beta^{(m)}$ in (2.6) does not have a closed form.

The implications of the process ϵ_t on the correlation structure in the y_t's and the regression model is as follows. Note first that $\epsilon_t^* = \exp(\epsilon_t)$ has a log-normal distribution with mean $\alpha = \exp(\frac{1}{2}\sigma^2)$ and variance $\nu^2 = \exp(2\sigma^2) - \exp(\sigma^2)$. The unobserved process ϵ_t allows for overdispersion and autocorrelation in y_t. In addition, the degree of overdispersion depends on the marginal mean of y_t. For the Poisson model, the autocorrelation in y_t must be less than or equal to that in ϵ_t, and the degree of autocorrelation in y_t relative to ϵ_t decreases as the marginal mean of y_t and ν^2 decrease.

2.2. The prior distributions

Informative prior elicitation is an important part of a Bayesian analysis. This is especially true for the problem of variable subset selection, since proper prior distributions are required for the computation of 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, as suggested in Section 1. Suppose there are N historical data sets and the sample size of the ith historical study is n_{0i}. Let y_{0i} denote the $n_{0i} \times 1$ vector of time series counts for the ith historical study, and let $X_{0i}^{(m)}$ denote the $n_{0i} \times k_m$ matrix of covariates corresponding to the ith historical study. In addition let ϵ_{0i} denote the latent process for the ith historical study, where ϵ_{0i} is an $n_{0i} \times 1$ vector, $i = 1, \ldots, N$, and ϵ_{0i} has an n_{0i} dimensional multivariate normal distribution with mean 0 and covariance matrix $\sigma^2\Sigma_{0i}$, where Σ_{0i} is an $n_{0i} \times n_{0i}$ matrix with $(j, j^*)th$ element equal to $\rho^{|j-j^*|}$. Finally let $y_0 = (y_{01}, \ldots, y_{0N})$ denote the response vector for all the historical studies. Throughout, we assume that given the model parameters, y_0 and y are independent.

We propose a prior distribution for $\beta^{(m)}$ taking the form

$$\begin{aligned}&\pi(\beta^{(m)} \mid \sigma^2, \rho, y_{0i}, a_{0i}, I = m)\\ &\propto \prod_{i=1}^{N} \int p(y_{0i} \mid \beta^{(m)}, \epsilon_{0i}, I = m)^{a_{0i}} \phi_{n_{0i}}(\epsilon_{0i}|0, \sigma^2\Sigma_{0i})\, d\epsilon_{0i},\end{aligned} \tag{2.7}$$

where a_{0i} is a scalar prior parameter that controls the weight of the ith historical study relative to the likelihood of the current study. Small values of a_{0i} give less weight whereas large values give more weight. It is most sensible to restrict a_{0i} to $0 \leq a_{0i} \leq 1$, since we do not want to weight the historical data more than the current data. The parameter a_{0i} can also be interpreted as a precision parameter which takes into account the between and within study variability in the historical data sets.

Let $a_0 = (a_{01}, \ldots, a_{0N})$. The prior specification is completed by specifying priors for (σ^2, ρ, a_0). We take these parameters to be independent a priori. We specify an inverse gamma prior for σ^2, denoted $IG(\delta_0, \gamma_0)$, a scaled beta prior for ρ, denoted scbeta(ν_0, ψ_0), and independent identically distributed beta priors for each a_{0i}, denoted beta(α_0, λ_0). Here, $(\delta_0, \gamma_0, \nu_0, \psi_0, \alpha_0, \lambda_0)$ are specified prior hyperparameters. Thus, we propose a joint prior distribution of the form

$$\pi(\beta^{(m)}, \sigma^2, \rho, a_0|y_0, I = m) \propto p_0^*(\beta^{(m)}, \sigma^2, \rho, a_0|y_0, I = m)\ , \tag{2.8}$$

where $p_0^*(\beta^{(m)}, \sigma^2, \rho, a_0|y_0, I = m)$ is the *unnormalized* prior density defined by

$$\begin{aligned}&p_0^*(\beta^{(m)}, \sigma^2, \rho, a_0|y_0, I = m)\\ &= \prod_{i=1}^{N} \left(\int p(y_{0i} \mid \beta^{(m)}, \epsilon_{0i}, I = m)^{a_{0i}} \phi_{n_{0i}}(\epsilon_{0i}|0, \sigma^2\Sigma_{0i}) d\epsilon_{0i} \right)\\ &\times \left(\prod_{i=1}^{N} a_{0i}^{\alpha_0 - 1}(1 - a_{0i})^{\lambda_0 - 1} \right) \times (\sigma^2)^{-(\delta_0+1)} \exp(-\sigma^{-2}\gamma_0)(1+\rho)^{\nu_0 - 1}(1-\rho)^{\psi_0 - 1}.\end{aligned} \tag{2.9}$$

Our joint prior for $(\beta^{(m)}, \sigma^2, \rho, a_0)$ clearly does not have a closed form in general. However, it has a natural motivation and several appealing interpretations. One motivation for the prior in (2.9) is that, by taking a_0 random, the tails of the marginal prior distribution for $\beta^{(m)}$ are heavier than those obtained by taking a_{0i} to be a fixed hyperparameter. In addition, a prior on a_0 provides flexibility and allows us to express our uncertainty about it. By allowing different a_{0i}'s for different historical studies, we are able to develop a more flexible prior that can weight each historical study differently. This would certainly be desirable if one historical study has a much larger sample size than others. Another motivation for (2.9) is that it mimics the marginal likelihood function of $\beta^{(m)}$ based on the historical data. If for example $a_{0i} = 1$, then (2.9) is precisely the marginal

likelihood function of $\beta^{(m)}$ based on the historical data. Thus, our prior can be viewed as a weighted marginal likelihood of $\beta^{(m)}$. This seems like a natural prior when historical data is available.

To show the propriety of the prior distribution given by (2.9), we first introduce a useful lemma.

Lemma 2.1. *Let $\alpha_0 > 0$, $\lambda_0 > 0$. There exists $K = K(\alpha_0, \lambda_0) > 0$ such that $\forall\ 0 \le \xi \le 1$,*

$$\int_0^1 \xi^{a_{0i}} a_{0i}^{\alpha_0 - 1}(1 - a_{0i})^{\lambda_0 - 1} da_{0i} \le K(1 + \ln(1/\xi))^{-\alpha_0}. \tag{2.10}$$

The proof of the lemma is given in the Appendix.

Let y_{0it} denote the tth component of y_{0i} and let $(x_{0it}^{(m)})'$ denote the tth row of $X_{0i}^{(m)}$. Using Lemma 2.1, we obtain the following result, which ensures the propriety of the joint prior distribution $\pi(\beta^{(m)}, \sigma^2, \rho, a_0 | y_0, I = m)$.

Theorem 2.1. *Assume that*

$$\exp\{(y_{0it}\theta_{0it} - q(\theta_{0it})) - c(y_{0it})\} \le M, \tag{2.11}$$

for $t = 1, \ldots, n_{0i}$, $i = 1, \ldots, N$, where M is some finite constant. Suppose there exist $y_{0it_{i1}}$, $y_{0it_{i2}}$, ..., $y_{0it_{ik_m}}$ $(1 \le t_{i1} \le t_{i2} \le \cdots \le t_{ik_m})$ such that

$$\int_{-\infty}^{\infty} e^{d_0|\eta|} \exp\left\{(y_{0it_j} h(\eta) - q(h(\eta))\right\} d\eta < \infty \tag{2.12}$$

for some $d_0 > 0$ and $j = 1, \ldots, k_m$, and that the corresponding design matrix $\left(x_{0it_1}^{(m)}, x_{0it_2}^{(m)}, \ldots, x_{0it_{k_m}}^{(m)}\right)'$ has full rank k_m. Then if $\alpha_0 > k_m/N$, $\lambda_0 > 0$, and (2.12) holds, the joint prior distribution $\pi(\beta^{(m)}, \sigma^2, \rho, a_0 | y_0, I = m)$ is proper.

The proof of Theorem 2.1 is given in the Appendix.

For elicitation purposes, it is easier to work with the prior mean and variance of a_{0i}, given by $\mu_{a_0} = \alpha_0/(\alpha_0 + \lambda_0)$ and $\sigma_{a_0}^2 = \mu_{a_0}(1 - \mu_{a_0})(\alpha_0 + \lambda_0 + 1)^{-1}$. From Theorem 2.1, a sufficient condition for the propriety of the prior distribution is that $\alpha_0 > (k+1)/N$ for the full model. Therefore, a reasonable starting point for the analysis is to choose $\alpha_0 = \lambda_0 = (k+2)/N$, which gives $\mu_{a_0} = 1/2$. Then we conduct several sensitivity analyses within a suitable range of the uniform prior, using various values of $(\mu_{a_0}, \sigma_{a_0})$. We do not recommend doing an analysis based on one set of proposed values of $(\mu_{a_0}, \sigma_{a_0})$.

2.3. Prior distribution on the model space

Let

$$p_0^*(\beta^{(m)} | y_0, I = m) = \int p_0^*(\beta^{(m)}, \sigma^2, \rho, a_0 | y_0, I = m)\ d\sigma^2\ d\rho\ da_0, \tag{2.13}$$

where $p_0^*(\beta^{(m)}, \sigma^2, \rho, a_0|y_0, I = m)$ is given by (2.9). We see that $p_0^*(\beta^{(m)}|y_0, I = m)$ is proportional to the marginal prior of $\beta^{(m)}$. We propose to take the prior probability of model m, denoted $p(I = m)$, as

$$p(I = m) = \frac{\int p_0^*(\beta^{(m)}|y_0, I = m)\ d\beta^{(m)}}{\sum_{j=1}^{\mathcal{K}} \int p_0^*(\beta^{(j)}|y_{0j}, I = j)\ d\beta^{(j)}}\ . \tag{2.14}$$

The choice of $p(I = m)$ in (2.14) is a natural one since the numerator is just the normalizing constant of the joint prior of $(\beta^{(m)}, a_0, \sigma^2, \rho)$ under model m. The prior model probabilities in (2.14) are based on coherent Bayesian updating and this fact has several attractive interpretations. First $p(I = m)$ in (2.14) corresponds to the posterior probability of model m based on the data y_0 under model m, using a uniform prior for the previous study, i.e., $p_0(I = m) = 2^{-k}$ for $m \in \mathcal{M}$ as $\alpha_0 \to \infty$. We also note that as $\alpha_0 \to \infty$, $a_{0i} \to 1$ with probability 1. Second as $\lambda_0 \to \infty$, $p(I = m)$ reduces to a uniform prior on the model space. Therefore, as $\lambda_0 \to \infty$, the historical data y_0 have a minimal impact in determining $p(m)$. In addition, as $\lambda_0 \to \infty$, $a_{0i} \to 0$ with probability 1. Finally, we mention that the choice of $p(I = m)$ in (2.14) greatly eases the computational burden for calculating posterior model probabilities. There is more detailed discussion in Section 3.

3. Posterior Model Probabilities

In this section we explore the theoretical properties of posterior model probabilities based on the choice of the prior model probabilities of (2.14), and then propose novel Monte Carlo implementation procedures to compute posterior model probabilities. A key result is a formula for the posterior model probability that does not depend directly on $p(I = m)$.

The posterior probability of model m is given by

$$p(I = m|y) = \frac{p(y|I = m)\ p(I = m)}{\sum_{j=1}^{\mathcal{K}} p(y|I = j)\ p(I = j)}\ , \tag{3.1}$$

where $p(y|I = m)$ denotes the marginal distribution of the data under model m for the current study, and $p(I = m)$ denotes the prior probability of model m in (2.14).

Theorem 3.1. *$p(I = m|y)$ in (3.1) is given by*

$$p(I = m|y) = \frac{p(\beta^{(-m)} = 0|y, y_0, I = \mathcal{K})}{\sum_{j=1}^{\mathcal{K}} p(\beta^{(-j)} = 0|y, y_0, I = \mathcal{K})}\ , \tag{3.2}$$

$m = 1, \ldots, \mathcal{K}$, where $p(\beta^{(-m)} = 0|y, y_0, I = \mathcal{K})$ is the marginal posterior density of $\beta^{(-m)}$ evaluated at $\beta^{(-m)} = 0$.

In (3.2), for notational convenience, we assume $p(\beta^{(-\mathcal{K})} = 0|y, y_0, I = \mathcal{K}) = 1$. The proof of Theorem 3.1 is given in the Appendix. We mention here that the derivation of (3.2) assumes that y_0 and y are independent given the model parameters. The result in (3.2) is very attractive since it shows that the posterior model probability $p(I = m|y)$ is simply a function of the marginal posterior density functions of $\beta^{(-m)}$ for the full model evaluated at $\beta^{(-m)} = 0$. This is an important feature since it allows us to compute the posterior model probabilities directly *without* numerically computing the prior model probabilities. 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 transformation is used quite often in practice to numerically stabilize the Gibbs sampler and adaptive rejection algorithms.

Due to the complexity of our model, the analytical evaluation of $p(\beta^{(-m)} = 0|y, y_0, I = \mathcal{K})$ does not appear possible. Therefore, we propose a novel Monte Carlo method to compute posterior model probabilities using a single MCMC sample from the full model. The hierarchical centering reparameterization technique of Gelfand, Sahu and Carlin (1996) is particularly suitable for the implementation of MCMC sampling for our problem. This is due to the fact that (2.9) leads to (3.2). From (3.2), it is easy to see that the posterior model probability is proportional to the marginal posterior density evaluated at 0. Thus, computing the posterior model probability is essentially equivalent to estimating the marginal posterior density. It is well known that the hierarchical centering technique is very useful in developing an efficient Monte Carlo method for estimating the marginal posterior density $p(\beta^{(-m)} = 0|y, y_0, I = \mathcal{K})$ and this leads to the efficient computation of (3.2). To the best of our knowledge, this is the first time that the hierarchical centering reparameterization technique has been used to ease the computational burden in Bayesian variable selection.

To this end, consider the following reparameterization:

$$\eta = \epsilon + X^{(\mathcal{K})}\beta^{(\mathcal{K})} \tag{3.3}$$

and

$$\eta_{0i} = \epsilon_{0i} + X_{0i}^{(\mathcal{K})}\beta^{(\mathcal{K})} \ , \tag{3.4}$$

for $i = 1, \ldots, N$. Let $\eta_0 = (\eta_{01}, \ldots, \eta_{0N})$. Then we write the reparameterized posterior for the full model, $p(\beta^{(\mathcal{K})}, \sigma^2, \rho, a_0, \eta, \eta_0|y, y_0, I = \mathcal{K})$. Using the hierarchical centering technique, we obtain an MCMC sample $\{(\beta_{(l)}^{(\mathcal{K})}, \sigma_{(l)}^2, \rho_{(l)}, a_{0(l)}, \eta_{(l)}, \eta_{0(l)}), \quad l = 1, \ldots, L\}$ from this reparameterized posterior. Then, following the

lines of Chen (1994), $p(\beta^{(-m)} = 0|y, y_0, I = \mathcal{K})$ can be estimated by the conditional marginal density estimation (CMDE) method. 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 $p(\beta^{(-m)} = 0|y, y_0, I = \mathcal{K})$ is given by

$$\hat{p}(\beta^{(-m)} = 0|y, y_0, I = \mathcal{K}) = \frac{1}{L}\sum_{l=1}^{L} N_{k+1-k_m}(\beta^{(-m)} = 0|\beta_{(l)}^{(-m)}, \hat{\beta}_{(l)}, (B_{(l)})^{-1}), \tag{3.5}$$

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

$$B_{(l)} = \frac{1}{\sigma_{(l)}^2}\left((X^{(\mathcal{K})})'\Sigma_{(l)}^{-1}X^{(\mathcal{K})} + \sum_{i=1}^{N}(X_{0i}^{(\mathcal{K})})'\Sigma_{0i(l)}^{-1}X_{0i}^{(\mathcal{K})}\right),$$

$$\hat{\beta}^{(\mathcal{K})} = \frac{1}{\sigma_{(l)}^2}\left\{B_{(l)}^{-1}\left((X^{(\mathcal{K})})'\Sigma_{(l)}^{-1}\eta_{(l)} + \sum_{i=1}^{N}(X_{0i}^{(\mathcal{K})})'\Sigma_{0i(l)}^{-1}\eta_{0i(l)}\right)\right\},$$

$\Sigma_{(l)}$ is an $n \times n$ matrix with $(j, j^*)th$ element equal to $\rho_{(l)}^{|j-j^*|}$, and $\Sigma_{0i(l)}$ is an $n_{0i} \times n_{0i}$ matrix with $(j, j^*)th$ element equal to $\rho_{(l)}^{|j-j^*|}$.

There are several advantages of the above Monte Carlo procedure. First, as previously mentioned, it is *not* required to compute $p(I = m)$ for each model. Second, we need only one random draw from $p(\beta^{(\mathcal{K})}, \sigma^2, \rho, a_0, \eta, \eta_0|y, y_0, I = \mathcal{K})$. Third, after we obtain an MCMC sample from the posterior distribution of the full model, calculating $\hat{p}(\beta^{(-m)} = 0|y, y_0, I = \mathcal{K})$ given by (3.5) is straightforward and almost free of computational time. Fourth, for the purposes of computing posterior model probabilities, one need only store a $(k+1)$-dimensional vector $\hat{\beta}_{(l)}$ and a $(k+1) \times (k+1)$ matrix $B_{(l)}$ for each MCMC sampling iteration. This becomes even more advantageous for cases where multiple previous studies are available and each n_{0i} is large. The above features of our Monte Carlo procedure make Bayesian variable selection feasible in the presence of a large number of covariates (say, $k > 20$).

We note here that Bayesian inference for non-Gaussian time series models has been considered by Shephard and Pitt (1997), but their model and computational development is quite different from what we do here. Similarly, there are several differences between our methodology and that of Chen, Ibrahim and Yiannoutsos (1999).

4. Examples

Example 1. Simulation study

In this example, we demonstrate variable subset selection with our proposed methodology using a simulated data set. We also demonstrate the computational feasibility of our methods.

The data for the current study is generated as follows. We generate $n = 50$ independent observations from a Poisson distribution each with mean $\lambda_t = \exp(\beta_0 + \beta_1 x_{t1} + \beta_2 x_{t2} + \epsilon_t)$, $t = 1, \ldots, 50$, and $\beta = (\beta_0, \beta_1, \beta_2) = (1, 1, -1)$. Also, $(x_{t1}, x_{t2})'$ are generated as i.i.d. bivariate normal random vectors with mean $(0.5, 0.5)'$ and covariance matrix $\begin{pmatrix} 0.5 & 0.25 \\ 0.25 & 0.5 \end{pmatrix}$. We take $\epsilon = (\epsilon_1, \ldots, \epsilon_{50})'$ to have a 50-dimensional multivariate normal distribution with mean 0 and covariance matrix $\sigma^2\Sigma$, where $\sigma^2 = 1$ and the (i, j)th element of Σ is $(0.5)^{|i-j|}$. In addition, we generate two other covariates (x_{t3}, x_{t4}) which are *i.i.d.* normal random variables each with mean 0 and variance .5, and independent of (x_{t1}, x_{t2}). Thus, the true model contains the two covariates (x_1, x_2), and the "full" model contains the four covariates $(x_1, \ldots, x_4)$. Therefore, our model space $\mathcal{M}$ contains 16 models, with an intercept included in each. The historical data were generated in a similar fashion. We take $n_0 = 50$, and $\sigma^2 = 1.2$, with all other parameters the same as for the current data. In addition, $(x_{0t1}, \ldots, x_{0t4})$ are generated in exactly the same way as the current data.

Table 1 shows posterior probabilities for the top model under various choices of μ_{a_0}, σ_{a_0} and N. The posterior model probability is denoted by $p(I = m \mid y)$ in Table 1. From Table 1, under each choice of $(\mu_{a_0}, \sigma_{a_0})$ and N, the true model (x_1, x_2) obtains the largest posterior probability. Although not shown in Table 1, the (x_1, x_2, x_3) model is consistently chosen as the second best model under the choices of $(\mu_{a_0}, \sigma_{a_0})$ and N given in Table 1. In addition, the order of the models with respect to their posterior probabilities is preserved as $(\mu_{a_0}, \sigma_{a_0})$ are varied according to Table 1. We do not see a dramatic change in the posterior model probabilities as $(\mu_{a_0}, \sigma_{a_0})$ are varied for a given N. It is apparent that the posterior model probabilities are quite robust with respect to changes in $(\mu_{a_0}, \sigma_{a_0})$. In addition, we observe a monotonic increase in the posterior model probabilities as we assign more weight to the historical data, i.e., as μ_{a_0} increases and/or σ_{a_0} decreases. Specifically, for $\mu_{a_0} = 0.5$ a monotonic increase in the posterior model probability is observed as σ_{a_0} is decreased. This is a desirable feature since it shows that a heavier weight given to the historical data results in an increase in the posterior probability of the true model. Also, a low weight for the historical data with $(\mu_{a_0}, \sigma_{a_0}) = (0.09, 0.04)$ still yields (x_1, x_2) as the top model with posterior probability of 0.349. In addition, more extreme values

of $(\mu_{a_0}, \sigma_{a_0})$ were used. For example, using an extremely low weight for the historical data with $(\mu_{a_0}, \sigma_{a_0}) = (0.05, 0.02)$, the true model, (x_1, x_2), obtains the largest posterior probability of 0.347 and the (x_1, x_2, x_3) model obtains the second largest posterior probability. We note that $(\mu_{a_0}, \sigma_{a_0}) = (0.05, 0.02)$ represents the smallest weight we can place on the historical data and still obtain proper priors (see Theorem 2.1). Similar results are obtained for other small values of μ_{a_0} and moderate to large values of σ_{a_0}. We note here that we cannot do Bayesian variable selection using $a_{0i} = 0$ with probability 1, since this would result in an improper prior for β and, in this case, the posterior model probabilities would not be well defined. When historical data is not available, one can specify a normal prior for β, and the specification of the hyperparameters would be based on elicitation from expert opinion or case-specific information.

Table 1 also indicates that a monotonic increase in the posterior probability of the true model occurs as N is increased. This is a solid feature of our methodology since it shows that increasing the number of historical studies provides more precise estimates of the posterior model probabilities. Similar results were obtained for other combinations for $(\mu_{a_0}, \sigma_{a_0})$. Finally, an analysis using $\rho = 0$ and $(\mu_{a_0}, \sigma_{a_0}) = (0.5, 0.06)$ was conducted, and results very similar to those of Table 1 were obtained. Specifically, the model obtaining the largest posterior probability is the (x_1, x_2) model with value 0.414, which is similar to the one using $\rho = 0.5$. In addition, the model with the second largest posterior probability is (x_1, x_2, x_4) with value 0.243, and the model with the third largest posterior probability is (x_1, x_2, x_3) with value 0.215.

Example 2. Pollen data

Ragweed pollen data were collected daily in Kalamazoo, Michigan from 1991 to 1994. Frequentist analyses of these data using standard Poisson regression methods have been conducted by Stark, Ryan, McDonald and Burge (1997). Our aim here is to demonstrate our Bayesian methodology for variable selection. The response variable y is the pollen count for a particular day in the season for a given year. Initially, we take the 1991, 1992, and 1993 data ($N = 3$) as the historical data and the 1994 data as the current data. The data for each year was collected roughly over a 3 month interval between the months of July and October. However, for each year, the first and last observations were collected on different days. For example, in 1991, the first observation was collected on July 28 and the last was collected on October 27th. In 1992, the first observation was collected on August 6th and the last observation on October 26th.

The full model contains an intercept and seven covariates, extensively discussed and motivated by Stark et al. (1997). These are x_1 = rain, (which is a binary variable taking the value 0 if there were at least three hours of steady rain,

and 1 otherwise), x_2 = day in the pollen season , $x_3 = \log(\text{day})$. Two covariates are functions of temperature. These are x_4 which is the lowess smoothed function of temperature constructed from a non-parametric estimate of the regression of pollen count on average temperature, and x_5, which denotes the deviation from the daily averages temperature to the lowess line. The final two covariates are x_6 = windspeed and x_7 = cold (a binary variable taking the value 0 if the overnight temperature dropped below 50 degrees Fahrenheit, and 1 otherwise).

The model space $\mathcal{M}$ contains 2^7 models. We specify noninformative priors for ρ and σ^2. Specifically, we take a uniform prior for ρ on $[-1, 1]$ (i.e. $\nu_0 = \psi_0 = 1$) and take $\sigma^2 \sim IG(.005, .005)$. Table 2 give results for the model with the largest posterior probability based on several values of $(\mu_{a_0}, \sigma_{a_0})$. The top model in each case is $(x_1, x_2, x_3, x_4, x_5)$, for all combinations of $(\mu_{a_0}, \sigma_{a_0})$ (and N). In addition, we see that the posterior model probabilities increase monotonically as more weight is given to the historical data. For example, using $N = 3$, when we put very small weight on the historical data, such as $(\mu_{a_0}, \sigma_{a_0}) = (0.009, 0.003)$, the $(x_1, x_2, x_3, x_4, x_5)$ model still obtains the largest posterior probability, with value 0.117. When we put extremely small weight on the historical data such as $(\mu_{a_0}, \sigma_{a_0}) = (0.0009, 0.0003)$, the $(x_2, x_3, x_4, x_5, x_7)$ model obtains the largest posterior probability, with value 0.122 and the $(x_1, x_2, x_3, x_4, x_5)$ model obtains the fourth largest posterior probability with value 0.101. When we put a small weight on the historical data using a moderate variance, $(\mu_{a_0}, \sigma_{a_0}) = (0.09, 0.027)$, the $(x_1, x_2, x_3, x_4, x_5)$ model obtains the largest posterior probability with value 0.130, and the $(x_2, x_3, x_4, x_5, x_7)$ model obtains the second largest posterior probability, with value 0.127. Thus we see that the model choice is reasonably robust to the choice of $(\mu_{a_0}, \sigma_{a_0})$, consistently yielding the $(x_1, x_2, x_3, x_4, x_5)$ model as the top model for a suitable range of $(\mu_{a_0}, \sigma_{a_0})$. Based on these analyses, it does not appear that the variables x_6 (windspeed) and x_7 (coldness of temperature) are important predictors of pollen counts.

An analysis was also conducted using $\rho = 0$ and $(\mu_{a_0}, \sigma_{a_0}) = (0.5, 0.05)$. In this case, the model that obtains the largest posterior probability is $(x_1, x_2, x_3, x_5, x_6)$ with value 0.272, and the model that obtains the second largest posterior probability is $(x_1, x_2, x_3, x_4, x_5, x_6)$ with value 0.267. These results are a bit different from those of Table 2. This can be partially explained by the fact that, for $(\mu_{a_0}, \sigma_{a_0}) = (0.5, 0.05)$, the posterior mean of ρ equals 0.87, implying a strong degree of correlation between the time measurements. Thus, posterior model probabilities can be sensitive to the choice of ρ if there is a high correlation in the data.

We also did a sensitivity analysis on the choice of N. We computed the posterior model probabilities for $N = 1, 2$ and these are shown in Table 2. The top model for $N = 1$ and $N = 2$ is $(x_1, x_2, x_3, x_4, x_5)$ for all combinations of

$(\mu_{a_0}, \sigma_{a_0})$. Here, $N = 1$ corresponds to using the 1993 data as historical data, and $N = 2$ corresponds to using the 1992 and 1993 data as historical data. We see the same behavior as in Example 1. For a given $(\mu_{a_0}, \sigma_{a_0})$, there is a monotonic increase in the posterior model probability as N is increased.

In this example, as well as in Example 1, 50,000 Gibbs iterations were used in all of the computations after a burn-in of 1,000 iterations. Convergence was checked using the methods discussed in Cowles and Carlin (1996). Specifically, trace plots, autocorrelations, and psr's were computed, and convergence was observed to occur before 500 iterations.

5. Discussion

The examples presented in Section 4 had posterior model probabilities quite robust to various choices of $(\mu_{a_0}, \sigma_{a_0})$, including choices that give high or low weight to the historical data. In Example 1, posterior model probabilities were not sensitive to the choice of ρ, $\rho = 0$ and $\rho = 0.5$ gave nearly identical results. This suggests that when there is low to moderate correlation in the data, the posterior model probabilities are not sensitive to the choice of ρ. However, in Example 2, the results for $\rho = 0$ were different from those based on a posterior mean of ρ equal to 0.87. This suggests that when there is high correlation in the data, the posterior model probabilities can be sensitive to the choice of ρ.

Table 1. Posterior model probabilities for simulated data.

$(\mu_{a_0}, \sigma_{a_0})$	$N = 1$	$N = 2$	$N = 3$
(0.5, 0.15)	0.356	0.357	0.359
(0.5, 0.11)	0.381	0.383	0.403
(0.5, 0.08)	0.411	0.427	0.467
(0.5, 0.06)	0.422	0.436	0.502
(0.98, 0.02)	0.443	0.474	0.571

Table 2. Posterior model probabilities for pollen data.

$(\mu_{a_0}, \sigma_{a_0})$	$N = 1$	$N = 2$	$N = 3$
(0.5, 0.11)	0.116	0.121	0.142
(0.5, 0.08)	0.154	0.206	0.290
(0.5, 0.06)	0.211	0.261	0.385
(0.5, 0.05)	0.262	0.279	0.420
(0.98, 0.02)	0.274	0.279	0.421

Acknowledgements

The authors wish to thank the Editor, Associate Editor and Referees for several suggestions which have greatly improved the paper. We would like to

thank Dr. Harriet Burge for providing us with the data in Example 2, whose work was partially supported by the Center Grant Es0002 from the NIEHS. Dr. Ibrahim's research was supported by NIH grants CA 70101 and CA 74015, and Dr. Chen's research was supported by NSF grant DMS-9702172 and NIH grant CA 74015.

Appendix

Proof of Lemma 2.1. When $1/2 \le \xi \le 1$, (2.10) is obviously true. It is easy to see that, for any $0 < \xi < 1/2$,

$$\begin{aligned}
&\int_0^1 \xi^{a_{0i}} a_{0i}^{\alpha_0 - 1}(1 - a_{0i})^{\lambda_0 - 1} da_{0i} \\
&= \int_0^{1/2} \xi^{a_{0i}} a_{0i}^{\alpha_0 - 1}(1 - a_{0i})^{\lambda_0 - 1} da_{0i} + \int_{1/2}^1 \xi^{a_{0i}} a_{0i}^{\alpha_0 - 1}(1 - a_{0i})^{\lambda_0 - 1} da_{0i} \\
&\le K^* \Big[\int_0^{1/2} \exp(-a_{0i} \ln(1/\xi)) a_{0i}^{\alpha_0 - 1} da_{0i} + \xi^{1/2} \Big] \\
&= K^* \Big[(\ln(1/\xi))^{-\alpha_0} \int_0^{\ln(1/\xi)/2} \exp(-a_{0i}) a_{0i}^{\alpha_0 - 1} da_{0i} + \xi^{1/2} \Big] \\
&\le K^* K^{**} (\ln(1/\xi))^{-\alpha_0} \le K(1 + \ln(1/\xi))^{-\alpha_0},
\end{aligned}$$

where $K^* > 0$, $K^{**} > 0$, and $K > 0$ are constants. This proves the lemma.

Proof of Theorem 2.1. We first show that if (2.12) holds, for $i = 1, \ldots, N$,

$$\int \exp\left(d_0^* ||\beta^{(m)}||\right) p(y_{0i} \mid \beta^{(m)}, \epsilon_{0i}, I = m) d\beta^{(m)} < K_1, \tag{A.1}$$

where $d_0^* > 0$, $K_1 > 0$ is a finite constant independent of ϵ_{0i}, and $||\beta^{(m)}|| = \sqrt{(\beta^{(m)})' \beta^{(m)}}$.

We have

$$\begin{aligned}
p(y_{0i} \mid \beta^{(m)}, \epsilon_{0i}, I = m) \le M^* \prod_{j=1}^{k_m} \exp \Big\{ &(y_{0it_j} h(x'_{0it_j} \beta^{(m)} + \epsilon_{0it_j}) \\
&- q(h(x'_{0it_j} \beta^{(m)} + \epsilon_{0it_j})) \Big\},
\end{aligned} \tag{A.2}$$

where $M^* > 0$ is a finite constant, and ϵ_{0it_j} is the t_j^{th} component of ϵ_{0i}. Now we make the transformation $u = (u_1, u_2, \ldots, u_{k_m})' = \left(x_{0it_1}^{(m)}, x_{0it_2}^{(m)}, \ldots, x_{0it_{k_m}}^{(m)}\right)' \beta^{(m)} + (\epsilon_{0it_1}, \epsilon_{0it_2}, \ldots, \epsilon_{0it_{k_m}})'$. This is a one-to-one linear transformation in k_m dimensions, since $\left(x_{0it_1}^{(m)}, x_{0it_2}^{(m)}, \ldots, x_{0it_{k_m}}^{(m)}\right)'$ has full rank k_m. Thus,

$$||\beta^{(m)}|| \le c_1 \sum_{j=1}^{k_m} |u_j|, \tag{A.3}$$

where $c_1 > 0$ is a constant. It is easy to see that (A.2) and (A.3) lead to

$$\int \exp\left(d_0^*||\beta^{(m)}||\right) p(y_{0i} \mid \beta^{(m)}, \epsilon_{0i}, I = m)\ d\beta^{(m)}$$

$$\leq M^{**} \prod_{j=1}^{k_m} \left\{ \int_{-\infty}^{\infty} \exp\left(d_0^* c_1 |u_j|\right) \exp\left(\left(y_{0it_j} h(u_j) - q(h(u_j))\right)\right) du_j \right\} = K_1 < \infty \quad \text{(A.4)}$$

by (2.12), where $M^{**} > 0$ is a constant. This proves (A.1).

Since (2.11) is true, without loss of generality, we assume that $p(y_{0i} \mid \beta^{(m)}, \epsilon_{0i}, I = m) \leq 1$ for $j = 1, \ldots, N$. Using Lemma 2.1, (A.1), and $\alpha_0 > k_m/N$, we have

$$\int \left\{ \prod_{i=1}^{N} \int_0^1 \left[p(y_{0i} \mid \beta^{(m)}, \epsilon_{0i}, I = m)\right]^{a_{0i}}\ a_{0i}^{\alpha_0 - 1} (1 - a_{0i})^{\lambda_0 - 1} da_{0i} \right\} d\beta^{(m)}$$

$$\leq \int K^N \prod_{i=1}^{N} \left(1 - \ln\left[p(y_{0i} \mid \beta^{(m)}, \epsilon_{0i}, I = m)\right]\right)^{-\alpha_0}\ d\beta^{(m)}$$

$$= K^N \int \prod_{i=1}^{N} \left(1 - \ln\left[p(y_{0i} \mid \beta^{(m)}, \epsilon_{0i}, I = m)\right]\right)^{-\alpha_0}$$

$$\times 1\left\{ \max_{1 \leq i \leq N} p(y_{0i} \mid \beta^{(m)}, \epsilon_{0i}, I = m) > e^{-d_0^*(||\beta^{(m)}||+1)} \right\} d\beta^{(m)}$$

$$+ K^N \int \prod_{i=1}^{N} \left(1 - \ln\left[p(y_{0i} \mid \beta^{(m)}, \epsilon_{0i}, I = m)\right]\right)^{-\alpha_0}$$

$$\times 1\left\{ \max_{1 \leq i \leq N} p(y_{0i} \mid \beta^{(m)}, \epsilon_{0i}, I = m) \leq e^{-d_0^*(||\beta^{(m)}||+1)} \right\}\ d\beta^{(m)}$$

$$\leq K^N \int 1\left\{ \max_{1 \leq i \leq N} p(y_{0i} \mid \beta^{(m)}, \epsilon_{0i}, I = m) > e^{-d_0^*(||\beta^{(m)}||+1)} \right\} d\beta^{(m)}$$

$$+ K^N \int \prod_{i=1}^{N} [d_0^*(||\beta^{(m)}|| + 1)]^{-\alpha_0} d\beta^{(m)}$$

$$\leq K^N \sum_{1 \leq i \leq N} \int p(y_{0i} \mid \beta^{(m)}, \epsilon_{0i}, I = m) e^{d_0^*(||\beta^{(m)}||+1)} d\beta^{(m)}$$

$$+ K^N (d_0^*)^{-\alpha_0 N} \int (||\beta^{(m)}|| + 1)^{-N\alpha_0} d\beta^{(m)}\ \leq K_2 < \infty, \quad \text{(A.5)}$$

where $K_2 > 0$ is a constant, independent of ϵ_{0i} for $i = 1, \ldots, N$. Finally, it follows directly from (A.5) that the normalizing constant of the prior is less than or equal to

$$K_2 \int_{-1}^{1} \int_0^{\infty} \prod_{i=1}^{N} \left(\int\ \phi_{n_{0i}}(\epsilon_{0i} | 0, \sigma^2 \Sigma_{0i})\ d\epsilon_{0i} \right)$$

$$\times(\sigma^2)^{-(\delta_0+1)}\exp(-\sigma^{-2}\gamma_0)\ \times(1+\rho)^{\nu_0-1}(1-\rho)^{\psi_0-1}\ d\sigma^2 d\rho$$
$$= K_2\int_{-1}^{1}\int_{0}^{\infty}(\sigma^2)^{-(\delta_0+1)}\exp(-\sigma^{-2}\gamma_0)\ \times(1+\rho)^{\nu_0-1}(1-\rho)^{\psi_0-1}\ d\sigma^2 d\rho<\infty,$$

since $(\sigma^2)^{-(\delta_0+1)}\exp(-\sigma^{-2}\gamma_0)$ and $(1+\rho)^{\nu_0-1}(1-\rho)^{\psi_0-1}$ are proper priors. This proves the theorem.

Proof of Theorem 3.1. Let $\pi(\beta^{(-m)}|y_0, I=\mathcal{K})$ and $p(\beta^{(-m)}|y, y_0, I=\mathcal{K})$ denote the respective marginal prior and posterior distributions of $\beta^{(-m)}$ obtained from the full model. The Savage-Dicky density ratio (see, for example, Verdinelli and Wasserman (1995)) directly yields

$$\frac{p(y|I=m)}{p(y|I=\mathcal{K})}=\frac{p(\beta^{(-m)}=0|y,y_0,I=\mathcal{K})}{\pi(\beta^{(-m)}=0|y_0,I=\mathcal{K})}\ , \qquad m=1,\ldots,\mathcal{K}. \tag{A.6}$$

Using (A.6) and (3.1), it suffices to show that $p(I=m)\propto\pi(\beta^{(-m)}=0|y_0,I=\mathcal{K})$.

It can be easily observed that $\int p_0^*(\beta^{(\mathcal{K})},\sigma^2,\rho|y_0,I=\mathcal{K})\ d\beta^{(\mathcal{K})}\ d\sigma^2\ d\rho=\frac{p_0^*(\beta^{(m)},\beta^{(-m)}=0,\sigma^2,\rho|y_0,I=\mathcal{K})}{\pi(\beta^{(m)},\beta^{(-m)}=0,\sigma^2,\rho|y_0,I=\mathcal{K})}$, and $\int p_0^*(\beta^{(m)}|y_0,I=m)\ d\beta^{(m)}=\frac{p_0^*(\beta^{(m)},\sigma^2,\rho|y_0,I=m)}{\pi(\beta^{(m)},\sigma^2,\rho|y_0,I=m)}$. Then we are led to $p_0^*(\beta^{(m)},\beta^{(-m)}=0,\sigma^2,\rho|y_0,I=\mathcal{K})=p_0^*(\beta^{(m)},\sigma^2,\rho|y_0,I=m)$ and $\pi(\beta^{(m)},\beta^{(-m)}=0,\sigma^2,\rho|y_0,I=\mathcal{K})=\pi(\beta^{(-m)}=0|y_0,I=\mathcal{K})\pi(\beta^{(m)},\sigma^2,\rho|y_0,I=m)$. The above two identities yield

$$\int p_0^*(\beta^{(\mathcal{K})},\sigma^2,\rho|y_0,I=\mathcal{K})\ d\beta^{(\mathcal{K})}\ d\sigma^2\ d\rho=\frac{\int p_0^*(\beta^{(m)}|y_0,I=m)\ d\beta^{(m)}}{\pi(\beta^{(-m)}=0|y_0,I=\mathcal{K})}.$$

Now note that $\frac{\int p_0^*(\beta^{(\mathcal{K})},\sigma^2,\rho|y_0,I=\mathcal{K})d\beta^{(\mathcal{K})}d\sigma^2 d\rho}{\sum_{j=1}^{\mathcal{K}}\int p_0^*(\beta^{(j)}|y_0,I=j)\ d\beta^{(j)}}$ is independent of the model index m. This completes the proof.

References

Bedrick, E. J., Christensen, R. and Johnson, W. (1996). A new perspective on priors for generalized linear models. *J. Amer. Statist. Assoc.* **91**, 1450-1460.

Chen, M.-H. (1994). Importance-weighted marginal Bayesian posterior density estimation. *J. Amer. Statist. Assoc.* **89**, 818-824.

Chen, M.-H., Ibrahim, J. G. and Yiannoutsos, C. (1999). Prior elicitation, variable selection, and and Bayesian computation for logistic regression models. *J. Roy. Statist. Soc. Ser. B* **61**, 223-242.

Chen, M.-H. and Shao, Q.-M. (1997). Performance study of marginal posterior density estimation via Kullback-Leibler divergence. *Test* **6**, 321-350.

Cowles, M. K. and Carlin, B. P. (1996). Markov chain Monte Carlo convergence diagnostics: a comparative review. *J. Amer. Statist. Assoc.* **91**, 883-904.

Gelfand, A. E., Sahu, S. K. and Carlin, B. P. (1996). Efficient parametrisations for generalized linear mixed models (with discussion). In *Bayesian Statistics 5* (Edited by J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith), 165-180. Oxford University Press, Oxford.

Gelfand, A. E., Smith, A. F. M. and Lee, T. M. (1992). Bayesian analysis of constrained parameter and truncated data problems using Gibbs sampling. *J. Amer. Statist. Assoc.* **87**, 523-532.

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. *J. Amer. Statist. Assoc.* **93**, 1282-1293.

Shephard, N. and Pitt, M. K. (1997). Likelihood analysis of non-Gaussian measurement time series. *Biometrika* **84**, 653-668.

Stark, P. C., Ryan, L. M., McDonald, J. L. amd Burge, H. A. (1997). Using meteorologic data to predict daily ragweed pollen levels. *Aerobiologia* **13**, 177-184.

Verdinelli, I. and Wasserman, L. (1995). Computing Bayes factors using a generalization of the Savage-Dickey density ratio. *J. Amer. Statist. Assoc.* **90**, 614-618.

Zeger, S. L., (1988). A regression model for time series of counts. *Biometrika* **75**, 621-629.

Department of Biostatistics, Harvard University, School of Public Health, 44 Binney St., Boston MA 02115, U.S.A.

E-mail: ibrahim@jimmy.harvard.edu

(Received June 1998; accepted November 1999)