In Search of Lost (Mixing) Time: Adaptive Markov chain Monte Carlo schemes for Bayesian variable selection with very large p

J. E. Griffin, K. Łatuszyński and M. F. J. Steel*

January 7, 2018

Abstract

The availability of data sets with large numbers of variables is rapidly increasing. The effective application of Bayesian variable selection methods for regression with these data sets has proved difficult since available Markov chain Monte Carlo methods do not perform well in typical problem sizes of interest. The current paper proposes new adaptive Markov chain Monte Carlo algorithms to address this shortcoming. The adaptive design of these algorithms exploits the observation that in large p small n settings, the majority of the p variables will be approximately uncorrelated a posteriori. The algorithms adaptively build suitable non-local proposals that result in moves with squared jumping distance significantly larger than standard methods. Their performance is studied empirically in high-dimensional problems (with both simulated and actual data) and speedups of up to 4 orders of magnitude are observed. The proposed algorithms are easily implementable on multi-core architectures and are well suited for parallel tempering or sequential Monte Carlo implementations.

Keywords: variable selection; spike-and-slab priors; high-dimensional data; large p, small n problems; linear regression: expected squared jumping distance; optimal scaling

*Jim Griffin, School of Mathematics, Statistics and Actuarial Science, University of Kent, Canterbury, CT2 7NF, U.K. (Email: J.E.Griffin-28@kent.ac.uk), Krys Łatuszyński, Department of Statistics, University of Warwick, Coventry, CV4 7AL, U.K. (Email: K.G.Latuszynski@warwick.ac.uk) and Mark Steel, Department of Statistics, University of Warwick, Coventry, CV4 7AL, U.K. (Email: m.steel@warwick.ac.uk).

1 Introduction

The availability of large data sets has led to an increasing interest in variable selection methods applied to regression models with many potential variables but few observations (*large* p, *small* n problems). Frequentist approaches have mainly concentrated on providing point estimates under assumptions of sparsity using penalized maximum likelihood procedures (Hastie et al., 2015). However, some recent work has considered constructing confidence intervals and taking into account model uncertainty (Shah and Samworth, 2013; Dezeure et al., 2015). Bayesian approaches to variable selection are an attractive and natural alternative and lead to a posterior distribution on all possible models which can be used to address model uncertainty for variable selection and prediction. A growing literature provides a theoretical basis for good properties of the posterior distribution in large p problems (see *e.g.* Castillo et al., 2015; Johnson and Rossell, 2012).

The posterior probabilities of all models can usually only be calculated or approximated if p is smaller than 30. If p is larger, Markov chain Monte Carlo methods are typically used to sample from the posterior distribution (George and McCulloch, 1997; Dellaportas et al., 2002; O'Hara and Sillanpää, 2009; Bottolo and Richardson, 2010; Clyde et al., 2011). García-Donato and Martínez-Beneito (2013) provide an interesting discussion of the benefits of a Markov chain Monte Carlo approach. The most widely used Markov chain Monte Carlo algorithm in this context is the Metropolis-Hastings sampler where new models are proposed using Add-Remove-Swap samplers (Brown et al., 1998). For example, this method is used by Nikooienejad et al. (2016) in a binary regression model with a non-local prior for the regression coefficients on a data set with 7129 genes. Some supporting theoretical understanding of convergence is available for the Add-Remove-Swap samplers, e.g. conditions for rapid mixing in linear regression model have been derived by Yang et al. (2016). However, the mixing of these methods is often very poor for practical run times, when applied to data sets with thousands of potential covariates. As an alternative, several authors have considered using more general shrinkage priors and developed MCMC algorithms that can work in high-dimensional problems (see *e.g.* Johndrow and Orenstein, 2017; Bhattacharya et al., 2016).

The challenge of performing Markov chain Monte Carlo for Bayesian variable selection in high dimensions has lead to several developments sacrificing exact posterior exploration. For example, van den Boom et al. (2015) consider using compressed sensing to define an approximation to the posterior distribution on model space. Liang et al. (2013) used the stochastic approximation Monte Carlo algorithm (Liang et al., 2007) to develop algorithms which can efficiently explore model space. In another direction, variable selection can be performed as a post-processing step after fitting a model including all variables (see *e.g.* Bondell and Reich, 2012; Hahn and Carvalho, 2015). Several authors develop algorithms

that focus on high probability a posteriori models. In particular Rockova and George (2014) propose a deterministic expectation-maximisation based algorithm for identifying posterior modes, while Papaspiliopoulos and Rossell (2017) develop an exact deterministic algorithm that returns the most probable model of any given size in the important special case of block-diagonal design models.

Alternatively, Markov chain Monte Carlo methods for variable selection can be tailored to the data to allow faster convergence and mixing using the adaptive Markov chain Monte Carlo framework (see *e.g.* Green et al., 2015, Section 2.4, and references therein). Several strategies have been developed in literature for both the Metropolis-type algorithms (Lamnisos et al., 2013; Ji and Schmidler, 2013) and Gibbs samplers (Nott and Kohn, 2005; Richardson et al., 2010). Our proposal is a Metropolis-Hastings kernel that learns the relative importance of the variables, unlike Ji and Schmidler (2013) who use an independent proposal, and unlike Lamnisos et al. (2013) who only tune the number of variables proposed to be changed. This leads to substantially more efficient algorithms than commonly-used methods in high-dimensional settings and for which the computational cost of one step scales linearly with p. The design of these algorithms utilize the observation that in large p, small n settings posterior correlations will be negligible for the vast majority of the p inclusion indicators. The algorithms adaptively build suitable non-local Metropolis-Hastings type proposals that result in moves with expected squared jumping distance (Gelman et al., 1996) significantly larger than standard methods. In idealized examples the limiting versions of our adaptive algorithms converge in $\mathcal{O}(1)$ and result in super-efficient sampling. They outperform independent sampling in terms of the expected squared jump distance and also in the sense of the central limit theorem asymptotic variance. This is in contrast to the behaviour of optimal local random walk Metropolis algorithms that on analogous idealized targets need at least $\mathcal{O}(p)$ samples to converge (Roberts et al., 1997). The performance of our algorithms is studied empirically in realistic high-dimension problems for both synthetic and real data. In particular, in Section 4.1, for a well studied synthetic data example, speedups of up to 4 orders of magnitude are observed compared to standard algorithms. Moreover, in Section 4.2, we show the efficiency of the method in the presence of multicollinearity on a real data example with $p = 100$ variables, and in Section 4.3, we present a real data gene expression example with $p = 22\ 576$, for which a fully Bayesian analysis has never been presented before, and reliably estimate the posterior inclusion probabilities for all variables.

2 Design of the Adaptive Samplers

2.1 The Setting and the Transition Kernel

Our approach is applicable to general regression settings but we now focus the discussion on the family of normal linear regression models. This will allow for clean efficiency comparisons independent of model specific sampling details, such as e.g. the details of a reversible jump implementation. Hence, consider the model specification

$$y = \alpha \mathbf{1} + X_\gamma \beta_\gamma + \epsilon, \qquad \epsilon \sim \mathrm{N}(0, \sigma^2 I_n) \tag{1}$$

where y is an $(n \times 1)$-dimensional vector of responses, $X = (x_1, \ldots, x_p)$ is an $(n \times p)$-dimensional data matrix and $\gamma = (\gamma_1, \ldots, \gamma_p) \in \Gamma = \{0, 1\}^p$ is a vector of indicator variables in which γ_i denotes whether the i-th variable is included in the model (when $\gamma_i = 1$). The matrix X_γ is formed by those columns of X corresponding to the included variables. Bayesian variable selection involves placing a prior on the parameters of the regression model in (1), $(\alpha, \beta_\gamma, \sigma^2)$, as well as on the model γ. For clarity of exposition and validity of comparisons we will assume the commonly used prior structure

$$p(\alpha, \sigma^2, \beta_\gamma, \gamma) \propto \sigma^{-2} p(\beta_\gamma \mid \sigma^2, \gamma) p(\gamma) \tag{2}$$

with $\beta_\gamma \mid \sigma^2, \gamma \sim \mathrm{N}(0, \sigma^2 V_\gamma)$, and $p(\gamma) = h^{p_\gamma}(1-h)^{p-p_\gamma}$ where $p_\gamma = \sum_{j=1}^p \gamma_j$. The hyperparameter $0 < h < 1$ is the prior probability that a particular variable is included in the model and V_γ is often chosen as proportional to $(X_\gamma^T X_\gamma)^{-1}$ (a g-prior) or to the identity matrix (implying conditional prior independence between the regression coefficients). For both priors, the marginal likelihood $p(y \mid \gamma)$ can be calculated analytically. The prior can be further extended with hyperpriors, for example, assuming that $h \sim \mathrm{Be}(a, b)$.

Our choice of the transition kernel to sample from the posterior distribution $\pi(\gamma \mid y) \propto p(y \mid \gamma)\, p(\gamma)$ is motivated by diffusion limits and associated mixing time results in well-understood settings related to model selection, and discussed below in Section 2.2. We will consider using a non-symmetric Metropolis-Hastings kernel with proposal parametrisation suitable for optimising the expected squared jumping distance on the model space. Let the probability of proposing to move from model γ to γ' be given in a product form

$$q_\eta(\gamma, \gamma') = \prod_{j=1}^p q_{\eta,j}(\gamma_j, \gamma_j') \tag{3}$$

where $\eta = (A, D) = (A_1, \ldots, A_p, D_1, \ldots, D_p)$, $q_{\eta,j}(\gamma_j = 0, \gamma_j' = 1) = A_j$ and $q_{\eta,j}(\gamma_j = 1, \gamma_j' = 0) = D_j$. The proposed model γ' is sampled using p independent Bernoulli trials which allows fast sampling and multiple variables to be added or deleted from the model in

one iteration, in contrast to the standard Add-Remove-Swap proposal. The proposed model is accepted using the standard Metropolis-Hastings acceptance probability

$$a_\eta(\gamma, \gamma') = \min\left\{1, \frac{\pi(\gamma' \mid y) q_\eta(\gamma', \gamma)}{\pi(\gamma \mid y) q_\eta(\gamma, \gamma')}\right\}. \tag{4}$$

2.2 In Search of Lost Mixing Time: Optimising the Sampler

In adaptive Markov chain Monte Carlo algorithms, the Markovian kernels used by the sampler are tuned on the fly to improve mixing using some computationally accessible performance criterion. Optimal scaling (Roberts et al., 1997) is a commonly used and mathematically well justified criterion for the random walk Metropolis algorithm. Suppose that μ_p is a p-dimensional probability density function which has the form

$$\mu_p = \prod_{j=1}^{p} f, \tag{5}$$

for some smooth enough f. The optimal scale of a random walk proposal for μ_p should be adjusted so that the mean acceptance rate is approximately 0.234. The underlying analysis also implies that the optimised random walk Metropolis will converge to stationarity in $\mathcal{O}(p)$ steps (Roberts and Rosenthal, 2016). In practical settings, targeting the 0.234 acceptance rate turns out to be a useful guide even in moderate dimensions and well beyond the restrictive independent, identically distributed product form assumption of Roberts et al. (1997).

It is therefore very tempting to follow this commonly used optimisation strategy in case of the Bayesian variable selection problems, where the state space is $\Gamma = \{0, 1\}^p$. It would amount to designing a symmetric random walk Metropolis kernel on the model space and tuning it to achieve the 0.234 acceptance rate in the hope that as p increases, the chain will mix in $\mathcal{O}(p)$ steps. However, there are several reasons to think the variable selection problem is a special case, and this approach might not work as expected.

Firstly, consider optimal scaling of the random walk Metropolis for the product measures

$$\mu_p(\gamma_1, \ldots, \gamma_p) = s^{\#\{\gamma_j = 1\}}(1-s)^{\#\{\gamma_j = 0\}} = s^{p_\gamma}(1-s)^{p-p_\gamma}$$

analysed by Roberts (1998). The results support the 0.234 optimal acceptance rate and $\mathcal{O}(p)$ mixing with fixed s close to $1/2$ as p tends to infinity. However, the picture is more complex if s tends to zero as p tends to infinity, which is relevant to most variable selection problems. The numerical results of Section 3 in Roberts (1998) indicate that in such a setting the optimally tuned random walk Metropolis proposes to change two γ_j's at a time but the acceptance rate deteriorates to zero resulting in the chain not moving. This suggests the actual mixing in this regime is slower than the $\mathcal{O}(p)$ observed for smooth continuous densities.

Secondly, Neal et al. (2012) established, under additional regularity conditions, that the optimally tuned random walk Metropolis for target measures of the form (5) with discontinuous target densities f mixes in $\mathcal{O}(p^2)$ steps, an order of magnitude slower than with smooth target densities f, and that the optimal acceptance rate is $e^{-2} \approx 0.1353$. Motivated by this result, and by practical difficulties of making the random walk Metropolis efficient for discontinuous densities in epidemic examples, Neal and Lee (2017) consider optimal scaling of the independence sampler. Rather surprisingly, the optimally tuned independence sampler recovers the $O(p)$ mixing and acceptance rate of 0.234 without any additional smoothness conditions,

Informed by these findings, we consider target densities on the vertices of the hypercube $\{0,1\}^p$ that have product form, but are not identically distributed,

$$\pi_p(\gamma) = \prod_{j=1}^{p} \pi_j^{\gamma_j} (1-\pi_j)^{1-\gamma_j} \tag{6}$$

where $0 < \pi_j < 1$ for $j = 1, \ldots, p$. In the algorithmic design, we shall exploit the simplicity of product measures on $\Gamma = \{0,1\}^p$ compared to product measures on more general spaces. Statistically, and in algorithmic performance, we shall benefit from the observation that in typical *large p, small n* variable selection problems, a posteriori correlations between γ_j's will be negligible for $j \in \mathcal{I}$, while strong correlations and multimodality will occur only for $j \in \mathcal{C}$, with $\mathcal{I} \cup \mathcal{C} = \{1, \ldots, p\}$ and $|\mathcal{I}| \gg |\mathcal{C}|$. Therefore, while formally optimising the performance of the sampler for a distribution given by (6), informally we can think of the target distribution as

$$\pi_p(\gamma) \approx \nu(\gamma_{\mathcal{C}}) \times \prod_{j \in \mathcal{I}} \pi_j^{\gamma_j} (1-\pi_j)^{1-\gamma_j}, \tag{7}$$

where $\nu(\gamma_{\mathcal{C}})$ encodes the joint distribution of the γ_j's for the small subset $\mathcal{C}$.

Consider now the non-symmetric Metropolis-Hastings algorithm with the product form proposal $q_\eta(\gamma, \gamma')$ given by (3) targeting the posterior distribution given by (6). The acceptance probability (4) becomes

$$a_\eta(\gamma, \gamma') = \min \left\{ 1, \prod_{j:\gamma_j=0, \gamma'_j=1} \frac{D_j}{A_j} \frac{\pi_j}{1-\pi_j} \prod_{j:\gamma_j=1, \gamma'_j=0} \frac{A_j}{D_j} \frac{1-\pi_j}{\pi_j} \right\}. \tag{8}$$

Note that $\alpha_\eta(\cdot, \cdot) \equiv 1$ for any choice of $\eta = (A, D)$ satisfying

$$\frac{A_j}{D_j} = \frac{\pi_j}{1-\pi_j}, \quad \text{for every } j. \tag{9}$$

To discuss optimal choices of η, we consider the several commonly used criteria for Markov chains with stationary distribution π and transition kernel P on a finite discrete state space Γ.

- *Mixing time* (Levin et al., 2009; Roberts and Rosenthal, 2004). The mixing time of a Markov chain is defined as $\rho := \min\{t : \max_{\gamma\in\Gamma} \|P^t(\gamma,\cdot) - \pi(\cdot)\|_{TV} < 1/2\}$ where $\|\mu_1(\cdot) - \mu_2(\cdot)\|_{TV} := \sup_{G\subset\Gamma} |\mu_1(G) - \mu_2(G)|$.
- *Expected squared jumping distance* (Gelman et al., 1996). Equip Γ with a metric $d(\cdot,\cdot)$. Let $\gamma^{(0)}$ and $\gamma^{(1)}$ be two consecutive values in a Markov chain trajectory and define the squared jumping distance as $\Delta^2 := d^2(\gamma^{(0)}, \gamma^{(1)})$. If $\Gamma = \{0,1\}^p$, the natural choice of metric is $\Delta^2 = \sum_{j=1}^p |\gamma_j^{(0)} - \gamma_j^{(1)}|^2$. The expected squared jumping distance is defined as $E_\pi[\Delta^2]$, where $\gamma^{(0)} \sim \pi$. In this setting, the expected squared jumping distance is equivalent to the average number of variables changed in one iteration.
- *Peskun ordering* (Peskun, 1973). Suppose that the Markov chain is ergodic, then, for any function $f : \Gamma \to \mathbb{R}$, $\frac{1}{\sqrt{n}} \sum_{k=0}^{n-1} f(\gamma^{(k)}) \xrightarrow{D} N(E_\pi f, \sigma^2_{P,f})$, where the constant $\sigma^2_{P,f}$ depends on the transition kernel P and function f. Consider transition kernels P_1 and P_2. If $\sigma^2_{P_1,f} \leq \sigma^2_{P_2,f}$ for every f, then P_1 dominates P_2 in Peskun ordering. If P_1 dominates all other kernels from a given class, then P_1 is optimal in this class with respect to Peskun ordering. Apart from toy examples, Peskun ordering can be rarely established without further restrictions. Hence, for the model selection problem, where inclusion probabilities are often of interest, we consider Peskun ordering for the class $\mathbb{L}(\Gamma)$ of linear combinations of univariate functions:

$$\mathbb{L}(\Gamma) := \{f : \Gamma \to \mathbb{R} : f(\gamma) = a_0 + \sum_{j=1}^p a_j f_j(\gamma_j)\}. \tag{10}$$

Proposition 1. *Consider the class of Metropolis-Hastings algorithms with target distribution given by* (6) *and proposal* $q_\eta(\gamma,\gamma')$ *given by* (3) *and satisfying* (9)*. Then*

(i) setting $A_j = 1 - D_j = \pi_j$*, results in*

(a) independent sampling and optimal mixing time $\rho = 1$;

(b) the following asymptotic variance:

$$\sigma^2_{P,f} = Var_\pi f \qquad \textit{for arbitrary } f; \tag{11}$$

$$\sigma^2_{P,f} = Var_\pi f = \sum_{j=1}^p a_j^2 Var_{\pi^{(j)}} f_j \qquad \textit{for } f \in \mathbb{L}(\Gamma), \tag{12}$$

where $Var_\pi f$ *is the stationary variance of* f *under* $\pi_p(\gamma)$ *in* (6) *and* $Var_{\pi^{(j)}} f_j$ *is the stationary variance of* f_j *under the marginal* $\pi^{(j)} := \{1-\pi_j, \pi_j\}$ *on* $\{0,1\}$;

(c) the following expected squared jumping distance (and equivalently the average number of variables being changed in one iteration):

$$E_\pi[\Delta^2] = 2\sum_{j=1}^p \pi_j(1-\pi_j); \tag{13}$$

(ii) setting $A_j = \min\{1, \frac{\pi_j}{1-\pi_j}\}$ *and* $D_j = \min\{1, \frac{1-\pi_j}{\pi_j}\}$*, results in*

(a) maximal expected squared jumping distance (and equivalently maximal average number of variables being changed in one iteration), which becomes

$$E_\pi[\Delta^2] = 2\sum_{j=1}^{p} \min\{1-\pi_j, \pi_j\}; \tag{14}$$

(b) optimality with respect to the Peskun ordering for the class of linear functions $\mathbb{L}(\Gamma)$ *defined in* (10)*; the asymptotic variance becomes*

$$\sigma^2_{P,f} = \sum_{j=1}^{p} \big(2\max\{1-\pi_j, \pi_j\} - 1\big) a_j^2 Var_{\pi^{(j)}} f_j \qquad \textit{for } f \in \mathbb{L}(\Gamma). \tag{15}$$

The proof is given in Appendix A.

Remark 1. The advantage in terms of expected squared jumping distance and central limit theorem asymptotic variance of the setting (ii) over independent sampling (i), is particularly substantial for important variables for which π_j is close to $1/2$.

Remark 2. In discrete spaces, moves which do not change the model can be proposed. Such moves have acceptance probability one but do not help mixing. Therefore, the average acceptance rate is inappropriate and Schäfer and Chopin (2013) suggest using the mutation rate:

$$\bar{a}_M = \int \mathrm{I}(\gamma \neq \gamma') a_\eta(\gamma, \gamma') q_\eta(\gamma, \gamma') \pi(\gamma) d\gamma' \, d\gamma.$$

Under setting (i), the mutation rate is

$$\bar{a}_M = 1 - \prod_{j=1}^{p} q_{\eta,j}(\gamma_j, \gamma_j) = 1 - \prod_{j=1}^{p} \left[(1-\pi_j)^2 + \pi_j^2\right]$$

and, under setting (ii), the mutation rate is

$$\bar{a}_M = 1 - \prod_{j=1}^{p} q_{\eta,j}(\gamma_j, \gamma_j) = 1 - \prod_{j=1}^{p} |2\pi_j - 1|.$$

Therefore, setting (ii) always leads to a higher mutation rate.

This suggests that the proposal in setting (ii) should be preferred over that in (i) when designing a Metropolis-Hastings sampler for idealized posteriors of the form in (6). In practice, this choice may be too greedy since the correlated set of variables $\mathcal{C}$ contributing $\nu(\gamma_\mathcal{C})$ in (7) may considerably decrease acceptance rate for such proposals, and a scaled proposal of the form

$$A_j = \zeta_j \min\left\{1, \frac{\pi_j}{1-\pi_j}\right\}, \qquad D_j = \zeta_j \min\left\{1, \frac{1-\pi_j}{\pi_j}\right\}, \tag{16}$$

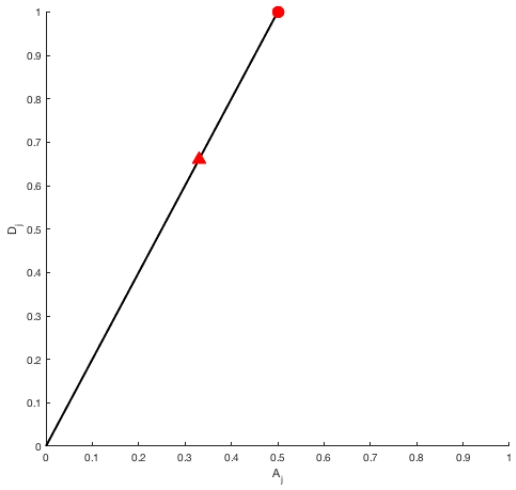


Figure 1: The solid black segment presents A_j's and D_j's corresponding to different choices of $\zeta_j \in [0, 1]$ in (16). Any point in the segment results in acceptance probability $= 1$ for the idealized target (6). The iid sampling (i), marked with a triangle, is a shrunk version of the superefficient sampling (ii), marked with a bullet.

may be preferred. Note that the independent sampling (i) corresponds to a particular choice of ζ_j's above, and the family of these proposals for all $\zeta_j \in [0, 1]$ is illustrated as the solid black segment in Figure 1.

In the next section, we devise adaptive MCMC algorithms that aim to tune proposals of the form (3) so that A_j's and D_j's lie approximately on the solid black segment in Figure 1. Their placement along the segment will be guided by how much the correlated set of variables $\mathcal{C}$ contributes to (7), and balance between attaining high acceptance rates and proposing large jumps.

2.3 Remembrance of Things Past: Adaptive Markov chain Monte Carlo Design

Adaptive Markov chain Monte Carlo aims to use past samples of the Markov chain to optimise the chosen performance criterion and consequently the transition kernel on the fly, as simulation progresses (see *e.g.* Andrieu and Thoms, 2008; Roberts and Rosenthal, 2009; Green et al., 2015). As discussed in the introduction, developing effective Markov chain Monte Carlo methods for sampling from the posterior model distribution has proved challenging for large p. Adaptive methods offer the potential to improve mixing on model spaces and the Monte Carlo estimation of posterior inclusion probabilities.

Based on Proposition 1, we use the adaptive Markov chain Monte Carlo approach to optimise the non-symmetric Metropolis-Hastings type algorithm with product form proposal $q_\eta(\gamma, \gamma')$ given by (3). Let $\gamma^{(i)}$ be the value of γ at the start of the i-th iteration, γ' be the subsequently proposed value and $\eta^{(i)} = (A^{(i)}, D^{(i)})$ be the value of the tuning parameters used at the i-th iteration. We define for $j = 1, \ldots, p$,

$$\gamma_j^{A\,(i)} = \begin{cases} 1 \text{ if } \gamma'_j \neq \gamma_j^{(i)} \text{ and } \gamma_j^{(i)} = 0 \\ 0 \text{ otherwise} \end{cases}, \qquad \gamma_j^{D\,(i)} = \begin{cases} 1 \text{ if } \gamma'_j \neq \gamma_j^{(i)} \text{ and } \gamma_j^{(i)} = 1 \\ 0 \text{ otherwise} \end{cases}$$

and the map $\text{logit}_\epsilon : (\epsilon, 1-\epsilon) \to \mathbb{R}$ by

$$\text{logit}_\epsilon(x) = \log(x - \epsilon) - \log(1 - x - \epsilon),$$

where $0 \leq \epsilon \leq 1/2$. This is the usual logit transform if $\epsilon = 0$.

We define two strategies for adapting η. The first adaptive strategy is a general purpose method that we term *Exploratory Individual Adaptation* (IA). It aims to find pairs (A_j, D_j) on the segment illustrated in Figure 1, balancing between large values of (A_j, D_j), and their possible negative effect on average acceptance probability. There are three types of updates for $A^{(i)}$ and $D^{(i)}$ which move towards the correct ratio A_j/D_j and then along the segment (note that the slope of the segment is not known in practice, as it depends on π_j):

1. *Expansion step*, which aims to increase the average jumping distance while maintaining the same average acceptance rate. If the acceptance rate for the proposed model $a_{\eta^{(i)}}(\gamma^{(i)}, \gamma')$ is above threshold τ_U, coordinates of $A^{(i)}$ and $D^{(i)}$ indicated by $\gamma^{A(i)}$ and $\gamma^{D(i)}$, i.e corresponding to all the variables proposed to change, increase approximately on the log scale, so that $A_j^{(i+1)}/D_j^{(i+1)} \approx A_j^{(i)}/D_j^{(i)}$.

2. *Shrinkage step*, which aims to increase the acceptance rate by proposing less ambitious moves. If $a_{\eta^{(i)}}(\gamma^{(i)}, \gamma')$ is below a critical threshold τ_L, the proposal may be too ambitious and the strategy is to decrease coordinates of $A^{(i)}$ that correspond to $\gamma^{A(i)}$ and decrease the coordinates of $D^{(i)}$ that correspond to $\gamma^{D(i)}$.

3. *Correction step*, which aims to increase the acceptance rate by correcting the ratio between A's and D's. If $a_{\eta^{(i)}}(\gamma^{(i)}, \gamma')$ is below τ_U, but above the critical threshold τ_L, the strategy is to improve acceptance rate (c.f. (8)) by increasing coordinates of $A^{(i)}$ that correspond to $\gamma^{D(i)}$ and decreasing those corresponding to $\gamma^{A(i)}$, and analogously, by increasing the coordinates of $D^{(i)}$ that correspond to $\gamma^{A(i)}$ and decreasing those corresponding to $\gamma^{D(i)}$.

The gradient fields of these updates are shown in Figure 2. These three moves can be

combined into the following adaptation of $A^{(i)}$ and $D^{(i)}$:

$$\begin{aligned}\text{logit}_\epsilon\left(A_j^{(i+1)}\right) &= \text{logit}_\epsilon\left(A_j^{(i)}\right) + \phi_i \times \Bigg(\gamma_j^{A(i)}\mathbb{I}_{\left\{a_{\eta^{(i)}}(\gamma^{(i)},\gamma') \geq \tau_U\right\}} \\ &\quad + \gamma_j^{D(i)}\mathbb{I}_{\left\{a_{\eta^{(i)}}(\gamma^{(i)},\gamma') \geq \tau_L\right\}} - \gamma_j^{A(i)}\mathbb{I}_{\left\{a_{\eta^{(i)}}(\gamma^{(i)},\gamma') \leq \tau_U\right\}}\Bigg),\end{aligned} \tag{17}$$

and

$$\begin{aligned}\text{logit}_\epsilon\left(D_j^{(i+1)}\right) &= \text{logit}_\epsilon\left(D_j^{(i)}\right) + \phi_i \times \Bigg(\gamma_j^{D(i)}\mathbb{I}_{\left\{a_{\eta^{(i)}}(\gamma^{(i)},\gamma') \geq \tau_U\right\}} \\ &\quad + \gamma_j^{A(i)}\mathbb{I}_{\left\{a_{\eta^{(i)}}(\gamma^{(i)},\gamma') \geq \tau_L\right\}} - \gamma_j^{D(i)}\mathbb{I}_{\left\{a_{\eta^{(i)}}(\gamma^{(i)},\gamma') \leq \tau_U\right\}}\Bigg),\end{aligned} \tag{18}$$

for $j = 1\ldots, p$ where $\phi_i = O(i^{-\lambda})$ for some constant $1/2 < \lambda \leq 1$ and $a_{\eta^{(i)}}\left(\gamma^{(i)}, \gamma'\right)$ represents the acceptance probability at the i-th iteration. The transformation implies that $\epsilon < A_j^{(i)} < 1 - \epsilon$ and $\epsilon < D_j^{(i)} < 1 - \epsilon$ and we assume that $0 < \epsilon < 1/2$. Based on several simulation studies, we suggest to take $\tau_U = 0.1$ and $\tau_L = 0.01$. As discussed in Section 2.2 targeting a low acceptance rate is often beneficial in irregular cases, so we expect this choice to be robust in real data applications. In all our simulations with this parameter setting, the resulting mean acceptance rate was between 0.15 and 0.35, i.e. in the high efficiency region identified in Roberts et al. (1997). We also suggest the initial choice of parameters such that $A_j^{(1)}/D_j^{(1)} \approx h/(1-h)$ as this summarises the prior information on $\pi_j/(1-\pi_j)$, and in particular $D_j^{(1)} \equiv 1$ and $A_j^{(1)} \equiv h$ often works well. The steps of the exploratory individual adaptation algorithm are shown in Algorithm 1.

for $i = 1$ to $i = M$
  sample $\gamma' \sim q_{\eta^{(i)}}(\gamma^{(i)}, \cdot)$ and $U \sim U(0,1)$;
  if $U < a_{\eta^{(i)}}(\gamma^{(i)}, \gamma')$ then
    $\gamma^{(i+1)} := \gamma'$
  else
    $\gamma^{(i+1)} := \gamma^{(i)}$
  endif
  update $A^{(i+1)}$ using (17) and $D^{(i+1)}$ using (18)
endfor

Algorithm 1: Exploratory Individual Adaptation (IA)

Algorithm 1 learns two parameters $A_j^{(i)}$ and $D_j^{(i)}$ for each variable which can be time-consuming if p is large. Alternatively, we could learn $\pi_1, \ldots, \pi_p$ to approximate the slope of the segment in Figure 1, and use the proposal (16) motivated by part (ii) of Proposition 1 and

the fact that the posterior distribution will not generally have a product form. However, to accelerate adaptation, we shall use the same scale parameter for all variables. We term this approach the *Adaptively Scaled Individual Adaptation* (ASI) proposal. In particular, we use

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

and

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

for $j = 1, \ldots, p$ where $0 < \zeta^{(i)} < 1$ is a tuning parameter and $\pi_j^{(i)}$ is a Rao-Blackwellised estimate of the posterior inclusion probability of variable j at the start of the i-th iteration. The value of $\zeta^{(i)}$ is updated using

$$\text{logit}_\epsilon\left(\zeta^{(i+1)}\right) = \text{logit}_\epsilon\left(\zeta^{(i)}\right) + \phi_i(a_{\eta^{(i)}}(\gamma^{(i)}, \gamma') - \tau), \tag{19}$$

where τ is a targeted acceptance rate. To avoid proposing to change no variable with high probability, we set $\zeta^{(i+1)} = 1/\Delta^{(i+1)}$ if $\zeta^{(i+1)}\Delta^{(i+1)} < 1$ where $\Delta^{(i+1)} = 2\sum_{j=1}^{p} \min\left\{\pi_j^{(i+1)}, 1 - \pi_j^{(i+1)}\right\}$. This ensures that the algorithm will propose to change at least one variable with high probability. The idea of using Rao-Blackwellised estimates of posterior inclusion probabilities has been considered before. We follow Ghosh and Clyde (2011) who work with the Rao-Blackwellised estimate conditional on the model, whilst integrating over the regression coefficients, rather than Guan and Stephens (2011) who condition on the regression coefficients for the included variables. This can be achieved using an $O(p)$ algorithm and the formulae are provided in Appendix B. The adaptively scaled individual adaptation algorithm is described in Algorithm 2.

Craiu et al. (2009) showed empirically that running multiple independent Markov chains with the same adaptive parameters improves the rate of convergence of adaptive algorithms towards their target acceptance rate in the context of the classical adaptive Metropolis algorithm of Haario et al. (2001) (see also Bornn et al. 2013) . Therefore, we will compare algorithms with different numbers of independent parallel chains and refer to this as *multiple chain acceleration*.

3 Ergodicity of the Algorithms

Since adaptive Markov chain Monte Carlo algorithms violate the Markov condition, the standard and well developed Markov chain theory can not be used to establish ergodicity and we need to derive appropriate results for our algorithms. We verify validity of our algorithms by establishing conditions introduced in Roberts and Rosenthal (2007), namely simultaneous uniform ergodicity and diminishing adaptation.

for $i = 1$ to $i = M$

sample $\gamma' \sim q_{\eta^{(i)}}(\gamma^{(i)}, \cdot)$ and $U \sim U(0,1)$;

if $U < a_{\eta^{(i)}}(\gamma^{(i)}, \gamma')$ then

$\gamma^{(i+1)} := \gamma'$

else

$\gamma^{(i+1)} := \gamma^{(i)}$

endif

Update $\pi_1^{(i+1)}, \ldots, \pi_p^{(i+1)}$ as in Appendix B and set $\tilde{\pi}_j^{(i+1)} = \epsilon + (1 - 2\epsilon)\, \pi_j^{(i+1)}$

Update $\zeta^{(i+1)}$ as in (19)

Calculate $A_j^{(i+1)} = \zeta^{(i+1)} \min\left\{1, \tilde{\pi}_j^{(i+1)} / \left(1 - \tilde{\pi}_j^{(i+1)}\right)\right\}$ for $j = 1, \ldots, p$

Calculate $D_j^{(i+1)} = \zeta^{(i+1)} \min\left\{1, \left(1 - \tilde{\pi}_j^{(i+1)}\right) / \tilde{\pi}_j^{(i+1)}\right\}$ for $j = 1, \ldots, p$

endfor

Algorithm 2: Adaptively Scaled Individual Adaptation (ASI)

Recall that $\pi(\gamma \mid y) \propto p(y|\gamma)p(\gamma)$ is the target posterior on the model space Γ and the vector of adaptive parameters at time i is

$$\eta^{(i)} = (A^{(i)}, D^{(i)}) \quad \in \quad [\varepsilon, 1-\varepsilon]^{2p} =: \Delta_\varepsilon \tag{20}$$

with specific update strategies being defined either by exploratory individual adaptation or adaptively scaled individual adaptation. By $P_\eta(\gamma, \cdot)$ denote the non-adaptive Markov chain kernel corresponding to the fixed choice of η. Under the dynamics of either algorithm, for $S \subseteq \Gamma$ we have

$$\begin{aligned} P_\eta(\gamma, S) &= \mathbb{P}\Big[\gamma^{(i+1)} \in S \,\Big|\, \gamma^{(i)} = \gamma, \eta^{(i)} = \eta\Big] \\ &= \sum_{\gamma' \in S} q_\eta(\gamma, \gamma') a_\eta(\gamma, \gamma') + \mathbb{I}_{\{\gamma \in S\}} \sum_{\gamma' \in \Gamma} q_\eta(\gamma, \gamma')\big(1 - a_\eta(\gamma, \gamma')\big). \end{aligned} \tag{21}$$

In the case of multiple chain acceleration, where r copies of the chain are run, the respective model state space is the product space and thus the current state of the algorithm at time i is $\gamma^{\otimes r,\,(i)} \in \Gamma^r$ and the stationary distribution is the product density $\pi^{\otimes r}$ on Γ^r. The single chain version corresponds to $r = 1$ and so all results apply to that case.

To assess ergodicity, we need to define the distribution of the adaptive algorithm at time i, and the associated total variation distance: for $S \subseteq \Gamma^r$

$$\mathcal{L}^{(i)}\big[(\gamma^{\otimes r}, \eta), S\big] := \mathbb{P}\Big[\gamma^{\otimes r,\,(i)} \in S \,\Big|\, \gamma^{\otimes r,\,(0)} = \gamma^{\otimes r}, \eta^{(0)} = \eta\Big],$$

and

$$T(\gamma^{\otimes r}, \eta, i) := \|\mathcal{L}^{(i)}\left[(\gamma^{\otimes r}, \eta), \cdot\right] - \pi^{\otimes r}(\cdot)\|_{TV}$$
$$= \sup_{S \in \Gamma^r} |\mathcal{L}^{(i)}\left[(\gamma^{\otimes r}, \eta), S\right] - \pi^{\otimes r}(S)|.$$

We show that all the considered algorithms are ergodic, and satisfy a weak law of large numbers *i.e.* for any starting point $\gamma^{\otimes r} \in \Gamma^r$ and any initial parameter value $\eta \in \Delta_\varepsilon$, we have

$$\lim_{i \to \infty} T(\gamma^{\otimes r}, \eta, i) = 0, \quad \text{and} \tag{22}$$

$$\frac{1}{k} \sum_{i=1}^{k} f(\gamma^{\otimes r,(i)}) \overset{k \to \infty}{\longrightarrow} \pi(f) \quad \text{in probability,} \quad \text{for every} \quad f : \Gamma^r \to \mathbb{R}, \tag{23}$$

$$\text{where} \quad \pi(f) = \sum_{\gamma^{\otimes r} \in \Gamma^r} f(\gamma^{\otimes r}) \pi^{\otimes r}(\gamma^{\otimes r}).$$

To this end we first establish the following result.

Lemma 1. *The family of Markov chains defined by transition kernels P_η of* (21) *is simultaneously uniformly ergodic for any $\varepsilon > 0$ in* (20)*,* i.e. *for all $\delta > 0$ there exists $N = N(\delta, \varepsilon) \in \mathbb{N}$, such that for any starting point $\gamma^{\otimes r} \in \Gamma^r$ and any parameter value $\eta \in \Delta_\varepsilon$*

$$\|P_\eta^N(\gamma^{\otimes r}, \cdot) - \pi^{\otimes r}(\cdot)\|_{TV} \leq \delta.$$

This result along with diminishing adaptation directly leads to the following

Theorem 1. *Assume that $p(y|\gamma)$ is available analytically for all $\gamma \in \Gamma$ and $\varepsilon > 0$ in* (17)*,* (18)*, or* (19)*. Then each of the algorithms (exploratory individual adaptation, adaptively scaled individual adaptation and their multiple chain acceleration versions) are ergodic and satisfy a weak law of large numbers.*

Proofs can be found in Appendix A. A comprehensive analysis of the algorithms for other generalised linear models or for linear models using non-conjugate prior distributions requires a case-by-case treatment, and is beyond the scope of this paper. However, we note that if the prior distributions of additional parameters are continuous, supported on a compact set and everywhere positive, establishing ergodicity for a specific model will typically be possible with some technical care.

4 Results

4.1 Simulated Data

We consider the simulated data example of Yang et al. (2016). They assume that there are n observations and p regressors and the data is generated from the model

$$Y = X\beta^\star + \epsilon$$

where $\epsilon \sim \mathrm{N}(0, \sigma^2 I)$ for $\sigma^2 = 1$. The first 10 regression coefficients are non-zero and we use

$$\beta^\star = \mathrm{SNR}\sqrt{\frac{\sigma^2 \log p}{n}}(2, -3, 2, 2, -3, 3, -2, 3, -2, 3, 0, \ldots, 0)^T \in \mathbb{R}^p.$$

The i-th vector of regressors is generated as $x_i \sim \mathrm{N}(0, \Sigma)$ where $\Sigma_{jk} = \rho^{|j-k|}$. In our examples, we use the value $\rho = 0.6$ which represents a relative large correlation between the regressors.

We are interested in the performance of the two adaptive algorithms (exploratory individual adaptation and adaptively scaled individual adaptation) relative to a simple Add-Delete-Swap algorithm. We define the ratio of the relative time-standardized effective sample size of algorithm A versus algorithm B to be

$$r_{A,B} = \frac{\mathrm{ESS}_A/t_A}{\mathrm{ESS}_B/t_B}$$

where ESS_A is the effective sample size for algorithm A. This is estimated by

$$\hat{r}_{A,B} = \frac{s_B^2 t_B}{s_A^2 t_A}$$

where t_A and t_B are the times taken for one run of algorithm A and algorithm B respectively, and s_A^2 and s_B^2 are the sample variances of the posterior inclusion probabilities for algorithm A and algorithm B respectively over different runs of the algorithms.

The posterior distribution changes substantially with the SNR and the size of the data set. All ten true non-zero coefficients are given posterior inclusion probabilities greater than 0.9 in the two high SNR scenarios (SNR=2 and SNR=3) for each value of n and p and no true non-zero coefficients are given posterior inclusion probabilities greater than 0.2 in the low SNR scenario (SNR=0.5) for each value of n and p. In the intermediate SNR scenario (SNR=1), the number of true non-zero coefficients given posterior inclusion probabilities greater than 0.9 are 4 ($n = 500, p = 500$), 8 ($n = 1000, p = 500$), 3 ($n = 500, p = 5000$) and 0 ($n = 1000, p = 5000$). Important variables are hard to identify with low SNR when the posterior is very dispersed. As SNR increases, the important variables becomer easier to identify and can be fairly well identified when SNR=3. Generally, the results are consistent

with our intuition that true non-zero regression coefficients should be detected with greater posterior probability for larger SNR, larger value of n and smaller value of p.

Table 1 shows the median relative time-standardized effective sample sizes for the exploratory individual adaptation and adaptively scaled individual adaptation algorithms with 5 or 25 multiple chains for different combinations of n, p and SNR. The median is taken across the estimated relative time-standardized effective sample sizes for all posterior inclusion probabilities. Clearly, the adaptively scaled individual adaptation algorithm outper-

Table 1: The median values of $\hat{r}_{A,B}$ for the posterior inclusion probabilities over all variables where B is the standard Metropolis-Hastings algorithm and A is either the exploratory individual adaptation (IA) or adaptively scaled individual adaptation (ASI) algorithm for the different simulated data sets

		5 chains				25 chains			
		SNR				SNR			
(n, p)		0.5	1	2	3	0.5	1	2	3
(500, 500)	IA	4.9	1.8	5.5	5.1	1.2	1.5	2.4	2.3
	ASI	1.7	21.3	31.8	7.5	2.0	36.0	42.7	12.6
(500, 5000)	IA	8.7	2.2	718.0	81.5	7.1	2.9	2267.2	147.2
	ASI	29.9	126.9	2053.1	2271.3	53.5	353.3	12319.5	7612.3
(1000, 500)	IA	5.9	16.3	7.7	4.2	1.6	80.7	4.4	1.8
	ASI	41.9	2.1	16.9	12.0	32.8	34.0	27.9	14.4
(1000, 5000)	IA	2.2	2.2	9167.2	11.3	5.6	2.5	15960.7	199.8
	ASI	15.4	37.0	4423.1	30.8	54.9	53.4	11558.2	736.4

forms the exploratory individual adaptation algorithm for most settings with either 5 or 25 multiple chains. The performance of the exploratory individual adaptation and, especially, the adaptive scaled individual adaptation algorithm with 25 chains is better than the corresponding performance with 5 chains for most cases. Concentrating on results with the adaptively scaled individual adaptation algorithm, the largest increase in performance compared to a simple Metropolis-Hastings algorithm occurs with SNR=2. In this case, there are three or four orders of magnitude improvements when $p = 5000$ and several orders of magnitude improvements for other SNR with $p = 5000$. In smaller problems with $p = 500$, there are still substantial improvements in efficiency over the simpler Metropolis-Hastings sampler.

The superior performance of the adaptively scaled individual adaptation algorithm (which has one tuneable parameter) over the exploratory individual adaptation algorithm (which has

$2p$ tuneable parameters) is due to the substantially faster convergence of the tuning parameters of the adaptively scaled individual adaptation algorithm to optimal values. Plotting posterior inclusion probabilities against A and D at the end of a run shows that, in most cases, the values of A for a variable are close to the corresponding posterior inclusion probabilities for both algorithms. However, the value of D are mostly close to 1 for adaptively scaled individual adaptation but not for exploratory individual adaptation. If D_j is close to 1, then variable j is highly likely to be proposed to be removed if already included in the model. This leads to improved mixing rates if the posterior inclusion probabilities for that variable is close to zero since it allows that variable to be included more often in a fixed run length. This is hard to learn through individual adaptation (since variables with low posterior inclusion probabilities will be rarely included in the model and so the algorithm learns the D_j slowly for those variables) whereas the Rao-Blackwellised estimates can often quickly determine which variables have low posterior inclusion probabilities.

4.2 Tecator Data

The tecator data contains 172 observations and 100 variables. They have been previously analysed using Bayesian linear regression techniques by Griffin and Brown (2010), who give a description of the data, and Lamnisos et al. (2013). The regressors show a high degree of multi-collinearity and so this is a challenging example for Bayesian variable selection algorithms. The prior used was (2) with $V_\gamma = 100I$ and $h = 5/100$. Even short simulations of the exploratory individual adaptation algorithm for this data, such as 5 multiple chains with 3000 burn in and 3000 recorded iterations afterwards, taking about 5 seconds on a laptop, show consistent convergence across runs.

Our purpose was to study the adaptive behaviour of the exploratory individual adaptation algorithm on this real data example, in particular to compare the idealized values of the A_j's and D_j's with the values attained by the algorithm.

We use multiple chain acceleration with 50 multiple chains over the total of 6000 iterations. The algorithm parameters were set to $\tau_L = 0.01$ and $\tau_U = 0.1$. The resulting mean acceptance rate was approximately 0.2 indicating close to optimal efficiency. The average number of variables proposed to be changed in a single accepted proposal was 23, approximately twice the average model size, meaning that in a typical move all of the current variables were deleted from the model, and a set of completely fresh variables was proposed.

Figure 3(a) shows how the exploratory individual adaptation algorithm approximates setting (ii) of Proposition 1, namely the super-efficient sampling from the idealized posterior (6). Figure 3(b) illustrates how the attained values of A_j's somewhat overestimate the idealized values $\min\{1, \pi_j/(1-\pi_j)\}$ of setting (ii) in Proposition 1. This indicates that the chosen parameter values $\tau_L = 0.01$ and $\tau_U = 0.1$ of the algorithm overcompensate the cor-

related part of the posterior $\nu(\gamma_{\mathcal{C}})$ in (7), which is not very pronounced for this dataset. To quantify the performance, we ran both algorithms with adaptation in the burn-in only and calculated the effective sample size. With a burn-in of 10 000 iterations and 30 000 draws, the effective sample per multiple chain was 2878 with exploratory individual adaptation and 6949 with adaptively scaled individual adaptation. This is an impressive performance for both algorithms given the multicollinearity in the regressors. The difference in performance can be explained by the time taken for the two algorithms to converge to optimal values for the proposal. To illustrate the effect, we re-ran the algorithms with the burn-in extended to 30 000 iterations, the effective sample per multiple chain was now 3851 with exploratory individual adaptation but 7044 with adaptively scaled individual adaptation.

This example shows that the simplified posterior (7) is a good fit with many datasets and can indeed be used to guide and design algorithms.

4.3 PCR Data

Bondell and Reich (2012) described a variable selection problem with 22 576 variables and 60 observations on two inbred mouse populations. The covariates are gender and gene expression measurements for 22 575 genes. Using quantitative real-time polymerase chain reaction (PCR) several physiological phenotypes are recorded. We consider one of these phenotypes, phosphoenopruvate carboxykinase (PEPCK) as the response variable. Bondell and Reich (2012) apply their method to both a subset of 2 000 variables (selected on the basis of marginal correlations with the response) and the full data set. We use our adaptively scaled individual adaptation algorithm on the full data set of 22 576 variables. In prior (2) we adopt $V_\gamma = 100I$ and a hierarchical prior was used for γ by assuming that $h \sim \text{Be}(1, (p-5)/5)$ which implies that the prior mean number of included variables is 5. The algorithm was run with $\tau = 0.234$, 25 multiple chains, a burn-in of 3 000 iterations and 8 000 subsequent iterations and no thinning. Rao-blackwellised updates of $\pi^{(i)}$ were only used in the burn-in and posterior inclusion probability for the j-th variable was estimated by the mean of the posterior sample of γ_j. This took about 2 hours and 30 minutes to run using MATLAB and an Intel i7 @ 3.60 GHz processor. Three independent runs of the algorithms were executed to compare convergence.

Figure 4 shows the estimated posterior inclusion probabilities for each run and the pairwise comparisons between the different runs. The estimates from each independent chain are very similar and indicate that the sampler is able to accurately represent the posterior distribution. To put these results in perspective, we ran three other algorithms on the same data. The algorithms were the Add-Delete-Swap algorithm and two recently proposed methods for high-dimensional variable selection: the Hamming Ball sampler (Titsias and Yau, 2017) and the Ji-Schmidler adaptive sampler (Ji and Schmidler, 2013). Each method was

run in MATLAB for 2 hours and 30 minutes to make a direct comparison with our proposed method. Some results are shown in figure 5 for two independent runs of each method. Clearly, the Hamming Ball and Ji-Schmidler samplers are not able to adequately characterise the posterior model distribution with the two runs leading to dramatically different results. The Add-Delete-Swap sampler performs better but provides substantially more variable estimates of the posterior inclusion probabilities than the adaptively scaled individual adaptation method.

5 Conclusion

This paper introduces two adaptive Markov chain Monte Carlo algorithms for variable selection problems with very large p and small n. We recommend the adaptively scaled individual adaptation proposal, which is able to quickly find good proposals. This method uses a Rao-Blackwellised estimate of the posterior inclusion probability for each variable in an independent proposal. On simulated data this algorithm shows orders of magnitude improvements in effective sample size compared to the standard Metropolis-Hastings algorithm. The method is also applied to PCR data with 22 576 variables and shows excellent agreement in the posterior inclusion probabilities across independent runs of the algorithm, unlike the existing methods we have tried. Tuning the algorithm is key to defining an effective sampling scheme. We find that multiple independent chains with a shared proposal lead to better convergence to the optimal parameter values. Code to run both algorithms is available from

`https://www.kent.ac.uk/smsas/personal/jeg28/Version3.0.zip`

There are a number of possible directions for future research. We have only considered serial implementations of our algorithms in this paper. However, the algorithms are naturally parallelizable across the multiple chains but work is needed on efficient updating of the shared adaptive parameters. The proposals developed in this paper are well-suited to work within standard computational techniques for highly correlated or multi-modal distributions such as parallel tempering (Geyer, 1991) or sequential Monte Carlo samplers (see Del Moral et al. (2006) or Schäfer and Chopin (2013), with application to variable selection) which use powered versions of the posterior distribution. In such implementations, further computational gains may be obtained for parallel tempering versions where multimodality and correlations lessen in high temperatures and the heated posteriors become increasingly more suitable for our adaptively designed chains to pursue super-efficient mixing. Finally, it will be interesting to apply these algorithms to more complicated data which may have a non-Gaussian likelihood or a more complicated prior distribution (for example, a linear model with interactions).

Acknowledgements

KŁ acknowledges support of the Royal Society through the Royal Society University Research Fellowship.

References

Andrieu, C. and J. Thoms (2008). A tutorial on adaptive MCMC. *Statistics and Computing 18*, 343–373.

Bhattacharya, A., A. Chakraborty, and B. K. Mallick (2016). Fast sampling with Gaussian scale mixture priors in high-dimensional regression. *Biometrika 4*, 985–991.

Bondell, H. D. and B. J. Reich (2012). Consistent high-dimensional variable selection via penalized credible regions. *Journal of the American Statistical Association 107*, 1610–1624.

Bornn, L., P. E. Jacob, P. Del Moral, and A. Doucet (2013). An adaptive interacting Wang-Landau algorithm for automatic density exploration. *Journal of Computational and Graphical Statistics 22*, 749–773.

Bottolo, L. and S. Richardson (2010). Evolutionary stochastic search for Bayesian model exploration. *Bayesian Analysis 5*, 583–618.

Brown, P. J., M. Vannucci, and T. Fearn (1998). Multivariate Bayesian variable selection and prediction. *Journal of the Royal Statistical Society, B 60*, 627–641.

Castillo, I., J. Schmidt-Hieber, and A. van der Vaart (2015). Bayesian linear regression with sparse priors. *The Annals of Statistics 43*, 1986–2018.

Clyde, M. A., J. Ghosh, and M. L. Littman (2011). Bayesian adaptive sampling for variable selection and model averaging. *Journal of Computational and Graphical Statistics 20*, 80–101.

Craiu, R. V., J. Rosenthal, and C. Yang (2009). Learn from thy neighbor: Parallel-chain and regional adaptive MCMC. *Journal of the American Statistical Association 104*, 1454–1466.

Del Moral, P., A. Doucet, and A. Jasra (2006). Sequential Monte Carlo samplers. *J. R. Stat. Soc. Ser. B Stat. Methodol. 68*(3), 411–436.

Dellaportas, P., J. J. Forster, and I. Ntzoufras (2002). On Bayesian model and variable selection using MCMC. *Statistics and Computing 12*, 27–36.

Dezeure, R., P. Buehlmann, L. Meier, and N. Meinshausen (2015). High-dimensional inference: Confidence intervals, p-values and R-Software hdi. *Statistical Science 30*, 533–558.

García-Donato, G. and M. A. Martínez-Beneito (2013). On sampling strategies for Bayesian variable selection problems with large model spaces. *Journal of the American Statistical Association 108*, 340–352.

Gelman, A., G. O. Roberts, and W. R. Gilks (1996). Efficient Metropolis jumping rules. In *Bayesian statistics, 5 (Alicante, 1994)*, Oxford Sci. Publ., pp. 599–607. Oxford Univ. Press, New York.

George, E. I. and R. E. McCulloch (1997). Approaches for Bayesian variable selection. *Statistica sinica 7*, 339–373.

Geyer, C. J. (1991). Markov chain monte carlo maximum likelihood. In *Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface*, pp. 156–163.

Ghosh, J. and M. A. Clyde (2011). Rao-Blackwellisation for Bayesian variable selection and model averaging in linear and binary regression: A novel data augmentation approach. *Journal of the American Statistical Association 106*, 1041–1052.

Green, P. J., K. Łatuszyński, M. Pereyra, and C. P. Robert (2015). Bayesian computation: a summary of the current state, and samples backwards and forwards. *Stat. Comput. 25*(4), 835–862.

Griffin, J. E. and P. J. Brown (2010). Inference with normal-gamma prior distributions in regression problems. *Bayesian Analysis 5*, 171–188.

Guan, Y. and M. Stephens (2011). Bayesian variable selection regression for genome-wide association studies and other large-scale problems. *The Annals of Applied Statistics 5*, 1780–1815.

Haario, H., E. Saksman, and J. Tamminen (2001). An adaptive Metropolis algorithm. *Bernoulli 7*, 223–242.

Hahn, P. R. and C. M. Carvalho (2015). Decoupling shrinkage and selection in Bayesian linear models: a posterior summary perspective. *Journal of the American Statistical Association 110*, 435–448.

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

Ji, C. and S. C. Schmidler (2013). Adaptive Markov chain Monte Carlo for Bayesian variable selection. *Journal of Computational and Graphical Statistics 22*, 708–728.

Johndrow, J. E. and P. Orenstein (2017). Scalable MCMC for Bayes shrinkage priors. Technical Report arXiv:1705.00841.

Johnson, V. E. and D. Rossell (2012). Bayesian model selection in high-dimensional settings. *Journal of the American Statistical Association 107*(498), 649–660.

Lamnisos, D. S., J. E. Griffin, and M. F. J. Steel (2013). Adaptive Monte Carlo for Bayesian variable selection in regression models. *Journal of Computational and Graphical Statistics 22*, 729–748.

Łatuszyński, K. and G. O. Roberts (2013). CLTs and asymptotic variance of time-sampled Markov chains. *Methodol. Comput. Appl. Probab. 15*(1), 237–247.

Levin, D. A., Y. Peres, and E. L. Wilmer (2009). *Markov chains and mixing times*. American Mathematical Society, Providence, RI. With a chapter by James G. Propp and David B. Wilson.

Liang, F., C. Liu, and R. J. Carroll (2007). Stochastic approximation in Monte Carlo computation. *Journal of the American Statistical Assocation 102*, 305–320.

Liang, F., Q. Song, and K. Yu (2013). Bayesian subset modeling for high-dimensional generalized linear models. *Journal of the American Statistical Association 108*, 589–606.

Neal, P., G. Roberts, and W. K. Yuen (2012). Optimal scaling of random walk Metropolis algorithms with discontinuous target densities. *Ann. Appl. Probab. 22*(5), 1880–1927.

Neal, P. J. and C. Lee (2017). Optimal scaling of the independence sampler: theory and practice. *Bernoulli, to appear*.

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

Nott, D. J. and R. Kohn (2005). Adaptive sampling for Bayesian variable selection. *Biometrika 92*, 747–763.

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

Papaspiliopoulos, O. and D. Rossell (2017). Bayesian block-diagonal variable selection and model averaging. arXiv:1606.03749v2.

Peskun, P. H. (1973). Optimum Monte-Carlo sampling using Markov chains. *Biometrika 60*, 607–612.

Richardson, S., L. Bottolo, and J. S. Rosenthal (2010). Bayesian models for sparse regression analysis of high dimensional data. *Bayesian Statistics 9*, 539–568.

Roberts, G. O. (1998). Optimal Metropolis algorithms for product measures on the vertices of a hypercube. *Stochastics Stochastic Rep. 62*(3-4), 275–283.

Roberts, G. O., A. Gelman, and W. R. Gilks (1997). Weak convergence and optimal scaling of random walk Metropolis algorithms. *Annals of Applied Probability 7*, 110–120.

Roberts, G. O. and J. S. Rosenthal (2004). General state space Markov chains and MCMC algorithms. *Probability Surveys 1*, 20–71.

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

Roberts, G. O. and J. S. Rosenthal (2009). Examples of adaptive MCMC. *Journal of Computational and Graphical Statistics 18*, 349–367.

Roberts, G. O. and J. S. Rosenthal (2016). Complexity bounds for Markov chain Monte Carlo algorithms via diffusion limits. *J. Appl. Probab. 53*(2), 410–420.

Rockova, V. and E. I. George (2014). EMVS: The EM approach to Bayesian variable selection. *Journal of the American Statistical Association 109*(506), 828–846.

Schäfer, C. and N. Chopin (2013). Sequential Monte Carlo on large binary sampling spaces. *Statistics and Computing 23*, 163–184.

Shah, R. D. and R. J. Samworth (2013). Variable selection with error control: Another look at Stability Selection. *Journal of the Royal Statistical Society, Series B 75*, 55–80.

Titsias, M. K. and C. Yau (2017). The Hamming ball sampler. *Journal of the American Statistical Association 112*, forthcoming.

van den Boom, W., G. Reeves, and D. B. Dunson (2015). Scalable approximations of marginal posteriors in variable selection. arxiv:1506.06629.

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

A Proofs

of Proposition 1. For both (i) and (ii) notice that since the proposal is of product form and probability of acceptance is one, each sequence of individual indicator variables $\{\gamma_j^{(t)}\}_{t=0,1,\ldots}$ evolves independently of other coordinates, and is a Markov chain on $\{0,1\}$ governed by, say, transition kernel P_j, with stationary distribution $\pi^{(j)} = \{1-\pi_j, \pi_j\}$.

Part (i):

(a) and formula (11) of (b) are immediate because the proposal samples from the stationary distribution and is accepted with probability 1.

To verify formula (12) of (b), use that individual coordinates are Markovian, and for $f \in \mathbb{L}(\Gamma)$ compute:

$$\sigma^2_{P,f} = \sum_{j=1}^{p} a_j^2 \sigma^2_{P,f_j} = \sum_{j=1}^{p} a_j^2 \sigma^2_{P_j,f_j}. \tag{24}$$

Now recall that P_j in (i) is independent sampling from $\pi^{(j)}$, i.e. $P_j = \Pi_j := \left[\begin{smallmatrix} 1-\pi_j & \pi_j \\ 1-\pi_j & \pi_j \end{smallmatrix}\right]$, hence $\sigma^2_{P_j,f_j} = Var_{\pi^{(j)}} f_j$.

To verify (i), formula (13) in (c), note that

$$E_\pi[\Delta^2] = E_\pi\left[\sum_{j=1}^{p} |\gamma_j^{(0)} - \gamma_j^{(1)}|^2\right] = \sum_{j=1}^{p} E_\pi |\gamma_j^{(0)} - \gamma_j^{(1)}|, \tag{25}$$

and for the independent sampling Markov chain $E_\pi|\gamma_j^{(0)} - \gamma_j^{(1)}| = 2\pi_j(1-\pi_j)$.

Part (ii):

For maximality in (a), recall (25), and it is enough to check (by simple algebra) that the transition kernel $P_j := \left[\begin{smallmatrix} 1-A_j & A_j \\ D_j & 1-D_j \end{smallmatrix}\right]$, resulting from this choice of (A, D), maximises $E_{1-\pi_j,\pi_j}|\gamma_j - \gamma_j'|$ over all possible Markov chains on $\{0,1\}$ with stationary distribution $\{1-\pi_j, \pi_j\}$.

For formula (14) of $(ii)(a)$, recall (25), and note that $E_\pi|\gamma_j^{(0)} - \gamma_j^{(1)}| = 2\max\{1-\pi_j, \pi_j\}$.

For Peskun optimality in (b), recall formula (24) and consider σ_{P_j,f_j} in this setting. It is enough to verify that for each j, the kernel $P_j = \left[\begin{smallmatrix} 1-A_j & A_j \\ D_j & 1-D_j \end{smallmatrix}\right]$ is optimal with respect to Peskun ordering among all Markov chains on $\{0,1\}$ with stationary distribution $\pi^{(j)}$. Indeed, by simple algebra, P_j maximises off-diagonal elements among all stochastic matrices with stationary distribution $\pi^{(j)}$ and by Theorem 2.1.1 of Peskun (1973), is optimal.

To recover formula (15) of $(ii)(b)$, recall (24), and consider asymptotic variance terms of

individual coordinates $\sigma^2_{P_j,f_j}$ for this case. These can be computed directly, but we take a shortcut noting that

$$P_j = \begin{bmatrix} 1-\min\{1,\frac{\pi_j}{1-\pi_j}\} & \min\{1,\frac{\pi_j}{1-\pi_j}\} \\ \min\{1,\frac{1-\pi_j}{\pi_j}\} & 1-\min\{1,\frac{1-\pi_j}{\pi_j}\} \end{bmatrix} \qquad \text{and} \qquad \Pi_j = \begin{bmatrix} 1-\pi_j & \pi_j \\ 1-\pi_j & \pi_j \end{bmatrix}$$

admit the representation

$$\Pi_j = \max\{1-\pi_j, \pi_j\}P_j + (1 - \max\{1-\pi_j, \pi_j\})\mathbb{I},$$

where $\mathbb{I}$ is identity matrix. Thus Π_j is a lazy version of P_j and, by Corollary 1 of Łatuszyński and Roberts (2013), their asymptotic variances are related by

$$Var_{\pi^{(j)}} f_j = \sigma^2_{\Pi_j,f_j} = \frac{1}{\max\{1-\pi_j,\pi_j\}}\sigma^2_{P_j,f_j} + \frac{1-\max\{1-\pi_j,\pi_j\}}{\max\{1-\pi_j,\pi_j\}} Var_{\pi^{(j)}} f_j.$$

Putting $\sigma^2_{P_j,f_j} = (2\max\{1-\pi_j,\pi_j\}-1)Var_{\pi^{(j)}} f_j$ into (24) concludes (15). □

of Lemma 1. To verify the result it is enough to check that the whole state space Γ^r is $1-$small with the same constant $b > 0$, (c.f. Roberts and Rosenthal (2004)), that is check, for example, that there exists $b > 0$ s.t. for every $\eta \in \Delta_\varepsilon$ and every $\gamma^{\otimes r}, \gamma'^{\otimes r} \in \Gamma^r$ we have

$$q_\eta(\gamma^{\otimes r}, \gamma'^{\otimes r}) \geq b \tag{26}$$

where $q_\eta(\gamma^{\otimes r}, \gamma'^{\otimes r})$ is the transition for the r chains. First decompose the move into proposal and acceptance

$$P_\eta(\gamma^{\otimes r}, \gamma'^{\otimes r}) \;=\; q_\eta(\gamma^{\otimes r}, \gamma'^{\otimes r}) \times a_\eta(\gamma^{\otimes r}, \gamma'^{\otimes r}),$$

and notice that by the proposal construction $q_\eta(\gamma^{\otimes r}, \gamma'^{\otimes r}) \geq \varepsilon^{rp}$ since $\text{diam}(\Gamma^r) = rp$. Similarly

$$\begin{aligned} a_\eta(\gamma^{\otimes r}, \gamma'^{\otimes r}) &= \min\left\{1, \frac{\pi^{\otimes r}(\gamma'^{\otimes r}) q_\eta(\gamma'^{\otimes r}, \gamma^{\otimes r})}{\pi^{\otimes r}(\gamma^{\otimes r}) q_\eta(\gamma^{\otimes r}, \gamma'^{\otimes r})}\right\} \\ &\geq \pi^{\otimes r}(\gamma'^{\otimes r}) q_\eta(\gamma'^{\otimes r}, \gamma^{\otimes r}) \;\geq\; \pi_m^r \times \varepsilon^{rp}, \end{aligned}$$

where $\pi_m := \min_{\gamma\in\Gamma} \pi(\gamma)$. Consequently in (26) we can take

$$b = \pi_m^r \times \varepsilon^{2rp},$$

and we have established simultaneous uniform ergodicity. □

of Theorem 1. Theorem 1 follows from Theorem 1 (ergodicity) and Theorem 5 (WLLN) of Roberts and Rosenthal (2007) using Lemma 1 to establish simultaneous uniform ergodicity for nonadaptive kernels. Clearly, both proposals have diminishing adaptation, *i.e.* the random variable

$$\mathcal{D}_i := \sup_{\gamma^{\otimes r}\in\Gamma^r} \|q_{\eta^{(i+1)}}(\gamma^{\otimes r}, \cdot) - q_{\eta^{(i)}}(\gamma^{\otimes r}, \cdot)\|$$

converges to 0 in probability as $i \to \infty$. □

B Rao-Blackwellisation

For notational simplicity, in the expressions below we use X to denote $X_{\gamma^{(i)}}$. Furthermore, we choose $V_\gamma = n_0 I$ and define $F = (X^T X + n_0 I)^{-1}$. The sequential update of the posterior inclusion probability is, if $\gamma_j = 0$,

$$\pi_j^{(i)} = \begin{cases} \frac{i-1}{i}\pi_j^{(i-1)} + \frac{1}{i}\frac{h\, n_0^{1/2} A^{-1/2}\exp\{\Psi\}}{h\, n_0^{1/2} A^{-1/2}\exp\{\Psi\}+1-h} & \text{if } h \text{ is fixed} \\ \frac{i-1}{i}\pi_j^{(i-1)} + \frac{1}{i}\frac{\frac{p\gamma+a}{p-1+a+b}\, n_0^{1/2} A^{-1/2}\exp\{\Psi\}}{\frac{p\gamma+a}{p-1+a+b}\, n_0^{1/2} A^{-1/2}\exp\{\Psi\}+1-\frac{p\gamma+a}{p-1+a+b}} & \text{if } h \sim \text{Be}(a,b) \end{cases},$$

where

$$\Psi = n/2\left[\log\left(y^T y - y^T X F X^T y\right) - \log(y^T y - B))\right]$$

$$B = y^T X F X^T y + y^T x_j x_j^T X F X^T y + y^T X F X^T x_j x_j^T y + y^T x_j S_j x_j^T y$$

$$S_j = 1/(x_j^T x_j + n_0 - x_j^T X F X^T x_j)$$

$$A = x_j^T x_j + n_0 - x_j^T X F X^T x_j.$$

If $\gamma_j = 1$ and the columns of X are permuted so that $X = [X^- \; x_j]$,

$$\pi_j^{(i)} = \begin{cases} \frac{i-1}{i}\pi_j^{(i-1)} + \frac{1}{i}\frac{h\, n_0^{1/2} A^{-1/2}\exp\{\Psi\}}{h\, n_0^{1/2} A^{-1/2}\exp\{\Psi\}+1-h} & \text{if } h \text{ is fixed} \\ \frac{i-1}{i}\pi_j^{(i-1)} + \frac{1}{i}\frac{\frac{p\gamma-1+a}{p-1+a+b}\, n_0^{1/2} A^{-1/2}\exp\{\Psi\}}{\frac{p\gamma-1+a}{p-1+a+b}\, n_0^{1/2} A^{-1/2}\exp\{\Psi\}+1-\frac{p\gamma-1+a}{p-1+a+b}} & \text{if } h \sim \text{Be}(a,b) \end{cases},$$

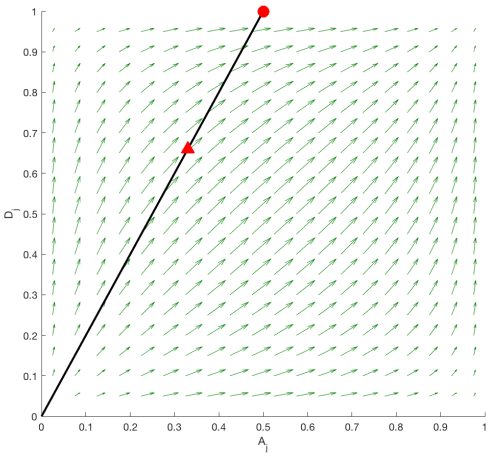
where

$$\Psi = n/2\left[\log(y^T y - B) - \log\left(y^T y - y^T X F X^T y\right)\right]$$

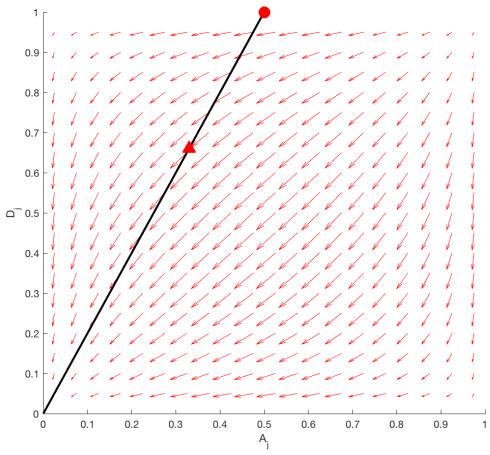
$$B = (y^T T X)_{1:(p-1)} F_{1:(p-1),1:(p-1)} - F_{1:(p-1),p} F^T_{1:(p-1),p} \times 1/F_{p,p} (X^T y)_{1:(p-1)}$$

$$A = x_j^T x_j + n_0 - x_j^T X^- B X^{-T} x_j.$$

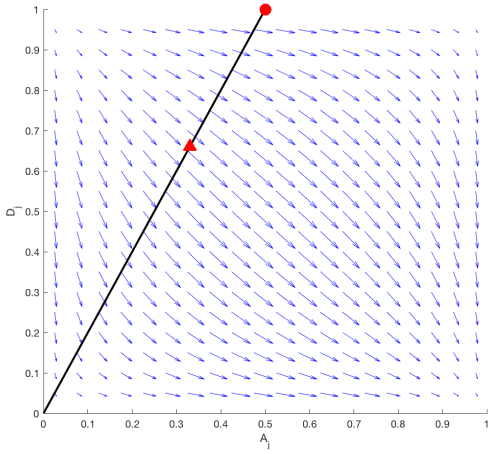
These results allow the Rao-Blackwellised estimates to be calculated. In particular, if the value of F and $\left(y^T y - y^T X F X^T y\right)$ (which are needed to calculate the marginal likelihood) are stored, these operations can be completed in $O(p)$ steps.



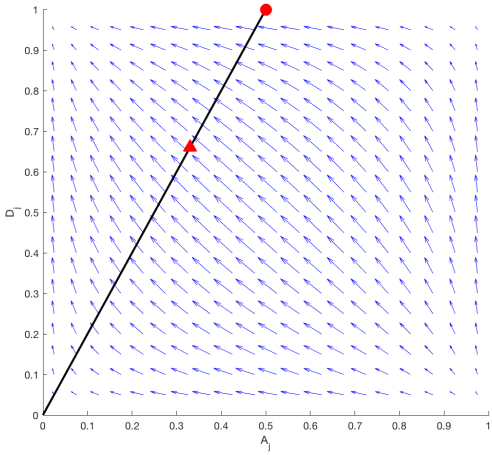
(a) Gradient field of the *expansion step* for $\gamma^{A(i)}$ and $\gamma^{D(i)}$



(b) Gradient field of the *shrinkage step* for $\gamma^{A(i)}$ and $\gamma^{D(i)}$

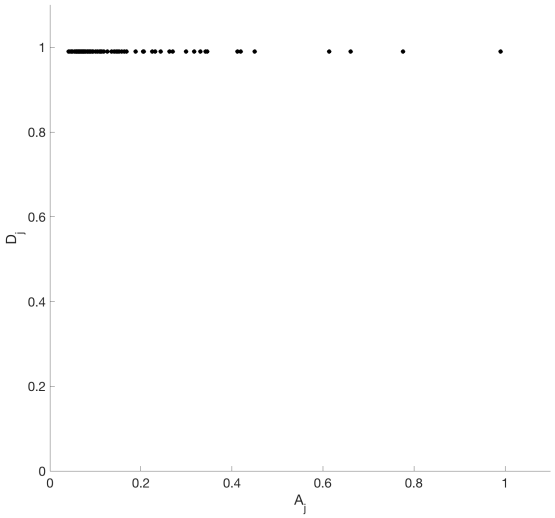


(c) Gradient field of the *correction step* for $\gamma^{D(i)}$

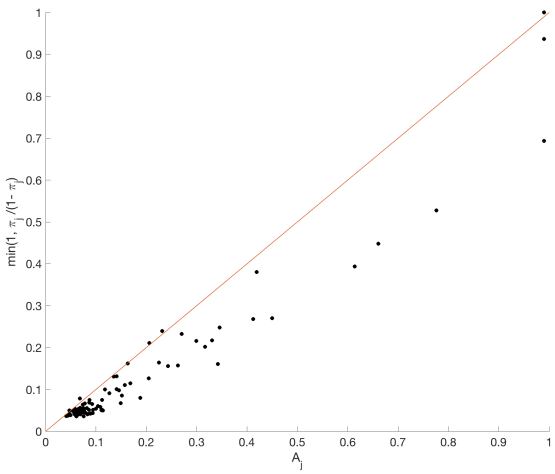


(d) Gradient field of the *correction step* for $\gamma^{A(i)}$

Figure 2: Gradient fields guiding parameter updates of the exploratory individual adaptation algorithm towards and along the segment.



(a) Limiting values of the (A_j, D_j) pairs align at the top ends of the segments of Figure 1, with D_J's close to 1, and correspond to the super-efficient setting (ii) of Proposition 1.



(b) The attained values of A_j's overestimate the idealized values $\min\{1, \frac{\pi_j}{1-\pi_j}\}$ of setting (ii) in Proposition 1, indicating low dependence in the posterior.

Figure 3: Behaviour of the adaptive parameter $\eta = (A, D)$ of the exploratory individual adaptation algorithm.

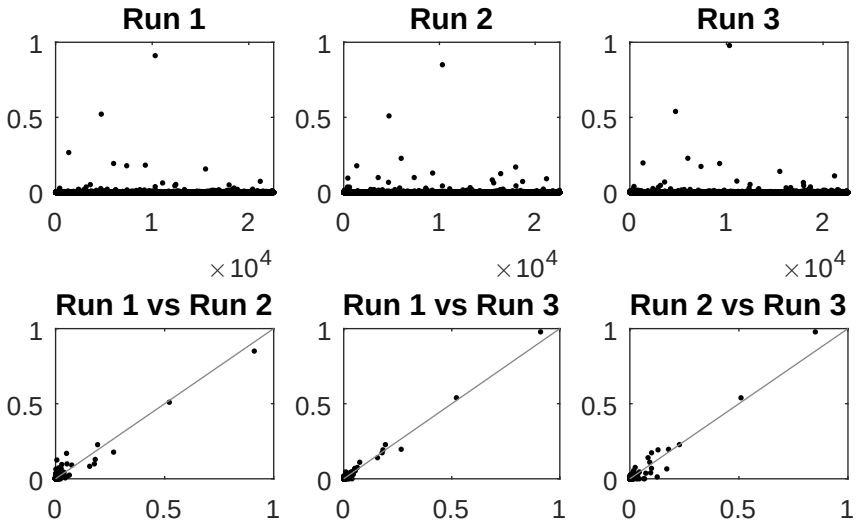


Figure 4: PCR data: The posterior inclusion probabilities from three independent runs of the adaptively scaled individual adaptation algorithm.

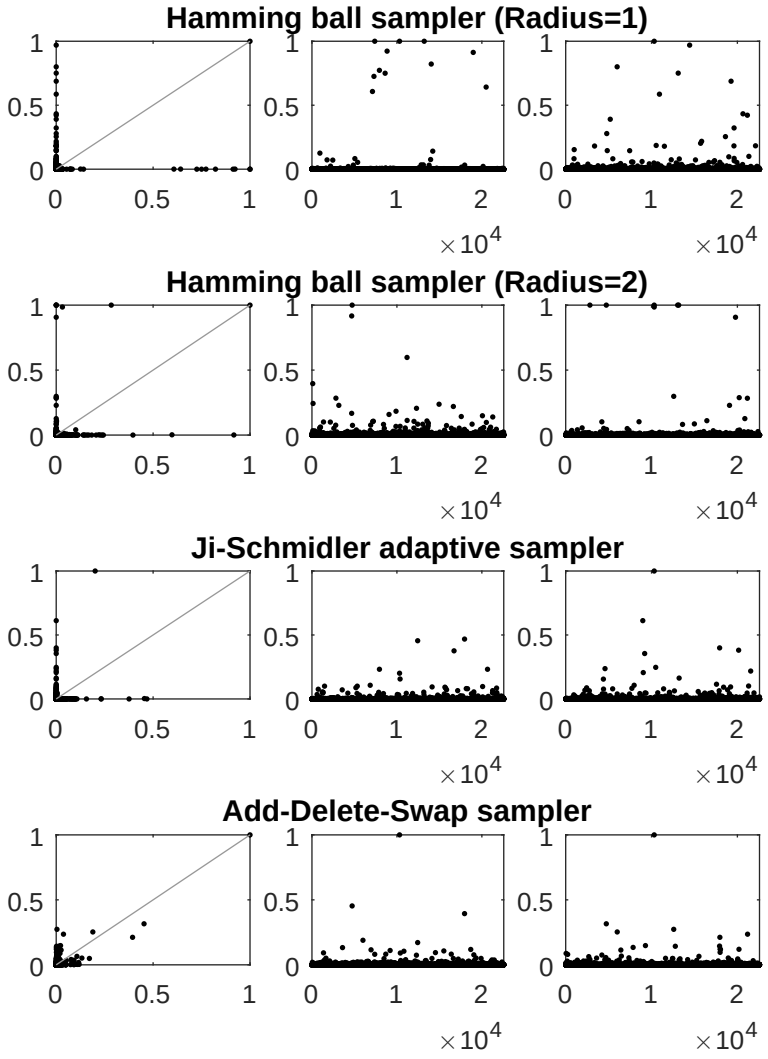


Figure 5: The first column plots the posterior inclusion probabilities from the two runs, the second and third columns show the posterior inclusion probabilities from each run.