Statistics and Computing (2021) 31:6
https://doi.org/10.1007/s11222-020-09974-2

An adaptive MCMC method for Bayesian variable selection in logistic and accelerated failure time regression models

Kitty Yuen Yi Wan[1] · **Jim E. Griffin**[2]

Received: 25 December 2019 / Accepted: 4 December 2020 / Published online: 12 January 2021

Abstract
Bayesian variable selection is an important method for discovering variables which are most useful for explaining the variation in a response. The widespread use of this method has been restricted by the challenging computational problem of sampling from the corresponding posterior distribution. Recently, the use of adaptive Monte Carlo methods has been shown to lead to performance improvement over traditionally used algorithms in linear regression models. This paper looks at applying one of these algorithms (the adaptively scaled independence sampler) to logistic regression and accelerated failure time models. We investigate the use of this algorithm with data augmentation, Laplace approximation and the correlated pseudo-marginal method. The performance of the algorithms is compared on several genomic data sets.

Keywords Pólya-gamma sampling · Correlated pseudo-marginal method · High-dimensional regression · Gene expression · Laplace approximation · Data augmentation

1 Introduction

The availability of large-scale data sets has led to interest in variable selection for regression models with large number of regressors. Typically, these variable selection problems are called "large p, small n" variable selection problems. Standard approaches to this problem include penalized maximum likelihood methods (Hastie et al. 2015) and Bayesian variable selection (Chipman et al. 2001; O'Hara and Sillanpää 2009; García-Donato and Martínez-Beneito 2013). Linear regression is the mostly widely studied problem in the literature, but other generalized linear models play an important role in answering challenging scientific questions. For example, Sanyal et al. (2017) use an iterative Bayesian procedure to analyse a case–control study with 55,000 SNPs and Nikooienejad et al. (2020) considered Bayesian variable selection applied to two cancer survival data sets with 13,267 and 13,335 genes, respectively.

In Bayesian variable selection, a joint prior distribution is placed on models defined by each possible subset of the potential regressors and the parameters of each of these models (the regression coefficients and other parameters such as dispersion parameters). This defines a posterior distribution on the parameters of the model and the models which can be used to investigate the importance of different variables or to make predictions for future observations. Working with this posterior distribution is challenging since (1) different models are defined on parameter spaces with different sizes and (2) there are 2^p possible models for p potential regressors which leads to a vast space if p is large. The first challenge can be circumvented in the linear regression model by working with the marginal likelihood of the models (which is available analytically for commonly used prior distributions), but this is not possible in other generalized linear models. As discussed by García-Donato and Martínez-Beneito (2013), these issues can be addressed by sampling from the posterior distribution using Markov chain Monte Carlo algorithms which provide unbiased estimates of quantities of interest such as the posterior inclusion probability (PIP) for the jth variable, which is the marginal posterior probability that the jth variable is included in the model, or Bayesian model-averaged predictions. Designing MCMC algorithms for Bayesian variable selection which mix well is a computationally challenging task if p is large and a large literature has developed around different approaches (see *e.g.*

✉ Jim E. Griffin
j.griffin@ucl.ac.uk

[1] Novartis Pharma AG, Basel, Switzerland

[2] Department of Statistical Science, University College London, London, UK

Schäfer and Chopin 2013; Titsias and Yau 2017; Shin et al. 2018; Zanella and Roberts 2019).

In this paper, we will concentrate on computational methods for Bayesian variable selection in logistic regression and accelerated failure time models. In this case, the marginal likelihood of the models is not available analytically. There are three main approaches to addressing this issue. Firstly, the marginal likelihood can be approximated, usually with a Laplace approximation, to define an approximated posterior which is sampled using MCMC. Secondly, data augmentation methods introduce latent variables, whose marginal distribution given the model only can be calculated analytically. This allows an MCMC scheme to be used where the latent variables are updated from their full conditional and the model is updated using methods for linear regression models. Thirdly, reversible jump MCMC can be used to work directly on the joint posterior distribution of the model and the parameters which moves between models by proposing a new set of regression coefficients for the proposed model.

In binary data, there has been extensive work on variable selection in probit regression (see *e.g.* Sha et al. 2003, 2004) using the data augmentation approach of Albert and Chib (1993), which was extended to logistic regression by Holmes and Held (2006). Nikooienejad et al. (2016) use non-local prior and a Laplace approximation to define a Gibbs sampler. In time-to-event data, work has been divided between accelerated failure time (AFT) models and Cox regression models. Sha et al. (2006) initially demonstrated how MCMC with data augmentation could be used for Bayesian variable selection when the AFT model has a normal or t distribution. Newcombe et al. (2017) consider an AFT model with a Weibull distribution and propose a reversible jump MCMC sampler. Zhang et al. (2018) use a Dirichlet process mixture of normals for the error distribution and use a Bayesian lasso to induce sparsity in the regression coefficients. Held et al. (2016) review previous work on Bayesian variable selection in Cox regression models and develop a method based on test Bayes factor to derive a posterior distribution on models. Many authors have concentrated on finding high probability models. Nikooienejad et al. (2020) use a non-local prior for the regression coefficients and use a Laplace approximation with the S5 algorithm (Shin et al. 2018). Annest et al. (2009) use an iterative screening algorithm. Duan et al. (2018) derive an EM algorithm to find high probability model using the framework of EMVS (Rockova and George 2014).

This paper makes three main contributions. Firstly, we extend the Adaptively Scaled Individual (ASI) adaptive MCMC algorithm of Griffin et al. (2020) to binary response (using logistic regression models) and time-to-event data (using accelerated failure time models). Secondly, we propose a correlated pseudo-marginal scheme for GLMs. Thirdly, we compare the performance of the adaptive MCMC algorithm to the Add–Delete–Swap Metropolis–Hastings sampler using correlated pseudo-marginal methods, data augmentation and the Laplace approximation in the large p setting for both logistic regression and accelerated failure time models.

The paper is organized in the following way. Section 2 reviews the Bayesian approach to variable selection in generalized linear models. Section 3 describes different computational strategies for Bayesian variable selection. Section 4 compares the performance of the methods on different real data sets with many regressors and few observations. Section 5 discusses the work and offers some guidelines for the use of the algorithm.

2 Generalized linear models and Bayesian variable selection

We assume that there are p variables available (for example, a list of SNPs or gene expression measurements) and that we wish to find a subset of these variables which explains the variation in a response. We define $\gamma = (\gamma_1, \ldots, \gamma_p) \in \Gamma$ to be a vector of indicator variables with $\gamma_j = 1$ if the jth variable is included in the model (and $\gamma_j = 0$ otherwise). Let y be an $(n \times 1)$-dimensional vector of responses which is modelled by

$$y_i \sim F(\mu_i, \phi) \tag{1}$$

where $F(\mu, \phi)$ is a distribution in the exponential family with mean μ and dispersion parameter ϕ. We define the linear predictor for the ith observation to be

$$\eta_i = z_i\alpha + x_i^{(\gamma)}\beta_\gamma \tag{2}$$

where z_i is a set of regressors which always appear in the model (and which will usually include an intercept and could include a treatment effect variable or other important variables), $x_i^{(\gamma)}$ is a vector which contains the value for the ith observation of the variables included in the model (*i.e.* for which $\gamma_j = 1$), and α and β_γ are vectors of regression coefficients. We will write X to be a matrix whose rows are the value for the ith observation of all variables, and Z and $X^{(\gamma)}$ to be matrix whose rows are z_i and $x_i^{(\gamma)}$, respectively. The linear predictor η_i is linked to the mean μ_i by $\eta_i = g(\mu_i)$ where g is a link function. We define p_α to be the dimension of α and $p_\gamma = \sum_{j=1}^{p} \gamma_j$.

The logistic regression model can be used to link a proportion of successes to the linear predictors using a generalized linear model. We let y_i be the proportion of success and assume that $n_i y_i \sim \text{Bin}(n_i, p_i)$. The logistic regression model assumes a linear relationship between the covariates and the probability of success (measured on the log-odds

scale),

$$\eta_i = \log\left(\frac{p_i}{1-p_i}\right) = z_i\alpha + x_i^{(\gamma)}\beta_\gamma, \quad i = 1, \ldots, n.$$

The accelerate failure time (AFT) can be used to model time-to-event data which may be censored. We assume that the time-to-event for the ith individual is t_i but that this time is only observed if $t_i < c_i$ for some (right) censoring time c_i and, otherwise, we observe c_i. We will define $\delta_i = \mathrm{I}(t_i > c_i)$ which is 1 if the ith observation is censored. The AFT model (on the log-scale) can be written as

$$y_i = \log t_i = z_i\alpha + x_i^{(\gamma)}\beta_\gamma + \sigma\epsilon_i, \qquad i = 1, \ldots, n \tag{3}$$

where the errors $\epsilon_i \overset{i.i.d.}{\sim} G$ for a standardized distribution G, such as the standard normal or t distribution. This is a parametric survival model that assumes that the individual survival time t_i depends on the multiplicative effect of an unknown function of covariates over a baseline survival time α. We will follow Sha et al. (2006) by using this model for Bayesian variable selection with censored outcomes.

In Bayesian variable selection, a prior distribution is assumed for the parameters of the model (α, β_γ, ϕ and γ) which defines a posterior distribution

$$p(\alpha, \beta_\gamma, \phi, \gamma|X, Z, y) \propto p(y|Z, X^{(\gamma)}, \alpha, \beta_\gamma, \phi, \gamma)\, p(\alpha, \beta_\gamma, \phi, \gamma).$$

In some probability models, such as the binomial distribution, the dispersion parameter is known and so ϕ does not appear in the prior or posterior distribution leading to

$$p(\alpha, \beta_\gamma, \gamma|X, Z, y) \propto p(y|Z, X^{(\gamma)}, \alpha, \beta_\gamma, \gamma)p(\alpha, \beta_\gamma, \gamma).$$

We will assume the commonly used prior structure

$$p(\alpha, \beta_\gamma, \gamma|\phi) \propto p(\beta_\gamma \mid \phi, \gamma)\, p(\gamma) \tag{4}$$

with $\beta_\gamma \mid \sigma^2, \gamma \sim \mathrm{N}(0, \phi V_\gamma)$, and $p(\gamma) = h^{p_\gamma}(1-h)^{p-p_\gamma}$. If the dispersion parameter is unknown, an additional prior distribution is placed on ϕ. The hyperparameter $0 < h < 1$ is the prior probability that a particular variable is included in the model and can be chosen by defining a prior expected model size, p_0, by $h = \frac{p_0}{p}$. The prior can be further extended with hyperpriors to define a heavier tailed prior distribution for p_γ by assuming that $h \sim \mathrm{Be}(1, \frac{p-p_0}{p_0})$ (Ley and Steel 2009). The scaled covariance matrix V_γ is often chosen as proportional to the identity matrix (implying conditional prior independence between the regression coefficients), $(X_\gamma^T X_\gamma)^{-1}$ (a g-prior) or mixtures of g-priors (Liang et al. 2008; Li and Clyde 2018). Many computational methods, and all methods described in this paper, can be easily adjusted to works with any of these priors.

3 Computational approaches

In this section, we will concentrate on computational methods for Bayesian variable selection in two generalized linear models: logistic regression (for binary and some ordinal responses) and accelerated failure time (AFT) models (for time to response). We will write θ_γ for the parameters of model $\gamma \in \Gamma$ (which will be α, β_γ and ϕ if the dispersion is unknown and α and β_γ otherwise). In the linear regression models with normal errors, the marginal likelihood $p(y|\gamma)$ is analytically available. This leads to a simple Metropolis–Hastings updating step where a proposed γ' is sampled from a proposal where the probability of proposing γ' given current value γ is $q(\gamma, \gamma')$. The proposed model is accepted with probability

$$\alpha = \min\left\{1, \frac{p(y|\gamma')p(\gamma')q(\gamma', \gamma)}{p(y|\gamma)p(\gamma)q(\gamma, \gamma')}\right\}. \tag{5}$$

In GLMs, the marginal likelihood is not directly analytically available. We will describe three approaches: the data augmentation methods where latent variables ω are introduced that allow $p(y|\omega, \gamma)$ to be calculated analytically, the Laplace approximation where $p(y|\gamma)$ is approximated leading to approximation error in the posterior and the correlated pseudo-marginal methods where $p(y|\gamma)$ is approximated but leads to no approximation error in the posterior. These methods can be used with any proposal on model space. We will then describe how the ASI proposal (Griffin et al. 2020) can be used with these approaches.

3.1 Algorithms

3.1.1 Data augmentation

Data augmentation (Tanner and Wong 1987) approaches introduce latent variables to make an MCMC sampler simpler to implement. In Bayesian variables, these schemes introduce a fixed-dimension latent variable ω which allows $p(y|\gamma, \omega)$ to be calculated analytically. This leads to a simple MCMC scheme where γ can be updated conditional on ω using a Metropolis–Hastings sampler with standard proposal distribution on model space and ω is updated conditional on γ (often by first simulating the parameters θ_γ).

In the logistic regression model, the Pólya-gamma data augmentation method (Polson et al. 2013) can be used (see

e.g. Griffin et al. 2019). This exploits the following identity

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

where $\kappa = a - b/2$, $\omega_i \sim \mathrm{PG}(b, 0)$ and PG represents a Pólya-gamma distribution, which is defined in Polson et al. (2013). An extended likelihood with additional latent variables $\omega = (\omega_1, \ldots, \omega_n)$ has the form

$$p(y, \omega|\theta_\gamma, \gamma) \propto \prod_{i=1}^{n}\left[2^{-n_i}\exp\left\{\kappa_i\left(z_i\,\alpha + x_i^{(\gamma)}\beta_\gamma\right)\right\}\right.$$
$$\left.\times \exp\left\{-\omega_i(z_i\,\alpha + x_i^{(\gamma)}\beta_\gamma)^2/2\right\}\, p(\omega_i)\right]$$

where $\kappa_i = y_i - n_i/2$ and $p(\omega_i)$ is the $\mathrm{PG}(n_i, 0)$ distribution. The identity can be used to show that marginalizing this distribution over ω leads to the likelihood of the logistic regression model. If we let

$$J^{(\gamma)} = \begin{pmatrix} Z \\ X^{(\gamma)} \end{pmatrix}$$

and assume that $p(\alpha, \beta_\gamma) \sim \mathrm{N}(\mu_\gamma, V_\gamma)$, the marginal likelihood is

$$p(y|\gamma, \omega) \propto |V_\gamma|^{-1/2}\left|\tilde{J}^{(\gamma)\,T}\tilde{J}^{(\gamma)} + V_\gamma^{-1}\right|^{-1/2}$$
$$\times \exp\left\{-\frac{1}{2}\mu_\gamma^T V_\gamma^{-1}\mu_\gamma + \frac{1}{2}A^T B^{-1}A\right\}$$

where $A = J^{(\gamma)\,T}K + V_\gamma^{-1}\mu_\gamma$, $B = \tilde{J}^{(\gamma)\,T}\tilde{J}^{(\gamma)} + V_\gamma^{-1}$, $\tilde{J}^{\gamma}$ is $(n \times (p_\alpha + p_\gamma))$-dimensional matrix with entries $\tilde{J}_{i,j}^{(\gamma)} = \sqrt{\omega_i}\tilde{X}_{i,j}^{(\gamma)}$ and K is a $(n \times 1)$-dimensional vector with entries $K_i = \kappa_i$. The latent variables $\omega = (\omega_1, \ldots, \omega_n)$ can be updated by first sampling α, β_γ according to

$$(\alpha, \beta_\gamma) \sim \mathrm{N}\left((\tilde{J}^{(\gamma)\,T}\tilde{J}^{(\gamma)} + V_\gamma^{-1})^{-1}(J^{(\gamma)\,T}K + V_\gamma^{-1}\mu_\gamma),\right.$$
$$\left.(\tilde{J}^{(\gamma)\,T}\tilde{J}^{(\gamma)} + V_\gamma^{-1})^{-1}\right)$$

and then sampling $\omega_i \sim \mathrm{PG}(n_i, z_i\,\alpha + x_i^{(\gamma)}\beta_\gamma)$. Polson et al. (2013) describe efficient algorithms for the generation of Pólya-gamma distributed random variables.

In the accelerated failure time model, we introduce the variables $\omega_i = y_i$. These are latent if $\delta_i = 1$, *i.e.* ω_i is the missing survival time if the ith observation is censored. Conditional on the data and $\omega = (\omega_1, \ldots, \omega_n)$, the accelerated failure time model in (3) is a linear regression. If $\epsilon_i \sim \mathrm{N}(0, 1)$, we have a linear regression with normal errors. If we define W to be an $(n \times 1)$-dimensional vector with ith entry ω_i and assume that $p(\alpha, \beta_\gamma) \sim \mathrm{N}(\mu_\gamma, \sigma^2 V_\gamma)$, the marginal likelihood is

$$p(y|\gamma, \omega_1, \ldots, \omega_n) = |V_\gamma|^{-1/2}\left|J^{(\gamma)\,T}J^{(\gamma)} + V_\gamma^{-1}\right|^{-1/2}$$
$$\times \exp\left\{-\frac{1}{2}\mu_\gamma^T V_\gamma^{-1}\mu_\gamma + \frac{1}{2}A^T B^{-1}A\right\}$$

where $A = (J^{(\gamma)\,T}W + V_\gamma^{-1}\mu_\gamma)$ and $B = J^{(\gamma)\,T}J^{(\gamma)} + V_\gamma^{-1}$

The latent variables ω_i for $\delta_i = 1$ can be updated by first sampling $(\alpha, \beta_\gamma, \sigma^2)$. If we assume that $\sigma^{-2} \sim \mathrm{Ga}(a/2, b/2)$, then

$$\sigma^{-2} \sim \mathrm{Ga}\left((a + n + p_\alpha + p_\gamma)/2, \left(b + \mu_\gamma^T V_\gamma^{-1}\mu_\gamma\right.\right.$$
$$\left.\left.+W^T W - A^T B^{-1}A\right)/2\right)$$

and

$$(\alpha, \beta_\gamma) \sim \mathrm{N}\left(B^{-1}A, \sigma^2 B^{-1}\right).$$

The latent variable ω_i for $\delta_i = 1$ can be sampled from its full conditional distribution which is $\mathrm{TN}_{[c_i,\infty)}(z_i\,\alpha + x_i^{(\gamma)}\beta_\gamma, \sigma^2)$ where $\mathrm{TN}_A(\mu, \sigma^2)$ represents a normal distribution with mean μ and variance σ^2 truncated to the set $A \in \mathbb{R}$.

3.1.2 Laplace approximation

The Laplace approximation has been widely used for variable selection in generalized linear models. The marginal likelihood is approximated by

$$p(y|\gamma) \propto p\left(y\left|\hat{\theta}_\gamma\right.\right) p\left(\hat{\theta}_\gamma\right) |-H_\gamma|^{-1/2}(2\pi)^{d/2}$$

where d is the dimension of θ_γ, $\hat{\theta}$ is the posterior mode of θ_γ and H_γ is the Hessian of $\log p\left(y|\theta_\gamma\right) + \log p\left(\theta_\gamma\right)$ evaluated at $\hat{\theta}_\gamma$. The approximation follows from assuming that the posterior distribution of θ_γ can be approximated by

$$p_{\mathrm{Laplace}}\left(\theta_\gamma\right) = \mathrm{N}\left(\hat{\theta}_\gamma, \Sigma_{\mathrm{Laplace}}^{(\gamma)}\right) \tag{6}$$

where $\Sigma_{\mathrm{Laplace}}^{(\gamma)} = \left(-H_\gamma\right)^{-1}$.

3.1.3 Correlated pseudo-marginal sampler

The pseudo-marginal sampler (Andrieu and Roberts 2009) targets a distribution where the prior is multiplied by a Monte Carlo approximation $\hat{p}(y|\gamma)$ of the intractable marginal likelihood $p(y|\gamma)$ and then runs a Metropolis–Hastings sampler on this target distribution, *i.e.* the acceptance rate of the

Metropolis–Hastings sampler is

$$\alpha = \min\left\{1, \frac{\hat{p}(y|\gamma')\, p(\gamma')\, q(\gamma', \gamma)}{\hat{p}(y|\gamma)\, p(\gamma)\, q(\gamma, \gamma')}\right\}.$$

This would be the usual acceptance rate if $\hat{p}(y|\gamma) = p(y|\gamma)$, but Andrieu and Roberts (2009) show that the method samples from the correct target distribution with the weaker condition that $\mathrm{E}[\hat{p}(y|\gamma)] = p(y|\gamma)$ (unbiased estimation). The method has been extended to the correlated pseudo-marginal method (Deligiannidis et al. 2018) were the random numbers used to calculate $\hat{p}(y|\gamma')$ are positively correlated with those used to calculate $\hat{p}(y|\gamma)$. They show that this can reduce the variance of the ratio $\hat{p}(y|\gamma')/\hat{p}(y|\gamma)$ and help the mixing of the Markov chain.

In Bayesian variable selection for GLMs, it is natural to use the Laplace approximation of $p(\theta_\gamma|\gamma, y)$ in (6) as an importance sampling distribution in an importance sampling approximation to $\hat{p}(y|\gamma)$. The approximation is

$$\hat{p}(y|\gamma) = \frac{1}{N}\sum_{i=1}^{N} \frac{p\left(y\left|\theta_\gamma^{(i)}, \gamma\right.\right)\, p\left(\theta_\gamma^{(i)}\right)}{p_{\text{Laplace}}\left(\theta_\gamma^{(i)}\right)}$$

where $\theta_\gamma^{(i)} \overset{i.i.d.}{\sim} p_{\text{Laplace}}$. The samples $\theta_\gamma^{(i)}$ can be written as $\theta_\gamma^{(i)} = \hat{\theta}_\gamma + C_\gamma \Gamma \nu_i$ where Γ is a $p_\gamma \times p$-dimensional matrix, where $\Gamma_{i,j} = 1$ if the i-th variable included in the model has index j and $\Gamma_{i,j} = 0$ otherwise, C_γ is the Cholesky decomposition of $\Sigma^{(\gamma)}_{\text{Laplace}}$ and $\nu_i \overset{i.i.d.}{\sim} \mathrm{N}\left(0, I_p\right)$. A correlated pseudo-marginal sampler can be implemented in the following way. At iteration k, suppose that $\nu_i, \ldots, \nu_N$ are the current values of the random variates before making the proposal, then propose $\nu_1', \ldots, \nu_N'$ by

$$\nu_{i,k}' = \rho \nu_{i,k} + \sqrt{1-\rho^2}\lambda_{i,k}$$

where $\lambda_{i,k} \overset{i.i.d.}{\sim} \mathrm{N}(0, 1)$. This implies that η_i follows an autoregressive process of lag 1 with AR parameter ρ and a standard normal stationary distribution. In practice, η_i' only need to be simulated if $\gamma_k' = 1$ and, if $\gamma_k = 0$ and $\gamma_k' = 1$, we can simulated $\nu_{i,k} \sim \mathrm{N}(0, 1)$. Lamnisos et al. (2009) suggest using the automatic generic sampler (Green 2003) for logistic regression model with their adaptive proposal. The correlated pseudo-marginal is equivalent to this sampler when $N = 1$ and $\rho = 1$ (completed dependence between successive ν_i) if an extra Gibbs step is introduced where ν_i is updated from its full conditional distribution. We use this additional step for all values of ρ and so this correlated pseudo-marginal sampler provides a generalization of the automatic generic sampler for GLMs.

3.2 Adaptively scaled individual proposal

Griffin et al. (2020) develop a parameterized proposal distribution for Bayesian variable selection in linear regression models and a method for adaptively tuning the parameters during the MCMC run. They demonstrate that the method can mix substantially faster than the standard Add–Delete–Swap sampler. They also show that the method can accurately estimate the PIPs of each variable for a range of problems up to thousands of regressors in a reasonable amount of time. For example, they run their algorithm on two large data sets. One data set had 22,575 variables and provided accurate results in 25 min, whereas the other data set had 79,748 variables and provided accurate results in 2.5 h, respectively.

The proposal on model space, Γ, is

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

where $\eta = (A, D) = (A_1, \ldots, A_p, D_1, \ldots, D_p)$, $q_{\eta,j}(\gamma_j = 0, \gamma_j' = 1) = A_j$ and $q_{\eta,j}(\gamma_j = 1, \gamma_j' = 0) = D_j$. Each dimension of γ' is independently proposed conditional on γ. The probability of proposing to add the jth variable if it is currently excluded from the model is A_j, and the probability of proposing to delete the jth variable if it is currently included in the model is D_j. They show that an effective choice for (A, D) is

$$A_j = \zeta \min\left\{1, \frac{\pi_j}{1-\pi_j}\right\}, \qquad D_j = \zeta \min\left\{1, \frac{1-\pi_j}{\pi_j}\right\}, \tag{7}$$

where π_j is the PIP of the jth variable and $0 < \zeta < 1$ is a tuning parameter. They suggest estimating π_j within the Gibbs sampler using a Rao–Blackwellised estimate and tuning ζ using a simple adaptation scheme. The proposal is

$$A_j^{(i)} = \zeta^{(i)} \min\left\{1, \frac{\pi_j^{(i)}}{1-\pi_j^{(i)}}\right\},$$
$$D_j^{(i)} = \zeta^{(i)} \min\left\{1, \frac{1-\pi_j^{(i)}}{\pi_j^{(i)}}\right\},$$

where $\pi_j^{(i)}$ is a Rao–Blackwellised estimate of the PIP of the jth variable calculate using the first i sample and $\zeta^{(i)}$ is updated using

$$\mathrm{logit}_\epsilon\left(\zeta^{(i+1)}\right) = \mathrm{logit}_\epsilon\left(\zeta^{(i)}\right) + \phi_i\left(a_{\eta^{(i)}}\left(\gamma^{(i)}, \gamma'\right) - \tau\right), \tag{8}$$

where $\text{logit}_\epsilon(x) = \log(x-\epsilon) - \log(1-x-\epsilon)$, $\phi_i = O(i^{-\lambda})$ for some constant $1/2 < \lambda \leq 1$ $a_{\eta^{(i)}}(\gamma^{(i)}, \gamma')$ is the Metropolis–Hastings acceptance probability at the ith iteration, and τ is a targeted acceptance rate. The algorithm is summarized in Algorithm 1.

for $i = 1$ to $i = M$
  sample $\gamma' \sim q_{\eta^{(i)}}(\gamma^{(i)}, \cdot)$ and $U \sim U(0,1)$;
  if $U < a_{\eta^{(i)}}(\gamma^{(i)}, \gamma')$ then
    $\gamma^{(i+1)} := \gamma'$
  else
    $\gamma^{(i+1)} := \gamma^{(i)}$
  endif
  Update $\pi_1^{(i+1)}, \ldots, \pi_p^{(i+1)}$ and set $\tilde{\pi}_j^{(i+1)} = \epsilon + (1-2\epsilon)\pi_j^{(i+1)}$
  Update $\zeta^{(i+1)}$ as in (8)
  Calculate $A_j^{(i+1)} = \zeta^{(i+1)} \min\left\{1, \tilde{\pi}_j^{(i+1)} / \left(1 - \tilde{\pi}_j^{(i+1)}\right)\right\}$
  for $j = 1, \ldots, p$
  Calculate $D_j^{(i+1)} = \zeta^{(i+1)} \min\left\{1, \left(1 - \tilde{\pi}_j^{(i+1)}\right) / \tilde{\pi}_j^{(i+1)}\right\}$
  for $j = 1, \ldots, p$
endfor

Algorithm 1: Adaptively Scaled Individual Adaptation (ASI)

The ASI algorithm uses a Rao–Blackwellised estimate of the PIP's $\pi_1, \ldots, \pi_p$ calculated in the run of the algorithm. Griffin et al. (2020) show how the update of the Rao–Blackwellised calculated in $O(p)$ operations in the linear regression model. This can be directly extended to the data augmentation approach. We assume that $V_\gamma = \begin{pmatrix} V_\alpha & \mathbf{0} \\ \mathbf{0}^T & g I_{p_\gamma} \end{pmatrix}$, where V_α is $(p_\alpha \times p_\alpha)$-dimensional matrix, I_q is the $q \times q$ identity matrix and $\mathbf{0}$ is a $(p_\alpha \times p_\gamma)$-dimensional matrix of 0's. After N posterior samples, $\gamma^{(1)}, \ldots, \gamma^{(N)}$, the Rao–Blackwellised estimate of $\pi_j = p(\gamma_j = 1|y)$ is

$$\hat{\pi_j} = \frac{1}{N} \sum_{k=1}^{N} \frac{\tilde{h}_j^{(k)} \, \text{BF}_j\left(\gamma_{-j}^{(k)}\right)}{1 - \tilde{h}_j^{(k)} + \tilde{h}_j^{(k)} \, \text{BF}_j\left(\gamma_{-j}^{(k)}\right)}$$

where $\tilde{h}_j^{(k)} = h$ if h is fixed or $\tilde{h}_j^{(k)} = \frac{\#\gamma_{-j}^{(k)} + 1 + a}{p + a + b}$ if $h \sim \text{Be}(a, b)$ and $B = \tilde{J}_\gamma^T \tilde{J}_\gamma + V_\gamma^{-1}$. If $\gamma_j = 0$,

$$\text{BF}_j(\gamma_{-j}) = {d_j^\uparrow}^{-1/2} g^{-1/2} \exp\left\{\frac{1}{d_j^\uparrow}(\kappa^T x_j - \kappa^T J_\gamma B^{-1} \tilde{J}_\gamma^T \tilde{x}_j)^2\right\}$$

with $d_j^\uparrow = \tilde{x}_j^T \tilde{x}_j + g^{-1} - (\tilde{x}_j^T \tilde{J}_\gamma) B^{-1} (\tilde{J}_\gamma^T \tilde{x}_j)$ and $\tilde{x_j}$ is a $(n \times 1)$-dimensional vector with ith entry $\tilde{x_{i,j}} = \sqrt{\omega_i} x_{i,j}$. In the case $\gamma_j = 1$, it is useful to define q_j to be ordered position of the included variables ($q_j = 1$ if j is the first included variable, etc.); then,

$$\text{BF}_j(\gamma_{-j}) = {d_j^\downarrow}^{-1/2} g^{-1/2} \exp\left\{-\frac{1}{2} d_j^\downarrow (\kappa^T J_\gamma (B^{-1})_{\cdot, q_j + p_\alpha})^2\right\}$$

where $d_j^\downarrow = 1/(B^{-1})_{q_j + p_\alpha, q_j + p_\alpha}$. Calculating the Rao–Blackwellised estimate using the Laplace approximation is time-consuming since this would involve an optimization step to a posterior mode for each potential variable at each iteration of the sampler. Therefore, in both the Laplace approximation and correlated pseudo-marginal approaches, the ASI algorithm with data augmentation is run for N_0 iteration to calculated Rao–Blackwellised estimates of the PIP's. After this initial phase has been run, $\pi_1^{(N_0)}, \ldots, \pi_p^{(N_0)}$ are used as the estimate PIPs at every iteration and only ζ is adaptively updated in the sampler (and often for only a finite number of iterations of the sampler).

Adaptive MCMC algorithms do not necessarily lead to ergodic Markov chains and so care is needed in their design. Griffin et al. (2020) show that the ASI algorithm leads to an ergodic chains in a linear regression model. In our algorithms, adaptation only occurs for a fixed number of samples during the burn-in phase and so the algorithms are ergodic by design. However, it is interesting to think about whether these samplers are ergodic if adaptation is allowed to continue indefinitely. Roberts and Rosenthal (2007) set out two conditions for the ergodicity of adaptive MCMC algorithms. Firstly, the algorithm must have diminishing adaptation that is the adaptation of parameters tends to decrease as the sampler runs. This property is established for the ASI algorithm in Griffin et al. (2020). The second is containment which means that the transition kernel (for any value of the adaptive parameter) reaches stationarity in bounded time. This property can be easily established if the MCMC algorithm is uniformly ergodic. This property can be easily established by restricting the state space of the adaptive MCMC sampler to a (large) bounded subset of $\mathbb{R}^{p_\gamma}$, for example, by bounding the regression coefficients in the Pólya-gamma sampler or the sampled values in the importance sampler in the correlated pseudo-marginal sampler. The property can also be easily established if the sampler for Bayesian variable selection is run on the posterior distribution of the model γ only since the finite state space implies uniform ergodicity. If the marginal likelihood $p(y|\gamma)$ is approximated using an importance sampler, then the pseudo-marginal sampler is also uniformly ergodic (Theorem 8, Andrieu and Roberts 2009) if the weights in the importance sampler are bounded. The Pólya-gamma sampling scheme is also uniformly ergodic (Choi and Hobert 2013) which implies that the adaptive scheme is uniformly adaptive.

4 Comparison of computational algorithms

The effective sample size of N MCMC draws of a parameter θ is defined to be

$$\text{ESS}_\theta = \frac{N}{1 + 2\sum_{k=1}^{t} \hat{\rho}_k}$$

where $\hat{\rho}_k$ is the estimated lag k autocorrelation of the chain for θ and t is a suitably chosen threshold (see *e.g.* Liu 2001). The ergodic average calculated using the N MCMC samples has the same Monte Carlo error as an independent sampler with ESS_θ samples and so larger values of ESS_θ for fixed N imply a better mixing chain. The use of the ESS in this adaptive context is justified since N_0 is chosen to be smaller than the burn-in phase and ζ is only adapted during the burn-in phase. We measure the performance of each algorithm by calculating the median of

$$\text{ESS} = \text{median}_{j=1,\ldots,p}(\text{ESS}_{\gamma_j}).$$

This gives an overall measure of the mixing of the MCMC chain but does not account for differences in the time taken by different algorithms. To include time in the performance measure, we define the time-normalized effective sample size of algorithm A to be ESS_A/T_A where ESS_A and T_A are the effective sample size and time taken to generate the MCMC samples, respectively.

The results compare various algorithms:

- ASI-DA—ASI proposal with data augmentation sampling.
- ADS-DA—Add–Delete–Swap proposal with data augmentation sampling.
- ASI-Laplace—ASI proposal with a Laplace approximation of the marginal likelihood used directly (the Rao–Blackwellised estimates in the ASI proposal are calculated using the Pólya-gamma sampling scheme).
- ADS-Laplace—Add–Delete–Swap proposal with a Laplace approximation of the marginal likelihood used directly
- ASI-CPM—ASI proposal with the correlated pseudo-marginal sampler with a multivariate normal importance distribution.
- ADS-CPM—ADS proposal with the correlated pseudo-marginal sampler with a multivariate normal importance distribution.

The ASI-DA, ADS-DA, ASI-CPM and ADS-CPM will provide samples for the true posterior, whereas the ASI-Laplace and ADS-Laplace will sample from the posterior defined by the Laplace approximation. The best ESS/Time for the method which samples from the true posterior distribution is the ASI-DA (6.6) with similar performance from the correlated pseudo-marginal with normal importance, $\rho = 1$ and $N = 5$. The methods performs about 3 times better than their ADS counterparts. The ASI-Laplace provides the best ESS/Time but only by a small amount.

The correlated pseudo-marginal method should become more efficient than the PG sampling method when the sample size becomes larger since: 1) ESS will decreases for PG sampling relative to correlated pseudo-marginal (due to the introduction of more latent variables – one for every observation) and 2) increased sampling costs for the PG sampling relative to the correlated pseudo-marginal (again, due to the introduction of more latent variables).

4.1 Logistic regression

We begin by considering a simulation study which converts the linear regression simulation study of Yang et al. (2016) to logistic regression. They assume a linear predictor $\eta = X\beta^\star$ and generated observations with normal errors. In this simulation study, we use a logistic link to give $y_i = \exp\{\eta_i\}/(1 + \exp\{\eta_i\})$. Only the first 10 regression coefficients are nonzero with values

$$\beta^\star = \text{SNR}(2, -3, 2, 2, -3, 3, -2, 3, -2, 3, 0, \ldots, 0)^T \in \mathbb{R}^p$$

where SNR controls the signal-to-noise ratio. The ith vector of regressors is generated as $x_i \sim \text{N}(0, \Sigma)$ where $\Sigma_{jk} = \rho^{|j-k|}$. In our examples, we use the value $\rho = 0.6$ which represents a relative large correlation between the regressors. We ran a simulation study to compare the performance of the exact samplers for all combinations of $n = 500, 1000$ and $p = 500, 5000$. Each sampler was run for 105,000 iterations with a burn-in of 5000. The adaptive parameters were only adapted during the burn-in phase. The results are presented in Table 1. The CPM results correspond to $N = 1$ and $\rho = 1$ which provided the largest time-normalized effective sample sizes. The ASI algorithm outperforms the ADS algorithm for both the DA and CPM updating schemes for all simulated data sets apart from the data set with $n = 500$, $p = 5000$ and $\text{SNR} = 1$. This is a low information data set since it combines a small sample size, large number of regressors and a low signal-to-noise ratio (the ASI only marginally outperforms ADS in the lower signal-to-noise case where $\text{SNR} = 0.5$). The CPM methods always outperform the DA methods. The difference between the ASI method and the ADS method with CPM updating tends to be larger for more informative data sets (those with larger sample size, smaller numbers of regressors and larger signal-to-noise ratios). The posterior distribution in these cases is better able to distinguish between informative and uninformative variables in the regression and the ASI method is better able to exploit this information than the ADS sampler.

Table 1 Time-normalized effective sample size for various computational methods for the simulated logistic regression data sets

(n, p)	Adapt. Alg.	MCMC Alg.	SNR			
			0.5	1	2	3
(500, 500)	ASI	DA	2.3	2.5	1.5	1.5
	ADS	DA	1	1	1	1
	ASI	CPM	**10.2**	**9.1**	**7.0**	**9.1**
	ADS	CPM	5.3	3.6	3.0	2.2
(500, 5000)	ASI	DA	1.6	1.5	1.6	1.0
	ADS	DA	1	1	1	1
	ASI	CPM	**2.7**	2.8	**4.2**	**4.9**
	ADS	CPM	2.6	**3.1**	1.9	1.7
(1000, 500)	ASI	DA	2.2	2.2	1.5	5.8
	ADS	DA	1	1	1	1
	ASI	CPM	**14.2**	**14.0**	**9.4**	**51.9**
	ADS	CPM	8.6	8.4	3.4	3.0
(1000, 5000)	ASI	DA	1.4	1.9	2.1	1.4
	ADS	DA	1	1	1	1
	ASI	CPM	**5.1**	**4.6**	**5.9**	**7.8**
	ADS	CPM	4.0	3.2	2.7	2.3

The best performing method for each simulated data set is shown in bold

We consider four data sets which have been previously analysed in the literature on computational methods for Bayesian variable selection in logistic regression models with a number of regressors (see *e.g.* Lamnisos et al. 2009): Arthritis (Sha et al. 2003), Colon Tumor (Alon et al. 1999), Leukemia (Armstrong et al. 2002) and Prostate (Singh et al. 2002). The sample size and number of regressors for each data set are shown in Table 2. All variables were scaled to have a sample standard deviation of one. The prior distribution has the form $\alpha \sim \mathrm{N}(0, 100)$, $p(\beta_\gamma|\gamma) \sim \mathrm{N}(0, I)$, $\gamma_j \overset{i.i.d.}{\sim} \text{Bernoulli}(h)$ for $j = 1, \ldots, p$ where $h \sim \text{Be}\left(1, \frac{p-5}{p}\right)$. The samplers were run for 105,000 iterations with a burn-in period of 5000 iterations. All samples after the burn-in were used in the inference.

The time-normalized effective sample size for each method and each data set is shown in Table 2. In the correlated pseudo-marginal sampler, we find that $\rho = 1$ leads to the largest time-normalized effective sample size for both the ASI and ADS sampler for all data sets and we choose the number of random samplers in the importance sampler, N, which gives the largest value for each data set (with that value of N shown in Table). The results show some clear patterns. The ASI correlated pseudo-marginal sampler is no worse than the ADS correlated pseudo-marginal sampler for all data sets. The Pólya-gamma ASI sampler outperforms the Pólya-gamma ADS sampler for the smaller data sets (Arthritis, Colon Tumor and Leukemia) but not the largest data sets (Prostate). The correlated pseudo-marginal ASI outperforms the Pólya-gamma ADS for all data sets. The Laplace approximation methods tend to have larger time-normalized ESS than the exact methods (all data sets apart from Leukemia).

To understand the effects, it is also useful to look at effective sample sizes which are shown in Table 3. The relative ESS for the ASI over the ADS method in the correlated pseudo-marginal is much larger than for the Pólya-gamma sampler. This reflects the effect of introducing latent variables in the data augmentation which slows the ASI methods ability to make large jumps in model space. The effective sample size for the correlated pseudo-marginal and Laplace is very similar for both ASI and ADS for all data sets. The differences in the time-normalized ESS reflect the additional overhead of updating the regression coefficients and calculating the importance sampling approximation.

4.2 Accelerated failure time modelling

Bayesian variable selection and associated computational methods for survival analysis have been less developed than approaches for logistic regression in the literature. Consequently, there is no commonly used set of the data that is used in previous work. We consider two data sets. The first looks at survival following chemotherapy for diffuse large-b-cell lymphoma (Rosenwald et al. 2002) (DLBCL). The second breast cancer van't Veer et al. (2002). The sample size, number of regressors and the percentage of censored observations are shown in Table 4. In both cases, there are a large number of variables and a large amount of censoring. The prior distribution has the form $p(\alpha, \sigma^{-2}) \propto 1$, $\gamma_j \overset{i.i.d.}{\sim} \text{Bernoulli}(h)$ for $j = 1, \ldots, p$ where $h \sim \text{Be}\left(1, \frac{p-5}{p}\right)$. The data sets were

Table 2 Sample size n, number of regressors p and time-normalized effective sample size for various computational methods for the four logistic regression data sets

	n	p	DA		CPM		Laplace	
			ASI	ADS	ASI	ADS	ASI	ADS
Arthritis	31	755	3.1	1	**3.3** ($N = 10$)	1.0 ($N = 1$)	3.6	1.4
Colon Tumor	62	1224	**1.6**	1	1.3 ($N = 2$)	1.3 ($N = 1$)	2.2	1.7
Leukemia	72	3571	1.2	1	**1.5** ($N = 2$)	0.8 ($N = 1$)	1.3	1.0
Prostate	136	10,150	0.7	1	**2.0** ($N = 4$)	0.6 ($N = 1$)	3.3	2.0

The best performing method for each data set is shown in bold
In the CPM methods, the best performing value of N is shown in brackets

Table 3 Effective sample size for various computational methods on the four logistic regression data sets

	DA		CPM		Laplace	
	ASI	ADS	ASI	ADS	ASI	ADS
Arthritis	73,487	60,893	226,874 ($N = 10$)	72,229 ($N = 1$)	227,171	74,259
Colon Tumor	57,863	59,454	113,629 ($N = 2$)	71,622 ($N = 1$)	110,815	72,242
Leukemia	55,581	64,887	118,348 ($N = 2$)	73,115 ($N = 1$)	123,032	74,748
Prostate	164,983	129,365	584,791 ($N = 1$)	131,426 ($N = 1$)	596,082	131,576

Table 4 Sample size n, number of regressors p, the percentage of censored observations (%) and the time-normalized effective sample size for various computational methods on the two survival data sets

	n	p	%	DA	CPM		Laplace	
				ASI	ADS	ASI	ADS	ASI	ADS
DLBCL	235	7399	43.4	8.1	1	**9.5**	0.8	13.3	1.0
						$N = 25$ $\rho = 1$	$N = 25$ $\rho = 0.7$		
Breast Cancer	614	3378	78.1	**1.45**	1	0.39	0.26	0.7	0.3
						$N = 5$ $\rho = 0.95$	$N = 5$ $\rho = 0.95$		

The best performing method for each data set is shown in bold

not standardized and the following priors were used for the regression coefficients: $p(\beta_\gamma|\gamma) \sim \mathrm{N}(0, 4\,I)$ for the DLBCL data and $p(\beta_\gamma|\gamma) \sim \mathrm{N}(0, 0.25\,I)$ for the breast cancer data. The samplers were run for 105,000 iterations with a burn-in period of 5000 iterations. All samples after the burn-in were used in the inference.

The time-normalized effective sample size for each method and each data set is shown in Table 4. Unlike the logistic regression examples, in the correlated pseudo-marginal sampler we find the optimal value of ρ differs between data sets and the optimal value of N differs between both data sets and algorithms. The results show that the ASI methods outperform the corresponding ADS methods for DA, CPM and Laplace in both data sets (with a substantial improvement for the DLBCL data set). The ASI-CPM and ASI-Laplace methods perform better than the ASI-DA method for the DLBCL but perform worse for the Breast Cancer data sets. Although we only have two data sets, the results illustrate an important trade-off between the data augmentation methods and the CPM and Laplace methods. The data augmentation scheme scales with number of censored observations. The CPM and Laplace methods effectively scale like the typical model size since most time is spent finding the posterior mode for each update. The mixing of the CPM and Laplace methods should be better than the data augmentation schemes since no latent variables are introduced. Since the performance is largely effected by the typical model size (which is often not known before the analysis), it is hard to provide criteria for deciding when to use which algorithm. Our recommendation is to use the DA method in problems with sample sizes in the hundreds since this works well in both cases. The performance of the DA will deteriorate relative to the CPM/Laplace methods for large values of n.

5 Discussion

In this paper, we have looked at several different schemes for Bayesian variable selection in logistic regression or accelerated failure time models. We find that the use of the adaptive Metropolis–Hastings sampler provides better mixing chains than the standard Add–Delete–Swap method in both models. The choice between data augmentation and the CPM or Laplace method is less clear-cut. In the logistic regression model, the CPM method outperforms the data augmentation scheme in 3 out of 4 data sets. The CPM method is exact and

performs similar to the (approximate) Laplace approximation method. Therefore, we suggest using the CPM with the ASI method for logistic regression. The performance in the accelerate failure time model is less clear. We find that the performance of the CPM and Laplace methods is strongly effected by the typical posterior model size. We recommend using the data augmentation method with ASI for data sets with sample sizes in the hundreds.

References

Albert, J., Chib, S.: Bayesian analysis of binary and polychotomous response data. J. Am. Stat. Assoc. **88**, 669–679 (1993)

Alon, U., Barkai, N., Notterman, D.A., Gish, K., Ybarra, D., Mack, D., Levine, A.J.: Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probe by oligonucleotide array. Proc. Natl. Acad. Sci. USA **96**, 6745–6750 (1999)

Andrieu, C., Roberts, G.O.: The pseudo-marginal approach for efficient Monte Carlo computations. Ann. Stat. **37**, 697–725 (2009)

Annest, A., Bumgarner, R.E., Raftery, A.E., Yeung, K.Y.: The iterative Bayesian model averaging algorithm for survival analysis: an improved method for gene selection and survival analysis on microarray data. BMC Bioinform. **10**, 72 (2009)

Armstrong, S.A., Staunton, J.E., Silverman, L.B., Pieters, R., den Boer, M.L., Minden, M.D., Sallan, S.E., Lander, E.S., Golub, T.R., Korsmeyer, S.J.: MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia. Nat. Genet. **30**, 41–47 (2002)

Chipman, H., George, E.I., McCulloch, R.E.: The practical implementation of Bayesian model selection. In: Lahiri, P. (ed.) Model Selection. Hayward, Maidston (2001)

Choi, H.M., Hobert, J.P.: The Polya-Gamma Gibbs sampler for Bayesian logistic regression is uniformly ergodic. Electron. J. Stat. **7**, 2054–2064 (2013)

Deligiannidis, G., Doucet, A., Pitt, M.K.: The correlated pseudo-marginal method. J. R. Stat. Soc. Ser. B **80**, 839–870 (2018)

Duan, W., Zhang, R., Zhao, Y., Shen, S., Wei, Y., Chen, F., Christiani, D.C.: Bayesian variable selection for parametric survival model with applications to cancer omics data. Hum. Genomics **12**, 49 (2018)

García-Donato, G., Martínez-Beneito, M.A.: On sampling strategies for Bayesian variable selection problems with large model spaces. J. Am. Stat. Assoc. **108**, 340–352 (2013)

Green, P.J.: Trans-dimensional Markov chain Monte Carlo. In: Green, P.J., Hjort, N.L., Richardson, S. (eds.) Highly Structured Stochastic Systems, pp. 179–198. Oxford University Press, Oxford (2003)

Griffin, J.E., Buxton, A.S., Matechou, E., Bormpoudakis, D., Griffiths, R.A.: Modelling environmental DNA data: Bayesian model selection accounting for false positive and false negative probabilities. J. R. Stat. Soc. Ser. C **69**, 377–392 (2019)

Griffin, J.E., Łatuszyński, K., Steel, M.F.J.: In search of lost (mixing) time: Adaptive Markov chain Monte Carlo schemes for Bayesian variable selection with very large p. Biometrika (2020). (**to appear**)

Hastie, T., Tibshirani, R., Wainwright, M.: Statistical Learning with Sparsity: The Lasso and Generalizations. Chapman & Hall / CRC, Boca Raton (2015)

Held, L., Gravestock, I., Bové, D.S.: Objective Bayesian model selection for Cox regression. Stat. Med. **35**, 5376–5390 (2016)

Holmes, C.C., Held, L.: Bayesian auxiliary variable models for binary and multinomial regression. Bayesian Anal. **1**, 145–168 (2006)

Lamnisos, D., Griffin, J.E., Steel, M.F.J.: Transdimensional sampling algorithms for Bayesian variable selection in classification problems with many more variables than observations. J. Comput. Graph. Stat. **18**, 592–612 (2009)

Ley, E., Steel, M.F.J.: On the effect of prior assumptions in Bayesian model averaging with applications to growth regression. J. Appl. Econ. **24**, 651–674 (2009)

Li, Y., Clyde, M.: Mixtures of g-priors in generalized linear models. J. Am. Stat. Assoc. **113**, 1828–1845 (2018)

Liang, F., Paulo, R., Molina, G., Clyde, M.A., Berger, J.O.: Mixtures of g priors for Bayesian variable selection. J. Am. Stat. Assoc. **103**, 410–423 (2008)

Liu, J.S.: Monte Carlo Strategies for Scientific Computing. Springer, Berlin (2001)

Newcombe, P.J., Raza Ali, H., Blows, F.M., Provenzano, E., Pharoah, P.D., Caldas, C., Richardson, S.: Weibull regression with Bayesian variable selection to identify prognostic tumor markers of breast cancer survival. Stat. Methods Med. Res. **26**, 414–436 (2017)

Nikooienejad, A., Wang, W., Johnson, V.E.: Bayesian variable selection for binary outcomes in high-dimensional genomic studies using non-local priors. Bioinformatics **32**, 1338–1345 (2016)

Nikooienejad, A., Wang, W., Johnson, V.E.: Bayesian variable selection for survival data using inverse moment priors. Ann. Appl. Stat. **14**, 809–828 (2020)

O'Hara, R.B., Sillanpää, M.J.: A review of Bayesian variable selection methods: what, how and which. Bayesian Anal. **4**, 85–117 (2009)

Polson, N.G., Scott, J.G., Windle, J.: Bayesian inference for logistic models using Pólya-Gamma latent variables. J. Am. Stat. Assoc. **108**, 1339–1349 (2013)

Roberts, G.O., Rosenthal, J.S.: Coupling and ergodicity of adaptive Markov chain Monte Carlo algorithms. J. Appl. Probab. **44**, 458–475 (2007)

Rockova, V., George, E.I.: EMVS: the EM approach to Bayesian variable selection. J. Am. Stat. Assoc. **109**(506), 828–846 (2014)

Rosenwald, A., Wright, G., Chan, W.C., Connors, J.M., Campo, E., Fisher, R.I., Gascoyne, R.D., Muller-Hermelink, H.K., Smeland, E.B., Giltnane, J.M., Hurt, E.M., Zhao, H., Averett, L., Yang, L., Wilson, W.H., Jaffe, E.S., Simon, R., Klausner, R.D., Powell, J., Duffey, P.L., Longo, D.L., Greiner, T.C., Weisenburger, D.D., Sanger, W.G., Dave, J.B., Lynch, J.C., Vose, J., Armitage, J.O., Montserrat, E., López-Guillermo, A., Grogan, T.M., Miller, T.P., LeBlanc, M., Ott, G., Kvaloy, S., Delabie, J., Holte, H., Krajci, P., Stokke, T., Staudt, L.M.: The use of molecular profiling to predict survival after chemotherapy for diffuse large-b-cell lymphoma. N. Engl. J. Med. **346**, 1937–1947 (2002)

Sanyal, N., Lo, M.-T., Kauppi, K., Djurovic, S., Andreassen, O.A., Johnson, V.E., Chen, C.-H.: Gwasinlps: non-local prior based iterative SNP selection tool for genome-wide association studies. Bioinformatics **35**, 1–11 (2017)

Schäfer, C., Chopin, N.: Sequential Monte Carlo on large binary sampling spaces. Stat. Comput. **23**, 163–184 (2013)

Sha, N., Vannucci, M., Brown, P., Trower, M., Amphlett, G., Falciani, F.: Gene selection in arthritis classification with large-scale microarray expression profiles. Comp. Funct. Genomics **4**, 171–181 (2003)

Sha, N., Vannucci, M., Tadesse, M.G., Brown, P., Dragoni, I., Davies, N., Roberts, T., Contestabile, A., Salmon, N., Buckley, C., Falciani, F.: Bayesian variable selection in multinomial probit models to identify molecular signatures of disease stage. Biometrics **60**, 812–819 (2004)

Sha, N., Tadesse, M.G., Vannucci, M.: Bayesian variable selection for the analysis of microarry data with censored outcomes. Bioinformatics **22**, 2262–2268 (2006)

Shin, M., Bhattacharya, A., Johnson, V.E.: Scalable Bayesian variable selection using nonlocal prior densities in ultrahigh-dimensional settings. Stat. Sinica **28**, 1053–1078 (2018)

Singh, D., Febbo, P.G., Ross, K., Jackson, D., Manola, J., Ladd, C., Tamayo, P., Renshaw, A., D'Amico, A.V., Richie, J.P., Lander, E.S., Loda, M., Kantoff, P., Golub, T., Sellers, W.: Gene expression correlates of clinical prostate cancer behaviour. Cancer Cell **1**, 203–209 (2002)

Tanner, M.A., Wong, W.H.: The calculation of posterior distributions by data augmentation. J. Am. Stat. Assoc. **82**, 528–540 (1987)

Titsias, M.K., Yau, C.: The Hamming ball sampler. J. Am. Stat. Assoc. **112**, 1598–1611 (2017)

van't Veer, L.J., Dai, H., van de Vijver, M.J., He, Y.D., Hart, A.A., Mao, M., Peterse, H.L., van der Kooy, K., Marton, M.J., Witteveen, A.T., Schreiber, G.J., Roberts, K.R.M.C., Linsley, P.S., Bernards, R., Friend, S.H.: Gene expression profiling predicts clinical outcome of breast cancer. Nature **415**, 530–536 (2002)

Yang, Y., Wainwright, M., Jordan, M.I.: On the computational complexity of high-dimensional Bayesian variable selection. Ann. Stat. **44**, 2497–2532 (2016)

Zanella, G., Roberts, G.O.: Scalable importance tempering and Bayesian variable selection. J. R. Stat. Soc. Ser. B **81**, 489–517 (2019)

Zhang, Z., Sinha, S., Maiti, T., Shipp, E.: Bayesian variable selection in the accelerated failure time model with an application to the surveillance, epidemiology, and end results breast cancer data. Stat. Methods Med. Res. **27**, 971–990 (2018)