Bayesian Analysis (2024) **19**, Number 1, pp. 261–291

A Metropolized Adaptive Subspace Algorithm for High-Dimensional Bayesian Variable Selection

Christian Staerk*, Maria Kateri†, and Ioannis Ntzoufras‡

Abstract. A simple and efficient adaptive Markov Chain Monte Carlo (MCMC) method, called the Metropolized Adaptive Subspace (MAdaSub) algorithm, is proposed for sampling from high-dimensional posterior model distributions in Bayesian variable selection. The MAdaSub algorithm is based on an independent Metropolis-Hastings sampler, where the individual proposal probabilities of the explanatory variables are updated after each iteration using a form of Bayesian adaptive learning, in a way that they finally converge to the respective covariates' posterior inclusion probabilities. We prove the ergodicity of the algorithm and present a parallel version of MAdaSub with an adaptation scheme for the proposal probabilities based on the combination of information from multiple chains. The effectiveness of the algorithm is demonstrated via various simulated and real data examples, including a high-dimensional problem with more than 20,000 covariates.

Keywords: adaptive MCMC, generalized linear models, high-dimensional data, sparsity, variable selection.

1 Introduction

Variable selection in regression models is one of the big challenges in the era of high-dimensional data where the number of explanatory variables might largely exceed the sample size. During the last two decades, many classical variable selection algorithms have been proposed which are often based on finding the solution to an appropriate optimization problem. As the most famous example, the Lasso (Tibshirani, 1996) relies on an ℓ_1-type relaxation of the original ℓ_0-type optimization problem. Convex methods like the Lasso are computationally very efficient and are therefore routinely used in high-dimensional statistical applications. However, such classical methods mainly focus on point estimation and do not provide a measure of uncertainty concerning the best model, per se, although recent works aim at addressing these issues as well (see e.g. Wasserman and Roeder, 2009, Meinshausen and Bühlmann, 2010 and Lee et al., 2016). On the other hand, a major advantage of a fully Bayesian approach is that it automatically accounts for model uncertainty. In particular, Bayesian model averaging (Raftery et al., 1997)

*Department of Medical Biometry, Informatics and Epidemiology, University Hospital Bonn, Venusberg Campus 1, 53127 Bonn, Germany, christian.staerk@uni-bonn.de

†Institute of Statistics, RWTH Aachen University, Pontdriesch 14-16, 52062 Aachen, Germany, maria.kateri@rwth-aachen.de

‡Department of Statistics, Athens University of Economics and Business, Trias 2, Athina 113 62, Greece, ntzoufras@aueb.gr

© 2024 International Society for Bayesian Analysis https://doi.org/10.1214/22-BA1351

and the median probability model (Barbieri and Berger, 2004) can be used for predictive inference. Furthermore, posterior inclusion probabilities of the individual covariates can be computed to quantify the Bayesian evidence.

Important ℓ_0-type criteria like the Bayesian Information Criterion (BIC, Schwarz, 1978) and the Extended Bayesian Information Criterion (EBIC, Chen and Chen, 2008) can be derived as asymptotic approximations to a fully Bayesian approach (compare e.g. Liang et al., 2013). It has been argued that ℓ_0-type methods posses favourable statistical properties in comparison to convex ℓ_1-type methods with respect to variable selection and prediction (see e.g. Raskutti et al., 2011 and Narisetty and He, 2014). Since solving the associated, generally NP-hard, discrete optimization problems by an exhaustive search is computationally prohibitive, there have been recent attempts in providing more efficient methods for resolving such issues, as for example, mixed integer optimization methods (Bertsimas et al., 2016) and Adaptive Subspace (AdaSub) methods (Staerk, 2018; Staerk et al., 2021).

The challenging practical issue of a fully Bayesian approach is similar to that of optimizing ℓ_0-type information criteria: computing (approximate) posterior model probabilities for all possible models is not feasible if the number of explanatory variables p is very large, since there are in general 2^p possible models which have to be considered. Often, Markov Chain Monte Carlo (MCMC) methods based on Metropolis-Hastings steps (e.g. Madigan et al., 1995), Gibbs samplers (e.g. George and McCulloch, 1993; Dellaportas et al., 2002) and "reversible jump" updates (e.g. Green, 1995) are used in order to obtain a representative sample from the posterior model distribution. However, the effectiveness of MCMC methods depends heavily on a sensible choice of the proposal distributions being used. Therefore, such methods may suffer from bad mixing resulting in a slow exploration of the model space, especially when the number of covariates is large. Moreover, tuning of the proposal distribution is often only feasible after manual "pilot" runs of the algorithm. Adaptive MCMC methods aim to address these issues by updating the proposal parameters "on the fly" during a single run of the algorithm so that the proposal distribution automatically adjusts according to the currently available information. Recently, a number of different adaptive MCMC algorithms have been proposed in the Bayesian variable selection context, see e.g. Nott and Kohn (2005), Lamnisos et al. (2013), Ji and Schmidler (2013), Griffin et al. (2014), Griffin et al. (2021) and Wan and Griffin (2021).

In this work we propose an alternative, simple and efficient adaptive independent Metropolis-Hastings algorithm for Bayesian variable selection, called the Metropolized Adaptive Subspace (MAdaSub) algorithm, and compare it to existing adaptive MCMC algorithms. In MAdaSub the individual proposal probabilities of the explanatory variables are sequentially adapted after each iteration. The employed updating scheme is inspired by the AdaSub method introduced in Staerk et al. (2021) and can itself be motivated in a Bayesian way, such that the individual proposal probabilities finally converge against the true respective posterior inclusion probabilities. In the limit, the algorithm can be viewed as a simple Metropolis-Hastings sampler using a product of independent Bernoulli proposals which is the closest to the unknown target distribution in terms of Kullback-Leibler divergence (among the distributions in the family of independent Bernoulli form). In comparison with alternative samplers based on "local" model

proposals, the adaptive independent proposals in MAdaSub encourage "global" moves towards regions near to the median probability model, which are particularly suited for high-dimensional settings with low signal-to-noise ratios and multimodal posterior distributions.

The paper is structured as follows. The considered setting of Bayesian variable selection in generalized linear models (GLMs) is briefly described in Section 2. The MAdaSub algorithm is motivated and introduced in Section 3. By making use of general results obtained by Roberts and Rosenthal (2007), it is shown that the MAdaSub algorithm is ergodic despite its continuing adaptation, i.e. that "in the limit" it samples from the targeted posterior model distribution (see Theorem 1). Alternative adaptive approaches are also briefly discussed and conceptually compared to the newly proposed algorithm. In Section 4, a parallel version of MAdaSub is presented where the proposal probabilities can be adapted using the information from all available chains, without affecting the ergodicity of the algorithm (see Theorem 3). Detailed proofs of the theoretical results of Sections 3 and 4 can be found in the Supplement to this paper (Staerk et al., 2022). The adaptive behaviour of MAdaSub and the choice of its tuning parameters are illustrated via low- and high-dimensional simulated data applications in Section 5, emphasizing that the speed of convergence against the targeted posterior depends on an appropriate choice of these parameters. In Section 6 various real data applications demonstrate that MAdaSub provides an efficient and stable way for sampling from high-dimensional posterior model distributions. The paper concludes with a discussion in Section 7. An R-implementation of MAdaSub is available at `https://github.com/chstaerk/MAdaSub`.

2 The setting

In this work we consider variable selection in univariate generalized linear models (GLMs), where the response variable Y is modelled in terms of p possible explanatory variables $X_1, \ldots, X_p$. More precisely, for a sample of size n, the components of the response vector $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$ are assumed to be independent with each of them having a distribution from a fixed exponential dispersion family with

$$g\big(E(Y_i \mid \boldsymbol{X}_{i,*})\big) = \mu + \sum_{j=1}^{p} \beta_j X_{i,j}, \quad i = 1, \ldots, n\,, \tag{1}$$

where g is a (fixed) link function, $\mu \in \mathbb{R}$ is the intercept and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T \in \mathbb{R}^p$ is the vector of regression coefficients. Here, $\boldsymbol{X} = (X_{i,j}) \in \mathbb{R}^{n \times p}$ is the design matrix; it's i-th row $\boldsymbol{X}_{i,*}$ corresponds to the i-th observation and it's j-th column $\boldsymbol{X}_{*,j} \equiv \boldsymbol{X}_j$ corresponds to the values of the j-th predictor. For a subset $S \subseteq \{1, \ldots, p\}$, the model induced by S is defined by a GLM of the form (1) but with design matrix $\boldsymbol{X}_S \in \mathbb{R}^{n \times |S|}$ in place of $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ and corresponding vector of coefficients $\boldsymbol{\beta}_S \in \mathbb{R}^{|S|}$, where $\boldsymbol{X}_S$ denotes the submatrix of the original design matrix $\boldsymbol{X}$ containing only the columns with indices in S. For brevity, we often simply refer to the model S. Without further notice, we assume that we always include an intercept μ in the corresponding GLM with design

matrix $\boldsymbol{X}_S$. We denote the set of labelled explanatory variables by $\mathcal{P} = \{1, \ldots, p\}$ and the full model space by $\mathcal{M} = \{S;\, S \subseteq \mathcal{P}\}$.

In a fully Bayesian approach we assign prior probabilities $\pi(S)$ to each of the considered models $S \in \mathcal{M}$ as well as priors $\pi(\mu, \psi, \boldsymbol{\beta}_S \,|\, S)$ for the parameters of each model $S \in \mathcal{M}$, where ψ denotes a possibly present dispersion parameter (e.g. the variance in a normal linear model). After observing data $\mathcal{D} = (\boldsymbol{X}, \boldsymbol{y})$, with $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ and $\boldsymbol{y} \in \mathbb{R}^n$, the posterior model probabilities are proportional to

$$\pi(S \,|\, \mathcal{D}) \propto \pi(\boldsymbol{y} \,|\, \boldsymbol{X}, S)\, \pi(S)\,, \;\; S \in \mathcal{M}\,, \tag{2}$$

where

$$\pi(\boldsymbol{y} \,|\, \boldsymbol{X}, S) = \int\!\!\int\!\!\int f(\boldsymbol{y} \,|\, \boldsymbol{X}, S, \mu, \psi, \boldsymbol{\beta}_S)\, \pi(\mu, \psi, \boldsymbol{\beta}_S \,|\, S)\, d\mu\, d\psi\, d\boldsymbol{\beta}_S \tag{3}$$

is the marginal likelihood of the data $\boldsymbol{y}$ under model S, while $f(\boldsymbol{y} \,|\, \boldsymbol{X}, S, \mu, \psi, \boldsymbol{\beta}_S)$ denotes the likelihood of the data $\boldsymbol{y}$ under model S given the parameter values $\mu, \psi, \boldsymbol{\beta}_S$ and the values of the explanatory variables $\boldsymbol{X}$. Note that the marginal likelihood $\pi(\boldsymbol{y} \,|\, \boldsymbol{X}, S)$ is generally only available in closed form when conjugate priors are used.

Remark 2.1. A prominent example in normal linear models is a conjugate prior structure, where the prior on the variance $\psi = \sigma^2$ is given by Jeffreys prior (independent of the model S) and the prior on the vector of coefficients $\boldsymbol{\beta}_S$ in model $S \in \mathcal{M}$ is given by a multivariate normal distribution, i.e.

$$\boldsymbol{\beta}_S \,|\, S, \sigma^2 \sim \mathcal{N}_{|S|}(\boldsymbol{\vartheta}_S, \sigma^2 g\, \boldsymbol{W}_S), \quad \pi(\sigma^2) \propto \frac{1}{\sigma^2}\,, \tag{4}$$

where $\boldsymbol{\vartheta}_S \in \mathbb{R}^{|S|}$, $g > 0$ and $\boldsymbol{W}_S \in \mathbb{R}^{|S| \times |S|}$ are hyperparameters. After centering each of the covariates $\boldsymbol{X}_j$, $j \in \mathcal{P}$, the improper prior $\pi(\mu) \propto 1$ is a common choice for the intercept μ (again, independent of the model S). With no specific prior information, the prior mean of $\boldsymbol{\beta}_S$ can be set to the zero vector ($\boldsymbol{\vartheta}_S = \boldsymbol{0}$). The matrix $\boldsymbol{W}_S$ is often chosen to be the identity matrix $\boldsymbol{I}_{|S|}$ of dimension $|S|$ or to be $\boldsymbol{W}_S = (\boldsymbol{X}_S^T \boldsymbol{X}_S)^{-1}$ yielding Zellner's g-prior (Zellner, 1986). The first choice corresponds to Ridge Regression and implies prior independence of the regression coefficients, while the second choice with $g = n$ corresponds to a unit information prior. In case no specific prior information is available about the possible regressors, a natural choice for the model prior is an independent Bernoulli prior of the form

$$\pi(S \,|\, \omega) = \omega^{|S|}(1-\omega)^{p-|S|},\; S \in \mathcal{M}\,, \tag{5}$$

where $\omega = \pi(j \in S)$ is the prior probability that variable X_j is included in the model, for all $j \in \mathcal{P}$. One can either set the prior inclusion probability ω to some fixed value or consider an additional hyperprior for ω, with the latter option yielding more flexibility. A convenient choice is the (conjugate) beta prior $\omega \sim \mathcal{B}e(a_\omega, b_\omega)$, where $a_\omega > 0$ and $b_\omega > 0$ can be chosen in order to reflect the prior expectation and prior variance of the model size $s = |S|$, $S \in \mathcal{M}$ (see Kohn et al., 2001 for details). In practice, one often imposes an a-priori upper bound $s_{\max}$ on the model size (with $s_{\max} \leq n$) by setting $\pi(S) = 0$ for $|S| > s_{\max}$ (cf. Liang et al., 2013; Rossell, 2021), while for fixed control variables X_j one can enforce the inclusion of such variables by setting $\pi(j \in S) = 1$.

In the general non-conjugate case the marginal likelihood is not readily computable and numerical methods may be used for deriving an approximation to the marginal likelihood. Laplace's method yields an asymptotic analytic approximation to the marginal likelihood (Kass and Raftery, 1995). Similarly, different information criteria like the Bayesian Information Criterion (BIC, Schwarz, 1978) or the Extended Bayesian Information Criterion (EBIC, Chen and Chen, 2008) can be used directly as asymptotic approximations to fully Bayesian posterior model probabilities under suitable choices of model priors. Under a uniform model prior, i.e. $\pi(S) = \frac{1}{2^p}$ for all $S \in \mathcal{M}$, the BIC can be derived as an approximation to $-2\log(\mathrm{BF}(S)) = -2\log(\mathrm{PO}(S))$, where $\mathrm{BF}(S) = \pi(\boldsymbol{y} \,|\, \boldsymbol{X}, S)/\pi(\boldsymbol{y} \,|\, \boldsymbol{X}, \emptyset)$ denotes the Bayes factor of model $S \in \mathcal{M}$ versus the null model $\emptyset \in \mathcal{M}$ and $\mathrm{PO}(S)$ denotes the corresponding posterior odds (Schwarz, 1978; Kass and Wasserman, 1995). In a high-dimensional but sparse situation, in which only a few of the many possible predictors contribute substantially to the response, a uniform prior on the model space is a naive choice since it induces severe overfitting. Therefore, Chen and Chen (2008) propose the prior

$$\pi(S) \propto \binom{p}{|S|}^{-\gamma}, \tag{6}$$

where $\gamma \in [0,1]$ is an additional parameter. If $\gamma = 1$, then $\pi(S) = \frac{1}{p+1}\binom{p}{|S|}^{-1}$, so the prior gives equal probability to each model size, and to each model of the same size; note that this prior does also coincide with the beta-binomial model prior discussed above when setting $a_\omega = b_\omega = 1$, providing automatic multiplicity correction (Scott and Berger, 2010). If $\gamma = 0$, then we obtain the uniform prior used in the original BIC. Similar to the derivation of the BIC one asymptotically obtains the EBIC with parameter $\gamma \in [0,1]$ as

$$\mathrm{EBIC}_\gamma(S) = -2\log\Big(f(\boldsymbol{y} \,|\, \boldsymbol{X}, S, \hat{\mu}_S, \hat{\psi}_S, \hat{\boldsymbol{\beta}}_S)\Big) + \Big(\log(n) + 2\gamma\log(p)\Big)|S|\,, \tag{7}$$

where $f(\boldsymbol{y} \,|\, \boldsymbol{X}, S, \hat{\mu}_S, \hat{\psi}_S, \hat{\boldsymbol{\beta}}_S)$ denotes the maximized likelihood under the model $S \in \mathcal{M}$ (compare Chen and Chen, 2012). Under the model prior (6) and a unit-information prior on the regression coefficients for each model $S \in \mathcal{M}$, one can asymptotically approximate the model posterior by

$$\pi(S \,|\, \mathcal{D}) \approx \frac{\exp\left(-\frac{1}{2} \times \mathrm{EBIC}_\gamma(S)\right)}{\sum_{S' \in \mathcal{M}} \exp\left(-\frac{1}{2} \times \mathrm{EBIC}_\gamma(S')\right)}\,,\ S \in \mathcal{M}\,. \tag{8}$$

In this work we consider situations where the marginal likelihood $\pi(\boldsymbol{y} \,|\, \boldsymbol{X}, S)$ is available in closed form due to the use of conjugate priors (see Remark 2.1) or where an approximation to the posterior $\pi(S \,|\, \mathcal{D})$ is used (e.g. via equation (8) with the EBIC or any other ℓ_0-type criteria such as the risk inflation criterion, cf. Foster and George, 1994; Rossell, 2021). This assumption allows one to focus on the essential part of efficient sampling in very large model spaces, avoiding challenging technicalities regarding sampling of model parameters for non-conjugate cases. It also facilitates empirical comparisons with other recent adaptive variable selection methods, which focus on conjugate priors (Zanella and Roberts, 2019; Griffin et al., 2021). Furthermore, conjugate priors

such as the g-prior as well as normalized ℓ_0-type selection criteria such as the EBIC in equation (8) have been shown to provide concentration of posterior model probabilities on the (Kullback-Leibler) optimal model under general conditions even in case of model misspecification (Rossell, 2021), as well as model selection consistency for the true model in GLMs without misspecification (Chen and Chen, 2012; Liang et al., 2013).

3 The MAdaSub algorithm

A simple way to sample from a given target distribution is to use an independent Metropolis-Hastings algorithm. Clearly, the efficiency of such an MCMC algorithm depends on the choice of the proposal distribution, which is in general not an easy task (see e.g. Rosenthal, 2011). In the ideal situation, the proposal distribution for an independence sampler should be the same as the target distribution $\pi(S \,|\, \mathcal{D})$, leading to an independent sample from the target distribution with corresponding acceptance probability of one. Adaptive MCMC algorithms aim to sequentially update the proposal distribution during the algorithm based on the previous samples such that, in case of the independence sampler, the proposal becomes closer and closer to the target distribution as the MCMC sample grows (see e.g. Holden et al., 2009, Giordani and Kohn, 2010). However, especially in high-dimensional situations, it is crucial that the adaptation of the proposal as well as sampling from the proposal can be carried out efficiently. For this reason, we restrict ourselves to proposal distributions which have an independent Bernoulli form: if $S \in \mathcal{M}$ is the current model, then we propose model $V \in \mathcal{M}$ with probability

$$q(V \,|\, S; \boldsymbol{r}) \equiv q(V; \boldsymbol{r}) = \prod_{j \in V} r_j \prod_{j \in \mathcal{P} \setminus V} (1 - r_j)\,, \tag{9}$$

for some vector $\boldsymbol{r} = (r_1, \ldots, r_p) \in (0,1)^p$ of individual proposal probabilities.

3.1 Serial version of the MAdaSub algorithm

The fundamental idea of the newly proposed MAdaSub algorithm (given below as Algorithm 1) is to sequentially update the individual proposal probabilities according to the currently "estimated" posterior inclusion probabilities. In more detail, after initializing the vector of proposal probabilities $\boldsymbol{r}^{(0)} = \left(r_1^{(0)}, \ldots, r_p^{(0)}\right) \in (0,1)^p$, the individual proposal probabilities $r_j^{(t)}$ of variables X_j are updated after each iteration t of the algorithm, such that $r_j^{(t)}$ finally converges to the actual posterior inclusion probability $\pi_j = \pi(j \in S \,|\, \mathcal{D})$, as $t \to \infty$ (see Corollary 2 below). Therefore, in the limit, we make use of the proposal

$$q(V; \boldsymbol{r}^*) = \prod_{j \in V} \pi_j \prod_{j \in \mathcal{P} \setminus V} (1 - \pi_j), \quad V \in \mathcal{M}\,, \quad \text{with } \boldsymbol{r}^* = (\pi_1, \ldots, \pi_p)\,, \tag{10}$$

which is the closest distribution (in terms of Kullback-Leibler divergence) to the actual target $\pi(S \,|\, \mathcal{D})$, among all distributions of independent Bernoulli form (9) (see Clyde

Algorithm 1 Metropolized Adaptive Subspace (MAdaSub) algorithm.

Input:

- Data $\mathcal{D} = (\boldsymbol{X}, \boldsymbol{y})$.
- (Approximate) kernel of posterior $\pi(S \,|\, \mathcal{D}) \propto \pi(\boldsymbol{y} \,|\, \boldsymbol{X}, S)\, \pi(S)$ for $S \in \mathcal{M}$.
- Vector of initial proposal probabilities $\boldsymbol{r}^{(0)} = \left(r_1^{(0)}, \ldots, r_p^{(0)}\right)^T \in (0,1)^p$.
- Parameters $L_j > 0$ for $j \in \mathcal{P}$, controlling the adaptation rate of the algorithm (e.g. $L_j = L = p$).
- Constant $\epsilon \in (0, 0.5)$ (chosen to be small, e.g. $\epsilon \leq \frac{1}{p}$).
- Number of iterations $T \in \mathbb{N}$.
- Starting point $S^{(0)} \in \mathcal{M}$ (optional).

Algorithm:

(1) If starting point $S^{(0)}$ not specified:
Sample $b_j^{(0)} \sim \text{Bernoulli}\left(r_j^{(0)}\right)$ independently for $j \in \mathcal{P}$.
Set $S^{(0)} = \{j \in \mathcal{P};\ b_j^{(0)} = 1\}$.

(2) For $t = 1, \ldots, T$:

(a) Truncate vector of proposal probabilities to $\tilde{\boldsymbol{r}}^{(t-1)} = \left(\tilde{r}_1^{(t-1)}, \ldots, \tilde{r}_p^{(t-1)}\right)^T$, i.e. for $j \in \mathcal{P}$ set

$$\tilde{r}_j^{(t-1)} = \begin{cases} r_j^{(t-1)} & \text{, if } r_j^{(t-1)} \in [\epsilon, 1-\epsilon]\,, \\ \epsilon & \text{, if } r_j^{(t-1)} < \epsilon\,, \\ 1-\epsilon & \text{, if } r_j^{(t-1)} > 1-\epsilon\,. \end{cases}$$

(b) Draw $b_j^{(t)} \sim \text{Bernoulli}\left(\tilde{r}_j^{(t-1)}\right)$ independently for $j \in \mathcal{P}$.

(c) Set $V^{(t)} = \{j \in \mathcal{P};\ b_j^{(t)} = 1\}$.

(d) Compute acceptance probability

$$\alpha^{(t)} = \min\left\{ \frac{\pi(\boldsymbol{y} \,|\, \boldsymbol{X}, V^{(t)})\, \pi(V^{(t)})\, q(S^{(t-1)}; \tilde{\boldsymbol{r}}^{(t-1)})}{\pi(\boldsymbol{y} \,|\, \boldsymbol{X}, S^{(t-1)})\, \pi(S^{(t-1)})\, q(V^{(t)}; \tilde{\boldsymbol{r}}^{(t-1)})}, 1 \right\}.$$

(e) Set $S^{(t)} = \begin{cases} V^{(t)} & \text{, with probability } \alpha^{(t)}, \\ S^{(t-1)} & \text{, with probability } 1-\alpha^{(t)}. \end{cases}$

(f) Update vector of proposal probabilities $\boldsymbol{r}^{(t)} = \left(r_1^{(t)}, \ldots, r_p^{(t)}\right)^T$ via

$$r_j^{(t)} = \frac{L_j r_j^{(0)} + \sum_{i=1}^{t} \mathbb{1}_{S^{(i)}}(j)}{L_j + t}\,, \quad j \in \mathcal{P}\,.$$

Output:

- Approximate sample $S^{(b+1)}, \ldots, S^{(T)}$ from posterior distribution $\pi(\cdot \,|\, \mathcal{D})$, after burn-in period of length b.

et al., 2011). Note that the median probability model (Barbieri and Berger, 2004; Barbieri et al., 2021), defined by $S_{\mathrm{MPM}} = \{j \in \mathcal{P};\, \pi_j \geq 0.5\}$, has the largest probability in the limiting proposal (10) of MAdaSub, i.e. $\arg\max_{V \in \mathcal{M}} q(V; \boldsymbol{r}^*) = S_{\mathrm{MPM}}$. Thus, MAdaSub can be interpreted as an adaptive algorithm which aims to adjust the proposal so that models in the region of the median probability model are proposed with increasing probability.

For $j \in \mathcal{P}$, the concrete update of $r_j^{(t)}$ after iteration $t \in \mathbb{N}$ is given by

$$r_j^{(t)} \;=\; \frac{L_j r_j^{(0)} + \sum_{i=1}^{t} \mathbb{1}_{S^{(i)}}(j)}{L_j + t} \;=\; \left(1 - \frac{1}{L_j + t}\right) r_j^{(t-1)} + \frac{\mathbb{1}_{S^{(t)}}(j)}{L_j + t}\,, \tag{11}$$

where, for $j \in \mathcal{P}$, $L_j > 0$ are additional parameters controlling the adaptation rate of the algorithm and $\mathbb{1}_{S^{(i)}}$ denotes the indicator function of the set $S^{(i)}$. If $j \in S^{(t)}$ (i.e. $\mathbb{1}_{S^{(t)}}(j) = 1$), then variable X_j is included in the sampled model in iteration t of the algorithm and the proposal probability $r_j^{(t)}$ of X_j increases in the next iteration $t+1$; similarly, if $j \notin S^{(t)}$ (i.e. $\mathbb{1}_{S^{(t)}}(j) = 0$), then the proposal probability decreases. The additional "truncation" step 2 (a) in the MAdaSub algorithm ensures that the truncated individual proposal probabilities $\tilde{r}_j^{(t)}$, $j \in \mathcal{P}$, are always included in the compact interval $\mathcal{I} = [\epsilon, 1-\epsilon]$, where $\epsilon \in (0, 0.5)$ is a pre-specified "precision" parameter. This adjustment simplifies the proof of the ergodicity of MAdaSub. Note that the mean size of the proposed model V from the proposal $q(V; \tilde{\boldsymbol{r}})$ in equation (9) with $\tilde{\boldsymbol{r}} \in [\epsilon, 1-\epsilon]^p$ is at least $E|V| \geq \epsilon \times p$; thus, in practice we recommended to set $\epsilon \leq \frac{1}{p}$, so that models of small size including the null model can be proposed with sufficiently large probability. On the other hand, if ϵ is chosen to be very small, then the MAdaSub algorithm may take a longer time to convergence in case proposal probabilities of informative variables are close to $\epsilon \approx 0$ during the initial burn-in period of the algorithm. Simulations and real data applications show that the choice $\epsilon = \frac{1}{p}$ works well in all considered situations (see Sections 5 and 6).

The updating scheme of the individual proposal probabilities is inspired by the AdaSub method proposed in Staerk (2018) and Staerk et al. (2021) and can itself be motivated in a Bayesian way: since we do not know the true posterior inclusion probability π_j of variable X_j for $j \in \mathcal{P}$, we place a beta prior on π_j with the following parametrization

$$\pi_j \sim \mathcal{B}e\left(L_j r_j^{(0)},\, L_j\left(1 - r_j^{(0)}\right)\right)\,, \tag{12}$$

where $r_j^{(0)} = E[\pi_j]$ is the prior expectation of π_j and $L_j > 0$ controls the variance of π_j via

$$\mathrm{Var}(\pi_j) = \frac{1}{L_j + 1} \times r_j^{(0)}\left(1 - r_j^{(0)}\right)\,. \tag{13}$$

If $L_j \to 0$, then $\mathrm{Var}(\pi_j) \to r_j^{(0)}\,(1 - r_j^{(0)})$, which is the variance of a Bernoulli random variable with mean $r_j^{(0)}$. If $L_j \to \infty$, then $\mathrm{Var}(\pi_j) \to 0$. Now, one might view the samples $S^{(1)}, \ldots, S^{(t)}$ obtained after t iterations of MAdaSub as "new" data and interpret the information learned about π_j as t approximately independent Bernoulli trials, where

$j \in S^{(i)}$ corresponds to "success" and $j \notin S^{(i)}$ corresponds to "failure". Then the (pseudo) posterior of π_j after iteration t of the algorithm is given by

$$\pi_j \,|\, S^{(1)}, \ldots, S^{(t)} \sim \mathcal{B}e\left(L_j r_j^{(0)} + \sum_{i=1}^{t} \mathbb{1}_{S^{(i)}}(j),\, L_j(1 - r_j^{(0)}) + \sum_{i=1}^{t} \mathbb{1}_{\mathcal{P}\setminus S^{(i)}}(j)\right), \quad (14)$$

with posterior expectation

$$E(\pi_j \,|\, S^{(1)}, \ldots, S^{(t)}) = \frac{L_j r_j^{(0)} + \sum_{i=1}^{t} \mathbb{1}_{S^{(i)}}(j)}{L_j + t} = r_j^{(t)} \quad (15)$$

and posterior variance

$$\mathrm{Var}(\pi_j \,|\, S^{(1)}, \ldots, S^{(t)}) = \frac{1}{L_j + t + 1} \times r_j^{(t)} \left(1 - r_j^{(t)}\right). \quad (16)$$

The interpretation of $r_j^{(0)}$ as the prior expectation for the posterior inclusion probability π_j motivates the choice of $r_j^{(0)} = \pi(j \in S)$ as the actual prior inclusion probability of variable X_j. If no particular prior information about specific variables is available, but the prior expected model size is equal to $q \in (0, p)$, then we recommend to set $r_j^{(0)} = \frac{q}{p}$ and $L = L_j = p$ for all $j \in \mathcal{P}$, corresponding to the prior $\pi_j \sim \mathcal{B}e(q, p - q)$ in equation (12). In this particular situation, equation (15) reduces to

$$E(\pi_j \,|\, S^{(1)}, \ldots, S^{(t)}) = \frac{q + \sum_{i=1}^{t} \mathbb{1}_{S^{(i)}}(j)}{p + t} = r_j^{(t)}. \quad (17)$$

Even though it seems natural to choose the parameters $r_j^{(0)}$ and L_j of MAdaSub as the respective prior quantities, this choice is not imperative. While the optimal choices of these parameters generally depend on the setting, various simulated and real data applications of MAdaSub indicate that choosing $r_j^{(0)} = \frac{q}{p}$ with $q \in [2, 10]$ and $L_j \in [p/2, 2p]$ for $j \in \mathcal{P}$ yields a stable algorithm with good mixing in sparse high-dimensional set-ups irrespective of the actual prior (see Sections 5 and 6). Furthermore, if one has already run and stopped the MAdaSub algorithm after a certain number of iterations T, then one can simply restart the algorithm with the already updated parameters $r_j^{(T)}$ and $L_j + T$ (compare equation (16)) as new starting values for the corresponding parameters.

Using general results for adaptive MCMC algorithms by Roberts and Rosenthal (2007), we show that MAdaSub is ergodic despite its continuing adaptation.

Theorem 1. *The MAdaSub algorithm (Algorithm 1) is ergodic for all choices of* $\boldsymbol{r}^{(0)} \in (0, 1)^p$, $L_j > 0$ *and* $\epsilon \in (0, 0.5)$ *and fulfils the weak law of large numbers.*

The proof of Theorem 1 can be found in Section A of the Supplement, where it is shown that MAdaSub satisfies both the simultaneous uniform ergodicity condition and the diminishing adaptation condition (cf. Roberts and Rosenthal, 2007). As an immediate consequence of Theorem 1 we obtain the following important result.

Corollary 2. *For all choices of* $\boldsymbol{r}^{(0)} \in (0,1)^p$*,* $L_j > 0$ *and* $\epsilon \in (0, 0.5)$*, the proposal probabilities* $r_j^{(t)}$ *of the explanatory variables* X_j *in MAdaSub converge (in probability) to the respective posterior inclusion probabilities* $\pi_j = \pi(j \in S \,|\, \mathcal{D})$*, i.e. for all* $j \in \mathcal{P}$ *it holds that* $r_j^{(t)} \xrightarrow{P} \pi_j$ *as* $t \to \infty$*.*

3.2 Comparison to related adaptive approaches

In this section we conceptually compare the proposed MAdaSub algorithm (Algorithm 1) with other approaches for high-dimensional Bayesian variable selection, focusing on adaptive MCMC algorithms most closely related to the new algorithm (see Section D of the Supplement for details on further related methods).

In a pioneering work, Nott and Kohn (2005) propose an adaptive sampling algorithm for Bayesian variable selection based on a Metropolized Gibbs sampler, showing empirically that the adaptive algorithm outperforms different non-adaptive algorithms in terms of efficiency per iteration. However, since their approach requires the computation of inverses of estimated covariance matrices, it does not scale well to very high-dimensional settings. Recently, several variants and extensions of the original adaptive MCMC sampler of Nott and Kohn (2005) have been developed, including an adaptive Metropolis-Hastings algorithm by Lamnisos et al. (2013), where the expected number of variables to be changed by the proposal is adapted during the algorithm. Zanella and Roberts (2019) propose a tempered Gibbs sampling algorithm with adaptive choices of components to be updated in each iteration. Furthermore, different individual adaptation algorithms have been developed in Griffin et al. (2014) as well as in the follow-up works of Griffin et al. (2021) and Wan and Griffin (2021), which are closely related to the proposed MAdaSub algorithm. These strategies are based on adaptive Metropolis-Hastings algorithms, where the employed proposal distributions are of the following form: if $S \in \mathcal{M}$ is the current model, then the probability of proposing the model $V \in \mathcal{M}$ is given by

$$\tilde{q}(V \,|\, S; \boldsymbol{\eta}) = \prod_{j \in V \setminus S} A_j \prod_{j \in S \setminus V} D_j \prod_{j \in \mathcal{P} \setminus (S \cup V)} (1 - A_j) \prod_{j \in S \cap V} (1 - D_j)\,, \tag{18}$$

where $\boldsymbol{\eta} = (A_1, \ldots, A_p, D_1, \ldots, D_p)^T \in (0,1)^{2p}$ is a vector of tuning parameters with the following interpretation: For $j \in \mathcal{P}$, A_j is the probability of adding variable X_j if it is not included in the current model S and D_j is the probability of deleting variable X_j if it is included in the current model S. An important difference is that the adaptation strategies in Griffin et al. (2021) specifically aim to guard against low acceptance rates of the proposal (18), while MAdaSub aims at obtaining a global independent proposal with the largest possible acceptance rate, focusing on regions close to the median probability model. Furthermore, the adaptation of the individual proposal probabilities in MAdaSub can be motivated in a Bayesian way, leading to a natural parallel implementation of the algorithm with an efficient joint updating scheme for the shared adaptive parameters (see Section 4). Finally, in contrast to MAdaSub, Griffin et al. (2021) make use of Rao-Blackwellized estimates of posterior inclusion probabilities to speed up convergence.

Schäfer and Chopin (2013) develop sequential Monte Carlo algorithms (cf. South et al., 2019) using model proposals which directly account for the non-independent posterior inclusion of covariates. In contrast, MAdaSub is an adaptive MCMC algorithm which is based on independent Bernoulli proposals. While similar extensions of MAdaSub might be desirable to better approximate the posterior distribution, this may come at the price of a larger computational cost for updating and sampling from the proposal. The simple independent Bernoulli proposals in MAdaSub can also be viewed as mean-field variational approximations to the full posterior model distribution. Despite its connection with variational Bayes approaches (e.g. Carbonetto and Stephens, 2012; Ormerod et al., 2017), MAdaSub samples from the full posterior distribution and the accuracy of the approximation only affects the efficiency of the sampler, as final acceptance rates are expected to be smaller for larger distances between the posterior and the closest independent Bernoulli proposal (cf. Neklyudov et al., 2019). Empirical results for MAdaSub (see Sections 5 and 6) indicate that even the simple independent Bernoulli proposals yield good mixing and sufficiently large acceptance rates in various settings.

Finally, MAdaSub is an extension of the Adaptive Subspace (AdaSub) method (Staerk et al., 2021), a stochastic search algorithm aiming to identify the best model according to a particular selection criterion (such as the EBIC) by adaptively solving low-dimensional sub-problems of the original problem. While the purpose of AdaSub is to obtain the solution to an optimization problem, its Metropolized version MAdaSub constitutes an adaptive MCMC algorithm which samples from the full posterior model distribution. Despite this difference, the adaptation scheme of AdaSub for the covariates' inclusion probabilities in the sub-problems can be similarly motivated in a Bayesian way (cf. Staerk, 2018). The adaptation in AdaSub and MAdaSub is also related to Thompson sampling for multi-armed bandits in reinforcement learning, which has recently been investigated in the context of non-parametric Bayesian variable selection (Liu and Ročková, 2021). In contrast to MAdaSub, Thompson Variable Selection (TVS) does not provide samples from the posterior distribution but is designed to minimize the regret (i.e. the difference between optimal and actual rewards); as a consequence, the sampling probabilities in TVS are not guaranteed to converge to the posterior inclusion probabilities.

4 Parallelization of the MAdaSub algorithm

In this section we present a parallel version of the MAdaSub algorithm which aims at increasing the computational efficiency and accelerating the convergence of the chains. The simplest approach to parallelization would be to independently run the MAdaSub algorithm in parallel on each of $K \in \mathbb{N}$ different workers, yielding K individual chains which, in the limit, sample from the posterior model distribution (see Theorem 1). However, it is desirable that the information learned about the adaptive parameters can be shared efficiently between the different chains, so that the convergence of the adaptive parameters to their optimal values can be accelerated, leading to a faster convergence of the chains to their common limiting distribution.

We propose a parallel version of MAdaSub, where the workers sample individual MAdaSub chains in parallel, but the acquired information is exchanged periodically be-

tween the chains and the adaptive proposal probabilities are updated together (see Algorithm 2 in Section B of the Supplement for full algorithmic details). More specifically, let $S^{(k,1)}, \ldots, S^{(k,T)}$ denote the models sampled by MAdaSub (see Algorithm 1) for the first T iterations on worker k, for $k \in \{1, \ldots, K\}$. Then, for each worker $k \in \{1, \ldots, K\}$, we define the jointly updated proposal probabilities after the first round ($m = 1$) of T iterations by

$$\bar{r}_j^{(k,1)} = \frac{L_j^{(k)} r_j^{(k,0)} + \sum_{t=1}^{T} \sum_{l=1}^{K} \mathbb{1}_{S^{(l,t)}}(j)}{L_j^{(k)} + TK}, \quad j \in \mathcal{P}\,, \tag{19}$$

where $r_j^{(k,0)}$ denotes the initial proposal probability for variable X_j and $L_j^{(k)}$ the corresponding adaptation parameter (both can be different across the chains).

After the joint update, each MAdaSub chain is resumed (with $\bar{r}_j^{(k,1)}$ as initial proposal probabilities and $L_j^{(k)} + TK$ as initial prior variance parameters for $j \in \mathcal{P}$) and is run independently on each of the workers for T additional iterations in a second round ($m = 2$); then the proposal probabilities are updated jointly again to $\bar{r}_j^{(k,2)}$, and so on (up to $m = R$ rounds in Algorithm 2 of the Supplement). The joint updates of the proposal probabilities after $m \in \mathbb{N}$ rounds of T iterations are given by

$$\bar{r}_j^{(k,m)} = \frac{L_j^{(k)} r_j^{(k,0)} + \sum_{t=1}^{mT} \sum_{l=1}^{K} \mathbb{1}_{S^{(l,t)}}(j)}{L_j^{(k)} + mTK}, \quad k \in \{1, \ldots, K\},\ j \in \mathcal{P}. \tag{20}$$

Similarly to the serial version of MAdaSub, the adaptive learning of its parallel version can be naturally motivated in a Bayesian way: each worker $k = 1, \ldots, K$ can be thought of as an individual subject continuously updating its prior belief about the true posterior inclusion probability π_j of variable X_j through new information from its individual chain; additionally, after a period of T iterations the subject updates its prior belief also by obtaining new information from the $K - 1$ other subjects. If the (possibly different) priors of subjects $k = 1, \ldots, K$ on π_j are

$$\pi_j \sim \mathcal{B}e\left(L_j^{(k)} r_j^{(k,0)},\, L_j^{(k)}\left(1 - r_j^{(k,0)}\right)\right),\, j \in \mathcal{P}\,, \tag{21}$$

where $r_j^{(k,0)} = E[\pi_j]$ is the prior expectation of subject k about π_j and $L_j^{(k)} > 0$ controls its prior variance, then the (pseudo) posterior of subject k about π_j after m rounds of T iterations of the parallel MAdaSub algorithm is given by (compare to equation (14))

$$\begin{aligned}\pi_j \,|\, S^{(1,1)}, \ldots, S^{(k,mT)} \sim \mathcal{B}e\Bigg(& L_j^{(k)} r_j^{(k,0)} + \sum_{i=1}^{mT} \sum_{l=1}^{K} \mathbb{1}_{S^{(l,i)}}(j), \\ & L_j^{(k)}(1 - r_j^{(k,0)}) + \sum_{i=1}^{mT} \sum_{l=1}^{K} \mathbb{1}_{\mathcal{P} \setminus S^{(l,i)}}(j)\Bigg)\end{aligned} \tag{22}$$

with posterior expectation (compare to equation (15))

$$E(\pi_j \,|\, S^{(1,1)}, \ldots, S^{(k,mT)}) = \bar{r}_j^{(k,m)}\,, \tag{23}$$

corresponding to the joint update in equation (20).

Although the individual chains in the parallel MAdaSub algorithm make use of the information from all the other chains in order to update the proposal parameters, the ergodicity of the chains is not affected.

Theorem 3. *Consider the parallel version of MAdaSub (see Algorithm 2 in the Supplement). Then, for each worker* $k \in \{1, \ldots, K\}$ *and all choices of* $\boldsymbol{r}^{(k,0)} \in (0,1)^p$, $L_j^{(k)} > 0$, $j \in \mathcal{P}$ *and* $\epsilon \in (0, 0.5)$, *each induced chain* $S^{(k,0)}, S^{(k,1)}, \ldots$ *of the workers* $k = 1, \ldots, K$ *is ergodic and fulfils the weak law of large numbers.*

Corollary 4. *For each worker* $k \in \{1, \ldots, K\}$ *and all choices of* $\boldsymbol{r}^{(k,0)} \in (0,1)^p$, $L_j^{(k)} > 0$, $j \in \mathcal{P}$ *and* $\epsilon \in (0, 0.5)$, *the proposal probabilities* $\bar{r}_j^{(k,m)}$ *of the explanatory variables* X_j *converge (in probability) to the respective posterior inclusion probabilities* $\pi_j = \pi(j \in S \,|\, \mathcal{D})$, *i.e. for all* $j \in \mathcal{P}$ *and* $k = 1, \ldots, K$ *it holds that* $\bar{r}_j^{(k,m)} \xrightarrow{P} \pi_j$ *as* $m \to \infty$.

Thus, the same convergence results hold for the parallel version as for the serial version of MAdaSub. The benefit of the parallel algorithm is that the convergence of the proposal probabilities against the posterior inclusion probabilities can be accelerated via the exchange of information between the parallel chains, so that the MCMC chains can converge faster against the full posterior distribution. There is a practical trade-off between the effectiveness regarding the joint update for the proposal probabilities and the efficiency regarding the communication between the different chains. If the number of rounds R is chosen to be small with a large number of iterations T per round, the available information from the multiple chains is not fully utilized during the algorithm; however, if the number of rounds R is chosen to be large with a small number of iterations T per round, then the computational cost of communication between the chains increases and may outweigh the benefit of the accelerated convergence of the proposal probabilities. If $T_{\max}$ denotes the maximum number of iterations, we observe that choosing the number of rounds $R \in [10, 100]$ with $T = T_{\max}/R$ iterations per round works well in practice (see Sections 5 and 6 as well as Table G.4 of the Supplement).

5 Simulated data applications

5.1 Illustrative example

We first illustrate the adaptive behaviour of the serial MAdaSub algorithm (Algorithm 1) in a relatively low-dimensional setting. In particular, we consider an illustrative simulated dataset $\mathcal{D} = (\boldsymbol{X}, \boldsymbol{y})$ with sample size $n = 60$ and $p = 20$ explanatory variables, by generating $\boldsymbol{X} = (X_{i,j}) \in \mathbb{R}^{n \times p}$ with i-th row $\boldsymbol{X}_{i,*} \sim \mathcal{N}_p(\boldsymbol{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma} = (\Sigma_{i,j}) \in \mathbb{R}^{p \times p}$ is the covariance matrix with entries $\Sigma_{k,l} = \rho^{|k-l|}$, $k, l \in \{1, \ldots, p\}$, corresponding to a Toeplitz correlation structure with $\rho = 0.9$. The true vector of regression coefficients is considered to be

$$\boldsymbol{\beta}_0 = (0.4, 0.8, 1.2, 1.6, 2.0, 0, \ldots, 0)^T \in \mathbb{R}^p \,,$$

with active set $S_0 = \{1, \ldots, 5\}$. The response $\boldsymbol{y} = (y_1, \ldots, y_n)^T$ is then simulated from the normal linear model via $y_i \overset{\text{ind.}}{\sim} N(\boldsymbol{X}_{i,*}\boldsymbol{\beta}_0, 1)$, $i = 1, \ldots, n$. We employ the g-prior with $g = n$ and an independent Bernoulli model prior with inclusion probability $\omega = 0.5$, resulting in a uniform prior over the model space (see Remark 2.1). In the MAdaSub algorithm we set $r_j^{(0)} = \frac{1}{2}$ for $j \in \mathcal{P}$, i.e. we use the prior inclusion probabilities as initial proposal probabilities. We first consider the choice $L_j = p$ (for $j \in \mathcal{P}$) for the variance parameters of MAdaSub, corresponding to equation (17). Furthermore, we set $\epsilon = \frac{1}{p}$ and run the MAdaSub algorithm for $T = 20{,}000$ iterations. To compare the results of MAdaSub with the true posterior model distribution, we have also conducted a full model enumeration using the Bayesian Adaptive Sampling (BAS) algorithm, which is implemented in the R-package `BAS` (Clyde, 2017).

To illustrate the efficient adaptation of MAdaSub, we present comparisons with independent Metropolis-Hastings algorithms where the individual proposal probabilities are *not* adapted during the algorithm, i.e. we set $r_j^{(t)} = r_j^{(0)}$ for all $t \in \mathbb{N}$ and $j \in \mathcal{P}$. In particular, we consider the choice $r_j^{(t)} = r_j^{(0)} = 0.5$, corresponding to the initial proposal distribution in MAdaSub, and the choice $r_j^{(t)} = r_j^{(0)} = \pi(j \in S \,|\, \mathcal{D})$, corresponding to the targeted proposal distribution, which is, as stated above, the closest independent Bernoulli proposal to the target $\pi(\cdot \,|\, \mathcal{D})$ in terms of Kullback-Leibler divergence (Clyde et al., 2011). Note that the non-adaptive independence sampler with posterior inclusion probabilities as proposal probabilities ($r_j^{(t)} = \pi(j \in S \,|\, \mathcal{D})$) is only considered as a benchmark and cannot be used in practice, since the true posterior probabilities are initially unknown and are to be estimated by the MCMC algorithms. Furthermore, we also present comparisons with a standard local "Markov chain Monte Carlo model composition" (MC^3) algorithm (Madigan et al., 1995), which in each iteration proposes to delete or add a single variable to the current model.

Figure 1 depicts the sizes $|V^{(t)}|$ of the proposed models and the sizes $|S^{(t)}|$ of the sampled models, while Figure 2 shows the evolution of the acceptance rates along the iterations t of the different MCMC algorithms. As might have been expected, the non-adaptive sampler with prior marginals as proposal probabilities performs poorly with a very slow exploration of the model space and a small acceptance rate which remains close to zero. On the other hand, the non-adaptive sampler with posterior marginals as proposal probabilities leads to fast mixing with corresponding acceptance rate of approximately 0.54. Even though the MAdaSub algorithm starts with exactly the same "initial configuration" as the non-adaptive sampler with prior marginals, it quickly adjusts the proposal probabilities accordingly, so that the resulting acceptance rate approaches the target value of 0.54 from the non-adaptive sampler with posterior marginals. In particular, when inspecting the evolution of the sampled model sizes in Figure 1, the MAdaSub algorithm is very difficult to distinguish from the sampler with posterior marginals after a very short burn-in period (see also Figure E.1 of the Supplement).

To illustrate the behaviour of the MAdaSub algorithm with respect to the variance parameters L_j, additionally to the choice $L_j = p$ we examine two further runs of MAdaSub with the same specifications as before, but with $L_j = p/n$ and with $L_j = 100p$, respectively. Figure 2 indicates that the original choice $L_j = p$ is favourable, yielding a

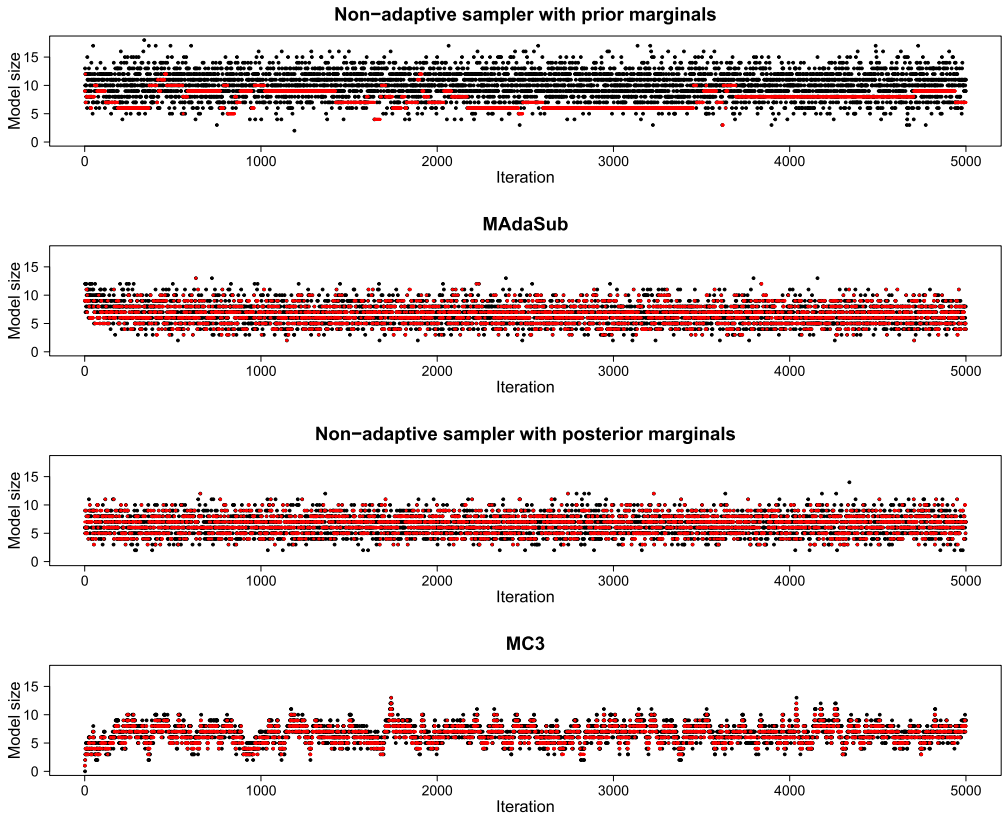

Figure 1: Illustrative example with g-prior. Evolution of the sizes $|V^{(t)}|$ of the proposed models (black) and of the sizes $|S^{(t)}|$ of the sampled models (red) along the first 5,000 iterations (t) for non-adaptive sampler with prior marginals as proposal probabilities, for MAdaSub (with $L_j = p$), for non-adaptive sampler with posterior marginals as proposal probabilities and for local add-delete MC^3 sampler (from top to bottom).

fast and "sustainable" increase of the acceptance rate (see also Figure E.2 of the Supplement for the evolution of proposal probabilities for the different L_j). On the other hand, for $L_j = 100p$ the proposal probabilities in MAdaSub are slowly adapted, while for $L_j = p/n$ the proposal probabilities are adapted very quickly, resulting in initially large acceptance rates; however, this increase is only due to a premature focus of the proposal on certain parts of the model space and thus the acceptance rate decreases at some point when the algorithm identifies other areas of high posterior probability that have not been covered by the proposal. This illustrative example shows that — despite the ergodicity of the MAdaSub algorithm for all choices of its tuning parameters (Theorem 1) — the speed of convergence against the target distribution crucially depends on an appropriate choice of these parameters. Regarding the variance parameters we observe that the choice $L_j = p$ for $j \in \mathcal{P}$ works well in practice (see also results below).

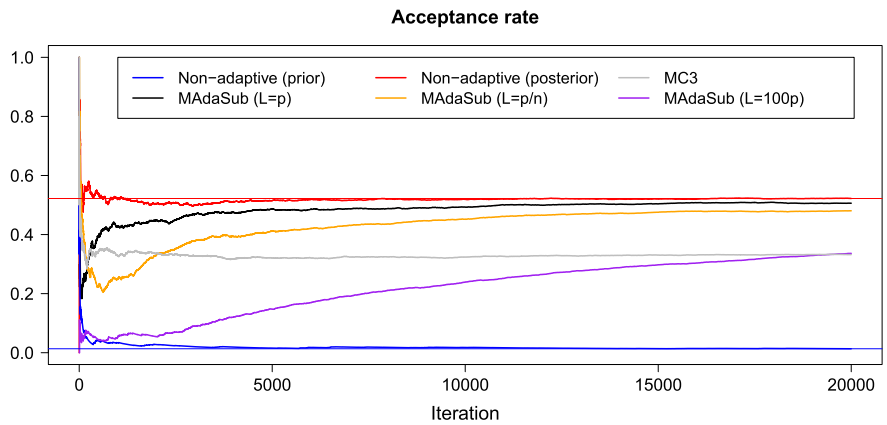

Figure 2: Illustrative example with g-prior. Evolution of acceptance rates along the iterations for non-adaptive independence sampler with prior marginals (blue) and posterior marginals (red) as proposal probabilities, for add-delete MC^3 sampler (gray), as well as for MAdaSub with $L_j = p$ (black), $L_j = p/n$ (orange) and $L_j = 100p$ (purple) for $j \in \mathcal{P}$.

The adaptive nature of MAdaSub entails the possibility for an automatic check of convergence of the algorithm: as the proposal probabilities $r_j^{(t)}$ are continuously adjusted towards the current empirical inclusion frequencies $f_j^{(t)} = \frac{1}{t}\sum_{i=1}^{t} \mathbb{1}_{S^{(i)}}(j)$ (see equation (11)), the algorithm may be stopped as soon as the individual proposal probabilities and empirical inclusion frequencies are within a prespecified distance $\delta \in (0,1)$ (e.g. $\delta = 0.005$, see Figure E.3 of the Supplement), i.e. the algorithm is stopped at iteration t_c if $\max_{j\in\mathcal{P}} |f_j^{(t_c)} - r_j^{(t_c)}| \leq \delta$. Even when automatic stopping may be applied, we additionally recommend to investigate the convergence of the MAdaSub algorithm via the diagnostic plots presented in this section and in Section E of the Supplement.

5.2 Low-dimensional simulation study

In this simulation study we further investigate the performance of the serial MAdaSub algorithm in relation to local non-adaptive and adaptive algorithms. In particular, we analyse how the algorithms are affected by high correlations between the covariates.

We consider a similar low-dimensional setting as in the illustrative data application with $p = 20$ covariates and sample size $n = 60$. To evaluate the performance in a variety of different data situations, for each simulated dataset the number s_0 of informative variables is randomly drawn from $\{0, 1, \ldots, 10\}$ and the true active set $S_0 \subseteq \mathcal{P}$ of size $|S_0| = s_0$ is randomly selected from the full set of covariates $\mathcal{P} = \{1, \ldots, p\}$; then, for each $j \in S_0$, the j-th component $\beta_{0,j}$ of the true coefficient vector $\boldsymbol{\beta}_0 \in \mathbb{R}^p$ is simulated from a uniform distribution $\beta_{0,j} \sim U(-2, 2)$. As before, the covariates are simulated using a Toeplitz correlation structure, while the response is simulated from a normal

linear model with error variance $\sigma^2 = 1$. We consider three different correlation settings by varying the correlation ρ between adjacent covariates in the Toeplitz structure: a low-correlated setting with $\rho = 0.3$, a highly-correlated setting with $\rho = 0.9$ and a very highly-correlated setting with $\rho = 0.99$. For each of the three settings, 200 different datasets are simulated as described above; in each case, we employ a g-prior with $g = n$ on the regression coefficients and a uniform prior on the model space.

For each simulated dataset we apply MAdaSub with 20,000 iterations, using $L_j = p$ for $j \in \mathcal{P}$ and $\epsilon = \frac{1}{p}$. In order to investigate the influence of the initial proposal probabilities $r_j^{(0)}$ in MAdaSub, two different choices for $r_j^{(0)}$ are considered: choice (a) based on prior inclusion probabilities $r_j^{(0)} = \frac{1}{2}$ and choice (b) based on (approximated) marginal posterior odds

$$\pi_j^{\text{marg}} = \frac{\text{PO}_j}{1 + \text{PO}_j} \quad \text{with} \quad \text{PO}_j = \frac{P(S = \{j\} \,|\, \mathcal{D})}{P(S = \emptyset \,|\, \mathcal{D})}, \; j \in \mathcal{P}, \tag{24}$$

and setting $r_j^{(0)} = \min\{\max\{\pi_j^{\text{marg}}, \frac{1}{p}\}, 0.9\}$ to prevent the premature focus of the algorithm on some covariates (if $\pi_j^{\text{marg}} \approx 1$) or the avoidance of other covariates (if $\pi_j^{\text{marg}} \approx 0$). Here, the marginal posterior odds PO_j are crude approximations to the true posterior odds, derived under the assumption of posterior independence of variable inclusion. The local MC^3 algorithm (Madigan et al., 1995) is applied as before as well as with additional swap moves to potentially improve the mixing (as in Griffin et al., 2021). Using the R-package `scaleBVS` (Zanella and Cabezas Gonzalez, 2020), we apply the adaptive weighted tempered Gibbs sampling algorithm of Zanella and Roberts (2019) to obtain (weighted) frequency estimates (as for the other algorithms) and Rao-Blackwellized estimates of posterior inclusion probabilities (PIPs). Exact PIPs are again derived using the BAS algorithm (Clyde, 2017). The algorithms are evaluated based on final acceptance rates and numbers of iterations for convergence of the estimates $\hat{f}_j^{(t)}$ to the true PIPs, where PIP convergence is defined to occur at the smallest iteration t_c for which $\max_{j \in \mathcal{P}} |\hat{f}_j^{(t_c)} - \pi_j| \leq 0.05$; if $t_c \geq 20{,}000$, then the number of iterations for convergence is displayed as 20,000 in Figure 3.

Figure 3 shows that the acceptance rates of the MAdaSub samplers tend to be substantially larger in comparison to the local MC^3 algorithms, while the acceptance rates of the weighted Tempered Gibbs Sampler (wTGS) are equal to one by construction. Nevertheless, for the MAdaSub samplers a decreasing trend of acceptance rates can be observed with increasing correlations. This observation reflects that for low-correlated situations the resulting posterior distribution is often closer to an independent Bernoulli form than for highly-correlated cases, and thus can be better approximated by the proposal distributions of MAdaSub, leading to larger acceptance rates. In the low-correlated setting ($\rho = 0.3$), the choice (b) for the initial proposal probabilities in MAdaSub based on marginal posterior odds leads to slightly larger acceptance rates and a faster PIP convergence compared to the MAdaSub sampler (a) based on the prior inclusion probabilities. However, in cases of high correlations among some of the covariates ($\rho = 0.9$ and $\rho = 0.99$), the prior choice (a) is clearly favourable yielding larger acceptance rates and a faster PIP convergence compared to the MAdaSub sampler (b) and the MC^3 algorithm.

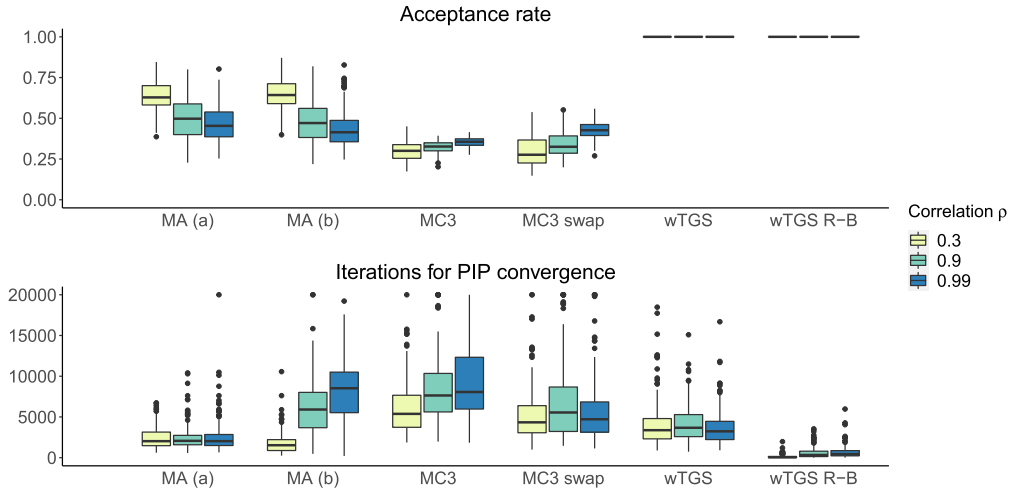

Figure 3: Low-dimensional simulation study with varying correlation $\rho \in \{0.3, 0.9, 0.99\}$ in Toeplitz structure. Performance of MAdaSub with initial proposal probabilities $r_j^{(0)} = 0.5$ based on prior inclusion probabilities (MA(a)), MAdaSub with $r_j^{(0)}$ being based on marginal posterior odds* (MA(b)), MC^3 samplers with and without "swap" moves, as well as adaptive weighted Tempered Gibbs Sampler based on weighted frequency estimates (wTGS) and Rao-Blackwellized estimates (wTGS R-B), in terms of acceptance rates and numbers of iterations for convergence of the estimates to the true posterior inclusion probabilities (PIP).

*The (approximated) marginal posterior odds are provided in equation (24).

Thus, while in low-correlated settings the marginal posterior odds yield reasonable first approximations to the true posterior odds, the prior inclusion probabilities are more robust and to be preferred as initial proposal probabilities in MAdaSub in situations with high correlations. Overall, the MAdaSub sampler (a) yields a well-mixing algorithm in all considered settings, which is also competitive to the adaptive wTGS algorithm based on weighted frequency estimates, while wTGS with Rao-Blackwellization (R-B) provides faster convergence. Note that the computational cost of R-B is small in this low-dimensional conjugate setting but increases for high-dimensional and non-conjugate settings with Laplace approximations (Zanella and Roberts, 2019; Wan and Griffin, 2021). An additional sensitivity analysis regarding different variance parameters L_j in MAdaSub (see Figure F.1 of the Supplement) supports the choice $L_j = p = 20$ in all considered correlation settings and indicates that results are very robust for $L_j \in [p/2, 2p]$.

5.3 High-dimensional simulation study

To investigate the performance of the serial and parallel versions of MAdaSub in high-dimensional settings, we consider the same simulation set-up as in Yang et al. (2016)

and Griffin et al. (2021): data are simulated from a sparse linear regression model with true coefficients

$$\boldsymbol{\beta}_0 = \text{SNR} \times \sqrt{\log(p)/n} \times (2, -3, 2, 2, -3, 3, -2, 3, -2, 3, 0, \ldots, 0)^T \in \mathbb{R}^p. \tag{25}$$

Similar to the low-dimensional simulations, covariates are generated from a Toeplitz correlation structure with $\rho = 0.6$ and the response is simulated via $y_i \overset{\text{ind.}}{\sim} N(\boldsymbol{X}_{i,*}\boldsymbol{\beta}_0, 1)$, $i = 1, \ldots, n$. As in Griffin et al. (2021), we consider the conjugate prior (4) with $g = 9$ and prior independence of the regression coefficients ($\boldsymbol{W}_S = \boldsymbol{I}_{|S|}$ for $S \in \mathcal{M}$), together with the model prior (5) with (fixed) prior inclusion probability $\omega = 10/p$. For each setting with $n \in \{500, 1000\}$, $p \in \{500, 5000\}$ and signal-to-noise ratio SNR $\in \{0.5, 1, 2, 3\}$, we simulate one dataset and apply each algorithm 200 times to assess the stability of estimated posterior inclusion probabilities. As in Griffin et al. (2021), each algorithm is based on 5 parallel chains using 5 CPUs. We consider the serial version of MAdaSub where the individual chains (Algorithm 1) are run in parallel but do not exchange any information and the parallel version (Algorithm 2 of the Supplement) where the chains exchange information regarding the proposal probabilities after each of $R = 50$ rounds (considering 25 burn-in rounds for both versions; each round consists of 1000 and 10,000 iterations for $p = 500$ and $p = 5000$, respectively). For the serial version, the initial proposal probabilities are set to the prior inclusion probabilities, i.e. $r_j^{(k,0)} = 10/p$, and the variance parameters $L_j^{(k)} = p$ are the same for all chains k. For the parallel version, we consider different random initializations of proposal probabilities $r_j^{(k,0)} = q^{(k)}/p$, $j \in \mathcal{P}$, with $q^{(k)} \sim U(2, 10)$ and variance parameters $L_j^{(k)} = L^{(k)}$, $j \in \mathcal{P}$, with $L^{(k)} \sim U(p/2, 2p)$ for each chain k. For all MAdaSub chains we set $\epsilon = 1/p$. Additional results of sensitivity analyses regarding different choices of the tuning parameters of MAdaSub can be found in Section G of the Supplement.

The performance of the MAdaSub algorithms (A) with serial and parallel updating schemes is assessed in terms of median acceptance rates, as well as in comparison to the add-delete-swap MC^3 algorithm (B) in terms of the median estimated ratio $\hat{r}_{A,B}^{(20)}$ of the relative time-standardized effective sample size of algorithm A versus algorithm B for the posterior inclusion probabilities (PIPs) over the 20 variables with the largest estimated PIPs (averaged over all algorithms). The estimated ratio of the relative time-standardized effective sample size is given by $\hat{r}_{A,B} = (s_B^2 t_B)/(s_A^2 t_A)$, with t_A and t_B the median computation times and s_A^2 and s_B^2 the variances of PIP estimates based on 200 independent runs of each algorithm (cf. Griffin et al., 2021). Here, we consider the median ratio $\hat{r}_{A,B}^{(20)}$ over the 20 variables with the largest estimated PIPs, as many variables receive very small posterior probability due to the sparsity-inducing prior and the sparse generating model with only 10 signal variables (in all settings the estimated PIPs for variables not among the top 20 are all below 0.5%, while the median estimated PIP over all variables is below 0.07%). Complementary results regarding the median of $\hat{r}_{A,B}$ over all variables are provided in Table G.1 of the Supplement, comparing the performance of MAdaSub also with the adaptive approaches in Griffin et al. (2021).

Table 1 shows that in all considered settings the median estimated time-standardized effective sample size for both MAdaSub versions is several orders larger than for the MC^3

(n, p)	MAdaSub	SNR = 0.5 $\hat{r}_{A,B}^{(20)}$ / Acc.	SNR = 1 $\hat{r}_{A,B}^{(20)}$ / Acc.	SNR = 2 $\hat{r}_{A,B}^{(20)}$ / Acc.	SNR = 3 $\hat{r}_{A,B}^{(20)}$ / Acc.
$(500, 500)$	serial	69.4 / 44.6%	23.0 / 31.9%	4.8 / 6.3%	8.3 / 9.3%
	parallel	22.9 / 45.3%	8.9 / 37.7%	7.5 / 18.1%	12.1 / 21.4%
$(500, 5000)$	serial	376.9 / 47.5%	50.3 / 46.6%	8.2 / 5.1%	17.9 / 9.5%
	parallel	474.4 / 48.0%	78.7 / 44.8%	82.8 / 17.5%	186.4 / 23.4%
$(1000, 500)$	serial	110.7 / 53.4%	13.7 / 39.0%	2.4 / 6.0%	8.7 / 9.0%
	parallel	62.0 / 54.2%	7.0 / 39.0%	7.3 / 17.7%	12.8 / 21.0%
$(1000, 5000)$	serial	657.3 / 45.3%	7.5 / 26.5%	23.9 / 9.4%	35.1 / 11.6%
	parallel	674.1 / 45.8%	6.2 / 10.7%	175.6 / 23.1%	281.7 / 24.7%

Table 1: Results of high-dimensional simulation study. Performance of MAdaSub algorithms (A) with serial and parallel updating schemes compared to add-delete-swap MC^3 algorithm (B) in terms of median estimated ratios $\hat{r}_{A,B}^{(20)}$ of the relative time-standardized effective sample size for PIPs over the 20 variables with the largest estimated PIPs. Median acceptance rates (Acc.) for MAdaSub are also provided.

algorithm. For low SNRs (e.g. SNR = 0.5), both MAdaSub versions tend to show larger improvements compared to the MC^3 algorithm than for high SNRs (e.g. SNR = 3). Note that for high SNRs, the posterior distribution tends to be more concentrated around the true model $S_0 = \{1, \ldots, 10\}$, so that local proposals of the add-delete-swap MC^3 algorithm may also be reasonable. On the other hand, for low SNR, the posterior tends to be less concentrated, so that global moves of MAdaSub have a larger potential to improve the mixing compared to the MC^3 algorithm. The acceptance rates of MAdaSub are also larger in small SNR scenarios, as the posterior model distribution tends to be better approximated by independent Bernoulli proposals. However, in all considered settings, the acceptance rates of MAdaSub are reasonably large with median acceptance rates between 5.1% and 54.2% (see Table 1) and are considerably larger compared to the MC^3 algorithm with median acceptance rates between 0.6% and 5.8% (detailed results not shown).

For low SNRs (SNR ≤ 1), serial updating in MAdaSub tends to yield larger (for $p = 500$) or similar (for $p = 5000$) time-standardized effective sample sizes compared to parallel updating, as both versions appear to have converged to stationarity with similar acceptance rates, while the parallel version tends to yield larger computation times as a result of communicating chains. For large SNRs (SNR ≥ 2), MAdaSub with parallel updating performs favourable since the proposal probabilities tend to converge faster than with serial updating, which leads to considerably larger acceptance rates and outweighs the computational cost of communicating chains. Previous results for the same simulation set-up indicate that the two alternative individual adaptation algorithms of Griffin et al. (2021) tend to yield the largest improvements compared to the MC^3 algorithm for higher SNR (particularly for SNR = 2). The proposal (18) of these algorithms allows for larger moves than the add-delete-swap proposal in MC^3, but — in contrast to the independence proposal of MAdaSub — the proposal (18) still locally depends on the previously sampled model. Overall, MAdaSub shows a competitive performance compared to the adaptive algorithms of Griffin et al. (2021), with advantages

of MAdaSub in low SNR settings and advantages of the adaptive algorithms of Griffin et al. (2021) in high SNR settings (see Table G.1 of the Supplement).

6 Real data applications

6.1 Tecator data

We first examine the Tecator dataset which has already been investigated in Griffin and Brown (2010), Lamnisos et al. (2013) and Griffin et al. (2021). The data has been recorded by Borggaard and Thodberg (1992) on a Tecator Infratec Food Analyzer and consists of $n = 172$ meat samples and their near-infrared absorbance spectra, represented by $p = 100$ channels in the wavelength range 850-1050nm (compare Griffin and Brown, 2010). The fat content of the samples is considered as the response variable. For comparison reasons, we choose the same conjugate prior set-up as in Lamnisos et al. (2013), i.e. we use the prior given in equation (4) with $g = 5$, $\boldsymbol{W}_S = \boldsymbol{I}_{|S|}$ for $S \in \mathcal{M}$ and we employ the independent Bernoulli model prior given in equation (5) with (fixed) prior inclusion probability $\omega = \frac{5}{100}$.

To investigate the stability of MAdaSub for different choices of its tuning parameters, we run 25 independent serial MAdaSub chains (Algorithm 1) with random initializations of the proposal probabilities $r_j^{(k,0)} = q^{(k)}/p$, $j \in \mathcal{P}$, with $q^{(k)} \sim U(2, 10)$ and of the variance parameters $L_j^{(k)} = L^{(k)}$, $j \in \mathcal{P}$, with $L^{(k)} \sim U(p/2, 2p)$, for each chain $k = 1, \ldots, 25$. Furthermore, we run 25 additional parallel MAdaSub chains (Algorithm 2) with the described random initializations, exchanging the information after each of $R = 58$ rounds of $T = 5{,}000$ iterations (yielding in total 290,000 iterations for each of the chains, cf. Lamnisos et al., 2013). Figure 4 shows the resulting empirical variable inclusion frequencies (as estimates of posterior inclusion probabilities) for the 25 serial and 25 parallel MAdaSub chains. From left to right, the first three plots of Figure 4 depict the development of the empirical inclusion frequencies for the first three rounds of 5,000 iterations each, while the rightmost plots depict the final empirical inclusion frequencies after 290,000 iterations (disregarding a burn-in period of 100,000 iterations, cf. Lamnisos et al., 2013). After the first 5,000 iterations, the empirical inclusion frequencies show a similar variability for the serial and parallel chains, as no communication between the parallel chains has yet occurred. After the second round of 5,000 further iterations, the benefit of the communication between the 25 parallel chains is apparent, leading to less variable estimates due to a faster convergence of the proposal probabilities against the posterior inclusion probabilities. Nevertheless, also the serial MAdaSub chains (with different initial tuning parameters) provide quite accurate estimates after only 10,000 iterations.

After 290,000 iterations, all of the serial and parallel MAdaSub chains yield very stable estimates of posterior inclusion probabilities, reproducing the results shown in Figure 1 of Lamnisos et al. (2013). Details on additional comparisons with Lamnisos et al. (2013) and computation times can be found in Section H of the Supplement. As the covariates represent 100 channels of the near-infrared absorbance spectrum, adjacent covariates are highly correlated and it is not surprising that they have similar

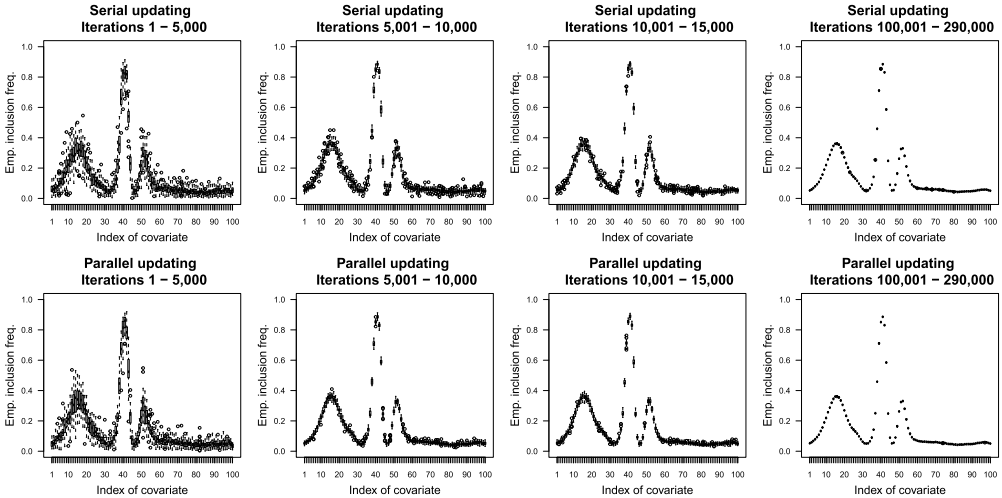

Figure 4: Tecator data application. Results of 25 independent serial MAdaSub chains (Algorithm 1) and of 25 parallel MAdaSub chains exchanging information after every 5,000 iterations (Algorithm 2) in terms of empirical variable inclusion frequencies f_j for $j \in \{1, \ldots, 100\}$.

posterior inclusion probabilities. If one is interested in selecting a final single model, the median probability model (which includes all variables with posterior inclusion probability greater than 0.5, see Barbieri and Berger, 2004) might not be the best choice in this particular situation, since then only variables corresponding to the "global mode" and no variables from the two other "local modes" in Figure 4 are selected. Alternatively, one may choose one or two variables from each of the three "local modes" or make use of Bayesian model averaging (Raftery et al., 1997) for predictive inference.

6.2 PCR and Leukemia data

We illustrate the effectiveness of MAdaSub for two further high-dimensional datasets. In particular, we consider the polymerase chain reaction (PCR) dataset of Lan et al. (2006) with $p = 22{,}575$ explanatory variables (expression levels of genes), sample size $n = 60$ (mice) and continuous response data (the dataset is available in JRSS(B) Datasets Vol. 77(5), Song and Liang, 2015). Furthermore, we consider the leukemia dataset of Golub et al. (1999) with 6817 gene expression measurements of $n = 72$ patients and binary response data (the dataset can be loaded via the R-package `golubEsets`, Golub, 2017). For the PCR dataset we face the problem of variable selection in a linear regression framework, while for the leukemia dataset we consider variable selection in a logistic regression framework. We have preprocessed the leukemia dataset as described in Dudoit et al. (2002), resulting in a restricted design matrix with $p = 3571$ columns (genes). Furthermore, in both datasets we have mean-centered the columns of the design matrix after the initial preprocessing.

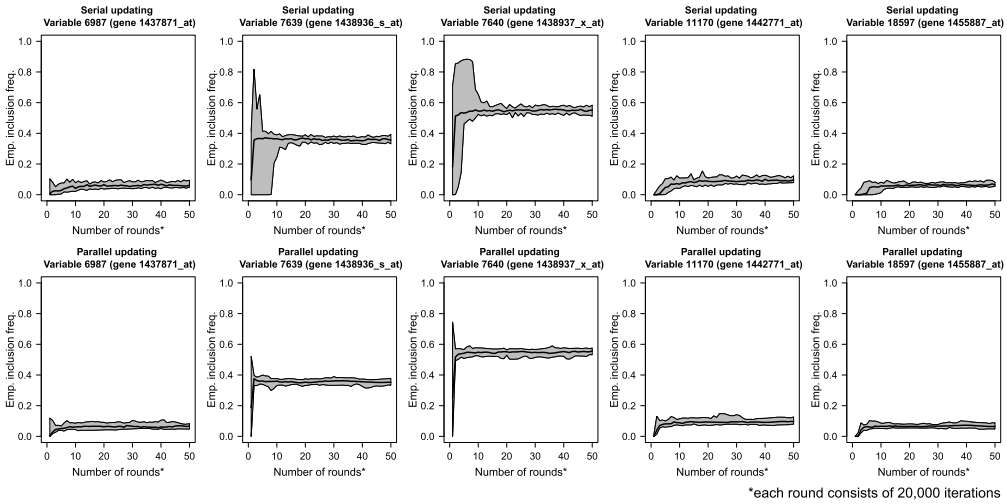

Figure 5: PCR data application. Evolution of empirical variable inclusion frequencies for 25 serial MAdaSub chains (Algorithm 1, top) and 25 parallel MAdaSub chains exchanging information after every round of 20,000 iterations (Algorithm 2, bottom). Bold lines represent median frequencies with 5%- and 95%-quantiles (shaded area) over the chains within each round, for most informative variables X_j (with final estimate $f_j \geq 0.05$ for at least one chain).

Here we adopt the posterior approximation induced by EBIC_γ with $\gamma = 1$ (see equation (8)), corresponding to a beta-binomial model prior with $a_\omega = b_\omega = 1$ as parameters in the beta distribution (see Section 2). For both datasets we run 25 independent serial MAdaSub chains with 1,000,000 iterations and 25 parallel MAdaSub chains exchanging information after each of $R = 50$ rounds of $T = 20{,}000$ iterations (yielding also 1,000,000 iterations for each parallel chain). For each serial and parallel chain $k = 1, \ldots, 50$, we set $\epsilon = \frac{1}{p}$ and randomly initialize the proposal probabilities $r_j^{(k,0)} = q^{(k)}/p$, $j \in \mathcal{P}$, with $q^{(k)} \sim U(2, 5)$ and the variance parameters $L_j^{(k)} = L^{(k)}$, $j \in \mathcal{P}$, with $L^{(k)} \sim U(p/2, 2p)$. For the leukemia dataset we make use of a fast C++ implementation for ML-estimation in logistic regression models via a limited-memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) algorithm, which is available in the R-package `RcppNumerical` (Qiu et al., 2016). For both datasets, the 50 MAdaSub chains are run in parallel on a computer cluster with 50 CPUs, yielding overall computation times of 2,836 seconds for the PCR data (2,310 seconds for a single chain) and 1,402 seconds for the leukemia data (995 seconds for a single chain).

Figures 5 and 6 show that, despite the high-dimensional model spaces and the different initializations of each chain, the parallel MAdaSub algorithm provides stable estimates of posterior inclusion probabilities for both datasets after a small number of rounds. In particular, the estimates from the parallel MAdaSub algorithm stabilize after only three rounds of 20,000 iterations (see also Figures I.3 and I.4 of the Sup-

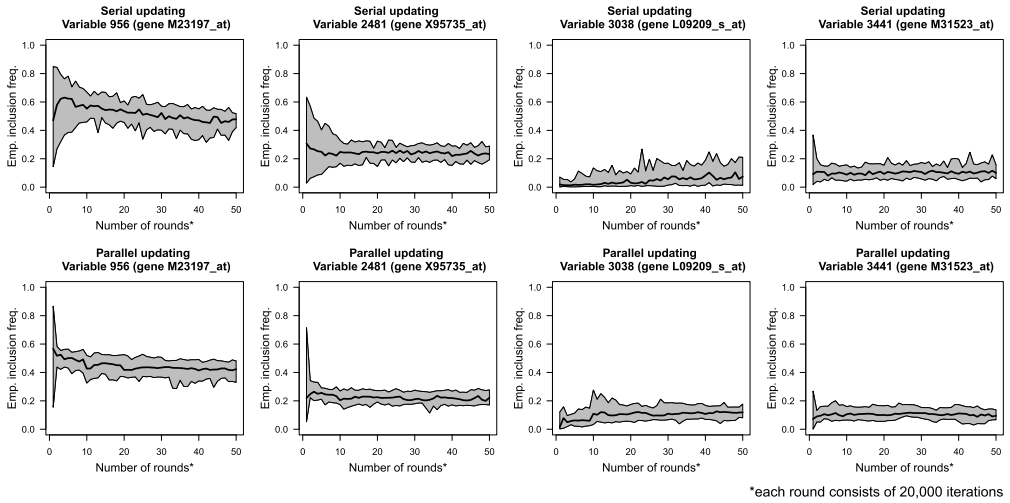

Figure 6: Leukemia data application. Evolution of empirical variable inclusion frequencies for 25 serial MAdaSub chains (Algorithm 1, top) and 25 parallel MAdaSub chains exchanging information after each round of 20,000 iterations (Algorithm 2, bottom) for most informative variables X_j (with final estimate $f_j \geq 0.1$ for at least one chain), cf. Figure 5.

plement). For the PCR data, all serial and parallel MAdaSub chains yield congruent estimates of posterior inclusion probabilities after 1,000,000 iterations (Figures 5, I.2 and I.3). The final acceptance rates of MAdaSub for the PCR dataset are between 20% and 22%, while the acceptance rates for the leukemia dataset are between 3% and 6%. The smaller acceptance rates for the leukemia dataset indicate that this corresponds to a more challenging scenario (i.e. the targeted posterior model distribution seems to be "further away" from an independent Bernoulli form). This observation is also reflected in the larger variability of the estimates from the MAdaSub chains without parallel updating (Figures 6, I.2 and I.4). The leukemia data application particularly illustrates the benefits of the parallel version of MAdaSub, where multiple chains with different initializations sequentially explore different regions of the model space, but exchange the information after each round of 20,000 iterations, increasing the speed of convergence of the proposal probabilities to the posterior inclusion probabilities.

Note that in very high-dimensional settings such as for the PCR data (with $p =$ 22,575), the classical MC^3 algorithm (Madigan et al., 1995) does not yield stable estimates due to slow mixing (cf. Griffin et al., 2021), while the BAS algorithm (Clyde, 2017) using sampling without replacement is computationally intractable. Further results in Griffin et al. (2021) show that several competing adaptive algorithms — including sequential Monte Carlo algorithms of Schäfer and Chopin (2013) and tempered Gibbs sampling algorithms of Zanella and Roberts (2019) — do not provide reliable estimates of posterior inclusion probabilities for the PCR data; only the adaptively scaled individual adaptation algorithm of Griffin et al. (2021) with proposals of the form (18) yields

stable results for the PCR data similarly to MAdaSub with a slightly different prior set-up (see Figures 10 and 11 of the Supplement of Griffin et al., 2021).

Due to the very large model spaces in both considered examples, posterior probabilities of individual models are generally small and corresponding MCMC estimates will typically not be very reliable. Therefore, as in similar studies (see Griffin et al., 2021), we have focused on the estimation of posterior inclusion probabilities (PIPs). For the PCR data two variables (genes) stand out with respect to the final estimates of their PIPs, namely the gene 1438937_x_at (covariate index $j = 7640$) with estimated PIP between 0.54 and 0.56, and the gene 1438936_s_at ($j = 7639$) with estimated PIP between 0.35 and 0.37. Similarly, for the leukemia data two genes stand out, namely the genes M23197_at ($j = 956$) with estimated PIP between 0.39 and 0.43 and X95735_at ($j = 2481$) with estimated PIP between 0.21 and 0.22 (considering final estimates from the 25 parallel chains only); these two genes are also among the four top scoring genes in a Bayesian probit regression analysis in Ai-Jun and Xin-Yuan (2009).

7 Discussion

We introduced the Metropolized Adaptive Subspace (MAdaSub) algorithm for sampling from high-dimensional posterior model distributions in situations where conjugate priors or approximations to the posterior are employed. We further developed an efficient parallel version of MAdaSub, where the information regarding the adaptive proposal probabilities of the variables can be shared periodically between the different chains. Simulated and real data applications illustrated that MAdaSub can efficiently sample from multimodal posterior model distributions, yielding stable estimates of posterior inclusion probabilities even for ten thousands of possible covariates.

The reliable estimation of posterior inclusion probabilities is particularly important for Bayesian inference, since the median probability model (MPM) — including all variables with posterior inclusion probability larger than 0.5 — has been shown to yield optimal predictions for uncorrelated covariates (Barbieri and Berger, 2004) and also a favourable performance for correlated designs (Barbieri et al., 2021), e.g. compared to the largest posterior probability model. MAdaSub provides a natural adaptive MCMC algorithm which focuses on the sequential adaptation of currently estimated inclusion probabilities, with the aim of driving the sampler quickly into regions near to the MPM; in the limit, the MPM itself is the model which receives the largest probability under the independent Bernoulli proposal of MAdaSub. Despite the continuing adaptation of the proposals, we have shown that MAdaSub constitutes a valid MCMC algorithm which samples from the full posterior model distribution. While the serial and parallel versions of MAdaSub are ergodic for all choices of their tuning parameters (see Theorem 1 and Theorem 3), in practice the speed of convergence against the targeted posterior depends crucially on a proper choice of their tuning parameters (see Section 5). Deriving theoretical results regarding the mixing time of the proposed algorithms is an important but challenging issue for further research.

Since MAdaSub is based on adaptive independent proposal distributions, in each iteration of the algorithm the proposed model is (almost) independent of the current

model, so that "distant" moves in the model space are encouraged. This can be advantageous in comparison to Gibbs samplers and Metropolis-Hastings algorithms based on local proposal distributions, which may yield larger acceptance rates but are more prone to be stuck in local modes of the posterior model distribution. In future work one may also consider combinations of the adaptive independent proposals in MAdaSub with adaptive local proposals as for example in Lamnisos et al. (2013) and Zanella and Roberts (2019). While MAdaSub yields competitive results without the use of Rao-Blackwellization compared to the related adaptive algorithms of Griffin et al. (2021), the incorporation of Rao-Blackwellized estimates of posterior inclusion probabilities in the burn-in phase or as initial proposal probabilities may further increase the speed of convergence of MAdaSub. Finally, the extension of MAdaSub to settings with non-conjugate priors is interesting to be investigated, for example by considering data augmentation approaches with additional latent variables or by incorporating reversible-jump moves (Green, 1995; Wan and Griffin, 2021).

Supplementary Material

Supplementary material for "A Metropolized adaptive subspace algorithm for high-dimensional Bayesian variable selection" (DOI: 10.1214/22-BA1351SUPP; .pdf). The Supplement includes detailed proofs of the ergodicity results, algorithmic details of the parallel version of MAdaSub, further details on related approaches, as well as additional results for the simulated and real data applications.

References

Ai-Jun, Y. and Xin-Yuan, S. (2009). "Bayesian variable selection for disease classification using gene expression data." *Bioinformatics*, 26(2): 215–222. doi: https://doi.org/10.1093/bioinformatics/btp638. 285

Barbieri, M. M. and Berger, J. O. (2004). "Optimal predictive model selection." *The Annals of Statistics*, 32(3): 870–897. MR2065192. doi: https://doi.org/10.1214/009053604000000238. 262, 268, 282, 285

Barbieri, M. M., Berger, J. O., George, E. I., and Ročková, V. (2021). "The median probability model and correlated variables." *Bayesian Analysis*, 16(4): 1085–1112. MR4381128. doi: https://doi.org/10.1214/20-BA1249. 268, 285

Bertsimas, D., King, A., and Mazumder, R. (2016). "Best subset selection via a modern optimization lens." *The Annals of Statistics*, 44(2): 813–852. MR3476618. doi: https://doi.org/10.1214/15-AOS1388. 262

Borggaard, C. and Thodberg, H. H. (1992). "Optimal minimal neural interpretation of spectra." *Analytical Chemistry*, 64(5): 545–551. doi: https://doi.org/10.1021/ac00029a018. 281

Carbonetto, P. and Stephens, M. (2012). "Scalable variational inference for Bayesian variable selection in regression, and its accuracy in genetic association studies."

Bayesian Analysis, 7(1): 73–108. MR2896713. doi: https://doi.org/10.1214/12-BA703. 271

Chen, J. and Chen, Z. (2008). "Extended Bayesian information criteria for model selection with large model spaces." *Biometrika*, 95(3): 759–771. MR2443189. doi: https://doi.org/10.1093/biomet/asn034. 262, 265

Chen, J. and Chen, Z. (2012). "Extended BIC for small-n-large-P sparse GLM." *Statistica Sinica*, 22(2): 555–574. MR2954352. doi: https://doi.org/10.5705/ss.2010.216. 265, 266

Clyde, M. (2017). *BAS: Bayesian Adaptive Sampling for Bayesian model averaging*. R package version 1.4.7. 274, 277, 284

Clyde, M. A., Ghosh, J., and Littman, M. L. (2011). "Bayesian adaptive sampling for variable selection and model averaging." *Journal of Computational and Graphical Statistics*, 20(1): 80–101. MR2816539. doi: https://doi.org/10.1198/jcgs.2010.09049. 266, 274

Dellaportas, P., Forster, J. J., and Ntzoufras, I. (2002). "On Bayesian model and variable selection using MCMC." *Statistics and Computing*, 12(1): 27–36. MR1877577. doi: https://doi.org/10.1023/A:1013164120801. 262

Dudoit, S., Fridlyand, J., and Speed, T. P. (2002). "Comparison of discrimination methods for the classification of tumors using gene expression data." *Journal of the American Statistical Association*, 97(457): 77–87. MR1963389. doi: https://doi.org/10.1198/016214502753479248. 282

Foster, D. P. and George, E. I. (1994). "The risk inflation criterion for multiple regression." *The Annals of Statistics*, 22(4): 1947–1975. MR1329177. doi: https://doi.org/10.1214/aos/1176325766. 265

George, E. I. and McCulloch, R. E. (1993). "Variable selection via Gibbs sampling." *Journal of the American Statistical Association*, 88(423): 881–889. doi: https://doi.org/10.1080/01621459.1993.10476353. 262

Giordani, P. and Kohn, R. (2010). "Adaptive independent Metropolis–Hastings by fast estimation of mixtures of normals." *Journal of Computational and Graphical Statistics*, 19(2): 243–259. MR2758304. doi: https://doi.org/10.1198/jcgs.2009.07174. 266

Golub, T. (2017). *golubEsets: ExprSets for Golub leukemia data*. R package version 1.20.0. 282

Golub, T., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., and Caligiuri, M. A. (1999). "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring." *Science*, 286(5439): 531–537. doi: https://doi.org/10.1126/science.286.5439.531. 282

Green, P. J. (1995). "Reversible jump Markov chain Monte Carlo computation and Bayesian model determination." *Biometrika*, 82(4): 711–732. MR1380810. doi: https://doi.org/10.1093/biomet/82.4.711. 262, 286

Griffin, J., Latuszynski, K., and Steel, M. (2014). "Individual adaptation: An adaptive MCMC scheme for variable selection problems." *arXiv preprint arXiv:1412.6760*. URL https://arxiv.org/abs/1412.6760 262, 270

Griffin, J., Latuszynski, K., and Steel, M. (2021). "In search of lost mixing time: adaptive Markov chain Monte Carlo schemes for Bayesian variable selection with very large p." *Biometrika*, 108(1): 53–69. MR4226189. doi: https://doi.org/10.1093/biomet/asaa055. 262, 265, 270, 277, 279, 280, 281, 284, 285, 286

Griffin, J. E. and Brown, P. J. (2010). "Inference with normal-gamma prior distributions in regression problems." *Bayesian Analysis*, 5(1): 171–188. MR2596440. doi: https://doi.org/10.1214/10-BA507. 281

Holden, L., Hauge, R., and Holden, M. (2009). "Adaptive independent Metropolis–Hastings." *The Annals of Applied Probability*, 19(1): 395–413. MR2498682. doi: https://doi.org/10.1214/08-AAP545. 266

Ji, C. and Schmidler, S. C. (2013). "Adaptive Markov chain Monte Carlo for Bayesian variable selection." *Journal of Computational and Graphical Statistics*, 22(3): 708–728. MR3173738. doi: https://doi.org/10.1080/10618600.2013.819178. 262

Kass, R. E. and Raftery, A. E. (1995). "Bayes factors." *Journal of the American Statistical Association*, 90(430): 773–795. MR3363402. doi: https://doi.org/10.1080/01621459.1995.10476572. 265

Kass, R. E. and Wasserman, L. (1995). "A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion." *Journal of the American Statistical Association*, 90(431): 928–934. MR1354008. doi: https://doi.org/10.2307/2291327. 265

Kohn, R., Smith, M., and Chan, D. (2001). "Nonparametric regression using linear combinations of basis functions." *Statistics and Computing*, 11(4): 313–322. MR1863502. doi: https://doi.org/10.1023/A:1011916902934. 264

Lamnisos, D., Griffin, J. E., and Steel, M. F. (2013). "Adaptive Monte Carlo for Bayesian variable selection in regression models." *Journal of Computational and Graphical Statistics*, 22(3): 729–748. MR3173739. doi: https://doi.org/10.1080/10618600.2012.694756. 262, 270, 281, 286

Lan, H., Chen, M., Flowers, J. B., Yandell, B. S., Stapleton, D. S., Mata, C. M., Mui, E. T.-K., Flowers, M. T., Schueler, K. L., and Manly, K. F. (2006). "Combined expression trait correlations and expression quantitative trait locus mapping." *PLoS Genetics*, 2(1): e6. doi: https://doi.org/10.1371/journal.pgen.0020006. 282

Lee, J. D., Sun, D. L., Sun, Y., and Taylor, J. E. (2016). "Exact post-selection inference, with application to the lasso." *The Annals of Statistics*, 44(3): 907–927. MR3485948. doi: https://doi.org/10.1214/15-AOS1371. 261

Liang, F., Song, Q., and Yu, K. (2013). "Bayesian subset modeling for high-dimensional generalized linear models." *Journal of the American Statistical Association*, 108(502): 589–606. MR3174644. doi: https://doi.org/10.1080/01621459.2012.761942. 262, 264, 266

Liu, Y. and Ročková, V. (2021). "Variable selection via Thompson sampling." *Journal of the American Statistical Association*. doi: https://doi.org/10.1080/01621459.2021.1928514. 271

Madigan, D., York, J., and Allard, D. (1995). "Bayesian graphical models for discrete data." *International Statistical Review / Revue Internationale de Statistique*, 63(2): 215–232. doi: https://doi.org/10.2307/1403615. 262, 274, 277, 284

Meinshausen, N. and Bühlmann, P. (2010). "Stability selection." *Journal of the Royal Statistical Society, Ser. B*, 72(4): 417–473. MR2758523. doi: https://doi.org/10.1111/j.1467-9868.2010.00740.x. 261

Narisetty, N. N. and He, X. (2014). "Bayesian variable selection with shrinking and diffusing priors." *The Annals of Statistics*, 42(2): 789–817. MR3210987. doi: https://doi.org/10.1214/14-AOS1207. 262

Neklyudov, K., Egorov, E., Shvechikov, P., and Vetrov, D. (2019). "Metropolis-Hastings view on variational inference and adversarial training." *arXiv preprint arXiv:1810.07151*. URL https://arxiv.org/abs/1810.07151 271

Nott, D. J. and Kohn, R. (2005). "Adaptive sampling for Bayesian variable selection." *Biometrika*, 92(4): 747–763. MR2234183. doi: https://doi.org/10.1093/biomet/92.4.747. 262, 270

Ormerod, J. T., You, C., and Müller, S. (2017). "A variational Bayes approach to variable selection." *Electronic Journal of Statistics*, 11(2): 3549–3594. MR3709863. doi: https://doi.org/10.1214/17-EJS1332. 271

Qiu, Y., Balan, S., Beall, M., Sauder, M., Okazaki, N., and Hahn, T. (2016). *RcppNumerical: 'Rcpp' integration for numerical computing libraries*. R package version 0.3-1. 283

Raftery, A. E., Madigan, D., and Hoeting, J. A. (1997). "Bayesian model averaging for linear regression models." *Journal of the American Statistical Association*, 92(437): 179–191. MR1436107. doi: https://doi.org/10.2307/2291462. 261, 282

Raskutti, G., Wainwright, M. J., and Yu, B. (2011). "Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls." *IEEE Transactions on Information Theory*, 57(10): 6976–6994. MR2882274. doi: https://doi.org/10.1109/TIT.2011.2165799. 262

Roberts, G. O. and Rosenthal, J. S. (2007). "Coupling and ergodicity of adaptive Markov chain Monte Carlo algorithms." *Journal of Applied Probability*, 44(2): 458–475. MR2340211. doi: https://doi.org/10.1239/jap/1183667414. 263, 269

Rosenthal, J. S. (2011). "Optimal proposal distributions and adaptive MCMC." *Handbook of Markov Chain Monte Carlo*, 4(10.1201). MR2858446. 266

Rossell, D. (2022). "Concentration of posterior model probabilities and normalized L_0 criteria." *Bayesian Analysis*, 17(2): 565–591. MR4483231. doi: https://doi.org/10.1214/21-ba1262. 264, 265, 266

Schäfer, C. and Chopin, N. (2013). "Sequential Monte Carlo on large binary sampling spaces." *Statistics and Computing*, 23(2): 1–22. MR3016936. doi: https://doi.org/10.1007/s11222-011-9299-z. 270, 284

Schwarz, G. (1978). "Estimating the dimension of a model." *The Annals of Statistics*, 6(2): 461–464. MR0468014. 262, 265

Scott, J. G. and Berger, J. O. (2010). "Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem." *The Annals of Statistics*, 38(5): 2587–2619. MR2722450. doi: https://doi.org/10.1214/10-AOS792. 265

Song, Q. and Liang, F. (2015). "A split-and-merge Bayesian variable selection approach for ultrahigh dimensional regression." *Journal of the Royal Statistical Society, Ser. B*, 77(5): 947–972. MR3414135. doi: https://doi.org/10.1111/rssb.12095. 282

South, L., Pettitt, A., and Drovandi, C. (2019). "Sequential Monte Carlo samplers with independent Markov chain Monte Carlo proposals." *Bayesian Analysis*, 14(3): 753–776. MR3960770. doi: https://doi.org/10.1214/18-BA1129. 271

Staerk, C. (2018). "Adaptive subspace methods for high-dimensional variable selection." Ph.D. thesis, RWTH Aachen University. URL http://doi.org/10.18154/RWTH-2018-226562 262, 268, 271

Staerk, C., Kateri, M., and Ntzoufras, I. (2021). "High-dimensional variable selection via low-dimensional adaptive learning." *Electronic Journal of Statistics*, 15(1): 830–879. MR4203346. doi: https://doi.org/10.1214/21-EJS1797. 262, 268, 271

Staerk, C., Kateri, M., and Ntzoufras, I. (2022). "Supplementary Material for "A Metropolized adaptive subspace algorithm for high-dimensional Bayesian variable selection"." *Bayesian Analysis*. doi: https://doi.org/10.1214/22-BA1351SUPP. 263

Tibshirani, R. (1996). "Regression shrinkage and selection via the lasso." *Journal of the Royal Statistical Society, Ser. B*, 58(1): 267–288. MR1379242. 261

Wan, K. Y. Y. and Griffin, J. E. (2021). "An adaptive MCMC method for Bayesian variable selection in logistic and accelerated failure time regression models." *Statistics and Computing*, 31(1): 1–11. MR4199466. doi: https://doi.org/10.1007/s11222-020-09974-2. 262, 270, 278, 286

Wasserman, L. and Roeder, K. (2009). "High dimensional variable selection." *The Annals of Statistics*, 37(5A): 2178–2201. MR2543689. doi: https://doi.org/10.1214/08-AOS646. 261

Yang, Y., Wainwright, M. J., and Jordan, M. I. (2016). "On the computational complexity of high-dimensional Bayesian variable selection." *The Annals of Statistics*, 44(6): 2497–2532. MR3576552. doi: https://doi.org/10.1214/15-AOS1417. 278

Zanella, G. and Cabezas Gonzalez, A. (2020). *scaleBVS: weighted Tempered Gibbs Sampling for Bayesian Variable Selection*. R package version 1.0. 277

Zanella, G. and Roberts, G. (2019). "Scalable importance tempering and Bayesian variable selection." *Journal of the Royal Statistical Society, Ser. B*, 81(3): 489–517.

MR3961496. doi: https://doi.org/10.1111/rssb.12316. 265, 270, 277, 278, 284, 286

Zellner, A. (1986). "On assessing prior distributions and Bayesian regression analysis with g-prior distributions." *Bayesian inference and decision techniques: Essays in Honor of Bruno De Finetti*, 6: 233–243. MR0881437. 264

Acknowledgments

We would like to thank the editor, the associate editor and two reviewers for their valuable comments and suggestions.