An adaptive MCMC scheme for Bayesian Variable Selection in Binary and Time-to-Event Endpoints via data augmentation approach

Author(s): Kitty Yuen Yi Wan[1], Douglas Robinson[2] and Jim Griffin[3]

Affiliations: [1]Novartis Pharma AG, Basel; [2]Novartis Pharmaceuticals Corporation, Cambridge, United States; [3]School of Mathematics, Statistics and Actuarial Science, University of Kent, Canterbury, United Kingdom.

Website to Download: https://www.kent.ac.uk/smsas/personal/jcg28/index.htm

Abstract

Advances in technology have enabled thousands of mutations to be sequenced simultaneously, hence providing a very powerful tool in biomarker detection and discovery. Such high-throughput data continues to be a challenge to statistical analyses because of their high-dimensionality. Modern statistical analysis such as regression models have been widely applied to determine which biomarkers are significantly correlated with outcome of interest. A powerful approach is Bayesian variable selection and many computational approaches have been developed for its use with many regressors. Most recently, Griffin et al (2018) developed a tuneable proposal distribution for Metropolis-Hastings sampling on model space and applied this to Bayesian variable in linear regression models. They demonstrated that their method could mix substantially faster than the standard Add-Delete-Swap sampler. We extend their work to logistic regression using Polya-Gamma latent variables [Polson et al (2013)] and Laplace approximation for binary endpoint, and accelerated failure time models [Sha et al (2006)] for time-to-event endpoint, via a simple data augmentation [Tanner et al (1987)] approach. These samplers are fast to run and mix quickly. The approach is demonstrated on two problems with many regressors.

Introduction

Griffin et al (2018) developed their methodology in the linear regression framework. To extend their work further to logistic regression models or accelerated failure time models is challenging, because the marginal likelihood is not available in closed form. Data augmentation techniques, where unobserved data or latent variables may be introduced, leads to a conditional Gaussian posterior distribution in these models. The method can then be directly applied to updates on model space. However, this potentially causes draw back of slow mixing due to the introduction of latent variable or slow run times due to time-consuming algorithms for simulation of the latent variables. To alleviate this problem in the logistic regression model, we consider a hybrid method which combines data augmentation with a method using an iteratively reweighted least squares approach [Gammerman (1997) and Lamnisos et al (2009)].

Methods

Griffin et al (2018) used a Metropolis-Hastings sampler to propose a move from model γ to γ' be given in a product form

$$q_\eta(\gamma,\gamma') = \prod_{j=1}^{p} q_{\eta,j}(\gamma_j,\gamma_j')$$

where $\eta = (A,D) = (A_1,\dots,A_p,D_1,\dots,D_p)$, $q_{\eta,j}(\gamma_j=0,\gamma_j'=1) = A_j$ and $q_{\eta,j}(\gamma_j=1,\gamma_j'=0) = D_j$. This implies that each dimension of a new value of each dimension of γ is independently proposed. The probability of proposing to add the j-th variable if it is currently excluded from the model is A_j and the probability of proposing to delete the j-th variable if it is currently included in the model is D_j. They show that an effective choice for (A,D) is

$$A_j = \zeta\min\left\{1,\frac{\pi_j}{1-\pi_j}\right\} \qquad D_j = \zeta min\left\{1,\frac{1-\pi_j}{\pi_j}\right\}$$

where π_j is the posterior inclusion probability for the j-th variable. The acceptance probability can be easily calculated since $p(y|\gamma)$ is available in analytically if conjugate priors are used in a linear regression model.

Logistic Regression model for binary outcome

The logistic regression model can be used to link a categorical outcome to the regressors using a generalized linear model. If $y_i \sim \text{Bin}(n_i,\tau_i)$, the logistic regression model stipulates a linear relationship between the regressors and the probability of success (measured on the log-odds scale),

$$\eta_i = \log\left(\frac{\tau_i}{1-\tau_i}\right) = \alpha + X_{\gamma,i}\beta_\gamma, \qquad i = 1,\dots,n$$

where α is the intercept term, $X_{\gamma,i}$ is a $(1\times p_\gamma)$–dimensional regressor vector, and β_γ is a $(p\times 1)$-dimensional regression parameter vector. We will write $\theta_\gamma = (\alpha,\beta_\gamma)'$. We assume that the prior for $\theta_\gamma \sim \text{N}(0,V_\gamma)$ for the regression coefficients and we let $y_i^* = \frac{y_i}{n_i}$.

Our extension of the algorithm due to Griffin et al (2018) to this problem uses the Polya-gamma data augmentation method [Polson et al (2013)] during the tuning period and an automatic generic sampling method [Green (2003)] after the tuning period. The Polya-gamma method leads to a conditionally Gaussian posterior which allow us to use efficient methods to compute the Rao-Blackwellised estimates of the posterior inclusion probabilities in the algorithm. However, the simulation of Polya-Gamma random variables is slow and the automatic generic sampling method is much faster after the tuning period. We will describe the two parts of the algorithm (during the tuning period and after the tuning period) separately.

During the tuning period, we use the Polya-Gamma method for sampling from the posterior of a logistic regression model. This exploits the following identity

$$\frac{(e^\psi)^a}{(1+e^\psi)^b} = 2^{-b}\exp\{\kappa\psi\}\int_0^\infty \exp\{-\omega\psi^2/2\}p(\psi)d\psi$$

where $\kappa = a-b/2$, $\omega_i \sim \text{PG}(b,0)$ and PG represents a Polya-Gamma distribution, which is defined in Polson et al (2013). The likelihood of the logistic regression model can be expressed as an extended with additional latent variables $\omega_1,\dots,\omega_n$ has the form

$$p(y,\omega_1,\dots,\omega_n|\theta_\gamma,\gamma) \propto \prod_{i=1}^{n}\left[2^{-n_i}\exp\{\kappa_i(\alpha+X_{\gamma,i}\beta_\gamma)\}\exp\{-\omega_i(\alpha+X_{\gamma,i}\beta_\gamma)/2\}p(\omega_i)\right]$$

where $\kappa_i = n_i - y_i$ and $p(\omega_i)$ is the $\text{PG}(n_i,0)$ distribution. The identity can be used to show that the marginal distribution on is likelihood of the logistic regression model. If we assume that $p(\theta_\gamma)\sim\text{N}(\mu_\gamma,V_\gamma)$, the marginal likelihood is

$$p(y|\gamma,\omega_1,\dots,\omega_n) = \prod_{i=1}^{n} 2^{-n_i}\left|V_\gamma\right|^{-1/2}\left|X_\gamma^{*\prime}X_\gamma^* + V_\gamma^{-1}\right|^{-1/2}$$
$$\times\exp\left\{-\frac{1}{2}\mu_\gamma'V_\gamma^{-1}\mu_\gamma + \frac{1}{2}\left(X_\gamma^{*\prime}K^* + V_\gamma^{-1}\mu_\gamma\right)'\left(X_\gamma^{*\prime}X_\gamma^* + V_\gamma^{-1}\right)^{-1}\left(X_\gamma^{*\prime}K^* + V_\gamma^{-1}\mu_\gamma\right)\right\}$$

where $X_{\gamma,i}^* = \sqrt{\omega_i}X_{\gamma,i}$ and $.K_i^* = \sqrt{\omega_i}\kappa_i$This is exactly the marginal likelihood for the linear regression mode in Griffin et al (2018) with X_γ^* and K^* playing the roles of X_γ and y in their notation. This allows us to calculated the Metropolis-Hastings acceptance probability for a new model and to calculate the Rao-Blackwellised estimates of π_i. The latent variables $\omega_1,\dots,\omega_n$ can be updated by first sampling β_γ according to

$$\beta_\gamma \sim \text{N}\left(\left(X_\gamma^{*\prime}X_\gamma^* + V_\gamma^{-1}\right)^{-1}\left(X_\gamma^{*\prime}K^* + V_\gamma^{-1}\mu_\gamma\right),\left(X_\gamma^{*\prime}X_\gamma^* + V_\gamma^{-1}\right)^{-1}\right)$$

and then sampling $\omega\sim\text{PG}(n_i,X_{\gamma,i}\beta_\gamma)$. Polson et al (2013) describe efficient algorithms for the generation of Polya-Gamma distributed random variables.

After the tuning period, we use the automatic generic sampling method [Green (2003)]. The method samples from the joint posterior distribution of θ_γ and γ using a reversible jump MCMC method with proposal

$$q\left(\left(\gamma,\theta_\gamma\right),\left(\gamma',\theta_\gamma'\right)\right) = \text{q}\left(\theta_\gamma,\theta_\gamma'\middle|\gamma,\gamma'\right)\text{q}(\gamma,\gamma')$$

Where $\text{q}\left(\theta_\gamma,\theta_\gamma'\middle|\gamma,\gamma'\right)$ is the automatic generic method proposal and $\text{q}(\gamma,\gamma')$ is the ASI proposal.

The automatic generic method proposal, $\text{q}\left(\theta_\gamma,\theta_\gamma'\middle|\gamma,\gamma'\right)$, assumes that there is a normal approximation to the posterior distribution of θ_γ with mean $\hat{\beta}_\gamma$ and variance-covariance matrix Σ_γ. The approximation is found using the iteratively reweighted least squares algorithm described in Gammerman (1997) . The starting point takes θ_γ and removes dimensions for which $\gamma_i = 1$ and $\gamma_i' = 0$ and adds zero in dimensions for which $\gamma_i = 0$ and $\gamma_i' = 1$. We will denote the mean of the approximation for γ' as $\hat{\beta}_{\gamma'}$ and the variance-covariance matrix as $\Sigma_{\gamma'}$. Let $\text{C}(\Sigma)$ represent the Cholesky decomposition of the matrix Σ and recall that $p_\gamma = \sum_{i=1}^{p}\gamma_i$ is the dimension of model γ_i. The automatic generic proposal is

$$\theta_\gamma' = \begin{cases} \mu' + C(\Sigma_{\gamma'})(v)_1^{p_{\gamma'}+1} & \text{if } p_{\gamma'} < p_\gamma \\ \mu' + C(\Sigma_{\gamma'})v & \text{if } p_{\gamma'} = p_\gamma \\ \mu' + C(\Sigma_{\gamma'})\begin{bmatrix} v \\ u \end{bmatrix} & \text{if } p_{\gamma'} > p_\gamma \end{cases}$$

where $v = \left[C(\Sigma_\gamma)\right]^{-1}(\theta_\gamma - \hat{\beta}_\gamma)$, $(.)_1^m$ represents the first m components of a vector and $u\sim\text{N}\left(0,I_{p_{\gamma'}-p_\gamma}\right)$. The use of this proposal leads to the acceptance probability

$$\min\left\{1,\frac{\pi(y|\gamma',\theta_{\gamma'})q(\gamma',\gamma)|c(\Sigma_{\gamma'})|\text{N}\left(u\middle|0,I_{p_\gamma-p_{\gamma'}}\right)}{\pi(y|\gamma,\theta_\gamma)q(\gamma,\gamma')|c(\Sigma_\gamma)|}\right\} \quad \text{if } p_{\gamma'} < p_\gamma$$

$$\min\left\{1,\frac{\pi(y|\gamma',\theta_{\gamma'})q(\gamma',\gamma)|c(\Sigma_{\gamma'})|}{\pi(y|\gamma,\theta_\gamma)q(\gamma,\gamma')|c(\Sigma_\gamma)|}\right\} \quad \text{if } p_{\gamma'} = p_\gamma$$

$$\min\left\{1,\frac{\pi(y|\gamma',\theta_{\gamma'})q(\gamma',\gamma)|c(\Sigma_{\gamma'})|}{\pi(y|\gamma,\theta_\gamma)q(\gamma,\gamma')|c(\Sigma_\gamma)|\text{N}\left(u\middle|0,I_{p_\gamma-p_{\gamma'}}\right)}\right\} \quad \text{if } p_{\gamma'} > p_\gamma$$

Accelerated failure time model for survival outcome

The accelerate failure time (AFT) can be used to model censored survival outcomes. This is a parametric survival model that assumes that the individual survival time depends on the multiplicative effect of an unknown function of regressors over a baseline survival time . We will follow Sha et al (2006) by using this model for Bayesian variable selection with censored outcomes, where missing times are easy to sample. Here, we consider parametric AFT models under normal and t distributional assumptions for ϵ_i. The AFT model (on the log-scale) can be written as,

$$\log(t_i) = \alpha + x_{\gamma,i}\beta_\gamma + \sigma\epsilon_i, \qquad i = 1,\dots,n$$

where t_i is the survival time, α is the intercept term, $x_{\gamma,i}$ is a $(p_\gamma\times 1)$–dimensional regressor vector, β_γ is a $(p_\gamma\times 1)$–dimensional vector of regression coefficients and the errors $\epsilon_i\sim$F where F is a standardized distribution such as the standard normal or t distribution.

We assume that some observations have been (right) censored at time $c_i < t_i$ so that we observe $t_i^* = \min(t_i,c_i)$ and $\delta_i = I(t_i > c_i)$. We define $w_i = \log(t_i)$. The AFT model can then be expressed using data augmentation [Tanner et al (1987)] by imputing the censored times $w_i = \log(t_i^*)$ if $\delta_i = 1$ and $w_i > \log(t_i^*)$ if $\delta_i = 0$.

This leads to the linear regression model

$$W = \alpha 1 + X_\gamma\beta_\gamma + \sigma\epsilon$$

where $\epsilon = (\epsilon_1,\dots,\epsilon_n)'$.

If $\epsilon_i\sim N(0\,1)$, we can update γ using the modified ASI algorithm. To update w_i, we simulate σ^2 and β_γ using

$$\sigma^{-2}\sim\text{Ga}\left(n/2,\left(W'W - W'X_\gamma\left(X_\gamma'X_\gamma + V_\gamma\right)^{-1}X_\gamma'W\right)/2\right)$$

and

$$\beta_\gamma\sim\text{N}\left(\left(X_\gamma'X_\gamma + V_\gamma\right)^{-1}X_\gamma'W,\sigma^2\left(X_\gamma'X_\gamma + V_\gamma\right)^{-1}\right).$$

This allows w_i's if $\delta_i = 0$ to be updated by $w_i\sim\text{TN}_{x>t_i^*}\left(x_{\gamma,i}\beta_\gamma,\sigma^2\right)$ where $\text{TN}_{x>a}(\mu,\sigma^2)$ represents a normal distribution with mean μ and variance σ^2 truncated to (a,∞).

If $\epsilon_i\sim t_\nu$, we can write

$$W = \alpha 1 + X_\gamma\beta_\gamma + \sigma S\epsilon$$

Where $s = \text{diag}(s_1^2,\dots,s_n^2)$ and $s_i^{-2}\sim\text{Ga}(\nu/2,\nu/2)$. We can update γ using the modified ASI algorithm with the marginal likelihood

$$p(y|\gamma,s_1^2,\dots,s_n^2)$$
$$= \prod_{i=1}^{n}\left|V_\gamma\right|^{-1/2}\left|X_\gamma'SX_\gamma + V_\gamma^{-1}\right|^{-1/2}\exp\left\{\frac{1}{2}W'SX_\gamma\left(X_\gamma'SX_\gamma + V_\gamma^{-1}\right)^{-1}X_\gamma'SW\right\}$$

To update w_i, we simulate $\sigma^2$and β_γ using

$$\sigma^{-2}\sim\text{Ga}\left(n/2,\left(W'SW - W'SX_\gamma\left(X_\gamma'SX_\gamma + V_\gamma\right)^{-1}X_\gamma'SW\right)/2\right)$$

and

$$\beta_\gamma\sim\text{N}\left(\left(X_\gamma'SX_\gamma + V_\gamma\right)^{-1}X_\gamma'SW,\sigma^2\left(X_\gamma'SX_\gamma + V_\gamma\right)^{-1}\right).$$

This allows w_i's if $\delta_i = 0$ to be updated by $w_i\sim\text{TN}_{x>t_i^*}\left(x_{\gamma,i}\beta_\gamma,s_i^2\sigma^2\right)$ where $\text{TN}_{x>a}(\mu,\sigma^2)$ represents a normal distribution with mean μ and variance σ^2 truncated to (a,∞). The scales s_i^2 can be updated using $s_i^{-2}\sim\text{Ga}\left((\nu+1)/2,\left[\nu + \left(w_i - \alpha - X_{\gamma,i}\beta_\gamma\right)^2\right]/2\right)$.

Simulation result

We generate some simulate data based on the logistic regression models (which will appear in the manuscript). We compared the performance of the Add-Delete-Swap (ADS) algorithm, ASI algorithm with IRLS and ASI algorithm with Polya-Gamma. The performance of the ADS algorithm is relatively consistent for all data sets . The ASI algorithm with IRLS updating is more efficient than the ADS algorithm for all data sets but generally performs better as n increases. The ASI algorithm with Polya-Gamma updating is always outperformed by the ASI algorithm with IRLS updating and sometimes by the ADS algorithm. This is due to the introduction of latent variables which slow the mixing of the algorithm.

Application

The methodologies were applied to two datasets: a Medulloblastoma data set for binary outcome, and a breast cancer data set for survival outcome.

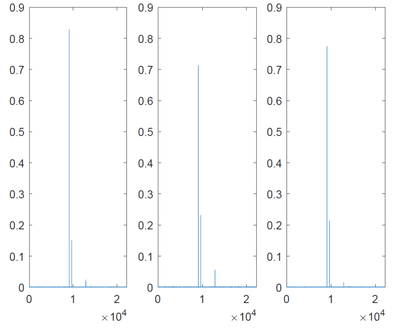
Medulloblastoma data

- The fresh frozen medulloblastoma tumor samples of 40 sonic hedgehog (SHH) and 147 non-SHH (NON) tumors were available, which consists of ~22,000 probes.

Table 1 – Medulloblastoma data: genes associated with group

probesetid	GeneSymbol	$p(y_j = 1\|X)$
209621_s_at	PDLIM3	0.772
210170_at	PDLIM3	0.198

Figure 1 – Medulloblastoma data: posterior inclusion probabilities with different MCMC runs



Conclusions

- We have extended the method by Griffin et al (2018) to binary and time-to-event endpoints framework via data argument approach.
- The algorithm is capable of identifying related regressors to clinical outcome with ~25,000 variables in a reasonable amount of time.
- The ASI algorithm with IRLS updating is more efficient than the ADS and ASI algorithm with Polya-Gamma updating in the simulation examples.

Acknowledgements

We thank Stuart Bailey and Matt Whiley for their discussions.

Poster presented at International Society for Bayesian Analysis (ISBA) 25 June - 29 June 2018, Edinburgh, United Kingdom.

References

1. Griffin JE, et al. In Search of Lost (Mixing) Time: Adaptive Markov chain Monte Carlo schemes for Bayesian variable selection with very large p. 2018 (In prep).
2. Polson NG, et al. Bayesian Inference for Logistic Models Using Polya-Gamma Latent Variables. JASA 2013; Vol. 108, No.504:1339–1349.

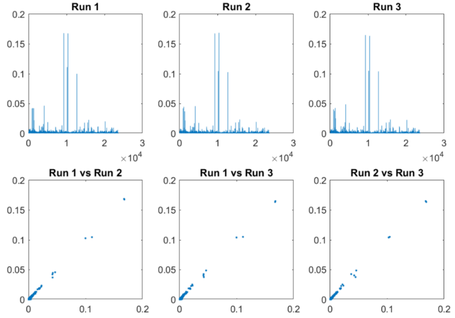
Breast cancer data

- We consider each patient's failure time as the outcome of interest as in Sha et al (2006). Patient who did not experience distant metastases within the five years constitute censored cases. This is differ from van't Veer et al. (2012), where they tackled the analysis as a classification problem.
- The gene expression levels were monitored using two-channel arrays with ~25,000 probes. Transcript abundance of genes were estimated using the intensity ratio with respect to a reference pool obtained by combining cRNA samples from all tissues.
- Two patients had several missing gene expression levels and were removed from the analysis.
- In van't Veer et al. (2002) a pre-processing of the data was conducted before they applied their method. Here, we considered all genes (~25,000) for the analysis.

Table 2 - Breast cancer data: genes associated with time to distant metastasis

Systematic name	Gene name	$p(y_j = 1\|X)$
Contig48328_RC		0.168
NM_020974	CEGP1	0.111
AL080059		0.168
NM_006681	NMU	0.100

Figure 2 – Breast cancer data: posterior inclusion probabilities with different MCMC runs



3. Sha N, et al. Bayesian variable selection for the analysis of microarray data with censored outcomes. Bioinformatics 2006;Vol. 22 no.18:2262–2268.
4. Tanner MA, et al. The Calculation of Posterior Distributions by Data Augmentation. JASA 1987; Vol. 82. No. 398: 528-540.
5. Gamerman, D. Sampling from the Posterior Distribution in Generalized Linear Models. Stat. and Comp. 1997; Vol. 7, 57-68.
6. Lamnisos, D. Transdimensional Sampling Algorithms for Bayesian Variable Selection in Classification Problems With Many More Variables Than Observations. Journal of Computational and Graphical Statistics. 2009;Vol.18. No.3 592-612.
7. Green, PJ. Trans-dimensional Markov chain Monte Carlo. Oxford University Press, 179-206.
8. van't Veer LJ, et al. Gene expression profiling predicts clinical outcome of breast cancer. Nature 2002; Vol. 415. 530-536.