entropy

IMPACT FACTOR 2.7

Indexed in: PubMed

CITESCORE 4.7

Article

Adaptive MCMC for Bayesian Variable Selection in Generalised Linear Models and Survival Models

Xitong Liang, Samuel Livingstone and Jim Griffin

Special Issue

Markov Chain Monte Carlo for Bayesian Inference

Edited by

Dr. Ricardo Sandes Ehlers

MDPI

https://doi.org/10.3390/e25091310

entropy

MDPI

Article

Adaptive MCMC for Bayesian Variable Selection in Generalised Linear Models and Survival Models

Xitong Liang *, Samuel Livingstone and Jim Griffin

Department of Statistical Science, University College London, London WC1E 6BT, UK; samuel.livingstone@ucl.ac.uk (S.L.); j.griffin@ucl.ac.uk (J.G.)
* Correspondence: xitong.liang.18@ucl.ac.uk

Abstract: Developing an efficient computational scheme for high-dimensional Bayesian variable selection in generalised linear models and survival models has always been a challenging problem due to the absence of closed-form solutions to the marginal likelihood. The Reversible Jump Markov Chain Monte Carlo (RJMCMC) approach can be employed to jointly sample models and coefficients, but the effective design of the trans-dimensional jumps of RJMCMC can be challenging, making it hard to implement. Alternatively, the marginal likelihood can be derived conditional on latent variables using a data-augmentation scheme (e.g., Pólya-gamma data augmentation for logistic regression) or using other estimation methods. However, suitable data-augmentation schemes are not available for every generalised linear model and survival model, and estimating the marginal likelihood using a Laplace approximation or a correlated pseudo-marginal method can be computationally expensive. In this paper, three main contributions are presented. Firstly, we present an extended *Point-wise implementation of Adaptive Random Neighbourhood Informed* proposal (PARNI) to efficiently sample models directly from the marginal posterior distributions of generalised linear models and survival models. Secondly, in light of the recently proposed approximate Laplace approximation, we describe an efficient and accurate estimation method for marginal likelihood that involves adaptive parameters. Additionally, we describe a new method to adapt the algorithmic tuning parameters of the PARNI proposal by replacing Rao-Blackwellised estimates with the combination of a warm-start estimate and the ergodic average. We present numerous numerical results from simulated data and eight high-dimensional genetic mapping data-sets to showcase the efficiency of the novel PARNI proposal compared with the baseline add–delete–swap proposal.

Keywords: Bayesian computation; Bayesian variable selection; spike-and-slab priors; adaptive Markov Chain Monte Carlo; generalised linear models; survival models

Citation: Liang, X.; Livingstone, S.; Griffin, J. Adaptive MCMC for Bayesian Variable Selection in Generalised Linear Models and Survival Models. *Entropy* **2023**, *25*, 1310. https://doi.org/10.3390/e25091310

Academic Editors: Geert Verdoolaege and Ricardo Sandes Ehlers

Received: 1 August 2023
Revised: 30 August 2023
Accepted: 5 September 2023
Published: 8 September 2023

1. Introduction

Variable selection is an automatic method for finding a small subset of covariates that explain most of the variation in the response of interest. In addition to identifying the most predictive covariates, there is a growing interest in exploring the low-rank structure between the covariates and the response, especially in genetic mapping problems where the objective is to find the expressed genes that are associated with a specific disease the most. In the frequentist framework, model selection is based on maximising the penalised log-likelihood [1] or minimising information criteria such as AIC [2] and BIC [3]. Other approaches, such as the deviance information criterion (DIC) [4] and widely applicable information criterion (WAIC) [5], which are generalisations of the AIC, are also popular in model selection.

A natural alternative to these frequentist approaches is Bayesian variable selection (BVS). In the Bayesian approach, a prior is imposed on all candidate models, and the resulting posterior distribution naturally captures model uncertainty. In this work, we consider a spike-and-slab prior [6,7], which introduces indicator variables denoting the

inclusion or exclusion of every covariate. Therefore, the spike-and-slab prior leads to a model posterior distribution that lies in a lattice with the same dimension as the number of covariates. We can understand the dependency between the importance of covariates and response using natural measures of the posterior distribution such as posterior model probability (PMP) and marginal posterior inclusion probability (PIP). The computation of the exact posterior distribution requires a full search over the whole model space, which is computationally infeasible when a high-dimensional data-set is analysed. In these settings, Markov Chain Monte Carlo (MCMC) algorithms are often used to explore the model space and estimate the posterior distribution. For "large n, large p" data-sets, which are now often encountered in some problems in genetics/genomics (such as genetic mapping studies), such algorithms must be carefully designed. In this work, we mainly consider Bayesian variable selection in generalised linear models and survival models and focus on three popular models: the logistic regression model [8,9], the Cox proportional hazards model with partial likelihood [10–14] and the Weibull regression model [15]. In each case, we illustrate how carefully designed algorithms can facilitate effective posterior computation.

A natural challenge of Bayesian variable selection methods in the above settings is that the marginal likelihood (or the integrated likelihood in [16]) is not analytically available. One set of solutions are Reversible Jump MCMC schemes (RJMCMC) [17], which sample from the joint space of models and regression coefficients by jointly proposing moves between models and regression coefficients. But it is often difficult to construct efficient proposals for these trans-dimensional jumps and design an MCMC scheme that mixes well [18]. For some specific models, data-augmentation methods [19] are available and result in closed-form marginal likelihood conditioned on latent variables, for instance, Pólya-gamma data augmentation [20] for logistic regression. For other models where no suitable data-augmentation scheme exists, the most popular approaches are the Laplace approximation and the correlated pseudo-marginal method [21], which rely on finding the maximum a posteriori (MAP) estimate of the regression coefficients. A novel scalable estimation method for marginal likelihood, approximate Laplace approximation (ALA), is introduced in [16] and relies on defining an initial value for the coefficient parameters. ALA can save computational time during the optimisation process of finding the MAP estimate, but it does not yield an asymptotically consistent estimate. A detailed discussion of these approaches will be given in Section 3.

Assuming that the marginal likelihood has been estimated, several MCMC algorithms can be used for simulation starting from the posterior distribution of BVS. The widely used add–delete–swap proposal [22] can be employed here. The add–delete–swap proposal generates a new model by randomly selecting one of three possible moves: addition/deletion of a covariate into/from the model or swapping one covariate that is included with another that is not. Although it has been proved in [23] that the add–delete–swap proposal can produce a rapidly mixing Markov Chain, the chains may still converge slowly, particularly when dealing with large-p problems. Adaptive MCMC schemes [24], which involve updating tuning parameters on the fly, are found to be valuable in addressing the issue of poor convergence. Lamnisos et al. [25] describe an adaptive add–delete–swap proposal that allows for simultaneous changes to multiple variables at a time. Griffin et al. [26] introduce the Adaptively Scaled Individual adaptation proposal (ASI), which simulates a new model with probability proportional to the product of PIPs. Wan and Griffin [27] extend the ASI proposal to logistic regression and accelerated failure time models. Other popular MCMC approaches include the Hamming ball sampler (HBS) [28], which proposes a new model within a Hamming neighbourhood using the PMPs as proposal weights, and the tempered Gibbs sampler [29,30], which uses tempering to efficiently sample from the multi-modal posteriors that commonly arise due to highly correlated covariates.

The recent work of [31] provides useful insights into the design of efficient MCMC schemes in discrete spaces. The work introduces the locally informed proposal, which re-weights a given non-informed base kernel with a function of the PMPs. It is shown in [31] that the locally informed proposal constructed with a balancing function that satisfies

certain functional properties is asymptotically optimal compared with other choices of function in terms of Peskun ordering. Building upon the idea of locally informed proposal, Zhou et al. [32] show that the locally informed and thresholded (LIT) proposal can achieve dimension-free mixing times under conditions similar to those mentioned in [23] for BVS in linear regression models. Recent work in [33] introduces a Point-wise implementation of the Adaptive Random Neighbourhood Informed proposal (PARNI), which combines the advantages of both adaptive schemes and locally informed proposals. The PARNI proposal outperforms other state-of-the-art algorithms in a wide range of high-dimensional data-sets for BVS in linear regression models.

Other computational approaches are also available for estimating the BVS posterior distribution. Hans et al. [34] introduce a novel Shotgun Stochastic Search (SSS) approach that also explores the "local neighbourhood" idea and targets very high-dimensional model spaces to find high-probability regions. The integrated nested Laplace approximations (INLAs) [35] can solve latent Gaussian models including generalised linear models and approximate the posterior marginals obtained from the continuous priors [36]. Sara et al. [37] view survival models as latent Gaussian models and also approximate the posterior marginals using INLAs. The posterior distribution can also be approximated using Variational Bayes (VB) [38]. Ray et al. [39] describe a scalable mean-field variational family to approximate the posterior distribution of BVS in linear regression and extended this VB approximation to the logistic regression model in [40]. Komodromos et al. [41] apply the Sparse Variational Bayes (SVB) method to approximate the posterior of proportional hazards models with partial likelihood. Other works develop a sampling strategy based on simulating piecewise deterministic Markov processes (PDMPs) [42,43], which directly target the posterior distribution obtained from a spike-and-slab prior.

In this paper, we extend the PARNI proposal to sampling from the BVS posterior distribution in generalised linear models and survival models. To avoid the overwhelming computational costs of approximating the marginal likelihood in the locally informed proposals and motivated by ALA [16], we introduce an ALA estimate of the marginal likelihood with a novel initial value. In contrast to the suggestion in [16], which initialises ALA at origin, the novel initial value is adaptively updated on the fly using previously sampled models. The new method is computationally less complex than the Laplace approximation or correlated pseudo-marginal scheme as a result of avoiding iterative optimisation and provides a more accurate estimate than the original ALA approach initialised at the origin. We also consider new approaches to adapt the tuning parameters in the PARNI proposal. The new adaptation scheme replaces the Rao-Blackwellised estimates of PIPs using the combination of a warm-start estimate and the ergodic average calculated using previously sampled models.

To illustrate the performance of the new PARNI scheme in real-life high-dimensional problems, we perform BVS on eight genetic mapping data-sets (four for the logistic regression model and four for survival analysis) and compare the output of the PARNI proposal with the add–delete–swap proposal as a baseline. For the logistic model with binary outcome, we consider the problem of finding expressed genes that are related the most to the presence of Systemic Lupus Erythematosus in a case-control study with 10,995 observations and various numbers of SNPs, from 5771 to 42,430, on four different chromosomes. In survival analysis, we consider four cancer-related data-sets (two for breast cancer and two for lung cancer), containing patients ranging from 130 to 1904 and genetic covariates varying from 662 to 54,675.

This paper is organised as follows: In Section 2, we review the model setup and prior specification for BVS in generalised linear models, Cox proportional hazards and Weibull survival models. In Section 3, we introduce four computational methods to estimate the marginal likelihood. Section 4 describes the PARNI proposal, highlighting the novelties in the adaption of algorithmic tuning parameters and the calculation of the accurate and efficient marginal likelihood estimates. We implement these MCMC algorithms in Section 5 and compare their performance with the add–delete–swap proposal on several real data-

sets. We include a discussion in Section 6, highlighting some possible future research directions.

2. Bayesian Variable Selection for Generalised Linear Models and Survival Models

2.1. Generic Model Setting

Suppose that p covariates are available in the data. Let $X = (x_1, \ldots, x_n)^T \in \mathbb{R}^{n \times p}$ be the full data matrix that contains n observations with rows $x_i = (X_{i1}, \ldots, X_{ip})$ and let $Z = (z_1, \ldots, z_n)^T \in \mathbb{R}^{n \times q}$ be the full data matrix that contains q variables that must be included in every model. Let binary vector $\gamma = (\gamma_1, \ldots, \gamma_p) \in \Gamma = \{0,1\}^p$ be a model indicator, where $\gamma_j = 1$ if the j-th variable is included in model $\mathcal{M}_\gamma$ and $\gamma_j = 0$ otherwise.

Let $y = (y_1, \ldots, y_n)$ be the vector of responses. The generalised linear model associated with model $\mathcal{M}_\gamma$ can be specified as

$$y_i \sim \mathrm{F}(\mu_{\gamma,i}, \phi) \tag{1}$$

where $\mathrm{F}(\mu, \phi)$ is a distribution that belongs to the exponential family with mean μ and dispersion ϕ. Linear predictor $\eta_{\gamma,i}$ is defined as

$$\eta_{\gamma,i} = z_i^T \alpha + x_{\gamma,i}^T \beta_\gamma \tag{2}$$

where $x_{\gamma,i}$ contains those variables j for which $\gamma_j = 1$. In addition, we define the size of model $\mathcal{M}_\gamma$ as $p_\gamma = \sum_{j=1}^{p} \gamma_j$. Linear predictor $\eta_{\gamma,i}$ is mapped to mean $\mu_{\gamma,i}$ using link function g as

$$\eta_{\gamma,i} = g(\mu_{\gamma,i}). \tag{3}$$

We consider the following setup of survival models: For the i-th patient, given hazard function $h_i(t)$ at time t, the probability that an event occurs at time T_i before a certain time t_i can be written as

$$F_{T_i}(t) = \mathbb{P}(T_i \leq t_i) = 1 - S_{T_i}(t_i)$$

where S_{T_i} is called the survival function and is defined by

$$S_{T_i}(t) = \exp\left\{ - \int_0^t h_i(u) du \right\}.$$

Data often involve censoring where the true time to event is not observed. Let t be the vector of the observed times, where each t_i denotes the minimum of censoring time C_i and survival time T_i. In the case of "right-censored" data, we define an n-dimensional event indicator vector d to denote, for each patient i, whether the event was observed during their follow-up ($d_i = 1$) or was censored ($d_i = 0$). In the case where the event was observed for patient i ($d_i = 1$), then t_i denotes their time to event; otherwise, we observed the length of their follow-up.

Given a model $\mathcal{M}_\gamma$ associated with linear predictor $\eta_{\gamma,i}$ as in (2), we consider the exponential hazard, $\lambda_{\gamma,i} = \exp(\eta_{\gamma,i})$, and assume that the hazard function conditioned on model $\mathcal{M}_\gamma$ has the form

$$h_{\gamma,i}(t) = h(t, \lambda_{\gamma,i}, k) \tag{4}$$

where k is an additional shape parameter if needed. We can conclude the following log-likelihood on $y = (t, d)$:

$$\log p(y | \alpha, \beta_\gamma, \gamma) = \sum_{i=1}^{n} d_i \log(h_{\gamma,i}(t_i)) - H_i(t_i)$$

where H_i is the cumulative hazard function for the i-th patient and is defined by

$$H_i(t) = \int_0^t -\frac{d \log S(t)}{dt}\bigg|_{t=u} du = -\log S_{T_i}(t).$$

2.2. Prior Elicitation

Recalling model indicator $\gamma \in \{0,1\}^p$, we consider the prior structure

$$p(\alpha, \beta_\gamma, \phi, \gamma) \propto p(\alpha)p(\beta_\gamma|\phi,\gamma)p(\phi|\gamma)p(\gamma). \tag{5}$$

For generalised linear models in which the dispersion parameter is known (e.g., the logistic regression model where $\phi = 1$) or some survival models that do not involve a dispersion parameter, the prior specification becomes

$$p(\alpha, \beta_\gamma, \gamma) \propto p(\alpha)p(\beta_\gamma|\gamma)p(\gamma), \tag{6}$$

which is equivalent to treating ϕ as a fixed parameter. In this work, we focus on the prior structure described in (6), and we assume that there is no additional dispersion parameter in the model.

We specify the following prior distribution for the coefficient parameters:

$$\begin{aligned} \alpha &\sim \mathrm{N}_q(0, \sigma_\alpha^2 I_q) \\ \beta_\gamma|\gamma &\sim \mathrm{N}_{p_\gamma}(0, g I_{p_\gamma}) \end{aligned} \tag{7}$$

where g is a positive scale parameter, σ_α^2 is the prior variance on the coefficients of the fixed covariates, I_p denotes a $p \times p$ identity matrix and $p_\gamma = \sum_{j=1}^p \gamma_j$ is the size of model $\mathcal{M}_\gamma$.

We consider the choice of model prior

$$p(\gamma) = h^{p_\gamma}(1-h)^{p-p_\gamma} \tag{8}$$

where hyper-parameter h denotes the prior inclusion probability for each variable.

It is possible to construct a fully Bayesian hierarchical model based on the prior specifications described above. We can impose the following hyper-priors on hyper-parameters g and h:

$$\begin{aligned} \sqrt{g} &\sim \mathrm{C}^+(0,1) \\ h &\sim \mathrm{Beta}(a,b) \end{aligned}$$

where $\mathrm{C}^+(0,1)$ denotes the standard half-Cauchy distribution and $\mathrm{Beta}(a,b)$ denotes the Beta distribution with parameters $a > 0$ and $b > 0$. The half-Cauchy hyper-prior is a generalisation of the horseshoe prior [44–46], employed on the global-scale parameter of a continuous mixture of normal priors, to BVS problems. Liang et al. [47] note that fixing g can lead to several paradoxes and problems of model mis-specification. For other possible choices of hyper-priors on g, see [47,48]. In the context of prior inclusion probability h, Ley et al. [49] advise against using a fixed h in the absence of strong prior knowledge about the number of important variables. Kohn et al. [50] propose a Beta-binomial model prior in which hyper-parameter h can be integrated out analytically, leading to

$$p(\gamma) = \frac{\mathrm{B}(a + p_\gamma, b + p - p_\gamma)}{\mathrm{B}(a,b)}$$

where $\mathrm{B}(\cdot,\cdot)$ denotes the Beta function.

2.3. Logistic Regression

Assume that $y_i \in \{0, 1\}$, with $y_i = 1$ indicating the success of an event and $y_i = 0$ indicating failure. Logistic regression links the proportion of successes to the linear predictor with a logistic link function g as

$$\eta_{\gamma,i} = \log\left(\frac{\mu_{\gamma,i}}{1-\mu_{\gamma,i}}\right) = z_i^T \alpha + x_{\gamma,i}^T \beta_\gamma, \quad i = 1, \ldots, n, \tag{9}$$

and the response variable is modelled as $y_i \sim \text{Bern}(\mu_{\gamma,i})$ under model $\mathcal{M}_\gamma$.

2.4. Cox Proportional Hazards (PHs) with Partial Likelihood

Starting with exponential hazard function $\lambda_{\gamma,i} = \exp(\eta_{\gamma,i})$ associated with model $\mathcal{M}_\gamma$, in the Cox proportional hazard function, the hazards are assumed to have the form

$$h_i(t) = h_0(t)\lambda_{\gamma,i} \tag{10}$$

where h_0 is some baseline hazard function. In this proportional model, all covariate effects are assumed to be multiplicative. The full likelihood is then given as

$$L(\alpha, \beta_\gamma, H_0 | y, \gamma) \propto \prod_{j=1}^{n} \left(\exp(\eta_{\gamma,i}) H_0'(t_i)\right)^{d_i} \exp\{-\exp(\eta_{\gamma,i}) H_0(t_i)\} \tag{11}$$

where $H_0(t) = \int_0^t h_0(u)du$ is the cumulative baseline hazard function. If we model the prior of H_0 with a prior process $p(H_0)$ on the cumulative hazard function, the resulting posterior distribution of α and β_γ is

$$p(\alpha, \beta_\gamma | y, \gamma) \propto \int L(\alpha, \beta_\gamma, H_0 | y, \gamma) \times p(\alpha) p(\beta_\gamma | \gamma) p(H_0) dH_0. \tag{12}$$

Alternatively, we can take the partial likelihood of Cox, which is given by

$$\text{PL}(\alpha, \beta_\gamma | y, \gamma) \propto \prod_{i=1}^{n} \left\{ \frac{\exp(\eta_{\gamma,i})}{\sum_{s \in \mathcal{R}(t_i)} \exp(\eta_{\gamma,s})} \right\}^{d_i} \tag{13}$$

where $\mathcal{R}(t) = \{i : t_i \geq t\}$ is the set of patients at risk at time t. Unlike the full likelihood formulated in (11), partial likelihood does not rely on the specification and estimation of baseline hazard function h_0. Partial likelihood and its variants are, therefore, popular alternatives to full likelihood in many survival studies [11,14,51]. It is highlighted in [52,53] that partial likelihood can be obtained by integrating out the baseline hazard function using a Gamma process prior. Bayesian inference with partial likelihood (13) relies on approximate posterior $p_{\text{PL}}(\alpha, \beta_\gamma | y, \gamma)$, which can be expressed as

$$p_{\text{PL}}(\alpha, \beta_\gamma | y, \gamma) \propto \text{PL}(\alpha, \beta_\gamma | y, \gamma) \times p(\alpha) p(\beta_\gamma | \gamma) \tag{14}$$

where baseline hazard function h_0 is eliminated.

2.5. Weibull Regression

In addition to the semi-parametric approach of Cox PHs with partial likelihood, we consider another commonly used parametric model for survival analysis, namely, the Weibull model. A Weibull model is obtained by extending the exponential model by raising the survival rate to a positive power k, giving

$$S_i(t) = \exp\left\{-(t\lambda_i)^k\right\}. \tag{15}$$

Parameter k is the shape parameter of a Weibull random variable. When $k < 1$, the hazard rate decreases over time. Conversely, when $k > 1$, the hazard rate increases over time. It is possible to recover the exponential survival model when $k = 1$, and it represents a constant hazard rate over time.

We can derive the hazard function

$$h_i(t) = -\frac{d}{dt}\log(S_i(t)) = \lambda_i k(\lambda_i t)^{k-1} \tag{16}$$

and the log-likelihood for parameters α, β_γ and k as

$$\log(L(\alpha, \beta_\gamma, k|y, \gamma)) = \sum_{i=1}^{n} d_i[\log(k) + k\log(\lambda_i) + (k-1)\log(t_i)] - (t_i\lambda_i)^k. \tag{17}$$

It should be noted that the Weibull distribution does not belong to the exponential family, unless shape parameter k is assumed to be fixed. In the Bayesian framework, we consider the prior $p(\log(k)) = N(0, \sigma_k^2)$ for some positive prior variance σ_k^2 as in [15]. To perform MCMC, we alternatively update $\gamma|k$ using the PARNI proposal and $k|\gamma$ using an adaptive random walk proposal as described in Appendix B.

3. Computation of Marginal Likelihood $p(y|\gamma)$

Let $\theta_\gamma = (\alpha, \beta_\gamma)$ be the collection of all coefficient parameters associated with model $\mathcal{M}_\gamma$. We are interested in simulating samples from the posterior distribution $\pi(\gamma) \propto p(y|\gamma)p(\gamma)$, where $p(y|\gamma)$ represents the marginal likelihood, given by

$$p(y|\gamma) \propto \int p(y|\theta_\gamma, \gamma)p(\theta_\gamma|\gamma)d\theta_\gamma. \tag{18}$$

In generalised linear models and survival analysis, a closed-form solution to (18) is typically not analytically available.

Assuming that an estimate of marginal likelihood $\hat{p}(y|\gamma)$ can be obtained, we consider MCMC algorithms with random neighbourhood proposals as described in [33], which is a sub-class of Metropolis–Hastings (MH) schemes [54,55]. The random neighbourhood proposal consists of the following three stages:

1. Around the current model, γ, randomly generate a neighbourhood $\mathcal{N} \sim p(\cdot|\gamma)$.
2. Propose a new model, γ', within random neighbourhood $\mathcal{N}$ according to $q_{\mathcal{N}}(\gamma, \cdot)$.
3. Accept the new proposal, γ', with the MH acceptance probability

$$\begin{aligned} \alpha(\gamma, \gamma') &= \min\left\{1, \frac{\pi(\gamma')p(\mathcal{N}'|\gamma')q_{\mathcal{N}'}(\gamma', \gamma)}{\pi(\gamma)p(\mathcal{N}|\gamma)q_{\mathcal{N}}(\gamma, \gamma')}\right\} \\ &= \min\left\{1, \frac{\hat{p}(y|\gamma')p(\gamma')p(\mathcal{N}'|\gamma')q_{\mathcal{N}'}(\gamma', \gamma)}{\hat{p}(y|\gamma)p(\gamma)p(\mathcal{N}|\gamma)q_{\mathcal{N}}(\gamma, \gamma')}\right\} \end{aligned} \tag{19}$$

where $\mathcal{N}'$ is the neighbourhood used in the reverse move of the MH scheme.

In this section, we will describe four methods commonly used to estimate marginal likelihood $p(y|\gamma)$: data augmentation, Laplace approximation, correlated pseudo-marginal and approximate Laplace approximation. Before introducing these methods, it is necessary to define the following terms for convenience. Let J_γ be a $n \times (q + p_\gamma)$ matrix which contains all necessary covariates for model $\mathcal{M}_\gamma$ and is given by $J_\gamma = (Z\ X_\gamma)$, and let V_γ be the variance–covariance matrix of the prior distribution of θ_γ, defined by

$$V_\gamma = \begin{pmatrix} \sigma_\alpha^2 I_q & 0 \\ 0 & gI_{p_\gamma} \end{pmatrix}. \tag{20}$$

3.1. Data Augmentation

The data-augmentation scheme [19] introduces latent variables ω into the model such that the posterior distribution of variables of interest becomes analytically tractable given ω. The Pólya-gamma data-augmentation scheme [20] can be utilised for the logistic regression model to evaluate the marginal likelihood. Given real numbers $\psi \in \mathbb{R}$, $a > 0$, $b > 0$ and a set of latent variables $\omega = (\omega_1, \ldots, \omega_n)$, in which each individual ω_i follows a Pólya-gamma distribution $\mathrm{PG}(b, 0)$, the application of Pólya-gamma data augmentation exploits the following identity:

$$\frac{(\exp(\psi))^a}{(1+\exp(\psi))^b} = 2^{-b}\exp(\kappa\psi)\int_0^\infty \exp\left(-\omega_i\psi^2/2\right)p(\omega_i)d\omega_i \tag{21}$$

where $\kappa = a - b/2$. The above identity implies that the posterior distribution of the coefficients can be represented as a multivariate normal distribution:

$$\theta_\gamma \sim \mathrm{N}\left(\Lambda_\gamma^{-1}\xi, \Lambda_\gamma^{-1}\right) \tag{22}$$

where $\xi = J_\gamma^T\kappa$, $\Lambda_\gamma = J_\gamma^T W J_\gamma + V_\gamma^{-1}$, κ is an n-dimensional vector with entries $\kappa_i = y_i - 1/2$ and W is a diagonal matrix with ω appearing along its diagonal. By integrating out coefficient θ_γ, analytically conditioned on Pólya-gamma random variables ω, we obtain the conditional marginal likelihood

$$p(y|\gamma,\omega) \propto |V_\gamma|^{-\frac{1}{2}}|\Lambda_\gamma|^{-\frac{1}{2}}\exp\left\{\frac{1}{2}\xi^T\Lambda_\gamma^{-1}\xi\right\}. \tag{23}$$

In each iteration of the MCMC algorithm, we update γ and ω alternatively. To refresh ω, we can perform a simulation directly from its posterior distribution, which also follows a Pólya-gamma distribution given by

$$\omega_i \sim \mathrm{PG}(1, \eta_{\gamma,i}). \tag{24}$$

where linear predictor $\eta_{\gamma,i}$ involves coefficient θ_γ simulated from (22). Efficient samples from the Pólya-gamma random variables can be simulated using the R package pgdraw (version 1.1) [56]. In addition, Zens et al. described the ultimate Pólya-gamma sampler [57] to address the slow mixing rate for categorical imbalanced data, as illustrated in [58].

In general, the data-augmentation schemes may not be applicable to all generalised linear models and survival models. Specifically, for the Cox proportional hazards with partial likelihood or the Weibull model, there is currently no suitable data augmentation to directly yield a parametric posterior distribution for the regression coefficients.

3.2. Laplace Approximation

Assuming a unimodal posterior distribution of the regression coefficients, the Laplace approximation estimates the marginal likelihood with a second-order Taylor approximation. This method leads to a Gaussian integral, with the solution of the marginal likelihood being given by

$$p_{\mathrm{LA}}(y|\gamma) = p(y|\hat{\theta}_\gamma,\gamma)p(\hat{\theta}_\gamma|\gamma)|\hat{H}_\gamma|^{-\frac{1}{2}}(2\pi)^{\frac{p_{\theta_\gamma}}{2}} \tag{25}$$

where $\hat{\theta}_\gamma$ is the posterior mode of θ_γ and H_γ is the negated Hessian of $\log p(y|\theta_\gamma,\gamma) + \log p(\theta_\gamma|\gamma)$ evaluated at mode $\hat{\theta}_\gamma$. Additionally, the Laplace approximation provides a normal approximation to the posterior distribution of coefficient θ_γ as

$$\pi_{\mathrm{LA}}(\theta_\gamma) = \mathrm{N}_{p_{\theta_\gamma}}(\hat{\theta}_\gamma, \hat{H}_\gamma^{-1}). \tag{26}$$

To incorporate the Laplace approximation in MH sampling, we replace marginal likelihood $p(y|\gamma)$ in (19) with the approximate $p_{\mathrm{LA}}(y|\gamma)$ as described above.

Laplace approximation has been shown to be asymptotically consistent for estimating Bayes factors [59] and Bayesian variable selection on generalised linear models [60]. In finite-sample problems, however, Laplace approximation introduces biases, so $p_{\mathrm{LA}}(y|\gamma)$ is not an unbiased estimate of true marginal likelihood $p(y|\gamma)$. The resulting MCMC scheme, which involves the step of Laplace approximation, targets a different distribution compared with the true posterior $\pi(\gamma)$. Instead, it targets the distribution $\pi_{\mathrm{LA}}(\gamma) \propto p_{\mathrm{LA}}(y|\gamma)p(\gamma)$.

3.3. Correlated Pseudo-Marginal Method

We can alternatively make use of normal approximation $\pi_{\mathrm{LA}}(\theta_\gamma)$ to derive an importance sampling estimate of marginal likelihood $p(y|\gamma)$. This estimator is unbiased and given by

$$\hat{p}(\gamma|y) = \frac{1}{N}\sum_{i=1}^{N} \frac{p(y|\theta_\gamma^{(i)},\gamma)p(\theta_\gamma^{(i)}|\gamma)}{\pi_{\mathrm{LA}}(\theta_\gamma^{(i)})} \tag{27}$$

where $\theta_\gamma^{(1)},\ldots,\theta_\gamma^{(N)}$ are N samples from $\pi_{\mathrm{LA}}(\theta_\gamma)$. As in Laplace approximation, we can replace marginal likelihood $p(y|\gamma)$ in (19) with estimated marginal likelihood $\hat{p}(y|\gamma)$. This leads to the pseudo-marginal scheme in [61,62]. Andrieu and Roberts [62] show that the resulting Markov Chain preserves π-reversibility as long as estimated marginal likelihood $\hat{p}(y|\gamma)$ is an unbiased estimator of the true marginal likelihood, $p(y|\gamma)$.

It is possible to extend a pseudo-marginal method to a correlated pseudo-marginal method [21], with the aim of reducing the estimation variance of the ratio of estimated marginal likelihoods $\hat{p}(y|\gamma')/\hat{p}(y|\gamma)$. The correlated pseudo-marginal method is applied to Bayesian variable selection for the logistic regression model in [27], which provides an implementation that we also adopt in this work.

3.4. Approximate Laplace Approximation

The above Laplace approximation and correlated pseudo-marginal methods are computationally intensive due to the optimisation process required to obtain the normal approximation in (26), especially for dealing with large-n data. To avoid the overwhelming computational cost associated with the optimisation process, Rossell et al. [16] introduce the approximate Laplace approximation method (ALA), which is more computationally tractable for large-n problems. In this work, we consider the alternative formula described in supplementary material S.1. of [16], as it offers better computational stability when inverting the Hessian under the independent prior in (7).

In ALA, a Taylor expansion of log-posterior density $\log p(y|\theta_\gamma,\gamma) + \log p(\theta_\gamma|\gamma)$ is performed at initial value θ_γ^0. Solving the resulting Gaussian integral leads to

$$p_{\mathrm{ALA}}(y|\gamma) = p(y|\theta_\gamma^0,\gamma)p(\theta_\gamma^0|\gamma)(2\pi)^{\frac{d}{2}}|H_\gamma^0|^{-\frac{1}{2}}\exp\left\{\frac{1}{2}g_\gamma^{0T}(H_\gamma^0)^{-1}g_\gamma^0\right\} \tag{28}$$

where g_γ^0 and H_γ^0 are the gradient and Hessian of the negative log-posterior density evaluated at θ_γ^0, respectively. It is suggested in [16] to set initial value θ_γ^0 to $\theta_\gamma^0 = 0$ for convenience.

By applying ALA to the MH acceptance probability in (19), we obtain an MCMC algorithm that targets the ALA posterior distribution $\pi_{\mathrm{ALA}}(\gamma) \propto p_{\mathrm{ALA}}(y|\gamma)p(\gamma)$ as the equilibrium distribution. Although Ref. [16] shows that ALA can recover the optimal model with respect to a mean squared loss, it is important to note that ALA is not consistent with respect to the marginal likelihood (in contrast to the classical Laplace approximation) and $p_{\mathrm{ALA}}(y|\gamma)$ is not an unbiased estimator of the true marginal likelihood, $p(y|\gamma)$.

4. Point-Wise Implementation of Adaptive Random Neighbourhood Informed Proposal

4.1. The PARNI Proposal

The PARNI proposal belongs to the class of random neighbourhood informed proposals, which typically involve the following two steps: (i) sampling a neighbourhood

$\mathcal{N} \sim p(\cdot|\gamma)$ and then (ii) proposing a model γ' within this neighbourhood $\mathcal{N}$ according to the informed proposal of [31]. In the PARNI proposal, we assume that the randomness in neighbourhood generation is characterised by an auxiliary variable $k \in \mathcal{K}$, with conditional distribution $p(k|\gamma)$, which leads to neighbourhood $\mathcal{N} = \mathcal{N}(\gamma, k)$, such that $p(k|\gamma) = p(\mathcal{N}|\gamma)$. By defining $\mathcal{K} = \{0,1\}^p$, the value of k indicates whether the change in the corresponding position in γ is included in neighbourhood $\mathcal{N}$. Specifically, for those positions j such that $k_j = 1$, the neighbourhood consists of models obtained by varying some or all of these positions in the current model, γ.

The conditional distribution of k takes the product form $p(k|\gamma) = \prod_{j=1}^{p} p(k_j|\gamma_j)$, where each k_j depends on the corresponding component γ_j in γ. This probability distribution is driven by a set of tuning parameters $(A_1, \ldots, A_p, D_1, \ldots, D_p)$, where $A_j, D_j \in (\epsilon, 1-\epsilon)$ for a small value of $\epsilon \in (0, 1/2)$. The probabilities of event $k_j = 1$ are then defined by

$$p(k_j = 1|\gamma_j = 0) = A_j, \quad p(k_j = 1|\gamma_j = 1) = D_j, \tag{29}$$

and the consequent neighbourhood is constructed as

$$\mathcal{N}(\gamma, k) = \Big\{\gamma^* \in \Gamma \mid \gamma_j^* = \gamma_j \ \forall j \text{ s.t. } k_j = 0\Big\}. \tag{30}$$

Neighbourhood $\mathcal{N}(\gamma, k)$ contains 2^{p_k}, models where p_k denotes the number of 1s in k. Performing a full enumeration over the entire neighbourhood is, therefore, computationally expensive when p_k is large. In fact, it becomes computationally infeasible to explore the whole neighbourhood when p_k is beyond 30. Liang et al. [33], therefore, consider a point-wise approximate implementation of this algorithm that dramatically reduces the number of model probability evaluations from $\mathcal{O}(2^{p_k})$ to $\mathcal{O}(2p_k)$.

The point-wise implementation proceeds by constructing a sequence of smaller neighbourhoods $\{\mathcal{N}_r\}$ such that each $\mathcal{N}_r$ is a subset of $\mathcal{N}$. A proposed model γ' is sequentially simulated from these neighbourhoods $\{\mathcal{N}_r\}$ according to locally informed proposals $q_{\mathcal{N}_r}$. This procedure requires us to define a sequence of intermediate models $\gamma = \gamma(0) \to \gamma(1) \to \cdots \to \gamma(p_k) = \gamma'$. We collect positions j such that $k_j = 1$ and define them as $j_1, \ldots, j_{p_k}$ (the order is random). Small neighbourhood $\mathcal{N}_r$ is then defined as follows:

$$\mathcal{N}_r = \mathcal{N}(\gamma(r-1), j_r) = \Big\{\gamma^* \in \Gamma \mid \gamma_j^* = \gamma(r-1)_j \ \forall j \neq j_r\Big\}. \tag{31}$$

Each small neighbourhood $\mathcal{N}_r$ only consists of two models, $\gamma(r-1)$ and γ^*, which only differ with $\gamma(r-1)$ at position j_r. The resulting proposal mass function is

$$q_k(\gamma, \gamma') = \prod_{r=1}^{p_k} q_{\mathcal{N}_r}(\gamma(r-1), \gamma(r)) \tag{32}$$

where $q_{\mathcal{N}}$ is the locally informed proposal over neighbourhood $\mathcal{N}$ and is defined by

$$q_{\mathcal{N}}(\gamma, \gamma') \propto \begin{cases} g\left(\frac{\pi(\gamma')p(k|\gamma')}{\pi(\gamma)p(k|\gamma)}\right)\left(\frac{\zeta}{1-\zeta}\right)^{d_H(\gamma,\gamma')}, & \text{if } \gamma' \in \mathcal{N} \\ 0, & \text{otherwise.} \end{cases} \tag{33}$$

Tuning parameter $\zeta \in (\epsilon, 1-\epsilon)$ denotes the non-informative jumping probability. Two different methods for adapting ζ are provided in [33]. One of the key factors influencing the performance of the informed proposal in (33) is the choice of weighting function g. Given a positive real number $x > 0$, a balancing function is defined as a function g that satisfies the condition $g(x) = xg(1/x)$. The locally informed proposal constructed with a balancing function is locally optimal in terms of Peskun ordering under mild conditions [31]. For the comparisons between different balancing functions, see Supplement B.1.3 of [31]. In this work, we exclusively focus on the Hastings' choice of balancing function given

by $g_H(x) = \min\{1, x\}$, as g_H has demonstrated better empirical performance in many problems (e.g., [33]).

To construct a π-reversible chain in the MH scheme, we define a collection of neighbourhoods $\{\mathcal{N}'_r\}$ for the reverse moves, where $\{\mathcal{N}'_r\}$ are identical to $\{\mathcal{N}_r\}$ but with reverse order. For a more detailed explanation of the PARNI proposal, we refer to Section 4.2.1 of [33]. The MH acceptance probability of the PARNI proposal is given by

$$\begin{aligned}\alpha(\gamma, \gamma') &= \min\left\{1, \frac{\pi(\gamma')p(k|\gamma')q_k(\gamma', \gamma)}{\pi(\gamma)p(k|\gamma)q_k(\gamma, \gamma')}\right\} \\ &= \min\left\{1, \frac{\hat{p}(y|\gamma')p(\gamma')p(k|\gamma')q_k(\gamma', \gamma)}{\hat{p}(y|\gamma)p(\gamma)p(k|\gamma)q_k(\gamma, \gamma')}\right\}. \end{aligned} \tag{34}$$

Remark 1. *The concept of a neighbourhood is also used in other schemes designed to estimate discrete posterior distributions, including the Shotgun Stochastic Search (SSS) approach [34] and Hamming ball sampler (HBS) [28]. The SSS method works on the same neighbourhood as that constructed with the add–delete–swap proposal [22]. Given the current model, γ, SSS constructs a neighbourhood $\mathcal{N}(\gamma)$ that comprises three disjoint sub-neighbourhoods: $\mathcal{N}_a(\gamma)$, $\mathcal{N}_d(\gamma)$ and $\mathcal{N}_s(\gamma)$. The "addition" neighbourhood, $\mathcal{N}_a(\gamma)$, is formed by adding a covariate into the model, and similarly, the "deletion" neighbourhood, $\mathcal{N}_d(\gamma)$, is formed by deleting a covariate from the model. Lastly, $\mathcal{N}_s(\gamma)$ is obtained by swapping an included covariate with an excluded one. On the contrary, the HBS constructs neighbourhoods based on the Hamming ball, $\mathcal{H}_d(\gamma)$, consisting of models that differ from γ by at most d positions. The typical example is the 1-Hamming ball, denoted by $\mathcal{H}_1(\gamma)$. It is worth mentioning that the SSS and HBS approaches construct neighbourhoods with sizes of $(p_\gamma + 1)p$ and p, respectively. By contrast, the PARNI proposal constructs neighbourhoods that are typically approximately of size p_{γ^*}, where γ^* denotes the true underlying model. Assuming that the size of the true underlying model is much smaller than p, as is typical in many applications, $p_{\gamma^*} \ll p$, and PARNI exhibits a higher level of scalability in handling the large-p data in comparison to SSS and HBS.*

In the remaining parts of this section, we will describe a novel scheme to estimate tuning parameters A and D, and a new method for efficiently estimating the marginal likelihood in the locally informed proposal of (33).

4.2. New Adaptation Scheme on Algorithmic Tuning Parameters

The performance of the PARNI proposal is largely dictated by the choice of algorithmic tuning parameters A and D. Griffin et al. [26] consider the informed proposal of the form

$$A_j = \min\left\{1, \frac{\pi_j}{1-\pi_j}\right\}, \quad D_j = \min\left\{1, \frac{1-\pi_j}{\pi_j}\right\} \tag{35}$$

where π_j denotes the PIP for the j-th covariate and is defined by $\pi_j = \pi(\gamma_j = 1)$. In their ASI scheme for BVS in the linear regression model, tuning parameters π are adaptively updated based on a Rao-Blackwellised estimate of the PIP given in Equation (10) of [26]. Wan and Griffin [27] extend ASI to the logistic regression model, in which they derive Rao-Blackwellised estimates of PIPs conditioned on the Pólya-gamma latent variables. Generalising this method to other generalised linear models and survival models that lack a suitable data-augmentation scheme is challenging. As the analytic marginal likelihood is inaccessible, it becomes intractable to derive Rao-Blackwellised estimates of PIPs. As an alternative, a simple Monte Carlo average over the output $\{\gamma^{(l)}\}_{l=1}^{L}$ can be taken, where L is the current iteration number. This ergodic average is calculated as

$$\tilde{\pi}_j^{(L)} = \frac{1}{L}\sum_{l=1}^{L} \mathbb{I}\left\{\gamma_j^{(l)} = 1\right\}. \tag{36}$$

The ergodic average tends to be broad and biased, and it often downweights the importance of highly correlated covariates. Using the ergodic average directly in the PARNI proposal, however, results in a feedback effect, wherein a poor ergodic average leads to inadequate exploration over the sample space, leading to a subsequent bad ergodic average. To combat this phenomenon, we consider the following composition of two measures: a "warm-start" approximation, $\tilde{\pi}^{(0)}$, and the ergodic average, $\tilde{\pi}^{(L)}$, obtained from the first L samples. This composite estimate is adaptively updated using the formula

$$\hat{\pi}_j^{(L)} = \phi_L \tilde{\pi}_j^{(0)} + (1 - \phi_L)\tilde{\pi}_j^{(L)} \tag{37}$$

where $\{\phi_l\}_{l=1}^{L}$ is a set of weights that control the trade-off between the warm-start approximation and the ergodic average.

Warm-start approximation $\tilde{\pi}_j^{(0)}$ is computed in the following way: Given the initial model of the Markov Chain, $\gamma^{(0)}$, and two related models, $\gamma^{j\uparrow} = (\gamma_j = 1, \gamma_{-j}^{(0)})$ and $\gamma^{j\downarrow} = (\gamma_j = 0, \gamma_{-j}^{(0)})$, for the j-th component, the Rao-Blackwellised estimate of the j-th PIP at model $\gamma^{(0)}$ is given by

$$\mathbb{P}(\gamma_j = 1 | \gamma_{-j} = \gamma_{-j}^{(0)}, y) = \frac{\pi(\gamma^{j\uparrow})}{\pi(\gamma^{j\uparrow}) + \pi(\gamma^{j\downarrow})} = \frac{\frac{p(y|\gamma^{j\uparrow})p(\gamma^{j\uparrow})}{p(y|\gamma^{j\downarrow})p(\gamma^{j\downarrow})}}{1 + \frac{p(y|\gamma^{j\uparrow})p(\gamma^{j\uparrow})}{p(y|\gamma^{j\downarrow})p(\gamma^{j\downarrow})}}.$$

We consider the ALA in (28) initialised at the origin to estimate the intractable Bayes factor, $p(y|\gamma^{j\uparrow})/p(y|\gamma^{j\downarrow})$, for including the j-th covariate. Let η_i be the i-th linear predictor, η_i^0 be the i-th linear predictor evaluated at the origin (i.e., $\eta_i^0 = 0$), X_j denote the j-th column of data matrix X, $\tilde{y}$ be a vector with i-th component equal to $\partial p(y|\theta_\gamma, \gamma)/\partial \eta_i$ evaluated at $\eta_i = \eta_i^0$ and W be a matrix such that $W_{il} = -\partial^2 p(y|\theta_\gamma, \gamma)/\partial \eta_i \partial \eta_l$ evaluated at $\eta_i = \eta_i^0$ and $\eta_l = \eta_l^0$. Thanks to the Schur complement, we can facilitate the computation of p Bayes factors as in [26,27]: when $\gamma^{(0)} = \gamma^{j\downarrow}$,

$$\frac{\tilde{p}(y|\gamma^{j\uparrow})}{\tilde{p}(y|\gamma^{j\downarrow})} = d_j^{\uparrow -\frac{1}{2}} g^{-\frac{1}{2}} \exp\left\{ \frac{1}{2d_j^{\uparrow}} (\tilde{y}^T X_\gamma \Lambda_\gamma^{-1} X_\gamma^T W X_j - \tilde{y}^T X_j) \right\} \tag{38}$$

where $\Lambda_\gamma = X_\gamma W X_\gamma + V_\gamma^{-1}$ and $d_j^{\uparrow} = X_j^T W X_j + 1/g - X_j^T W X_\gamma \Lambda_\gamma^{-1} X_\gamma^T W X_j$; when $\gamma^{(0)} = \gamma^{j\uparrow}$

$$\frac{\tilde{p}(y|\gamma^{j\uparrow})}{\tilde{p}(y|\gamma^{j\downarrow})} = d_j^{\downarrow -\frac{1}{2}} g^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2} d_j^{\downarrow} \left(\tilde{y}^T X_\gamma (\Lambda_\gamma^{-1})_{\cdot, q+p_j} \right)^2 \right\} \tag{39}$$

where $d_j^{\downarrow} = 1/(\Lambda_\gamma^{-1})_{q+p_j, q+p_j}$ and p_j is the ordered position of the j-th variable. When working with data that involve a high level of collinearity, it is also possible to increase the number of the ALA Rao-Blackwellised estimates.

The last building block of adapting $\hat{\pi}_j^{(L)}$ is defining weight ϕ_l. We employ a straightforward construction of ϕ_l given by

$$\phi_l = \begin{cases} 1 - \frac{1}{2}(N_b - l + 1)^{-0.5} & \text{if } l \leq N_b \\ \frac{1}{2}(l - N_b)^{-0.5} & \text{if } l > N_b. \end{cases} \tag{40}$$

where N_b denotes the length of the burn-in period. This choice results in a weight that exceeds $1/2$ during the period of burn-in and drops below $1/2$ afterwards. Consequently, the PARNI proposal initially relies on the warm-start approximation to explore the model space. As the chain converges to the high-probability region and the ergodic average stabilises, the PARNI proposal gradually uses more information from the ergodic average. After running for a longer time, the PARNI proposal completely relies on the ergodic average.

4.3. The Adaptive ALA Informed Proposal

In each iteration of the PARNI proposal, the locally informed proposal in (33) relies on computing the posterior model probabilities. Using the estimates from the Laplace approximation or correlated pseudo-marginal scheme in PARNI can be computationally intractable in "large-p, large-n" situations due to the use of an optimisation algorithm that most run many times in one iteration. It should be noted, however, that the model probabilities in the locally informed proposal do not need to precisely match the true posterior model probabilities, and the PARNI proposal can still generate samples that preserve π-reversibility as long as the correct (or proper estimate) π is used in the MH acceptance probability of (34). In the locally informed proposal, one can incorporate the approximate Laplace approximation initialised at the origin to design the proposal distribution. In the MH acceptance probability, we can then use the estimates obtained from the Laplace approximation or correlated pseudo-marginal method. Based on empirical observations, however, this ALA informed proposal may not always mix well. One reason for this is the phenomenon of downweighting the model probabilities of non-null models in favour of the null model, resulting in an informed proposal that is less informative than the true likelihood. The simulated chain is, therefore, more likely to get stuck and becomes less effective in exploring the model space.

Alternatively, we can note that the ALA estimate coincides with the Laplace approximation when the initial value is chosen to be posterior mode $\hat{\theta}_\gamma$ under model $\mathcal{M}_\gamma$. Therefore, the accuracy of the ALA estimate is crucially influenced by the choice of initial value θ^0_γ. We employ the ALA informed proposal with an adaptive initial value for ALA (adaptive ALA), which aims to reduce the estimation errors and thus improve the overall performance of the MCMC algorithm.

For each model in the neighbourhood, the adaptive ALA starts with an initial guess of linear predictor η and proceeds with the following steps:

1. Calculate a "guess" estimate from linear predictor η:

$$\theta^0_\gamma = (J^T_\gamma J_\gamma)^{-1} J^T_\gamma \eta. \tag{41}$$

2. Perform one step of Newton's method and obtain an updated estimate of the coefficient:

$$\tilde{\theta}_\gamma = \theta^0_\gamma - \left(H^0_\gamma\right)^{-1} g^0_\gamma \tag{42}$$

   where g^0_γ and H^0_γ are the gradient and Hessian of the negated log-posterior density evaluated at θ^0_γ, respectively.
3. Use the ALA estimate in (28) with $\tilde{\theta}_\gamma$ as the initial value to estimate marginal likelihood $p(y|\gamma)$.

Practically speaking, we can skip step 1 with the matrix inverse operation in (41) and obtain $\tilde{\theta}_\gamma$ directly from the initial guess of linear predictor η. This simplification is followed by [63] and is given in Appendix C. This approach leads to a coherent computational scheme that is easy to implement. We adaptively update the initial-guess η according to

$$\hat{\eta}^{(L)}_i = \frac{1}{L} \sum_{l=1}^{L} \hat{\eta}_{\gamma^{(l)},i} \tag{43}$$

where $\hat{\eta}_{\gamma,i} = J_\gamma \hat{\theta}_\gamma$ is the "optimal" i-th linear predictor obtained from MAP estimate $\hat{\theta}_\gamma$ under model $\mathcal{M}_\gamma$. By storing the MAP estimate obtained from the Laplace approximation or correlated pseudo-marginal scheme, we can compute the linear predictor without introducing additional computational costs.

In addition, we experimented with adapting coefficients $\hat{\beta}$ from the posterior samples and using the coefficients of the covariates selected by γ to navigate the ALA. This approach did not work well, however, because the posterior distribution of β differs significantly from

the posterior distribution of β conditioned on model γ. In contrast, the linear predictors offer more stability, in the sense that they do not vary as much across different models.

Combining all of the above components, we have the PARNI proposal. The complete algorithm is outlined in Algorithm 1.

Algorithm 1 The algorithmic pseudo-code of the Point-wise Adaptive Random Neighbourhood Sampler with Informed proposal (PARNI)

Initialise the chain at $\gamma^{(0)}$ and compute $\{\tilde{\pi}_j^{(0)}\}_{j=1}^p$;
for $i = 1$ to $i = N$ **do**
  Sample $k \sim p(\ \cdot\ |\gamma^{(i-1)})$ as in (29);
  Set $\gamma(0) = \gamma^{(i-1)}$, $p_k = \sum_{j=1}^p k_j$ and define $j_1, \ldots, j_{p_k}$;
  for $r = 1$ to $r = p_k$ **do**
    Construct $\mathcal{N}_r$ as in (31) and estimate $p(y|\gamma^*)$ for all $\gamma^* \in \mathcal{N}_r$ as in Section 4.3;
    Sample $\gamma(r) \sim q_{\mathcal{N}_r}(\gamma(r-1),\ \cdot\ )$ as in (33);
  end for
  Set $\gamma' = \gamma(p_k)$, estimate $p(y|\gamma')$ by LA or CPM and sample $U \sim \text{Unif}(0,1)$;
  If $U < \alpha(\gamma^{(i-1)}, \gamma')$ as in (34), then $\gamma^{(i)} = \gamma'$, else $\gamma^{(i)} = \gamma^{(i-1)}$;
  for $j = 1$ to $j = p$ **do**
    Update $\tilde{\pi}_j^{(i)}$ as in (36) and $\hat{\pi}_j^{(i)}$ as in (37);
    Update $A_j^{(i)} = \min\left\{1, \hat{\pi}_j^{(i)}/(1-\hat{\pi}_j^{(i)})\right\}$;
    Update $D_j^{(i)} = \min\left\{1, (1-\hat{\pi}_j^{(i)})/\hat{\pi}_j^{(i)}\right\}$;
  end for
  Update $\omega^{(i)}$ using the selected adaption scheme and $\hat{\eta}^{(i)}$ as in (43);
end for

5. Experiments

5.1. Simulated Data-Sets with Adaptive ALA Informed Proposal

In this subsection, we study the mixing behaviour of different versions of the PARNI proposal for the logistic regression model, Cox PHs and Weibull survival models. We simulate two data-sets with 500 covariates and 500 observations as described in Appendix D and compare the following four algorithms:

- **PARNI-adaptiveALA:** The PARNI proposal with adaptive approximate Laplace approximation in the informed proposal and the correlated pseudo-marginal scheme in the MH acceptance probability.
- **PARNI-LA:** The PARNI proposal with Laplace approximation in the informed proposal and the correlated pseudo-marginal scheme in the MH acceptance probability.
- **PARNI-ALA:** The PARNI proposal with approximate Laplace approximation in the informed proposal and the correlated pseudo-marginal scheme in the MH acceptance probability.
- **ADS (thinned):** The PARNI proposal with approximate Laplace approximation in the informed proposal and the correlated pseudo-marginal scheme in the MH acceptance probability.

The first three algorithms were run for 10,000 iterations, with the first 2000 iterations being discarded as burn-in, whereas the ADS (thinned) proposal was run for a CPU time similar to that of PARNI-adaptiveALA and PARNI-ALA, with all collected samples being thinned to 10,000 values.

Figure 1 presents trace plots of the log-posterior model probability and bar plots of CPU time for the PARNI-adaptiveALA, PARNI-LA, PARNI-ALA and ADS (thinned) proposals in the logistic model, and the Cox PHs and Weibull models. In all three models, the PARNI-adaptiveALA proposal mixes as well as the PARNI-LA proposal and performs much better than the PARNI-ALA proposal. The result of the ADS (thinned) proposal provides the benchmark performance of a simple add–delete–swap MCMC scheme on

these data-sets for comparison purposes. As illustrated in Figure 1, the adaptive ALA informed proposal is computationally much cheaper than the LA informed proposal. In comparison to the ALA informed proposal, the adaptive ALA informed proposal is also computationally competitive, and it only introduces the additional computational costs of updating linear predictor $\hat{\eta}^{(L)}$ and computing initial value $\tilde{\theta}_\gamma$ from the estimate of linear predictor $\hat{\eta}^{(L)}$ as in (42). In addition, the PARNI-adaptiveALA proposal demonstrates improved mixing behaviour in comparison to the baseline add–delete–swap proposal in all three models with similar CPU time. Therefore, we conclude that the PARNI-adaptiveALA proposal is more computationally efficient than the informed proposals constructed using Laplace approximation or ALA initialised at the origin.

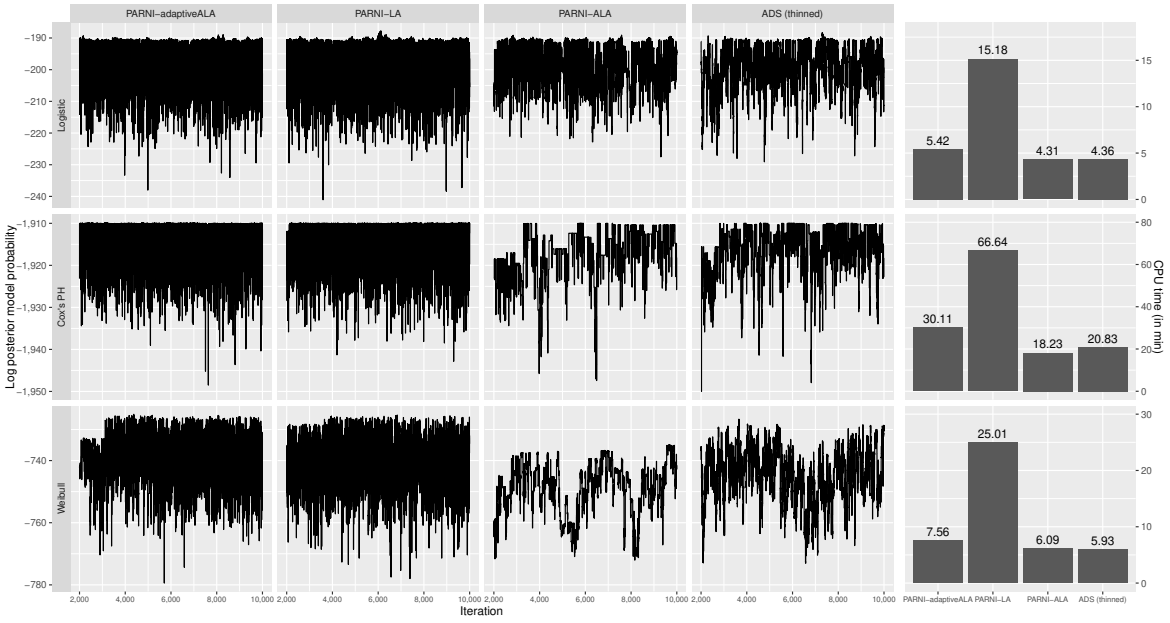

Figure 1. Left four columns: Trace plots of the log-posterior model probability from runs of the PARNI-adaptiveALA, PARNI-LA, PARNI-ALA and ADS (thinned) algorithms on simulated data-sets. Right column: Bar plots of the CPU time of simulating 10,000 samples on simulated data-sets with the PARNI-adaptiveALA, PARNI-LA, PARNI-ALA and ADS (thinned) algorithms.

5.2. Logistic Regression: Genetic Mapping for Systemic Lupus Erythematosus

Genetic mapping is a process of locating a specific gene or genetic variant within a particular genomic region and has the objective to find the precise genetic elements responsible for a particular trait or disease phenotype. One common application is to study whether an individual has a particular disease. In this scenario, one can use a logistic regression model with the response variable based on the case/control design and explanatory variables consisting of single-nucleotide polymorphisms (SNPs).

We consider a problem of identifying the SNPs that play a crucial role in predicting Systemic Lupus Erythematosus using a case/control study. It consists of genotypes from a genome-wide genetic case/control association study involving 4035 cases and 6959 controls, where the cases are SLE patients and the controls are from a public repository of European ancestry. These data were previously studied in [64] using step-wise logistic regression in a meta-analysis. In Chapter 5 of [65], Griffin and Steel apply Bayesian variable selection to analyse these SLE data but only focus on exploring the relationship between disease and SNPs on Chromosome 1. In addition to their work, we extend the study by including a total of four chromosomes. We consider a different number of SNPs for each chromosome, with Chromosome 1 having 5771 SNPs, Chromosome 3 having 42,430 SNPs, Chromosome 11 having 32,290 SNPs and Chromosome 21 having 9306 SNPs. We consider the prior specification in Section 2.2 with hyper-parameter $g = 1/4$, $\sigma_\alpha^2 = 1$ and assume the hyper-

prior of $h \sim \text{Beta}(1, (p-5)/5)$, where p denotes the number of SNPs. The full details of the data-set are provided in Table 1, including the five fixed covariates (gender and top four principal components of expressed genes) that are mandatory in all models.

Table 1. Details of Systemic Lupus Erythematosus data on Chromosomes 1, 3, 11 and 21.

Data-Set	Observations	Cases	Fixed Covariates	Genetic Covariates
Chromosome 1	10,995	4036	Gender, PC1–PC4	5771
Chromosome 3				42,430
Chromosome 11				32,290
Chromosome 21				9306

We implement the following four algorithms:

- **PARNI-DA**: PARNI proposal with Pólya-gamma data augmentation in both the informed proposal and the MH acceptance probability.
- **PARNI-CPM**: PARNI proposal with adaptive ALA informed proposal and correlated pseudo-marginal method in the MH acceptance probability.
- **ADS-DA**: Add–delete–swap proposal with Pólya-gamma data augmentation in the MH acceptance probability.
- **ADS-CPM**: Add–delete–swap proposal with the correlated pseudo-marginal method in the MH acceptance probability.

These MCMC algorithms all simulate samples from the exact posterior distribution, π. We treat the ADS-DA proposal as the baseline to showcase the rapid mixing of the PARNI proposals. Each algorithm was run for 1 h with 10 repetitions, and we recorded the estimates of PIPs. Firstly, we calculated the mean squared errors of the estimates of p PIPs compared with the "gold standard" estimates taken from the PARNI-CPM proposal, which was run for roughly 12 h. Then, we took the average over p mean squared errors to obtain the average mean squared error (average MSE). To compare the computational efficiency of the PARNI proposals with the baseline ADS-DA proposal, we provide the relative efficiency (in brackets) as the ratio of their average MSE.

The average MSEs and relative efficiency values are presented in Table 2. The PARNI proposals consistently outperform the ADS proposals in terms of the average MSE. The PARNI proposals show at least twofold improvements over the add–delete–swap proposal and lead to much larger improvements in most cases, such as in Chromosome 21, where the PARNI proposals perform 78 times better than the add–delete–swap proposal. On the other hand, both the PARNI-DA and ADS-DA proposals consistently result in smaller average MSEs compared with the PARNI-CPM and ADS-CPM proposals due to their computational advantages. Firstly, data augmentation can evaluate the conditional marginal likelihood without finding the posterior mode using iteratively re-weighted least squares. Secondly, the Pólya-gamma latent variables are drawn using the R package pgdraw (version 1.1) [56] implemented using Rcpp (version 1.0.10) [66].

Table 2. Systemic Lupus Erythematosus data: The average mean squared errors of the ADS-DA, ADS-CPM, PARNI-DA and PARNI-CPM proposals in estimating the posterior inclusion probabilities of all SNPs (*smaller is better*). The relative efficiency as the ratio of the average MSE between algorithm A and the ADS-DA proposal presented in brackets (*larger is better*). The best performance is presented in **bold**.

Data-Set	Algorithms			
	ADS-DA	ADS-CPM	PARNI-DA	PARNI-CPM
Chromosome 1	1.84×10^{-5} (1)	4.29×10^{-5} (0.43)	$\mathbf{5.14 \times 10^{-6}}$ (**3.58**)	7.34×10^{-6} (2.51)
Chromosome 3	2.01×10^{-4} (1)	2.37×10^{-4} (0.85)	8.76×10^{-5} (2.30)	$\mathbf{5.11 \times 10^{-5}}$ (**3.94**)
Chromosome 11	7.09×10^{-5} (1)	1.07×10^{-4} (0.66)	$\mathbf{9.73 \times 10^{-6}}$ (**7.29**)	9.89×10^{-6} (7.15)
Chromosome 21	1.18×10^{-5} (1)	1.67×10^{-5} (0.71)	$\mathbf{1.51 \times 10^{-7}}$ (**78.08**)	1.79×10^{-7} (65.93)

5.3. Survival Analysis: Variable Selection for Five Large Cancer-Related Gene Expression Data-Sets

We consider a total of four cancer-related real data-sets, where the first two data-sets are for breast cancer and the remaining data-sets are for lung cancer. NKI Breast Cancer Data (https://data.world/deviramanan2016/nki-breast-cancer-data, accessed on 4 September 2023) contain patient info, treatment, survival time and the 1554 most varying genes of 272 breast cancer patients. These data were analysed in [67,68] with the aim of reducing the mortality rates from this disease. The METABRIC breast cancer data-set is derived from the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) database. The METABRIC data-set was analysed in [69,70] and is publicly available in [71] (https://www.cbioportal.org/study/summary?id=brca_metabric, accessed on 4 September 2023). The data contain 1907 patients with the gene expression for 331 genes and mutations for 175 genes. Gene mutation variables are encoded as 1 if a mutation exists and 0 otherwise. For both data-sets, we include some clinical covariates, including the age of the patients and the stage of the cancer, as suggested by [72]. We also consider the treatment variables (such as chemotherapy and surgery type), which also influence survival time. The last two lung cancer data-sets, "GSE31210" and "GSE4573", were previously studied in [72], and they are publicly available in the Gene Expression Omnibus repository [73]. See Figure 1 in [72] for the estimated survival functions of these three data-sets. We provide the full details of these four real data-sets in Table 3.

Table 3. Details of 4 real data-sets for survival analysis.

Data-Set	Cancer Type	Observations	Events	Fixed Covariates	Genetic Covariates
NKI	Breast	272	77	Age, chemo, hormone, surgery, stage	1554
METABRIC	Breast	1903	622	Age, chemo, hormone, radio, surgery, stage	662
GSE31210	Lung	226	30	Age, gender, smoker, stage	54,675
GSE4573	Lung	130	63	Age, gender, stage	22,283

We consider two computational algorithms used in previous studies for logistic models, the PARNI-CPM and ADS-CPM proposals, as a data augmentation scheme is not available for the Cox PHs or Weibull model. We consider the hyper-prior of $h \sim \text{Beta}(1, (p-5)/5)$, where p denotes the number of genetic covariates, and impose a half-Cauchy hyper-prior on $\sqrt{g}$, where a Gibbs update is taken on g conditioned on the model (see Appendix A for more details). In addition, we assume $\sigma_\alpha^2 = 10^5$ and $\sigma_k^2 = 10^5$ (only for the Weibull model).

The average MSEs and relative efficiency values of the PARNI-CPM and ADS-CPM proposals on these four survival data-sets are shown in Table 4. For the Weibull model, the PARNI proposal consistently exhibits better computational efficiency compared with add–delete–swap on all four data-sets. In the case of the NKI and METABRIC data-sets, which have a relatively small number of covariates, the PARNI-CPM proposal produces PIP estimates that are seven times more accurate compared with ADS-CPM. For high-dimensional data-sets, we can obtain PIP estimates from the PARNI-CPM proposal that are two times better than ADS-CPM. The lesser improvement observed in the high-dimensional examples can be attributed to the increasing number of unimportant covariates, where both algorithms are good at excluding these unimportant covariates from the models.

The Bayesian variable selection in the Cox PHs model with partial likelihood is generally more challenging compared with the Weibull model. The primary reason is that the inclusion of the non-parametric setup introduces additional complexities in evaluating the log-likelihood functions and its Hessian matrices. The PARNI-CPM proposal provides roughly two times better estimates on the NKI and GSE4573 data-sets compared with ADS-CPM. However, the ADS-CPM proposal shows better performance on the remaining two data-sets. In the "small-p, large-n" METABRIC data, the add–delete–swap proposal shows greater computational efficiency compared with the PARNI proposal, as the informed

proposal needs to evaluate many computationally expensive Hessian matrices. In fact, the computation of evaluating the Hessian matrix scales with the order of $\mathcal{O}(n^2)$, in contrast with parametric models, where the computation of the Hessian matrix scales linearly with n. In the GSE31210 data with few strong signals, the posterior distribution on model space is relatively flat, and both algorithms have smaller average MSEs compared with the other data-sets. In particular, the add–delete–swap proposal can run for more iterations; it is, therefore, more computationally efficient compared with the PARNI-CPM proposal.

Table 4. Survival analysis data: The average mean squared errors of the ADS-CPM and PARNI-CPM proposals in estimating posterior inclusion probabilities of all genetic covariates (*smaller is better*). The relative efficiency as the ratio of the average MSE between algorithm A and the ADS-CPM proposal presented in brackets (*larger is better*). The best performance is presented in **bold**.

Data-Set	Cox PHs		Weibull Model	
	ADS-CPM	**PARNI-CPM**	**ADS-CPM**	**PARNI-CPM**
NKI	5.54×10^{-4} (1)	$\mathbf{3.19 \times 10^{-4}}$ (**1.74**)	3.21×10^{-5} (1)	$\mathbf{4.33 \times 10^{-6}}$ (**7.40**)
METABRIC	$\mathbf{1.27 \times 10^{-3}}$ (**1**)	3.80×10^{-3} (0.33)	1.36×10^{-3} (1)	$\mathbf{1.73 \times 10^{-4}}$ (**7.89**)
GSE31210	$\mathbf{1.26 \times 10^{-6}}$ (**1**)	3.56×10^{-6} (0.35)	3.91×10^{-5} (1)	$\mathbf{2.16 \times 10^{-5}}$ (**1.81**)
GSE4573	4.57×10^{-5} (1)	$\mathbf{2.54 \times 10^{-5}}$ (**1.83**)	3.56×10^{-5} (1)	$\mathbf{8.37 \times 10^{-6}}$ (**4.26**)

6. Discussion

In this work, we apply the PARNI proposal to Bayesian variable selection problems in generalised linear models and survival models. We find that the informed proposal obtained from the approximate Laplace approximation with our new adaptive initial point yields improved efficiency and accuracy in posterior sampling. We compare the performance of the PARNI proposal with the baseline add–delete–swap proposal in numerous "large-p, large-n" real-world data-sets, and the PARNI proposal with the correlated pseudo-marginal method provides PIP estimates with smaller mean squared errors than the add–delete–swap proposal in most of the problems. The numerical results from the Cox PHs also provide useful insights to improve the PARNI proposal in the future. Code to run the PARNI proposal on the logistic regression model, and the Cox PHs and Weibull models is available at https://github.com/XitongLiang/The-PARNI-scheme.git (accessed on 4 September 2023).

In addition to the three models described in the paper, the proposed technique can be extended to other generalised linear models and survival models. Two possible extensions are the Gamma generalised linear model [74] and various Bayesian non-parametric approaches to survival analysis [75]. It is still a challenging problem to reduce the computational cost of simulating samples when a data-set contains a large number of observations. As highlighted in [76], simple sub-sampling strategies may not lead to a substantial improvement in the computational efficiency of posterior sampling. It would be interesting, therefore, to design an efficient PARNI scheme specifically tailored for large-n data-sets.

Author Contributions: Conceptualization, X.L., S.L. and J.G.; methodology, X.L., S.L. and J.G.; software, X.L.; formal analysis, X.L., S.L. and J.G.; data curation, X.L.; writing—original draft preparation, X.L.; writing—review and editing, S.L. and J.G.; visualization, X.L., S.L. and J.G.; supervision, S.L. and J.G. All authors have read and agreed to the published version of the manuscript.

Funding: S.L. is supported by a UK Engineering and Physical Sciences Research Council grant (EP/V055380/1).

Institutional Review Board Statement: Not applicable.

Data Availability Statement: Systemic Lupus Erythematosus was described in [64]. The NKI data-set is publicly available at https://data.world/deviramanan2016/nki-breast-cancer-data (accessed on 4 September 2023). The METABRIC data-set is publicly available at https://www.cbioportal.org/study/summary?id=brca_metabric (accessed on 4 September 2023). The GSE310210 AND GSE4573

are publicly available in the Gene Expression Omnibus repository [73]. The experiments were performed using R software, version 4.2.3 [77].

Acknowledgments: The authors would like to acknowledge Tim Vyse and David Morris of King's College London for providing the data for the fine-mapping example.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

ALA	Approximate Laplace approximation
BVS	Bayesian variable selection
Cox PHs	Cox proportional hazards
CPM	Correlated pseudo-marginal
DA	Data augmentation
HBS	Hamming ball sampler
INLAs	Integrated nested Laplace approximations
LA	Laplace approximation
MCMC	Markov Chain Monte Carlo
MH	Metropolis–Hastings
PARNI	Point-wise implementation of Adaptive Random Neighbourhood Informed proposal
PDMP	Piece-wise deterministic Markov process
PIP	Posterior inclusion probability
RJMCMC	Reversible Jump Markov Chain Monte Carlo
SSS	Shotgun Stochastic Search
SVB	Sparse Variational Bayes
VB	Variational Bayes

Appendix A. Updating g in Hierarchical Model

We impose a standard half-Cauchy hyper-prior on $\sqrt{g}$, which defines the following density with $s = \sqrt{g}$:

$$p_s(s) = \frac{2}{\pi}\frac{1}{1+s^2}. \tag{A1}$$

As s can only take the non-negative and is defined on $(0, +\infty)$, we consider an MH update on the projection $\nu = \log(g) = 2\log(s)$, which is defined on $\mathbb{R}$. The transformed density on ν is

$$p_\nu(\nu) = \frac{2}{\pi}\frac{1}{1+\exp(\nu)} \times \frac{1}{2}\exp\left(\frac{1}{2}\nu\right). \tag{A2}$$

We can express p_ν in terms of g as

$$p_\nu(g) = \frac{1}{\pi}\frac{\sqrt{g}}{1+g}. \tag{A3}$$

Given that $\nu = \log(g)$ is the current value, we consider a random walk Metropolis with variance σ_g^2. By sampling $Z \sim N(0,1)$, we propose $\nu' = \nu + \sigma_g Z$ (equivalent to $g' = g\exp(\sigma_g Z)$). The MH acceptance probability of accepting this new proposal $\nu' = \log(g')$ is given by

$$\alpha(g, g') = \min\left\{1, \frac{p(y|\gamma, g')p_\nu(g')}{p(y|\gamma, g)p_\nu(g)}\right\}. \tag{A4}$$

At the i-th iteration, the variance of the random walk Metropolis proposal is adaptively updated according to the formula

$$\log((\sigma_g^2)^{(i+1)}) = \log((\sigma_g^2)^{(i)}) + i^{-0.7} \times (\alpha(g, g') - \tau) \tag{A5}$$

where τ is the optimal acceptance probability and is often set to 0.234.

Appendix B. Updating k in Weibull Model

Recall from Section 2.5 that a normal prior $N(0, \sigma_k^2)$ is assigned to log-transformed scale parameter k in the Weibull model. Let $s = \log(k)$; we have

$$p_s(s) = \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left\{-\frac{s^2}{2\sigma_k^2}\right\}. \tag{A6}$$

We consider a Gaussian random walk Metropolis with variance σ_{rw}^2 on s. After sampling $Z \sim \text{N}(0,1)$, we propose $s' = s + \sigma_{\text{rw}} Z$. The MH acceptance probability of accepting this new proposal $s' = \log(g')$ is given by

$$\alpha(s, s') = \min\left\{1, \frac{p(y|\gamma, g' = \exp(s'))p_s(s')}{p(y|\gamma, g = \exp(s))p_s(s)}\right\}. \tag{A7}$$

Similar to Appendix A, at the i-th iteration, the variance of the random walk Metropolis proposal is adaptively updated according to the formula

$$\log((\sigma_{\text{rw}}^2)^{(i+1)}) = \log((\sigma_{\text{rw}}^2)^{(i)}) + i^{-0.7} \times (\alpha(s, s') - \tau) \tag{A8}$$

where τ is the optimal acceptance probability and is often set to 0.234.

Appendix C. From Newton's Method to IRLS in Bayesian Modelling

Under model $\mathcal{M}_\gamma$, Newton's method leads to the update on the coefficients

$$\theta_\gamma^{(n+1)} = \theta_\gamma^{(n)} - \left(X_\gamma^T W_\gamma^{(n)} X_\gamma + V_\gamma^{-1}\right)^{-1} \left(X_\gamma^T \tilde{y}_\gamma^{(n)} + V_\gamma^{-1}\theta_\gamma^{(n)}\right) \tag{A9}$$

where $\eta^{(n)} = X_\gamma \theta_\gamma^{(n)}$ is the linear predictor; $W_\gamma^{(n)}$ is the negative second derivative of log-likelihood with respect to the linear predictor, $(W_\gamma^{(n)})_{il} = -\partial^2 / \partial\eta_{\gamma,i}\partial\eta_{\gamma,l}(p(y|\theta_\gamma, \gamma))$, evaluated at $\eta_{\gamma,i}^{(n)}$ and $\eta_{\gamma,l}^{(n)}$; and $\tilde{y}_\gamma^{(n)}$ is the negative first derivative of log-likelihood with respect to the linear predictor, $(\tilde{y}_\gamma^{(n)})_i = -\partial / \partial\eta_{\gamma,i}(p(y|\theta_\gamma, \gamma))$, evaluated at $\eta_{\gamma,i}^{(n)}$.

Multiplying both sides by $(X_\gamma^T W_\gamma^{(n)} X_\gamma + V_\gamma^{-1})$ yields

$$\left(X_\gamma^T W_\gamma^{(n)} X_\gamma + V_\gamma^{-1}\right)\theta_\gamma^{(n+1)} = \left(X_\gamma^T W_\gamma^{(n)} X_\gamma + V_\gamma^{-1}\right)\theta_\gamma^{(n)} - X_\gamma^T \tilde{y}_\gamma^{(n)} - V_\gamma^{-1}\theta_\gamma^{(n)}. \tag{A10}$$

We can simplify the RHS as

$$\left(X_\gamma^T W_\gamma^{(n)} X_\gamma + V_\gamma^{-1}\right)\theta_\gamma^{(n+1)} = X_\gamma^T W_\gamma^{(n)} X_\gamma \theta_\gamma^{(n)} - X_\gamma^T \tilde{y}_\gamma^{(n)} \tag{A11}$$

$$\Rightarrow \quad \left(X_\gamma^T W_\gamma^{(n)} X_\gamma + V_\gamma^{-1}\right)\theta_\gamma^{(n+1)} = X_\gamma^T W_\gamma^{(n)} \left(\eta_\gamma^{(n)} - (W_\gamma^{(n)})^{-1} \tilde{y}_\gamma^{(n)}\right) \tag{A12}$$

as $\eta_\gamma^{(n)} = X_\gamma \theta_\gamma^{(n)}$. We multiply both sides by $\left(X_\gamma^T W_\gamma^{(n)} X_\gamma + V_\gamma^{-1}\right)^{-1}$ and obtain the update in the form of IRLS and linear predictor

$$\theta_\gamma^{(n+1)} = \left(X_\gamma^T W_\gamma^{(n)} X_\gamma + V_\gamma^{-1}\right)^{-1} X_\gamma^T W_\gamma^{(n)} \left(\eta_\gamma^{(n)} - (W_\gamma^{(n)})^{-1} \tilde{y}^{(n)}\right). \tag{A13}$$

Appendix D. Data Simulation

We take the same strategy as in [23,27] to simulate logistic regression data. Assume that the linear predictor is defined by $\eta = X\beta$, where X are generated from the multivariate normal distribution; we map mean value μ_i with linear predictor η_i using a logistic link function given by $\mu_i = \exp(\eta_i)/(1 + \exp(\eta_i))$ and simulate $y_i \sim \text{Bern}(\mu_i)$. We conduct the same AR design for the covariates, where each observation (row) of data design matrix X

follows a multivariate normal distribution with mean zero and covariance Σ, with entries $\Sigma_{ij} = 0.6^{|i-j|}$. In terms of coefficient β, only the first 10 values are taken to be non-zero, and β is defined by

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

For the survival model, we take the same construction on data design matrix X, coefficient β and linear predictor $\eta = X\beta$ as in the logistic model. The survival time of each individual is simulated from a flexible generalised gamma parametric survival model [78] as suggested in [15]. The generalised gamma parametric survival model encompasses four commonly used survival models, the exponential, Weibull, log-normal and gamma survival models, as special cases. We adopt a similar hyper-parameter specification for the generalised gamma parametric survival model as presented in [15] with $\sigma = 0.8$ and $q = -2$. In addition, we consider hyper-parameters $\sigma_\alpha^2 = 100$, $g = 1$ and $h = 10/500$ for all these three models.

References

1. Hastie, T.; Tibshirani, R.; Wainwright, M. *Statistical Learning with Sparsity: The Lasso and Generalizations*; CRC Press: Boca Raton, FL, USA, 2015.
2. Akaike, H. Information theory and an extension of the maximum likelihood principle. In *Selected Papers of Hirotugu Akaike*; Springer: New York, NY, USA, 1998; pp. 199–213.
3. Schwarz, G. Estimating the dimension of a model. *Ann. Stat.* **1978**, *6* , 461–464. [CrossRef]
4. Spiegelhalter, D.J.; Best, N.G.; Carlin, B.P.; Van Der Linde, A. Bayesian measures of model complexity and fit. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **2002**, *64*, 583–639. [CrossRef]
5. Watanabe, S.; Opper, M. Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *J. Mach. Learn. Res.* **2010**, *11*, 3571–3594.
6. Mitchell, T.J.; Beauchamp, J.J. Bayesian variable selection in linear regression. *J. Am. Stat. Assoc.* **1988**, *83*, 1023–1032. [CrossRef]
7. Chipman, H.; George, E.I.; McCulloch, R.E.; Clyde, M.; Foster, D.P.; Stine, R.A. The practical implementation of Bayesian model selection. *Lect. Notes Monogr. Ser.* **2001**, *38*, 65–134.
8. Tian, Y.; Bondell, H.D.; Wilson, A. Bayesian variable selection for logistic regression. *Stat. Anal. Data Min. ASA Data Sci. J.* **2019**, *12*, 378–393. [CrossRef]
9. Chen, M.H.; Huang, L.; Ibrahim, J.G.; Kim, S. Bayesian variable selection and computation for generalized linear models with conjugate priors. *Bayesian Anal.* **2008**, *3*, 585. [CrossRef]
10. Cox, D.R. Partial likelihood. *Biometrika* **1975**, *62*, 269–276. [CrossRef]
11. Ibrahim, J.G.; Chen, M.H.; MacEachern, S.N. Bayesian variable selection for proportional hazards models. *Can. J. Stat.* **1999**, *27*, 701–717. [CrossRef]
12. Ibrahim, J.G.; Chen, M.H.; Kim, S. Bayesian variable selection for the Cox regression model with missing covariates. *Lifetime Data Anal.* **2008**, *14*, 496–520. [CrossRef]
13. Held, L.; Gravestock, I.; Sabanés Bové, D. Objective Bayesian model selection for Cox regression. *Stat. Med.* **2016**, *35*, 5376–5390. [CrossRef] [PubMed]
14. Rossell, D.; Rubio, F.J. Additive Bayesian variable selection under censoring and misspecification. *Stat. Sci.* **2023**, *38*, 13–29. [CrossRef]
15. 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 tumour markers of breast cancer survival. *Stat. Methods Med. Res.* **2017**, *26*, 414–436. [CrossRef] [PubMed]
16. Rossell, D.; Abril, O.; Bhattacharya, A. Approximate Laplace approximations for scalable model selection. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **2021**, *83*, 853–879. [CrossRef]
17. Green, P.J. Trans-dimensional Markov chain Monte Carlo. In *Highly Structured Stochastic Systems* ; Oxford Statistical Science Series; Oxford University Press: Oxford, UK, 2003; pp. 179–198.
18. Jasra, A.; Stephens, D.A.; Holmes, C.C. Population-based reversible jump Markov chain Monte Carlo. *Biometrika* **2007**, *94*, 787–807. [CrossRef]
19. Tanner, M.A.; Wong, W.H. The calculation of posterior distributions by data augmentation. *J. Am. Stat. Assoc.* **1987**, *82*, 528–540. [CrossRef]
20. Polson, N.G.; Scott, J.G.; Windle, J. Bayesian inference for logistic models using Pólya–Gamma latent variables. *J. Am. Stat. Assoc.* **2013**, *108*, 1339–1349. [CrossRef]
21. Deligiannidis, G.; Doucet, A.; Pitt, M.K. The correlated pseudomarginal method. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **2018**, *80*, 839–870. [CrossRef]

22. Brown, P.J.; Vannucci, M.; Fearn, T. Bayesian wavelength selection in multicomponent analysis. *J. Chemom. J. Chemom. Soc.* **1998**, *12*, 173–182. [CrossRef]
23. Yang, Y.; Wainwright, M.J.; Jordan, M.I. On the computataional complexity of high-dimensional Bayesian variable selection. *Ann. Stat.* **2016**, *44*, 2497–2532. [CrossRef]
24. Andrieu, C.; Thoms, J. A tutorial on adaptive MCMC. *Stat. Comput.* **2008**, *18*, 343–373. [CrossRef]
25. Lamnisos, D.; Griffin, J.E.; Steel, M.F. Transdimensional sampling algorithms for Bayesian variable selection in classification problems with many more variables than observations. *J. Comput. Graph. Stat.* **2009**, *18*, 592–612. [CrossRef]
26. Griffin, J.; Łatuszyński, K.; Steel, M. In search of lost mixing time: Adaptive Markov chain Monte Carlo schemes for Bayesian variable selection with very large p. *Biometrika* **2021**, *108*, 53–69. [CrossRef]
27. Wan, K.Y.Y.; Griffin, J.E. An adaptive MCMC method for Bayesian variable selection in logistic and accelerated failure time regression models. *Stat. Comput.* **2021**, *31*, 6. [CrossRef]
28. Titsias, M.K.; Yau, C. The Hamming ball sampler. *J. Am. Stat. Assoc.* **2017**, *112*, 1598–1611. [CrossRef]
29. Zanella, G.; Roberts, G. Scalable importance tempering and Bayesian variable selection. *J. R. Stat. Soc. Ser. Stat. Methodol.* **2019**, *81*, 489–517. [CrossRef]
30. Jankowiak, M. Fast Bayesian Variable Selection in Binomial and Negative Binomial Regression. *arXiv* **2021**, arXiv:2106.14981.
31. Zanella, G. Informed proposals for local MCMC in discrete spaces. *J. Am. Stat. Assoc.* **2020**, *115*, 852–865. [CrossRef]
32. Zhou, Q.; Yang, J.; Vats, D.; Roberts, G.O.; Rosenthal, J.S. Dimension-free mixing for high-dimensional Bayesian variable selection. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **2022**, *84*, 1751–1784. [CrossRef]
33. Liang, X.; Livingstone, S.; Griffin, J. Adaptive random neighbourhood informed Markov chain Monte Carlo for high-dimensional Bayesian variable selection. *Stat. Comput.* **2022**, *32*, 84. [CrossRef]
34. Hans, C.; Dobra, A.; West, M. Shotgun stochastic search for "large p" regression. *J. Am. Stat. Assoc.* **2007**, *102*, 507–516. [CrossRef]
35. Rue, H.; Martino, S.; Chopin, N. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *J. R. Stat. Soc. Ser. Stat. Methodol.* **2009**, *71*, 319–392. [CrossRef]
36. Griffin, J.E.; Brown, P.J. Bayesian global-local shrinkage methods for regularisation in the high dimension linear model. *Chemom. Intell. Lab. Syst.* **2021**, *210*, 104255. [CrossRef]
37. Martino, S.; Akerkar, R.; Rue, H. Approximate Bayesian inference for survival models. *Scand. J. Stat.* **2011**, *38*, 514–528. [CrossRef]
38. Blei, D.M.; Kucukelbir, A.; McAuliffe, J.D. Variational inference: A review for statisticians. *J. Am. Stat. Assoc.* **2017**, *112*, 859–877. [CrossRef]
39. Ray, K.; Szabo, B.; Clara, G. Spike and slab variational Bayes for high dimensional logistic regression. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 14423–14434.
40. Ray, K.; Szabó, B. Variational Bayes for high-dimensional linear regression with sparse priors. *J. Am. Stat. Assoc.* **2022**, *117*, 1270–1281. [CrossRef]
41. Komodromos, M.; Aboagye, E.O.; Evangelou, M.; Filippi, S.; Ray, K. Variational Bayes for high-dimensional proportional hazards models with applications within gene expression. *Bioinformatics* **2022**, *38*, 3918–3926. [CrossRef]
42. Bierkens, J.; Grazzi, S.; Meulen, F.v.d.; Schauer, M. Sticky PDMP samplers for sparse and local inference problems. *Stat. Comput.* **2023**, *33*, 8. [CrossRef]
43. Chevallier, A.; Fearnhead, P.; Sutton, M. Reversible jump PDMP samplers for variable selection. *J. Am. Stat. Assoc.* **2022**, 1–13. [CrossRef]
44. Carvalho, C.M.; Polson, N.G.; Scott, J.G. The horseshoe estimator for sparse signals. *Biometrika* **2010**, *97*, 465–480. [CrossRef]
45. Polson, N.G.; Scott, J.G. On the half-Cauchy prior for a global scale parameter. *Bayesian Anal.* **2012**, *7*, 887–902. [CrossRef]
46. Peltola, T.; Havulinna, A.S.; Salomaa, V.; Vehtari, A. Hierarchical Bayesian Survival Analysis and Projective Covariate Selection in Cardiovascular Event Risk Prediction. *BMA@UAI* **2014**, *27*, 79–88.
47. Liang, F.; Paulo, R.; Molina, G.; Clyde, M.A.; Berger, J.O. Mixtures of g priors for Bayesian variable selection. *J. Am. Stat. Assoc.* **2008**, *103*, 410–423. [CrossRef]
48. Li, Y.; Clyde, M.A. Mixtures of g-priors in generalized linear models. *J. Am. Stat. Assoc.* **2018**, *113*, 1828–1845. [CrossRef]
49. Ley, E.; Steel, M.F. On the effect of prior assumptions in Bayesian model averaging with applications to growth regression. *J. Appl. Econom.* **2009**, *24*, 651–674. [CrossRef]
50. Kohn, R.; Smith, M.; Chan, D. Nonparametric regression using linear combinations of basis functions. *Stat. Comput.* **2001**, *11*, 313–322. [CrossRef]
51. Nikooienejad, A.; Wang, W.; Johnson, V.E. Bayesian variable selection for survival data using inverse moment priors. *Ann. Appl. Stat.* **2020**, *14*, 809. [CrossRef] [PubMed]
52. Kalbfleisch, J.D. Non-parametric Bayesian analysis of survival time data. *J. R. Stat. Soc. Ser. B (Methodol.)* **1978**, *40*, 214–221. [CrossRef]
53. Sinha, D.; Ibrahim, J.G.; Chen, M.H. A Bayesian justification of Cox's partial likelihood. *Biometrika* **2003**, *90*, 629–641. [CrossRef]
54. Metropolis, N.; Rosenbluth, A.W.; Rosenbluth, M.N.; Teller, A.H.; Teller, E. Equation of state calculations by fast computing machines. *J. Chem. Phys.* **1953**, *21*, 1087–1092. [CrossRef]
55. Hastings, W.K. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **1970**, *57*, 97–109. [CrossRef]

56. Makalic, E.; Schmidt, D. High-Dimensional Bayesian Regularised Regression with the BayesReg Package. *arXiv* **2016**, arXiv:1611.06649v3.
57. Zens, G.; Frühwirth-Schnatter, S.; Wagner, H. Ultimate Pólya Gamma Samplers—Efficient MCMC for possibly imbalanced binary and categorical data. *arXiv* **2020**, arXiv:2011.06898.
58. Johndrow, J.E.; Smith, A.; Pillai, N.; Dunson, D.B. MCMC for imbalanced categorical data. *J. Am. Stat. Assoc.* **2018**, *114*, 1394–1403. [CrossRef]
59. Kass, R.E.; Tierney, L.; Kadane, J.B. The validity of posterior expansions based on Laplace's method. In *Bayesian and Likelihood Methods in Statistics and Econometrics*; Geissner, S., Hodges, J.S., Press, S.J., Zellner, A., Eds.; University of Minnesota: Minneapolis, MN, USA, 1990; pp. 473–487.
60. Barber, R.F.; Drton, M.; Tan, K.M. Laplace approximation in high-dimensional Bayesian regression. In Proceedings of the Statistical Analysis for High-Dimensional Data: The Abel Symposium 2014, Lofoten, Norway, 5–9 May 2014 ; Springer: Cham, Switzerland, 2016; pp. 15–36.
61. Beaumont, M.A. Estimation of population growth or decline in genetically monitored populations. *Genetics* **2003**, *164*, 1139–1160. [CrossRef]
62. Andrieu, C.; Roberts, G.O. The pseudo-marginal approach for efficient Monte Carlo computations. *Ann. Stat.* **2009**, *37*, 697–725. [CrossRef]
63. Gamerman, D. Sampling from the posterior distribution in generalized linear mixed models. *Stat. Comput.* **1997**, *7*, 57–68. [CrossRef]
64. Morris, D.L.; Sheng, Y.; Zhang, Y.; Wang, Y.F.; Zhu, Z.; Tombleson, P.; Chen, L.; Cunninghame Graham, D.S.; Bentham, J.; Roberts, A.L.; et al. Genome-wide association meta-analysis in Chinese and European individuals identifies ten new loci associated with systemic lupus erythematosus. *Nat. Genet.* **2016**, *48*, 940–946. [CrossRef]
65. Tadesse, M.G.; Vannucci, M. *Handbook of Bayesian Variable Selection*; CRC Press: Boca Raton, FL, USA, 2021.
66. Eddelbuettel, D.; François, R. Rcpp: Seamless R and C++ Integration. *J. Stat. Softw.* **2011**, *40*, 1–18. [CrossRef]
67. Lum, P.Y.; Singh, G.; Lehman, A.; Ishkanov, T.; Vejdemo-Johansson, M.; Alagappan, M.; Carlsson, J.; Carlsson, G. Extracting insights from the shape of complex data using topology. *Sci. Rep.* **2013**, *3*, 1236. [CrossRef] [PubMed]
68. Nicolau, M.; Levine, A.J.; Carlsson, G. Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival. *Proc. Natl. Acad. Sci. USA* **2011**, *108*, 7265–7270. [CrossRef] [PubMed]
69. Pereira, B.; Chin, S.F.; Rueda, O.M.; Vollan, H.K.M.; Provenzano, E.; Bardwell, H.A.; Pugh, M.; Jones, L.; Russell, R.; Sammut, S.J.; et al. The somatic mutation profiles of 2433 breast cancers refine their genomic and transcriptomic landscapes. *Nat. Commun.* **2016**, *7*, 11479. [CrossRef] [PubMed]
70. Mukherjee, A.; Russell, R.; Chin, S.F.; Liu, B.; Rueda, O.; Ali, H.; Turashvili, G.; Mahler-Araujo, B.; Ellis, I.; Aparicio, S.; et al. Associations between genomic stratification of breast cancer and centrally reviewed tumour pathology in the METABRIC cohort. *NPJ Breast Cancer* **2018**, *4*, 5. [CrossRef]
71. Cerami, E.; Gao, J.; Dogrusoz, U.; Gross, B.E.; Sumer, S.O.; Aksoy, B.A.; Jacobsen, A.; Byrne, C.J.; Heuer, M.L.; Larsson, E.; et al. The cBio cancer genomics portal: An open platform for exploring multidimensional cancer genomics data. *Cancer Discov.* **2012**, *2*, 401–404. [CrossRef] [PubMed]
72. Lang, M.; Kotthaus, H.; Marwedel, P.; Weihs, C.; Rahnenführer, J.; Bischl, B. Automatic model selection for high-dimensional survival analysis. *J. Stat. Comput. Simul.* **2015**, *85*, 62–76. [CrossRef]
73. Clough, E.; Barrett, T., The gene expression omnibus database. In *Statistical Genomics: Methods and Protocols*; Mathé, E., Davis, S., Eds.; Humana: New York, NY, USA, 2016; pp. 93–110.
74. Ng, V.K.; Cribbie, R.A. Using the gamma generalized linear model for modeling continuous, skewed and heteroscedastic outcomes in psychology. *Curr. Psychol.* **2017**, *36*, 225–235. [CrossRef]
75. Riva-Palacio, A.; Leisen, F.; Griffin, J. Survival regression models with dependent Bayesian nonparametric priors. *J. Am. Stat. Assoc.* **2022**, *117*, 1530–1539. [CrossRef]
76. Johndrow, J.E.; Pillai, N.S.; Smith, A. No free lunch for approximate MCMC. *arXiv* **2020**, arXiv:2010.12514.
77. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2013.
78. Cox, C.; Chu, H.; Schneider, M.F.; Munoz, A. Parametric survival analysis and taxonomy of hazard functions for the generalized gamma distribution. *Stat. Med.* **2007**, *26*, 4352–4374. [CrossRef]