Bayesian Subset Selection for Reproductive Dispersion Linear Models

Yuanying ZHAO

College of Mathematics and Information Science, Guiyang University, Guiyang 550005, China
E-mail: zhaoyuanying_@126.com

Dengke XU

Department of Statistics, Zhejiang Agriculture and Forest University, Lin'an 311300, China

Xingde DUAN

Department of Mathematics, Chuxiong Normal College, Chuxiong 675000, China

Yicheng PANG

School of Mathematics and Statistics, Guizhou University of Finance and Economics, Guiyang 550025, China

Abstract We propose a full Bayesian subset selection method for reproductive dispersion linear models, which bases on expanding the usual link function to a function that incorporates all possible subsets of predictors by adding indicators as parameters. The vector of indicator variables dictates which predictors to delete. An efficient MCMC procedure that combining Gibbs sampler and Metropolis-Hastings algorithm is presented to approximate the posterior distribution of the indicator variables. The promising subsets of predictors can be identified as those with higher posterior probability. Several numerical examples are used to illustrate the newly developed methodology.

Keywords Bayesian subset selection; Gibbs sampler; Metropolis-Hastings algorithm; reproductive dispersion linear models

1 Introduction

Reproductive dispersion distribution family proposed by Jorgensen^[1], which is a general distribution family and includes normal distribution, Gumbel distribution and exponential family distribution as its special case, has been received a lot of attention in recent years. For example, Tang et al.^[2] proposed a nonlinear reproductive dispersion model and studied its influence diagnostics; Chen et al.^[3] discussed influence diagnostics of semiparametric nonlinear reproductive dispersion models; Fu et al.^[4] investigated local influence analysis for nonlinear structural equation models with non-ignorable missing outcomes from a reproductive dispersion model.

Received November 27, 2013, accepted December 26, 2013

Supported by Grants from the National Natural Science Foundation of China (Grant No. 11301485), Scientific Research Project for Introduced Talents of Guiyang University, Scientific Research Project for Introduced Talents of Guizhou University of Finance and Economics in 2013

In particular, Chen and Tang^[5] developed a Bayesian approach to analyze semiparametric reproductive dispersion mixed-effects models on the basis of P-spline estimates of nonparametric components by using the Gibbs sampler^[6] and the Metropolis-Hastings algorithm^[7–8]. Recently, Tang and Zhao^[9] presented semiparametric Bayesian analysis of nonlinear reproductive dispersion mixed models for longitudinal data.

Various Bayesian methods have been proposed for selecting suitable subset models, those classical methods include Bayes information criterion^[10], asymptotic information criterion^[11], Bayes factor^[12] and pseudo-Bayes factor^[13], etc. In particular, Mitchell and Beauchamp^[14] presented 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. Furthermore, George and McCulloch^[15] developed a variable selection procedure via Gibbs sampling. The reversible jump MCMC^[16] was proposed for investigating model uncertainty. Nott and Leonte^[17] proposed sampling schemes for Bayesian selection in generalized linear models based on the Swendsen-Wang algorithm for the Ising model^[18]. Hans^[19] addressed model uncertainty for the Bayesian lasso regression by computing the marginal posterior probabilities for small model space. Recently, Lykou and Ntzoufras^[20] studied how to select variables using Bayesian lasso. However, to our knowledge, there is little work done on Bayesian subset selection for reproductive dispersion distribution family. Motivated by the idea of Kuo and Mallick^[21], the main purpose of this paper is to develop a novel Bayesian approach to select promising subsets under reproductive dispersion linear models by using MCMC technology.

This article is structured as follows. A reproductive dispersion linear model which includes model uncertainty is defined in Section 2. In Section 3, we explicitly describe a MCMC procedure that can identify the promising subsets of predictors under reproductive dispersion linear models. Several numerical examples are used to illustrate our proposed methodology in Section 4. A discussion is given in the final section. Technical details are given in the Appendix.

2 Model and notation

Suppose y_i ($i = 1, 2, \dots, n$) is the observation for the i th individual, we assume that $y_1, y_2, \dots, y_n$ are independently distributed and has the following probability density function

$$p(y_i; \mu_i) = a(y_i; \sigma^2) \exp \left\{ -\frac{1}{2\sigma^2} d(y_i; \mu_i) \right\} \quad (2.1)$$

where μ_i is the location parameter and may represent the mean of y_i , $\sigma^2 \triangleq \tau^{-1} \in \mathbb{B}(\mathbb{B} \subset \mathbb{R}^+)$ is referred to as the dispersion parameter which is known or can be estimated separately; $a(\cdot; \cdot)$ is some suitable known function; $d(y; \mu)$ is a unit deviance on $\mathbb{C} \times \Omega$ ($\Omega \subseteq \mathbb{C} \subseteq \mathbb{R}$ are open intervals) and satisfies $d(y; y) = 0$, $\forall y \in \mathbb{C}$ and $d(y; \mu) > 0$, $\forall y \neq \mu$, and is twice continuously differentiable with respect to (y, μ) on $\mathbb{C} \times \Omega$. The model (2.1) is referred to as a reproductive dispersion distributional family^[1] and includes normal distribution, double exponential distribution, Gumbel distribution and exponential family distribution as its special case. Motivated by the idea of Kuo and Mallick^[21], we explore a simple method of subset selection. Instead of building a hierarchical model, we embed indicator variable in predictors that incorporates all 2^p submodels. Specifically, let γ_j ($j = 1, 2, \dots, p$) be the indicator variable supported at two points 1 and 0. When $\gamma_j = 1$ we include the predictor in the predictor model.

When $\gamma_j = 0$, we omit the predictor when building the model. Therefore we assume that the systematic part of the model is specified by

$$\eta_i = g(\mu_i) = \sum_{j=1}^p x_{ij} \beta_j \gamma_j \quad (2.2)$$

where $g(\cdot)$ is a known continuously differentiable link function. For simplicity, we only consider the case that $g(\cdot)$ is the canonical link in this article, i.e., $g(\mu_i) = \mu_i$, then (2.2) reduces to

$$\mu_i = \sum_{j=1}^p \mathbf{x}_{ij} \beta_j \gamma_j \quad (2.3)$$

The model defined by equations (2.1) and (2.3) is referred to as a reproductive dispersion linear model which includes model uncertainty.

Let $\boldsymbol{\beta} = (\beta_1, \beta_2, \dots, \beta_p)^T$, $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \dots, \gamma_p)^T$, $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \tau)^T$ and $\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{ip})^T$. We assume that the prior of $\boldsymbol{\theta}$ is independent on $\boldsymbol{\gamma}$ and γ_j is mutually independent for $j = 1, 2, \dots, p$. Under the above assumption, the prior density $p(\boldsymbol{\theta}, \boldsymbol{\gamma})$ of $(\boldsymbol{\theta}, \boldsymbol{\gamma})$ can be expressed as

$$p(\boldsymbol{\theta}, \boldsymbol{\gamma}) = p(\boldsymbol{\beta}, \tau, \boldsymbol{\gamma}) \propto p(\tau) p(\boldsymbol{\beta} | \tau) \prod_{j=1}^p p(\gamma_j).$$

Based on the recommendations given by Chen and Tang^[5] and Kuo and Mallick^[21], we consider the following prior distributions for parameters τ , $\boldsymbol{\beta}$ and $\gamma_j (j = 1, 2, \dots, p)$:

$$p(\tau) \sim \Gamma(a_1, a_2), \quad p(\boldsymbol{\beta} | \tau) \sim N(\boldsymbol{\beta}_0, \tau^{-1} \mathbf{H}_0), \quad p(\gamma_j) \sim B(1, p_j) \quad (2.4)$$

where $a_1, a_2, \boldsymbol{\beta}_0, \mathbf{H}_0$ and p_j are hyperparameters whose values are assumed to be given, $\Gamma(a, b)$ denotes the gamma distribution with parameters a and b , $N(\boldsymbol{\beta}, \boldsymbol{\Sigma})$ denotes the multivariate normal distribution with mean $\boldsymbol{\beta}$ and covariance $\boldsymbol{\Sigma}$, $B(1, q)$ is the Bernoulli distribution with parameter q . In this paper, we set $a_1 = a_2 = 0.01$, $\boldsymbol{\beta}_0 = \mathbf{0}$, $\mathbf{H}_0 = 100 \mathbf{I}_p$ and $p_j = 0.5$ in all numerical examples which may present noninformation priors, where $\mathbf{I}_p$ is the $p \times p$ identity matrix.

3 Bayesian subset selection

In this section, we propose and develop a MCMC procedure that can identify the promising subsets of predictors under reproductive dispersion linear models. Denote $\mathbf{y} = \{y_1, y_2, \dots, y_n\}$, $\mathbf{x} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}$. The Bayesian inference on $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$ can be via the posterior distribution $\log p(\boldsymbol{\theta}, \boldsymbol{\gamma} | \mathbf{y}, \mathbf{x}) \propto \log p(\mathbf{y}, \mathbf{x} | \boldsymbol{\theta}, \boldsymbol{\gamma}) + \log p(\boldsymbol{\theta}, \boldsymbol{\gamma})$, where $p(\mathbf{y}, \mathbf{x} | \boldsymbol{\theta}, \boldsymbol{\gamma})$ is the likelihood function of the observed data $\{\mathbf{y}, \mathbf{x}\}$. The likelihood $p(\mathbf{y}, \mathbf{x} | \boldsymbol{\theta}, \boldsymbol{\gamma})$ involves intractable multiple high dimension integrals. Thus, it is rather difficult and tedious to work directly with this posterior density. Inspired by recent development in statistical computing, the Gibbs sampler^[6] is used to generate a sequence of random observations from the joint posterior distribution $p(\boldsymbol{\theta}, \boldsymbol{\gamma} | \mathbf{y}, \mathbf{x})$, which is proportional to $\prod_{i=1}^n p(y_i, x_i | \boldsymbol{\theta}, \boldsymbol{\gamma}) p(\boldsymbol{\theta}, \boldsymbol{\gamma})$, i.e.,

$$\prod_{i=1}^n a(y_i; \tau^{-1}) \exp \left\{ -\frac{\tau}{2} \sum_{i=1}^n d(y_i; \mu_i) \right\} p(\boldsymbol{\theta}, \boldsymbol{\gamma}) \quad (3.1)$$

In this algorithm, observations $\{\boldsymbol{\theta}, \boldsymbol{\gamma}\}$ are iteratively simulated from the following conditional distributions: $p(\boldsymbol{\theta} | \mathbf{y}, \mathbf{x}, \boldsymbol{\gamma})$ and $p(\boldsymbol{\gamma} | \mathbf{y}, \mathbf{x}, \boldsymbol{\theta})$. Then we identify the promising subset of

predictors with high posterior probabilities of γ .

3.1 Conditional distributions

In this subsection, we briefly present the conditional distributions required in implementing the Gibbs sampler.

For the conditional distribution of γ , we have

$$p(\gamma|\mathbf{y}, \mathbf{x}, \tau, \boldsymbol{\beta}) \propto \prod_{j=1}^p p(\gamma_j) \prod_{i=1}^n a(y_i; \tau^{-1}) \exp \left\{ -\frac{\tau}{2} \sum_{i=1}^n d(y_i; \mu_i) \right\} \quad (3.2)$$

For the conditional distribution $p(\tau|\mathbf{y}, \mathbf{x}, \boldsymbol{\theta})$, it follows from equations (3.1) and (2.4) that

$$p(\tau|\mathbf{y}, \mathbf{x}, \boldsymbol{\theta}) \propto \tau^{a_1 + \frac{p}{2} - 1} a(\mathbf{y}; \tau) \exp \left\{ -\tau \left[a_2 + \frac{1}{2} \sum_{i=1}^n d(y_i; \mu_i) + \frac{1}{2} (\boldsymbol{\beta} - \boldsymbol{\beta}_0)^T \mathbf{H}_0^{-1} (\boldsymbol{\beta} - \boldsymbol{\beta}_0) \right] \right\} \quad (3.3)$$

where $a(\mathbf{y}, \tau) = \prod_{i=1}^n a(y_i; \tau)$. Clearly, if $a(y_i; \tau) = f(y_i) \tau^q$ in which $f(y_i)$ is a function of y_i , then

$$p(\tau|\mathbf{y}, \mathbf{x}, \boldsymbol{\theta}) \sim \Gamma(\nu_\tau, g_\tau) \quad (3.4)$$

where $\nu_\tau = a_1 + \frac{p}{2} + nq$, and $g_\tau = a_2 + \frac{1}{2} \{ \sum_{i=1}^n d(y_i; \mu_i) + (\boldsymbol{\beta} - \boldsymbol{\beta}_0)^T \mathbf{H}_0^{-1} (\boldsymbol{\beta} - \boldsymbol{\beta}_0) \}$.

For the conditional distribution $p(\boldsymbol{\beta}|\mathbf{y}, \mathbf{x}, \tau, \gamma)$, it follows from equations (3.1) and (2.4) that

$$p(\boldsymbol{\beta}|\mathbf{y}, \mathbf{x}, \tau, \gamma) \propto \exp \left\{ -\frac{\tau}{2} \left[\sum_{i=1}^n d(y_i; \mu_i) + (\boldsymbol{\beta} - \boldsymbol{\beta}_0)^T \mathbf{H}_0^{-1} (\boldsymbol{\beta} - \boldsymbol{\beta}_0) \right] \right\} \quad (3.5)$$

3.2 Implementation

Now, we describe how to update the parameter γ . We update the variate $\gamma_j (j = 1, 2, \dots, p)$ given $\mathbf{y}, \mathbf{x}, \gamma_{-j}, \boldsymbol{\beta}, \tau$ by a Bernoulli distribution $B(1, \frac{c_j}{c_j + d_j})$, where γ_{-j} is γ with the j th element deleted,

$$c_j = p_j \prod_{i=1}^n p(y_i, x_i | \boldsymbol{\beta}, \tau, \gamma^*),$$

and

$$d_j = (1 - p_j) \prod_{i=1}^n p(y_i, x_i | \boldsymbol{\beta}, \tau, \gamma^{**}),$$

in which the vector $\gamma^* (\gamma^{**})$ is obtained from γ with the j th element replaced by 1(0). Clearly, sampling observations from conditional distributions given in (3.4) are straightforward and fast because the conditional distribution is a familiar standard distribution. However, it is rather difficult to sample observations from the conditional distributions given in (3.3) and (3.5) because these conditional distributions are unfamiliar standard and complex distributions. Therefore, the Metropolized independence sampler (MIS) and Metropolis-Hastings (MH) algorithm are employed to solve these difficulties. The technical details are given in the Appendix.

4 Numerical examples

In this section, several numerical examples are used to illustrate the preceding proposed method.

4.1 Simulation studies

In this subsection, numerical results with responses coming from the double exponential distribution and Gumbel distribution are used to investigate the performance of the proposed methodology.

We consider the first situation, the response y_i ($i = 1, 2, \dots, n$) is distributed as the double exponential distribution with probability density function

$$p(y_i; \mu_i) = \frac{\tau}{2} \exp\{-\tau|y_i - \mu_i|\} \quad (4.1)$$

and the location parameter μ_i is given by

$$\mu_i = x_{i1}\beta_1\gamma_1 + x_{i2}\beta_2\gamma_2 + x_{i3}\beta_3\gamma_3 + x_{i4}\beta_4\gamma_4 + x_{i5}\beta_5\gamma_5 \quad (4.2)$$

In other words, we consider this variable selection problem with $p = 5$. It is assumed that $\mathbf{x}_i = (x_{i1}, x_{i2}, x_{i3}, x_{i4}, x_{i5})^T$ is independently generated from the standard normal distribution $N(0, I_5)$. Furthermore, we assume that the true value of $\tau = 1.0$ and $\boldsymbol{\vartheta} = (\vartheta_1, \vartheta_2, \vartheta_3, \vartheta_4, \vartheta_5)^T = (\beta_1\gamma_1, \beta_2\gamma_2, \beta_3\gamma_3, \beta_4\gamma_4, \beta_5\gamma_5)^T = (0, 0, 0, 1, 1.2)^T$, i.e., (4.2) reduces to

$$\mu_i = x_{i4} + 1.2x_{i5} \quad (4.3)$$

then the above specified model (4.1) and (4.3) is a special case of reproductive dispersion linear models which includes model uncertainty. In all simulation studies, we consider that the sample size n equals 100. In the first simulation study, a data set is generated from (4.1) and (4.3), then we apply the preceding proposed method to the data set. Some results based on a sample of 15000 Gibbs samplers are displayed the frequencies of the four highest frequency models in Table 1. From Table 1 we can easily see that our analysis predicts the right model as expected. From the Gibbs samplers, we can also calculate the posterior mean and standard deviation of $\boldsymbol{\vartheta}$. The posterior means are 1.1127 and 1.2727 for ϑ_4 and ϑ_5 with standard deviations 0.1212 and 0.1215, each standard deviation incorporates our measure of model uncertainty.

The second simulation is identical to the first simulation study except that x_{i3} ($i = 1, 2, \dots, n$) is replaced by $x_{i3}^* = x_{i5} + 0.15z$, where $z \sim N(0, 1)$, yielding the correlation coefficient $\text{corr}(x_{i3}, x_{i5}) = 0.989$. This simulation is used to illustrate how our procedure performs in the presence of extreme collinearity. We conduct the same analysis as the first simulation study. Summaries of results are reported in Table 2, which implies that $\mathbf{x}_4$ and either one of $\mathbf{x}_3$ or $\mathbf{x}_5$ are identified as the promising right models.

Table 1 High frequency models in the first simulation study

Model variables	Proportion
$\mathbf{x}_4 \ \mathbf{x}_5$	88.29%
$\mathbf{x}_3 \ \mathbf{x}_4 \ \mathbf{x}_5$	4.11%
$\mathbf{x}_1 \ \mathbf{x}_4 \ \mathbf{x}_5$	3.80%
$\mathbf{x}_2 \ \mathbf{x}_4 \ \mathbf{x}_5$	3.11%

Table 2 High frequency models in the second simulation study

Model variables	Proportion
$\mathbf{x}_4 \ \mathbf{x}_5$	33.43%
$\mathbf{x}_3 \ \mathbf{x}_4$	57.54%
$\mathbf{x}_3 \ \mathbf{x}_4 \ \mathbf{x}_5$	4.65%
$\mathbf{x}_1 \ \mathbf{x}_4 \ \mathbf{x}_5$	3.22%

To demonstrate the practical potential of our procedure for data sets with a relatively large number of covariates, we conduct the following third simulation study. The responses are

coming from the double exponential distribution same as the first simulation study, but the location parameter μ_i is given as follows:

$$\mu_i = x_{i1}\beta_1\gamma_1 + x_{i2}\beta_2\gamma_2 + \cdots + x_{i15}\beta_{15}\gamma_{15} \quad (4.4)$$

The covarites $\mathbf{x}_j (j = 1, 2, \dots, 15)$ are obtained as $\mathbf{x}_j = \mathbf{x}_j^* + U$, where $\mathbf{x}_j^* \stackrel{i.i.d}{\sim} N(\mathbf{0}, \mathbf{I}_{100})$ and $U \sim N(\mathbf{0}, \mathbf{I}_{100})$, which leads to that the pairwise correlation is about 0.5. The coefficients $\boldsymbol{\vartheta} = (\vartheta_1, \dots, \vartheta_{15})^T$ are set at $(\vartheta_1, \dots, \vartheta_5) = (0, \dots, 0)$, $(\vartheta_6, \dots, \vartheta_{10}) = (1, \dots, 1)$ and $(\vartheta_{11}, \dots, \vartheta_{15}) = (2, \dots, 2)$ in which $\vartheta_j = \beta_j\gamma_j (j = 1, 2, \dots, 15)$. A data set is generated by equation (4.1) and (4.4). Summaries of results are presented in Table 3 based on 40000 Gibbs samplers, where False choice is defined to be at least one of the predictors in 1 to 5 is included or at least one of the predictors in 6 to 15 is deleted. From Table 3, 6.87% of the Gibbs samplers contains false choice with $\mathbf{x}_1$ included; 4.12% of the Gibbs samplers contains false choice with $\mathbf{x}_5$ included; 2.09% of the Gibbs samplers contains false choice with $\mathbf{x}_3$ included, etc. So high frequency 82.45% of the Gibbs samplers capture the right promising model, some variation is allowed for lower frequency models too.

Table 3 High frequency models for false choices in the third simulation study

Relative frequency	False choice
82.45%	None
6.87%	$\mathbf{x}_1$
4.12%	$\mathbf{x}_5$
2.09%	$\mathbf{x}_3$

Table 4 High frequency models in the fourth simulation study

Model variables	Proportion
$\mathbf{x}_4 \ \mathbf{x}_5$	96.45%
$\mathbf{x}_3 \ \mathbf{x}_4 \ \mathbf{x}_5$	1.13%
$\mathbf{x}_1 \ \mathbf{x}_4 \ \mathbf{x}_5$	1.08%
$\mathbf{x}_2 \ \mathbf{x}_4 \ \mathbf{x}_5$	0.88%

To investigate the performance under various response distributions, we consider another situation and conduct the following simulation studies. In the fourth simulation study, we assume that response $y_i (i = 1, 2, \dots, n)$ is distributed as the Gumbel distribution, which has the following probability density function

$$p(y_i; \mu_i) = \exp\{y_i - \mu_i - \exp(y_i - \mu_i)\} \quad (4.5)$$

and belongs to reproductive dispersion family^[1]. The fourth simulation study is conducted by using other settings same as given in the first simulation study. The corresponding results are reported in Table 4. Furthermore, we do the fifth (sixth) simulation study which is identical to the second (third) simulation study except data set is generated by equation (4.5). Summaries of these results are presented in Tables 5 and 6. All these results show that the same conclusion

Table 5 High frequency models in the fifth simulation study

Model variables	Proportion
$\mathbf{x}_4 \ \mathbf{x}_5$	41.12%
$\mathbf{x}_3 \ \mathbf{x}_4$	26.35%
$\mathbf{x}_2 \ \mathbf{x}_3 \ \mathbf{x}_4$	10.13%
$\mathbf{x}_2 \ \mathbf{x}_4 \ \mathbf{x}_5$	9.75%

Table 6 High frequency models for false choices in the sixth simulation study

Relative frequency	False choice
80.05%	None
8.40%	$\mathbf{x}_1$
5.56%	$\mathbf{x}_5$
4.15%	$\mathbf{x}_4$

as given in the first situation are drawn, which indicates that the response has little effect on our procedure.

4.2 Real examples

In this subsection, two real data examples are used to illustrate our preceding proposed methodology.

4.2.1 Hald data

Our first real data example is the familiar Hald data, which has been reported by Draper and Smith^[22] and been used by various authors to illustrate variable selection procedures. The data consist of $n = 13$ responses and $p = 4$ independent variables $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4$. Therefore, $2^4 = 16$ possible subset models are under consideration. Equation (4.1) and

$$\mu_i = x_{i1}\beta_1\gamma_1 + x_{i2}\beta_2\gamma_2 + x_{i3}\beta_3\gamma_3 + x_{i4}\beta_4\gamma_4 \quad (4.6)$$

are applied to fit the data. 10000 Gibbs samplers are simulated and higher frequency models are tabulated in Table 7. From Table 7 we can easily see that the results capture the $\mu_i = f(x_{i1}, x_{i2}, x_{i4})$ model, which is quite similar to the results of those forward selection in Draper and Smith^[22].

Table 7 High frequency models in the hald data analysis

Model variables	Proportion
$\mathbf{x}_1 \ \mathbf{x}_2 \ \mathbf{x}_4$	90.35%
$\mathbf{x}_2 \ \mathbf{x}_3 \ \mathbf{x}_4$	4.99%
$\mathbf{x}_1 \ \mathbf{x}_2 \ \mathbf{x}_3 \ \mathbf{x}_4$	4.22%
$\mathbf{x}_2 \ \mathbf{x}_4$	0.44%

4.2.2 Fitness data

The data for our second real example is fitness data, which has been studied by SAS User's Guide: Statistics^[23]. The data consist of $n = 31$ responses and $p = 6$ independent variables. Therefore, $2^6 = 64$ possible subset models are under consideration. We employ equation (4.1) and

$$\mu_i = x_{i1}\beta_1\gamma_1 + x_{i2}\beta_2\gamma_2 + x_{i3}\beta_3\gamma_3 + x_{i4}\beta_4\gamma_4 + x_{i5}\beta_5\gamma_5 + x_{i6}\beta_6\gamma_6 \quad (4.7)$$

to analyze the data, in which $x_{i1}, \dots, x_{i6}$ denote age (years), weight (kg), time to run 1.5 miles in minutes (runtime), heart rate while resting (restpulse), heart rate while running (runpulse) and maximum heart rate records while running (maxpulse), respectively. 15000 Gibbs samplers are generated and higher frequency models are tabulated in Table 8. From Table 8 we can easily

Table 8 High frequency models in the fitness data analysis

Model variables	Proportion
age runtime runpulse maxpulse	58.56%
runtime runpulse maxpulse	37.19%
age runtime restpulse runpulse maxpulse	2.91%
age runpulse maxpulse	0.38%

conclude that the results capture the $\mu_i = f(x_{i1}, x_{i3}, x_{i5}, x_{i6})$ model, which agrees with SAS on the variables to be deleted.

5 Conclusion and discussion

In this article, we extend the Bayesian variable selection methods proposed in [21] to analyze reproductive dispersion linear models which include model uncertainty. A full Bayesian approach is developed to investigate these models' uncertainty by using Gibbs sampler and MH algorithm. Several numerical examples are used show the efficiency of the proposed Bayesian approach. The results show that the developed Bayesian procedure is highly efficient and computationally fast. Other Bayesian variable selection approaches under reproductive dispersion distribution family will be of great research interest. The authors intend to investigate it in the near future.

Acknowledgements

The authors are grateful to professor Niansheng Tang for many valuable suggestions during the preparation of the manuscript and to the Editor-in-Chief of Journal of Systems Science and Information.

Appendix

The Metropolized independence sampler (MIS) is used to simulate observations from the conditional distribution $p(\tau|\mathbf{y}, \mathbf{x}, \boldsymbol{\beta}, \boldsymbol{\gamma})$ given in (3.3). Following the arguments of Liu^[24], the Metropolized independence sampler is implemented as follows: Given the current state $\tau^{(m)}$, a candidate τ is drawn from $h(\tau)$ and is accepted with probability

$$\min \left\{ 1, \frac{p(\tau|\mathbf{y}, \mathbf{x}, \boldsymbol{\beta}, \boldsymbol{\gamma})h(\tau^{(m)})}{p(\tau^{(m)}|\mathbf{y}, \mathbf{x}, \boldsymbol{\beta}, \boldsymbol{\gamma})h(\tau)} \right\},$$

where $h(\tau)$ is some function with respect to τ . As is pointed out by Liu^[24], the efficiency of MIS depends heavily on how close the trial density $g(\tau)$ is to the target density $p(\tau|\mathbf{y}, \mathbf{x}, \boldsymbol{\beta}, \boldsymbol{\gamma})$.

Similarly, the MH algorithm for simulating observations from the conditional distribution $p(\boldsymbol{\beta}|\mathbf{y}, \mathbf{x}, \tau, \boldsymbol{\gamma})$ given in (3.5) is implemented as follows: Given the current value $\boldsymbol{\beta}^{(m)}$, a new candidate $\boldsymbol{\beta}$ is generated from $N(\boldsymbol{\beta}^{(m)}, \sigma_{\boldsymbol{\beta}}^2 \boldsymbol{\Omega}_{\boldsymbol{\beta}})$ and is accepted with probability

$$\min \left\{ 1, \frac{p(\boldsymbol{\beta}|\mathbf{y}, \mathbf{x}, \tau, \boldsymbol{\gamma})}{p(\boldsymbol{\beta}^{(m)}|\mathbf{y}, \mathbf{x}, \tau, \boldsymbol{\gamma})} \right\},$$

where

$$\boldsymbol{\Omega}_{\boldsymbol{\beta}} = \tau^{-1} \left(\frac{1}{2} \sum_{i=1}^n \mathbf{x}_i^T \mathbf{V}_i \mathbf{x}_i + \mathbf{H}_0^{-1} \right)^{-1}$$

with $\mathbf{V}_i = E_{\mathbf{y}_i}(\partial^2 d(\mathbf{y}_i; \boldsymbol{\mu}_i) / \partial \boldsymbol{\mu}_i \partial \boldsymbol{\mu}_i^T) |_{\boldsymbol{\beta}=\mathbf{0}}$. $\sigma_{\boldsymbol{\beta}}^2$ is chosen such that the average acceptance rate is about 0.25 or more.

References

- [1] Jorgensen B. The theory of dispersion models. Chapman and Hall, London, 1997.
- [2] Tang N S, Wei B C, Wang X R. Influence diagnostics in nonlinear reproductive dispersion models. Statistics & Probability Letters, 2000, 46: 59–68.

- [3] Chen X D, Tang N S, Wang X R. Influence diagnostics of semiparametric nonlinear reproductive dispersion models. *Communications in Statistics — Theory and Methods*, 2010, 39: 3021–3040.
- [4] Fu Y Z, Tang N S, Chen X. Local influence analysis of non-linear structural equation models with non-ignorable missing outcomes from reproductive dispersion models. *Computational Statistics and Data Analysis*, 2009, 53: 3671–3684.
- [5] Chen X D, Tang N S. Bayesian analysis of semiparametric reproductive dispersion mixed-effects models. *Computational Statistics and Data Analysis*, 2010, 54: 2145–2158.
- [6] Geman S, Geman D. Stochastic relaxation, Gibbs distribution, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1984, 6: 721–741.
- [7] Metropolis N, Rosenbluth A W, Rosenbluth M N, et al. Equations of state calculations by fast computing machine. *Journal of Chemical Physics*, 1953, 21: 1087–1091.
- [8] Hastings W K. Monte Carlo sampling methods using Markov chains and their application. *Biometrika*, 1970, 57: 97–109.
- [9] Tang N S, Zhao Y Y. Semiparametric Bayesian analysis of nonlinear reproductive dispersion mixed models for longitudinal data. *Journal of Multivariate Analysis*, 2013, 115: 68–83.
- [10] Schwarz G. Estimating the dimension of a model. *The Annals of Statistics*, 1978, 6: 461–464.
- [11] Akaike H. A new look at the statistical identification model. *IEEE Transactions on Automatic Control*, 1974, 19: 716–723.
- [12] Kass R E, Raftery A E. Bayes factors. *Journal of the American Statistical Association*, 1995, 90: 773–795.
- [13] Geisser S, Eddy W F. A predictive approach to model selection. *Journal of the American Statistical Association*, 1979, 74: 153–160.
- [14] Mitchell T J, Beauchamp J J. Bayesian variable selection in linear regression (with discussion). *Journal of the American Statistical Association*, 1988, 83: 1023–1036.
- [15] George E L, McCulloch R E. Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, 1993, 88: 881–889.
- [16] Green P. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 1995, 82: 711–732.
- [17] Nott D J, Leonte D. Sampling schemes for Bayesian selection in generalized linear models. *Journal of Statistical Computation and Simulation*, 2004, 13(2): 362–382.
- [18] Swendsen R H, Wang J S. Nonuniversal critical dynamics in Monte Carlo simulations. *Physical Review Letters*, 1987, 58: 86–88.
- [19] Hans C. Model uncertainty and variable selection in Bayesian lasso regression. *Statistics and Computing*, 2010, 20(2): 221–229.
- [20] Lykou A, Ntzoufras I. On Bayesian lasso variable selection and the specification of the shrinkage parameter. *Statistics and Computing*, 2013, 23(3): 361–390.
- [21] Kuo L, Mallick B. Variable selection for regression models. *Sankhyā*, 1998, B60: 65–81.
- [22] Draper N, Smith H. *Applied regression analysis*. 2nd ed. John Wiley, New York, 1981.
- [23] SAS Institute. *SAS User's Guide: Statistics ver.5*. SAS Institute Inc, 1985.
- [24] Liu J S. *Monte Carlo strategies in scientific computing*. Springer-Verlag, New York, 2001.