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*

May 3, 2019

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, Department of Statistical Science, University College London, WC1E 6BT, U.K. (Email: j.griffin@ucl.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; O'Hara and Sillanpää, 2009; Clyde et al., 2011). García-Donato and Martínez-Beneito (2013) discuss the benefits of such methods. The most widely used Markov chain Monte Carlo algorithm in this context is the Metropolis-Hastings sampler where new models are proposed using add-delete-swap samplers (Brown et al., 1998; Chipman et al., 2001). For example, this approach 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-delete-swap samplers, *e.g.* conditions for rapid mixing in linear regression model have been derived by Yang et al. (2016). Others have considered more targeted moves in model space. For example, Titsias and Yau (2017) introduce the Hamming Ball sampler which more carefully selects the proposed model in a Metropolis-Hastings sampler (in a similar way to shotgun variable selection, Hans et al. (2007)) and Schäfer and Chopin (2013) develop a sequential Monte Carlo method that uses a sequence of annealed posterior distributions. However, the mixing of these methods is often thought to be poor, when applied to data sets with thousands of potential covariates. As an alternative, several authors use more general shrinkage priors and develop suitable MCMC algorithms for high-dimensional problems (see *e.g.* Bhattacharya et al., 2016). Nonlocal priors (Johnson and Rossell, 2012) are adopted in Shin et al. (2018), who use screening for high dimensions. Zanella and Roberts (2019) combine Markov chain Monte Carlo and importance sampling ideas in their tempered Gibbs sampler.

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, Liang et al. (2013) used the stochastic approximation Monte Carlo algorithm

(Liang et al., 2007) to 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 posterior probability 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 to find the most probable model of any given size in 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 the 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 utilizes 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-dimensional 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 4.4, we present real data gene expression examples with $p = 22\,576$ and with $p = 79\,748$, and reliably estimate the posterior inclusion probabilities for all variables. All proofs are grouped in the Supplementary Material.

2 Design of the Adaptive Samplers

2.1 The Setting

Our approach is applicable to general regression settings but we will focus on normal linear regression models. This will allow for clean efficiency comparisons independent of model-specific sampling details (e.g. of a reversible jump implementation). We define $\gamma = (\gamma_1, \ldots, \gamma_p) \in \Gamma = \{0,1\}^p$ to be a vector of indicator variables with $\gamma_i = 1$ if the i-th variable is included in the model and $p_\gamma = \sum_{j=1}^{p} \gamma_j$. We consider the model specification

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

where y is an $(n \times 1)$-dimensional vector of responses, $\mathbf{a}_q$ represents a q-dimensional column vector with entries a, and X_γ is a $(n \times p_\gamma)$-dimensional data matrix formed using the included variables. We consider Bayesian variable selection and, 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 $p(\gamma) = h^{p_\gamma}(1-h)^{p-p_\gamma}$. 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 $h \sim \mathrm{Be}(a, b)$.

We will consider sampling from the target distribution $\pi_p(\gamma) = p(\gamma|y)$ using a non-symmetric Metropolis-Hastings kernel. Let the probability of proposing to move from model γ to γ' be

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

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 proposal can be quickly sampled, the parametrisation allows optimisation of the expected squared jumping distance, and multiple variables can be added to or deleted from the model in one iteration. The proposed model is accepted using the standard Metropolis-Hastings acceptance probability

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

2.2 In Search of Lost Mixing Time: Optimising the Sampler

The transition kernel in (3) is highly parameterised with $2p$ parameters and these will be tuned using adaptive Markov chain Monte Carlo methods (see *e.g.* Andrieu and Thoms, 2008; Roberts and Rosenthal, 2009; Green et al., 2015). These methods allow the tuning of parameters on the fly to improve mixing using some computationally accessible performance criterion whilst maintaining the ergodicity of the chain. Suppose that μ_p is a p-dimensional probability density function which has the form $\mu_p = \prod_{j=1}^{p} f$. A commonly used result is that the optimal scale of a random walk proposal for μ_p leads to a mean acceptance rate of 0.234 as $p \to \infty$ for some smooth enough f. The underlying analysis also implies that the optimised random walk Metropolis will converge to stationarity in $\mathcal{O}(p)$ steps. This is a useful guide even in moderate dimensions and well beyond the restrictive independent, identically distributed product form assumption of Roberts et al. (1997). Lamnisos et al. (2013) show that this rule can be effectively used to tune a Metropolis-Hastings sampler for Bayesian variable selection. However, other results suggest that other optimal scaling rules could work well in Bayesian variable selection problems. Firstly, Neal et al. (2012) established, under additional regularity conditions, that if f is discontinuous, the optimal mean acceptance rate for a Metropolis-Hastings random walk is $e^{-2} \approx 0.1353$ and the chain mixes in $\mathcal{O}(p^2)$ steps, an order of magnitude slower than with smooth target densities f. Rather surprisingly, Lee and Neal (2018) show that the optimally tuned independence sampler in this settings recovers the $O(p)$ mixing and acceptance rate of 0.234 without any additional smoothness conditions. Secondly, Roberts (1998) considered optimal scaling of the random walk Metropolis-Hastings algorithm on $\Gamma = \{0,1\}^p$ for the product measures

$$\mu_p(\gamma_1, \ldots, \gamma_p) = s^{p_\gamma}(1-s)^{p-p_\gamma}, \qquad \gamma = (\gamma_1, \ldots, \gamma_p) \in \Gamma, \quad 0 < s < 1.$$

If s is close to $1/2$, the optimal $\mathcal{O}(p)$ mixing rate occurs as p tends to infinity if the mean acceptance rate is 0.234. If $s \to 0$ as $p \to \infty$, the numerical results of Section 3 in Roberts (1998) indicate that the optimally tuned random walk Metropolis proposes to change two γ_j's at a time but that 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.

In Bayesian variable selection, it is natural to assume that the variables differ in posterior inclusion probabilities and so we consider target densities that have the form

$$\pi_p(\gamma) = \prod_{j=1}^{p} \pi_j^{\gamma_j}(1-\pi_j)^{1-\gamma_j}, \qquad \gamma \in \Gamma \tag{5}$$

where $0 < \pi_j < 1$ for $j = 1, \ldots, p$. Consider the non-symmetric Metropolis-Hastings algorithm with the product form proposal $q_\eta(\gamma, \gamma')$ given by (3) targeting the posterior dis-

tribution given by (5). Note that $\alpha_\eta(\cdot,\cdot) \equiv 1$ for any choice of $\eta = (A, D)$ satisfying

$$\frac{A_j}{D_j} = \frac{\pi_j}{1-\pi_j}, \quad \text{for every } j. \tag{6}$$

To discuss optimal choices of η, we consider several commonly used criteria for Markov chains with stationary distribution π and transition kernel P on a finite discrete state space Γ. The *mixing time* of a Markov chain (Roberts and Rosenthal, 2004) is $\rho := \min\{t : \max_{\gamma\in\Gamma} \|P^t(\gamma,\cdot) - \pi(\cdot)\|_{TV} < 1/2\}$ where $\|\cdot\|_{TV}$ is the total variational norm. If $\Gamma = \{0,1\}^p$, it is natural to define the *expected squared jumping distance* (Gelman et al., 1996) as $E_\pi\left[\sum_{j=1}^p |\gamma_j^{(0)} - \gamma_j^{(1)}|^2\right]$ where $\gamma^{(0)}$ and $\gamma^{(1)}$ are two consecutive values in a Markov chain trajectory, which is the average number of variables changed in one iteration. 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 $\sigma^2_{P_1,f} \le \sigma^2_{P_2,f}$ for every f, then P_1 dominates P_2 in *Peskun ordering* (Peskun, 1973). 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 variable selection problem, where posterior inclusion probabilities are often of interest, we consider Peskun ordering for the class $\mathbb{L}(\Gamma)$ of linear combinations of univariate functions,

$$\mathbb{L}(\Gamma) := \left\{ f : \Gamma \to \mathbb{R} : f(\gamma) = a_0 + \sum_{j=1}^p a_j f_j(\gamma_j) \right\}. \tag{7}$$

We consider two proposals which satisfy (6). The *independent proposal* for which $A_j = 1 - D_j = \pi_j$ and the *random walk proposal* for which $A_j = \min\{1, \frac{\pi_j}{1-\pi_j}\}$ and $D_j = \min\{1, \frac{1-\pi_j}{\pi_j}\}$. The following proposition shows that the random walk proposal has more desirable properties.

Proposition 1 *Consider the class of Metropolis-Hastings algorithms with target distribution given by* (5) *and proposal* $q_\eta(\gamma,\gamma')$ *given by* (3) *with the independent or random walk proposal. Let* $Var_\pi f$ *be the stationary variance of* f *under* $\pi_p(\gamma)$ *and* $\pi^{(j)} := \{1-\pi_j, \pi_j\}$. *Then,*

(i) the independent proposal leads to

- *(a) independent sampling and optimal mixing time* $\rho = 1$;
- *(b) the expected squared jumping distance is* $E_\pi[\Delta^2] = 2\sum_{j=1}^p \pi_j(1-\pi_j)$;
- *(c) the asymptotic variances is* $\sigma^2_{P,f} = Var_\pi f$ *for arbitrary* f *and* $\sigma^2_{P,f} = Var_\pi f = \sum_{j=1}^p a_j^2 Var_{\pi^{(j)}} f_j$ *for* $f \in \mathbb{L}(\Gamma)$;

(ii) the random walk proposal leads to

(a) the expected squared jumping distance is $E_\pi[\Delta^2] = 2\sum_{j=1}^p \min\{1-\pi_j, \pi_j\}$, *which is maximal;*

(b) the asymptotic variance is $\sigma^2_{P,f} = \sum_{j=1}^p \big(2\max\{1-\pi_j, \pi_j\} - 1\big) a_j^2 Var_{\pi(j)} f_j$ *for* $f \in \mathbb{L}(\Gamma)$ *and it is optimal with respect to the Peskun ordering for the class of linear functions* $\mathbb{L}(\Gamma)$ *defined in* (7).

Remark 1 *The differences of the expected squared jumping distance and asymptotic variance for the two proposals is largest when* π_j *is close to* $1/2$.

Remark 2 *In discrete spaces, Schäfer and Chopin (2013) argue that the mutation rate*

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

which excludes moves which do not change the model, is more appropriate than average acceptance rate. The mutation rate is $\bar{a}_M = 1 - \prod_{j=1}^p [(1-\pi_j)^2 + \pi_j^2]$ *with independent sampling and is* $\bar{a}_M = 1 - \prod_{j=1}^p |2\pi_j - 1|$ *with the random walk proposal. Therefore, the random walk proposal always leads to a higher mutation rate.*

These results suggest that the random walk proposal should be preferred to the independent proposal when designing a Metropolis-Hastings sampler for idealised posteriors of the form in (5). In practice, the posterior distribution will not have a product form but can anything be said about its form when p is large? The following result sheds some light on this issue. We define $\text{BF}_j(\gamma_{-j})$ to be the Bayes factor of including the j-th variable given the values of $\gamma_{-j} = (\gamma_1, \ldots, \gamma_{j-1}, \gamma_{j+1}, \ldots, \gamma_p)$ and denote by γ_0 the vector γ without γ_j and γ_k.

Proposition 2 *Let* $a = \frac{BF_j(\gamma_k=1,\gamma_0)}{BF_j(\gamma_k=0,\gamma_0)}$. *If (i)* $a \to 1$ *or (ii)* $a \to A < \infty$ *and* $BF_j(\gamma_k = 0, \gamma_0)h \to 0$ *then* $p(\gamma_j = 1|\gamma_k = 1, \gamma_0) \to p(\gamma_j = 1|\gamma_k = 0, \gamma_0)$.

This result gives condition under which γ_j and γ_k are approximately independent. Condition (ii) is interesting in large p settings: γ_j and γ_k are approximately independent if p is large (and so h is small) and $\text{BF}_j(\gamma_k = 0, \gamma_0)$ is not large, *i.e.* the evidence in favour of including γ_j is not large. This will be the case for all variables apart from the most important. Although this result provides some reassurance, there will be some posterior correlation in many problems and the random walk proposal may propose to change too many variables leading to low acceptance rates. This can be addressed by using a scaled proposal of the form

$$A_j = \zeta_j \min\left\{1, \frac{\pi_j}{1-\pi_j}\right\}, \qquad D_j = \zeta_j \min\left\{1, \frac{1-\pi_j}{\pi_j}\right\}. \tag{8}$$

The family of these proposals for $\zeta_j \in [0,1]$ form a line segment for (A_j, D_j) between $(0,0)$ and $\left(\min\left\{1, \frac{\pi_j}{1-\pi_j}\right\}, \min\left\{1, \frac{1-\pi_j}{\pi_j}\right\}\right)$, illustrated in Figure 1 (Supplementary Material D). The independent proposal corresponds to the point on this line where $\zeta_j = \max\{\pi_j, 1-\pi_j\}$.

In the next section, we devise adaptive MCMC algorithms to tune proposals of the form (3) so that A_j's and D_j's lie approximately on this line. Larger values of ζ_j tend to lead to larger jumps whereas smaller values of ζ_j tend to increase acceptance. These algorithms aim to find a point which balances this trade-off. We define two strategies for adapting η: *Exploratory Individual Adaptation* and *Adaptively Scaled Individual Adaptation*.

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 consider algorithms with different numbers of independent parallel chains (but the same parameters of the proposal) and refer to this as multiple chain acceleration. To avoid the algorithms becoming trapped in well separated modes, we also consider parallel tempering versions of the algorithms, following Miasojedow et al. (2013) as explained in Supplementary Material B.

At this point, it is helpful to define some notation. Let $\eta^{(i)} = (A^{(i)}, D^{(i)})$ and $\gamma^{(i)}$ be the values of η and γ at the start of the i-th iteration, and γ' be the subsequently proposed value. Let $a_i = a_{\eta^{(i)}}(\gamma^{(i)}, \gamma')$ be the acceptance probability 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 reduces to the usual logit transform if $\epsilon = 0$.

2.3 Remembrance of Things Past: Exploratory Individual Adaptation

The first adaptive strategy is a general purpose method that we term *Exploratory Individual Adaptation* (EIA). It aims to find pairs (A_j, D_j) on the line segment defined by (6) which lead to good mixing. Proposals with larger values of A_j and D_j will tend to propose more changes to the included variables but will also tend to reduce the average acceptance probability or mutation rate. The method introduces two tuning parameter τ_L and τ_U. 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). Unless otherwise stated, $A_j^{(i+1)} = A_j^{(i)}$ and $D_j^{(i+1)} = D_j^{(i)}$:

1. Both the *expansion step* and the *shrinkage step* change $A_j^{(i+1)}$ and $D_j^{(i+1)}$ for j in $\gamma^{A(i)}$ and $\gamma^{D(i)}$ to adjust the average squared jumping distance whilst maintaining that $A_j^{(i+1)}/D_j^{(i+1)} \approx A_j^{(i)}/D_j^{(i)}$. The expansion step is used if a promising move is proposed (if $a_i > \tau_U$) and sets $A_j^{(i+1)}$ and $D_j^{(i+1)}$ larger than $A_j^{(i)}$ and $D_j^{(i)}$ respectively.

Similarly, the shrinkage step is used if an unpromising move has been proposed (if $a_i < \tau_L$) and $A_j^{(i+1)}$ and $D_j^{(i+1)}$ are set smaller than $A_j^{(i)}$ and $D_j^{(i)}$.

2. The *correction step* aims to increase the average acceptance rate by correcting the ratio between A's and D's. If $\tau_L < a_i < \tau_U$, we set $A_j^{(i+1)} > A_j^{(i)}$ and $D_j^{(i+1)} < D_j^{(i}$ if $\gamma_j^{D(i)} = 1$ and $A_j^{(i+1)} < A_j^{(i)}$ and $D_j^{(i+1)} > D_j^{(i)}$ if $\gamma_j^{A(i)} = 1$.

The gradient fields of these updates are shown in Figure 2 (Supplementary Material D). These three moves can be combined into the following adaptation of $A^{(i)}$ and $D^{(i)}$

$$\text{logit}_\epsilon A_j^{(i+1)} = \text{logit}_\epsilon A_j^{(i)} + \phi_i \left(\gamma_j^{A(i)} d_i(\tau_U) + \gamma_j^{D(i)} d_i(\tau_L) - \gamma_j^{A(i)}(1 - d_i(\tau_U))\right), \quad (9)$$

$$\text{logit}_\epsilon D_j^{(i+1)} = \text{logit}_\epsilon D_j^{(i)} + \phi_i \left(\gamma_j^{D(i)} d_i(\tau_U) + \gamma_j^{A(i)} d_i(\tau_L) - \gamma_j^{D(i)}(1 - d_i(\tau_U))\right), \quad (10)$$

for $j = 1 \ldots, p$ where $d_i(\tau) = \mathbb{I}\{a_i \geq \tau\}$ and $\phi_i = O(i^{-\lambda})$ for some constant $1/2 < \lambda \leq 1$. The transformation implies that $\epsilon < A_j^{(i)} < 1 - \epsilon$ and $\epsilon < D_j^{(i)} < 1 - \epsilon$ and we assume that $0 < \epsilon < 1/2$. It also implies diminishing adaptation (essentially since the derivative of the inverse logit is bounded, see Lemma 2). Based on several simulation studies, we suggest to take $\tau_L = 0.01$ and $\tau_U = 0.1$. 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 parameter ϵ controls the minimum and maximum values of A_i and D_i. In the large p setting, $A_i \approx \epsilon$ for unimportant variables and the expected number of those unimportant variables proposed to be included at each iteration will be approximately $p\epsilon$ (since the number of excluded, unimportant variables will be close to p). This expected value can be controlled by choosing $\epsilon = 0.1/p$. The EIA algorithm is described in Algorithm 1 and we indicate its transition kernel at time i as $P_{\eta^{(i)}}^{\text{EIA}}$.

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

update $A^{(i+1)}$ using (9) and $D^{(i+1)}$ using (10)

endfor

Algorithm 1: Exploratory Individual Adaptation (EIA)

2.4 Remembrance of Things Past: Adaptively Scaled Individual Adaptation

Algorithm 1 learns two parameters $A_j^{(i)}$ and $D_j^{(i)}$ for each variable and can be slow to converge to optimal values if p is large. Alternatively, we could learn $\pi_1, \ldots, \pi_p$ from the chain to approximate the slope of the line defined by (6) and use the proposal (8) with the same scale parameter for all variables. We term this approach the *Adaptively Scaled Individual Adaptation* (ASI) proposal. In particular, we use

$$A_j^{(i)} = \zeta^{(i)} \min\left\{1, \hat{\pi}_j^{(i)} / \left(1 - \hat{\pi}_j^{(i)}\right)\right\} \quad \text{and} \quad D_j^{(i)} = \zeta^{(i)} \min\left\{1, \left(1 - \hat{\pi}_j^{(i)}\right) / \hat{\pi}_j^{(i)}\right\}, \tag{11}$$

for $j = 1, \ldots, p$ where $0 < \zeta^{(i)} < 1$ is a tuning parameter and $\hat{\pi}_j^{(i)}$ is a Rao-Blackwellised estimate of the posterior inclusion probability of variable j at the i-th iteration. Like Ghosh and Clyde (2011), we work with the Rao-Blackwellised estimate conditional on the model, marginalizing out α, β_γ and σ^2, in contrast to Guan and Stephens (2011) who condition on the model parameters. We assume that $V_\gamma = gI_{p_\gamma}$, where I_q is the $q \times q$ identity matrix. After N posterior samples, $\gamma^{(1)}, \ldots, \gamma^{(N)}$, the Rao-Blackwellised estimate of $\pi_j = p(\gamma_j = 1|y)$ is

$$\hat{\pi}_j = \frac{1}{N} \sum_{k=1}^{N} \frac{\tilde{h}_j^{(k)} \, \mathrm{BF}_j\left(\gamma_{-j}^{(k)}\right)}{1 - \tilde{h}_j^{(k)} + \tilde{h}_j^{(k)} \, \mathrm{BF}_j\left(\gamma_{-j}^{(k)}\right)} \tag{12}$$

where $\tilde{h}_j^{(k)} = h$ if h is fixed or $\tilde{h}_j^{(k)} = \frac{\#\gamma_{-j}^{(k)} + 1 + a}{p + a + b}$ if $h \sim \mathrm{Be}(a, b)$. Let $Z_\gamma = [\mathbf{1}_n \; X_\gamma]$, $\Lambda_\gamma = \begin{pmatrix} 0 & \mathbf{0}_{p_\gamma}^T \\ \mathbf{0}_{p_\gamma} & V_\gamma^{-1} \end{pmatrix}$, $F = (Z_\gamma^T Z_\gamma + \Lambda_\gamma)^{-1}$ and $A = y^T y - y^T Z_\gamma F Z_\gamma^T y$. If $\gamma_j = 0$,

$$\mathrm{BF}_j(\gamma_{-j}) = {d_j^\uparrow}^{-1/2} g^{-1/2} \left(\frac{A - \frac{1}{d_j^\uparrow}(y^T x_j - y^T Z_\gamma F Z_\gamma^T x_j)^2}{A} \right)^{-n/2}$$

with $d_j^\uparrow = x_j^T x_j + g^{-1} - (x_j^T Z_\gamma) F (Z_\gamma^T x_j)$. If $\gamma_j = 1$, we define z_j to be ordered position of the included variables ($z_j = 1$ if j is the first included variable, etc.), then

$$\mathrm{BF}_j(\gamma_{-j}) = {d_j^\downarrow}^{-1/2} g^{-1/2} \left(\frac{A}{A + d_j^\downarrow (y^T Z_\gamma F_{\cdot, z_j+1})^2} \right)^{-n/2}$$

where $d_j^\downarrow = 1/F_{z_j+1, z_j+1}$. These results allow the contribution to the Rao-Blackwellised estimates for all values of j to be calculated in $O(p)$ operations at each iteration if the values of F and A (which are needed for calculating the marginal likelihood) are stored. Derivations are provided in Supplementary Material C. The value of $\zeta^{(i)}$ is updated using

$$\mathrm{logit}_\epsilon \zeta^{(i+1)} = \mathrm{logit}_\epsilon \zeta^{(i)} + \phi_i (a_i - \tau), \tag{13}$$

where τ is a targeted acceptance rate. We use $\epsilon = 0.1/p$ as in Algorithm 1. We shall see (in Lemma 2) that ASI also satisfies diminishing adaptation by verifying that the Rao-Blackwellised estimate in (12) evolves at the rate $1/i$ and reiterating the argument about inverse logit derivatives. 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 $\Delta^{(i+1)} = 2\sum_{j=1}^{p} \min\{\pi_j^{(i+1)}, 1 - \pi_j^{(i+1)}\}$. This ensures that the algorithm will propose to change at least one variable with high probability. The ASI algorithm is described in Algorithm 2 and we indicate its transition kernel at time i as $P^{\text{ASI}}_{\eta^{(i)}}$. We use $\kappa = 0.001$ to avoid the estimated probabilities becoming very small.

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

Update $\hat{\pi}_1^{(i+1)}, \ldots, \hat{\pi}_p^{(i+1)}$ as in (12) and set $\tilde{\pi}_j^{(i+1)} = \kappa + (1 - 2\kappa)\,\hat{\pi}_j^{(i+1)}$

Update $\zeta^{(i+1)}$ as in (13)

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)

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.

The target posterior specified in Section 2.1 on the model space Γ is

$$\pi_p(\gamma) = \pi_p(\gamma \mid y) \propto p(y|\gamma)p(\gamma) \tag{14}$$

with $p(y|\gamma)$ available analytically, and the vector of adaptive parameters at time i is

$$\eta^{(i)} = (A^{(i)}, D^{(i)}) \quad \in \quad [\epsilon, 1-\epsilon]^{2p} =: \Delta_\epsilon, \quad \text{with} \quad 0 < \epsilon < 1/2, \tag{15}$$

with the update strategies in Algorithm 1 or 2. $P_\eta(\gamma, \cdot)$ denotes the non-adaptive Markov chain kernel corresponding to a fixed choice of η. Under the dynamics of either algorithm,

for $S \subseteq \Gamma$ we have

$$\begin{aligned} P_\eta(\gamma, S) &= \mathbb{P}\Big[\gamma^{(i+1)} \in S \,\Big|\, \gamma^{(i)} = \gamma, \eta^{(i)} = \eta\Big] \\ &= \sum_{\gamma' \in S} q_\eta(\gamma, \gamma') a_\eta(\gamma, \gamma') + \mathbb{I}\{\gamma \in S\} \sum_{\gamma' \in \Gamma} q_\eta(\gamma, \gamma')\big(1 - a_\eta(\gamma, \gamma')\big). \quad (16) \end{aligned}$$

In the case of multiple chain acceleration, where L copies of the chain are run, the model state space becomes the product space and the current state of the algorithm at time i is $\gamma^{\otimes L,\,(i)} = (\gamma^{1,\,(i)}, \ldots, \gamma^{L,\,(i)}) \in \Gamma^L$. The single chain version corresponds to $L = 1$ and all results apply.

To assess ergodicity, we need to define the distribution of the adaptive algorithm at time i, and the associated total variation distance: for the l-th copy of the chain $\{\gamma^{l,(i)}\}_{i=0}^{\infty}$ and $S \subseteq \Gamma$ define

$$\begin{aligned} \mathcal{L}^{l,(i)}\big[(\gamma^l, \eta), S\big] &:= \mathbb{P}\Big[\gamma^{l,\,(i)} \in S \,\Big|\, \gamma^{l,\,(0)} = \gamma^l, \eta^{(0)} = \eta\Big], \quad \text{and} \\ T^l(\gamma^l, \eta, i) &:= \|\mathcal{L}^{l,(i)}\big[(\gamma^l, \eta), \cdot\big] - \pi_p(\cdot)\|_{TV} = \sup_{S \in \Gamma} |\mathcal{L}^{l,(i)}\big[(\gamma^l, \eta), S\big] - \pi_p(S)|. \end{aligned}$$

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

$$\text{ergodicity:} \qquad \lim_{i \to \infty} T^l(\gamma^l, \eta, i) = 0, \quad \text{for any } l = 1, \ldots, L; \qquad \text{and} \qquad (17)$$

$$\text{SLLN:} \qquad \frac{1}{L} \sum_{l=1}^{L} \frac{1}{k} \sum_{i=1}^{k} f(\gamma^{l,\,(i)}) \overset{k \to \infty}{\longrightarrow} \pi_p(f) \quad \text{almost surely, for any } f : \Gamma \to \mathbb{R} \qquad (18)$$

To this end we first establish the following lemmas.

Lemma 1 (Simultaneous Uniform Ergodicity) *The family of Markov chains defined by transition kernels P_η in* (16)*, targeting $\pi_p(\gamma)$ in* (14)*, is simultaneously uniformly ergodic for any $\epsilon > 0$ in* (15)*, and so is its multichain version. That is, for any $\delta > 0$ there exists $N = N(\delta, \epsilon) \in \mathbb{N}$, such that for any starting point $\gamma^{\otimes L} \in \Gamma^L$ and any parameter value $\eta \in \Delta_\epsilon$*

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

Lemma 2 (Diminishing Adaptation) *Recall the constant $1/2 \leq \lambda \leq 1$ defining the adaptation rate $\phi_i = O(i^{-\lambda})$ in* (9)*,* (10)*, or* (13)*, and the parameter $\kappa > 0$ in Algorithm 2. Then both algorithms: EIA and ASI satisfy diminishing adaptation. More precisely, their transition kernels satisfy*

$$\sup_{\gamma \in \Gamma} \|P^\bullet_{\eta^{(i+1)}}(\gamma, \cdot) - P^\bullet_{\eta^{(i)}}(\gamma, \cdot)\| \leq C i^{-\lambda}, \quad \textit{for some } C < \infty, \qquad (19)$$

where $\bullet$ stands for EIA or ASI.

Simultaneous uniform ergodicity together with diminishing adaptation leads to the following

Theorem 1 (Ergodicity and SLLN) *Consider the target $\pi_p(\gamma)$ of* (14)*, the constants $1/2 \leq \lambda \leq 1$ and $\epsilon > 0$ defining respectively the adaptation rate $\phi_i = O(i^{-\lambda})$ and region in* (9)*,* (10)*, or* (13)*, and the parameter $\kappa > 0$ in Algorithm 2. Then ergodicity* (17) *and the strong law of large numbers* (18) *hold for each of the algorithms: EIA, ASI and their multiple chain acceleration versions.*

Remark 3 *Lemma 2 and Theorem 1 remain true with any $\lambda > 0$, however $\lambda > 1$ results in finite adaptation (see e.g. Roberts and Rosenthal (2007)), and $\lambda < 1/2$ is rarely used in practice for finite sample stability concerns.*

Proofs can be found in Supplementary Material 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, if the prior distributions of additional parameters are continuous, supported on a compact set and everywhere positive, establishing ergodicity 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

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

where $e \sim \mathrm{N}(0, \sigma^2 I)$ 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 (EIA and ASI) relative to an add-delete-swap algorithm. We define the ratio of the relative time-standardized effective sample size of algorithm A versus algorithm B to be $r_{A,B} = (\mathrm{ESS}_A/t_A)/(\mathrm{ESS}_B/t_B)$ where ESS_A is the effective sample size for algorithm A. This is estimated by making 200

runs of each algorithm and calculating $\hat{r}_{A,B} = (s_B^2 t_B)/(s_A^2 t_A)$, where t_A and t_B are the median run-times and s_A^2 and s_B^2 are the sample variances of the posterior inclusion probabilities for algorithms A and B.

We use the prior in (2) with $V_\gamma = 9I$ and $h = 10/p$, implying a prior mean model size of 10. 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 and 8 for $p = 500$ and 3 and 0 for $p = 5000$. 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 EIA and ASI 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 ASI algorithm outperforms the

Table 1: Simulated data: 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 (EIA) or adaptively scaled individual adaptation (ASI) algorithm

		5 chains				25 chains			
		SNR				SNR			
(n, p)		0.5	1	2	3	0.5	1	2	3
(500, 500)	EIA	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)	EIA	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)	EIA	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)	EIA	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

EIA algorithm for most settings with either 5 or 25 multiple chains. The performance of the EIA 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 ASI 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 ASI algorithm (which has one tuneable parameter) over the EIA algorithm (which has $2p$ tuneable parameters) is due to the substantially faster convergence of the tuning parameters of the ASI 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_j are close to the corresponding posterior inclusion probabilities for both algorithms. However, the values of D_j are mostly close to 1 for ASI but not for EIA. 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 is consistent with the idealized super-efficient setting (ii) in Proposition 1 for $\pi_j < 0.5$ and leads to improved mixing rates for small π_j 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-Blackwellized estimates can often quickly determine which variables have low posterior inclusion probabilities.

4.2 Behaviour of the exploratory individual adaptation algorithm on the 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 runs of the EIA algorithm for this data, such as 5 multiple chains with 3000 burn in and 3000 recorded iterations, taking about 5 seconds on a laptop, show consistent convergence across runs.

Our purpose was to study the adaptive behaviour of the EIA 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 (without thinning). 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) in the Supplementary Material shows how the EIA algorithm approximates setting (ii) of Proposition 1, namely the super-efficient sampling from the idealized posterior (5). 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 overcompensates for dependence in the posterior, 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 4015 with EIA and 6673 with ASI. This is an impressive performance for both algorithms given the multicollinearity in the regressors. The difference in performance can be explained by the speed of convergence to optimal values for the proposal. To illustrate this, we re-ran the algorithms with the burn-in extended to 30 000 iterations: the effective sample per multiple chain was now 4503 with EIA but 6533 with ASI, indicating that the first algorithm had caught up somewhat. As a comparison, the effective sample size was 1555 for add-delete-swap and 15039 for the Hamming ball sampler with a burnin of 10 000 iterations. However, the Hamming ball sampler required 34 times the run time of the EIA sampler, rendering the latter nine times more efficient in terms of time-standardized effective sample size.

This example and the previous one show that the simplified posterior (5) is a good fit with many datasets and can indeed be used to guide and design algorithms.

4.3 Performance on problems with moderate p

We consider three more data sets with relatively small values for p (around 100) and high dependencies between the covariates, used to showcase the sequential Monte Carlo method proposed in Schäfer and Chopin (2013). They are the Boston Housing data ($n = 506$, $p = 104$), the concrete data ($n = 1030$, $p = 79$) and the protein data ($n = 96$, $p = 88$), which were constructed by Schäfer and Chopin (2013) to lead to challenging, multi-modal posterior mass functions. Further details about the data can be found in Schäfer and Chopin (2013). We focus here on the comparison of the ASI and the EIA algorithms with the add-delete-swap algorithm and the sequential Monte Carlo algorithm of Schäfer and Chopin (2013), also considering parallel tempering versions of the first three algorithms. In addition, we consider 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). Unlike Schäfer and Chopin (2013), we adopt the prior (2) with $V_\gamma = 100I$ and $h = 5/100$, while allowing for any combination of main effects and interactions. We use the method of Schäfer and Chopin (2013) for data visualization of the variation in the posterior marginal inclusion probabilities using boxplots. All algorithms were run for the same amount of time and we run them 200 times for each data set. For each variable, the white box contains the central 80% of results and the black boxes show the upper and lower 10% most extreme values. The coloured bars cover 0 up to the smallest recorded posterior inclusion probability across all runs. The results (also including other algorithms) are shown in the Supplementary Figure 4 (Boston housing), Figure 5 (concrete) and Figure 6 (protein). In the Supplementary material (Figure 7), we also include results for the Tecator data.

There are clear variations across the data sets with the protein and Tecator data leading to very consistent results whereas there were much greater variations in the inclusion probabilities for the Boston Housing and concrete data sets. Parallel tempering is most helpful for the ASI, a bit less so for the EIA algorithms while the add-delete-swap sampler only benefits somewhat from this modification.

The ASI, EIA, Schäfer-Chopin, Hamming Ball and add-delete-swap algorithms all provide similar levels of accuracy, whereas the parallel tempering versions of the ASI and add-delete-swap algorithms provide the most accurate results. This is likely due to the multimodality in the posterior distribution which is better addressed by parallel tempering than the annealing in the Schäfer-Chopin algorithm. For all cases, the Ji-Schmidler sampler performs the worst by some margin.

4.4 Performance on problems with very large p

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) three physiological phenotypes are recorded, and used as the response variable in the three data sets called PCR$i, i = 1, \ldots, 3$. We use prior (2) with $V_\gamma = gI$ where g is given a half-Cauchy hyper-prior distribution 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.

A fourth data set (SNP data) relates to genome-wide mapping of a complex trait (Carbonetto et al., 2017). The data are body and weight measurements for 993 outbred mice and 79 748 single nucleotide polymorphisms (SNPs) recorded for each mouse. The testis weight is the response, the body weight is a regressor which is always included in the model and variable selection is performed on the 79 748 SNPs. The high dimensionality makes this a difficult problem and Carbonetto et al. (2017) use a variational inference algorithm (varbvs)

for their analysis. We have used various prior specifications in (2), and present results for a half-Cauchy hyper-prior on g and $h = 5/p$.

For all four datasets, the individual adaptation algorithms were run with $\tau_L = 0.05$ and $\tau_U = 0.23$, and $\tau = 0.234$. The EIA algorithm had a burn-in of 2 150 iterations and 10 750 subsequent iterations and no thinning, and the ASI had 500 burn-in and 2 500 recorded iterations and no thinning (giving very similar run times). 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.

In addition, we use the add-delete-swap algorithm, with starting models chosen from the prior as well as a version of this algorithm started in the model suggested by the least absolute shrinkage and selection operator, the Hamming Ball sampler with radius 1 and the Schäfer-Chopin algorithm. Three independent runs of all algorithms were executed to gauge the degree of agreement across runs. Using MATLAB and an Intel i7 @ 3.60 GHz processor, each algorithm took approximately 25 minutes to run for the PCR data and around 2.5 hours for the SNP data.

Figures 8-14 in Supplementary Material E show the pairwise comparisons between the different runs for all data sets. The estimates from each independent chain for the ASI algorithm and for its parallel tempered version are very similar and indicate that the sampler is able to accurately represent the posterior distribution. The EIA algorithm does not seem to converge rapidly enough to effectively deal with these very high-dimensional model spaces in the relatively modest running time allocated. Clearly, the other samplers are not able to adequately characterise the posterior model distribution with runs leading to dramatically different results, especially for the PCR data. For the SNP data, the add-delete-swap method does not do too badly, but provides substantially more variable estimates of the posterior inclusion probabilities than the ASI method. Starting the add-delete-swap algorithm in the model selected by the least absolute shrinkage and selection operator never helps, and can actually harm the performance.

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 genetic data with 22 576 and 79 748 variables and shows excellent

agreement in the posterior inclusion probabilities across independent runs of the algorithm, unlike the existing methods we have tried. We find that multiple independent chains with a shared proposal lead to better convergence to the optimal parameter values and parallel tempering helps to deal with multimodal posteriors. For smaller data sets (say $p < 500$), the exploratory individual adaptation algorithm also performs very well. Code to run both algorithms is available from
`https://warwick.ac.uk/go/msteel/steel_homepage/software/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. 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.

Acknowledgements

KŁ acknowledges support of the Royal Society through the Royal Society University Research Fellowship and of EPSRC. The authors thank two anonymous referees and an associate editor for their insightful comments that helped improve the paper.

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.

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.

Carbonetto, P., X. Zhou, and M. Stephens (2017). varbvs: Fast variable selection for large-scale regression. Technical report.

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

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

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.

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.

Fort, G., E. Moulines, and P. Priouret (2011). Convergence of adaptive and interacting Markov chain Monte Carlo algorithms. *Ann. Statist. 39*, 3262–3289.

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.

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.

Hans, C., A. Dobra, and M. West (2007). Shotgun stochastic search for "large p" regression. *Journal of the American Statistical Association 102*, 507–516.

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.

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

Łatuszyński, K., G. O. Roberts, and J. S. Rosenthal (2013). Adaptive Gibbs samplers and related MCMC methods. *The Annals of Applied Probability 23*, 66–98.

Lee, C. and P. J. Neal (2018). Optimal scaling of the independence sampler: theory and practice. *Bernoulli 24*, 1636–1652.

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.

Miasojedow, B., E. Moulines, and M. Vihola (2013). An adaptive parallel tempering algorithm. *Journal of Computational and Graphical Statistics 22*, 649–664.

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.

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. *Biometrika 104*, 343–359.

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.

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.

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

Titsias, M. K. and C. Yau (2017). The Hamming ball sampler. *Journal of the American Statistical Association 112*(520), 1598–1611.

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.

Zanella, G. and G. O. Roberts (2019). Scalable importance tempering and Bayesian variable selection. *Journal of the Royal Statistical Society, Series B 81*, forthcoming.

A Proofs

A.1 Proof 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 $\{\gamma_j^{(t)}\}_{t=0,1,\dots}$ 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 the first part of (c) are immediate because the proposal samples from the stationary distribution and is accepted with probability 1.
To verify the second part of (c), use that individual coordinates are Markovian, and for $f \in \mathbb{L}(\Gamma)$ compute:

$$\sigma^2_{P,f} = \sum_{j=1}^{p} a_j^2 \sigma^2_{P,f_j} = \sum_{j=1}^{p} a_j^2 \sigma^2_{P_j,f_j}. \tag{20}$$

Now recall that P_j in (i) is independent sampling from $\pi^{(j)}$, i.e. $P_j = \Pi_j := \begin{bmatrix} 1-\pi_j & \pi_j \\ 1-\pi_j & \pi_j \end{bmatrix}$, hence $\sigma^2_{P_j,f_j} = Var_{\pi^{(j)}} f_j$.
To verify (b), note that

$$E_\pi[\Delta^2] = E_\pi \left[\sum_{j=1}^{p} |\gamma_j^{(0)} - \gamma_j^{(1)}|^2 \right] = \sum_{j=1}^{p} E_\pi |\gamma_j^{(0)} - \gamma_j^{(1)}|, \tag{21}$$

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 (21), and it is enough to check (by simple algebra) that the transition kernel $P_j := \begin{bmatrix} 1-A_j & A_j \\ D_j & 1-D_j \end{bmatrix}$, 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 (a), recall (21), and note that $E_\pi|\gamma_j^{(0)} - \gamma_j^{(1)}| = 2\min\{1-\pi_j, \pi_j\}$.
For Peskun optimality in (b), recall formula (20) and consider σ_{P_j,f_j} in this setting. It is enough to verify that for each j, the kernel $P_j = \begin{bmatrix} 1-A_j & A_j \\ D_j & 1-D_j \end{bmatrix}$ 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 (b), recall (20), 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

$$\Pi_j = \max\{1-\pi_j, \pi_j\} P_j + (1 - \max\{1-\pi_j, \pi_j\}) I_2.$$

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 (20) concludes the proof.

A.2 Proof of Proposition 2

$$
\begin{aligned}
&p(\gamma_j = 1|\gamma_k = 1, \gamma_0, y) - p(\gamma_j = 1|\gamma_k = 0, \gamma_0, y) \\
=&\frac{\text{BF}(\gamma_j = 1|\gamma_k = 1, \gamma_0)h}{(1-h) + \text{BF}(\gamma_j = 1|\gamma_k = 1, \gamma_0)h} - \frac{\text{BF}(\gamma_j = 1|\gamma_k = 0, \gamma_0)h}{(1-h) + \text{BF}(\gamma_j = 1|\gamma_k = 0, \gamma_0)h} \\
=&\frac{a\text{BF}(\gamma_j = 1|\gamma_k = 0, \gamma_0)h}{(1-h) + a\text{BF}(\gamma_j = 1|\gamma_k = 0, \gamma_0)h} - \frac{\text{BF}(\gamma_j = 1|\gamma_k = 0, \gamma_0)h}{(1-h) + \text{BF}(\gamma_j = 1|\gamma_k = 0, \gamma_0)h} \\
=&\text{BF}(\gamma_j = 1|\gamma_k = 0, \gamma_0)h \left[\frac{a}{(1-h) + a\text{BF}(\gamma_j = 1|\gamma_k = 0, \gamma_0)h} - \frac{1}{(1-h) + \text{BF}(\gamma_j = 1|\gamma_k = 0, \gamma_0)h} \right]
\end{aligned}
$$

This converges to 0 if (i) $a \to 1$ or if (ii) $a \to A < \infty$ and $\text{BF}(\gamma_j = 1|\gamma_k = 0, \gamma_0)h \to 0$.

A.3 Proof of Lemma 1

To verify the result it is enough to check that the whole state space Γ^L 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 $\eta \in \Delta_\epsilon$ and every $\gamma^{\otimes L}, \gamma'^{\otimes L} \in \Gamma^L$ we have

$$P_\eta(\gamma^{\otimes L}, \gamma'^{\otimes L}) \geq b \tag{22}$$

where $P_\eta(\gamma^{\otimes L}, \gamma'^{\otimes L})$ is the transition for the L chains. Decompose the move into proposal and acceptance

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

and notice that by the proposal construction

$$
\begin{aligned}
q_\eta(\gamma^{\otimes L}, \gamma'^{\otimes L}) &\geq \epsilon^{pL} \quad \text{for EIA;} \\
q_\eta(\gamma^{\otimes L}, \gamma'^{\otimes L}) &\geq (\kappa\epsilon)^{pL} \quad \text{for adaptively scaled individual adaptation,}
\end{aligned}
$$

since $\text{diam}(\Gamma^L) = pL$. Similarly for the acceptance term, since $\pi^{\otimes L}(\gamma_1^{\otimes L}) q_\eta(\gamma_1^{\otimes L}, \gamma_2^{\otimes L}) \leq 1$ for every $\gamma_1^{\otimes L}, \gamma_2^{\otimes L} \in \Gamma^L$, we can write

$$
\begin{aligned}
a_\eta(\gamma^{\otimes L}, \gamma'^{\otimes L}) &= \min\left\{1, \frac{\pi_p^{\otimes L}(\gamma'^{\otimes L}) q_\eta(\gamma'^{\otimes L}, \gamma^{\otimes L})}{\pi_p^{\otimes L}(\gamma^{\otimes L}) q_\eta(\gamma^{\otimes L}, \gamma'^{\otimes L})}\right\} \\
&\geq \pi_p^{\otimes L}(\gamma'^{\otimes L}) q_\eta(\gamma'^{\otimes L}, \gamma^{\otimes L}) \;\geq\; \pi_m^L \times (\kappa\epsilon)^{pL},
\end{aligned}
$$

where $\pi_m := \min_{\gamma \in \Gamma} \pi_p(\gamma)$. Consequently in (22) we can take

$$b = \pi_m^L \times (\kappa\epsilon)^{2pL},$$

and we have established simultaneous uniform ergodicity.

A.4 Proof of Lemma 2

First, observe that if we prove that the proposals of EIA or adaptively scaled individual adaptation satisfy the following diminishing property

$$\sup_{\gamma\in\Gamma} \|q_{\eta^{(i+1)}}(\gamma,\cdot) - q_{\eta^{(i)}}(\gamma,\cdot)\| \leq C a_i \to 0, \tag{23}$$

then by Lemma 4.21(ii) of Łatuszyński et al. (2013) (and precisely by inspection of their proof), also the transition kernels satisfy

$$\sup_{\gamma\in\Gamma} \|P_{\eta^{(i+1)}}(\gamma,\cdot) - P_{\eta^{(i)}}(\gamma,\cdot)\| \leq C_1 a_i \to 0, \tag{24}$$

for some C_1.

To establish (23), recall the proposal form (3), and compute

$$\begin{aligned}
\sup_{\gamma\in\Gamma} \|q_{\eta^{(i+1)}}(\gamma,\cdot) - q_{\eta^{(i)}}(\gamma,\cdot)\| &= \frac{1}{2}\sup_{\gamma\in\Gamma}\left\{\sum_{\gamma'\in\Gamma} |q_{\eta^{(i+1)}}(\gamma,\gamma') - q_{\eta^{(i)}}(\gamma,\gamma')|\right\} \\
&= \frac{1}{2}\sup_{\gamma\in\Gamma}\left\{\sum_{\gamma'\in\Gamma}\left|\prod_{j=1}^{p} q_{\eta^{(i+1)},j}(\gamma_j,\gamma'_j) - \prod_{j=1}^{p} q_{\eta^{(i)},j}(\gamma_j,\gamma'_j)\right|\right\} \\
&\leq \frac{1}{2}\sup_{\gamma\in\Gamma}\left\{\sum_{\gamma'\in\Gamma}\sum_{j=1}^{p}\left|q_{\eta^{(i+1)},j}(\gamma_j,\gamma'_j) - q_{\eta^{(i)},j}(\gamma_j,\gamma'_j)\right|\right\} =: \spadesuit_1, \quad (25)
\end{aligned}$$

where the last inequality follows from $|\prod_{j=1}^p a_j - \prod_{j=1}^p b_j| \leq \sum_{j=1}^p |a_j - b_j|$ for $a_j, b_j \in [0,1]$. From

$$q_{\eta^{(i+1)},j}(\gamma_j,\gamma'_j) = \left(A_j^{(i+1)}\right)^{(1-\gamma_j)\gamma'_j}\left(1-A_j^{(i+1)}\right)^{(1-\gamma_j)(1-\gamma'_j)}\left(D_j^{(i+1)}\right)^{\gamma_j(1-\gamma'_j)}\left(1-D_j^{(i+1)}\right)^{\gamma_j\gamma'_j},$$

we obtain

$$\begin{aligned}
\spadesuit_1 &\leq \frac{1}{2}\sup_{\gamma\in\Gamma}\left\{\sum_{\gamma'\in\Gamma}\sum_{j=1}^{p}\max\left\{|A_j^{(i+1)} - A_j^{(i)}|, |D_j^{(i+1)} - D_j^{(i)}|\right\}\right\} \\
&\leq \frac{1}{2}\left\{\sum_{\gamma'\in\Gamma}\sum_{j=1}^{p}\max\left\{\max_j\left\{|A_j^{(i+1)} - A_j^{(i)}|\right\}, \max_j\left\{|D_j^{(i+1)} - D_j^{(i)}|\right\}\right\}\right\} \\
&\leq p2^{p-1}\max\left\{\max_j\left\{|A_j^{(i+1)} - A_j^{(i)}|\right\}, \max_j\left\{|D_j^{(i+1)} - D_j^{(i)}|\right\}\right\} \\
&= C\max\left\{\max_j\left\{|A_j^{(i+1)} - A_j^{(i)}|\right\}, \max_j\left\{|D_j^{(i+1)} - D_j^{(i)}|\right\}\right\} =: \spadesuit_2.
\end{aligned}$$

For EIA the difference $|A_j^{(i+1)} - A_j^{(i)}|$ comes from the logit$_\epsilon$ update (9), and analogously the difference $|D_j^{(i+1)} - D_j^{(i)}|$ from the logit$_\epsilon$ update (10). Recall that logit$_\epsilon : (\epsilon, 1-\epsilon) \to \mathbb{R}$

and noticing $\frac{\partial \text{logit}_\epsilon(x)}{\partial x} > 4$, yields $\frac{\partial \text{logit}_\epsilon^{-1}(y)}{\partial y} < 1/4$. Consequently, updating the logit_ϵ by $\phi_i = O(i^{-\lambda})$ is equivalent to updating $A_j^{(i)}$ or $D_j^{(i)}$ by a term of at most the same order $O(i^{-\lambda})$. Hence for EIA

$$\spadesuit_2 \;=\; O(i^{-\lambda}). \tag{26}$$

For adaptively scaled individual adaptation recall (11) and its use in Algorithm 2. As the components are in $[0,1]$, we apply the triangle inequality to obtain:

$$|A_j^{(i+1)} - A_j^{(i)}| \;\leq\; |\zeta^{(i+1)} - \zeta^{(i)}| + \left|\min\left\{1, \frac{\tilde{\pi}_j^{(i+1)}}{1-\tilde{\pi}_j^{(i+1)}}\right\} - \min\left\{1, \frac{\tilde{\pi}_j^{(i)}}{1-\tilde{\pi}_j^{(i)}}\right\}\right| \tag{27}$$

$$|D_j^{(i+1)} - D_j^{(i)}| \;\leq\; |\zeta^{(i+1)} - \zeta^{(i)}| + \left|\min\left\{1, \frac{1-\tilde{\pi}_j^{(i+1)}}{\tilde{\pi}_j^{(i+1)}}\right\} - \min\left\{1, \frac{1-\tilde{\pi}_j^{(i)}}{\tilde{\pi}_j^{(i)}}\right\}\right| \tag{28}$$

The update equation for $\zeta^{(i)}$ is (13), hence

$$|\zeta^{(i+1)} - \zeta^{(i)}| = O(i^{-\lambda}) \tag{29}$$

by the same argument that led to (26). The term $\tilde{\pi}^{(i)}$ introduced in Algorithm 2 is

$$\tilde{\pi}_j^{(i)} = \kappa + (1-2\kappa)\hat{\pi}_j^{(i)} \tag{30}$$

where $\hat{\pi}_j^{(i)}$ is the Rao-Blackwellised estimate of $\pi_j = p(\gamma_j = 1|y)$ defined in (12). It remains to show that the second terms in (27) and (28) are at most $O(i^{-\lambda})$. We shall see that the terms are $O(i^{-1})$. To this end, first note that $\hat{\pi}_j^{(i)}$ using posterior samples $\gamma^{(1)}, \ldots, \gamma^{(i)}$ is

$$\hat{\pi}_j^{(i)} = \frac{1}{i}\sum_{k=1}^{i} p\left(\gamma_j = 1 \,\middle|\, \gamma_{-j}^{(k)}, y\right).$$

and therefore

$$\begin{aligned} |\hat{\pi}_j^{(i+1)} - \hat{\pi}_j^{(i)}| \;&=\; \left|\frac{i}{i+1}\hat{\pi}_j^{(i)} + \frac{1}{i+1} p\left(\gamma_j = 1 \,\middle|\, \gamma_{-j}^{(i+1)}, y\right) - \hat{\pi}_j^{(i)}\right| \\ &\leq\; \left|\frac{i}{i+1}\hat{\pi}_j^{(i)} - \hat{\pi}_j^{(i)}\right| + \left|\frac{1}{i+1} p\left(\gamma_j = 1 \,\middle|\, \gamma_{-j}^{(i+1)}, y\right)\right| \\ &\leq\; \frac{2}{i+1}. \end{aligned} \tag{31}$$

Next, consider the function $f_\kappa(x) = \frac{\kappa + (1-2\kappa)x}{\kappa + (1-2\kappa)(1-x)}$ and compute its derivative to see that it is Lipshitz with constant $\frac{1}{\kappa}$. Consequently, so is $g_\kappa(x) := \min\{1, f_\kappa(x)\}$, hence

$$\begin{aligned} \left|\min\left\{1, \frac{\tilde{\pi}_j^{(i+1)}}{1-\tilde{\pi}_j^{(i+1)}}\right\} - \min\left\{1, \frac{\tilde{\pi}_j^{(i)}}{1-\tilde{\pi}_j^{(i)}}\right\}\right| \;&=\; |g_\kappa(\hat{\pi}_j^{(i+1)}) - g_\kappa(\hat{\pi}_j^{(i)})| \;\leq\; \frac{1}{\kappa}|\hat{\pi}_j^{(i+1)} - \hat{\pi}_j^{(i)}| \\ &\leq\; \frac{2}{\kappa}\frac{1}{i+1} \;=\; O(i^{-1}). \end{aligned} \tag{32}$$

Combining (29) and (32) shows that the right-hand side of (27) is $O(i^{-\lambda})$ and a symmetric reasoning yields the result for (28), which completes the proof.

A.5 Proof of Theorem 1

We shall use Theorem 1 of Roberts and Rosenthal (2007) to obtain ergodicity (17). This requires simultaneous uniform ergodicity of the transition kernels, established in Lemma 1 and diminishing adaptation. Lemma 2 verifies diminishing adaptation in single chain implementations. For multiple chain implementations there will be up to L updates of the adaptive parameter between consecutive updates of chain number l, hence applying the triangle inequality and Lemma 2 yields diminishing adaptation with the same rate (and a constant multiplied by L).

The strong law of large numbers (18) will be demonstrated by first establishing it for each chain in the multiple chain implementation separately, i.e.

$$\frac{1}{k}\sum_{i=1}^{k} f(\gamma^{l,(i)}) \quad \stackrel{k\to\infty}{\longrightarrow} \quad \pi(f) \quad \text{almost surely,} \tag{33}$$

and then combining it into (18) by averaging the L chains. To obtain (33) we use Corollary 2.8 following Theorem 2.7 in Fort et al. (2011). This requires establishing the following conditions:

C1. (Drift condition A3 in Fort et al. (2011)) There exists a function $V : \Gamma \to [1, +\infty)$ and for any η there exist some constants $b_\eta < \infty$, $\delta_\eta \in (0, 1)$, $\lambda_\eta \in (0, 1)$ and a probability measure ν_η on Γ, such that

$$\begin{aligned} P_\eta V &\leq \lambda_\eta V + b_\eta; \\ P_\eta(\gamma, \cdot) &\geq \delta_\eta \nu_\eta(\cdot)\mathbb{I}\{V \leq c_\eta\}(\gamma), \qquad c_\eta := 2b_\eta(1-\lambda_\eta)^{-1} - 1. \end{aligned}$$

Indeed, it is immediate to check that the above condition is met with $V \equiv 1$, $\lambda_\eta = 1/2$, $b_\eta = 1$, the measure ν_η uniform on Γ and, following the proof of Lemma 1, with $\delta_\eta = \pi_m(\kappa\epsilon)^{2p}$, where $\pi_m := \min_{\gamma\in\Gamma} \pi_p(\gamma)$.

C2. (Condition A4 in Fort et al. (2011) after specialising to the parameters in C1 above)

$$\sum_{i=1}^{\infty} i^{-1} \sup_{\gamma\in\Gamma} \|P_{\eta^{(i+1)}}(\gamma, \cdot) - P_{\eta^{(i)}}(\gamma, \cdot)\| \quad < \quad \infty.$$

The condition indeed holds for EIA, adaptively scaled individual adaptation, and their multiple chain implementations, by the diminishing adaptation rate established in Lemma 2.

C3. Condition A5 in Fort et al. (2011) which is trivially satisfied with the parameters chosen in C1.

Since we have established C1, C2, and C3, by Corollary 2.8 in Fort et al. (2011), (33) holds, and so does (18).

B Parallel Tempering implementations

We will consider an adaptive parallel tempering scheme to avoid the algorithms becoming trapped in very well separated modes. A sequence of distributions $\pi_1, \ldots, \pi_m$ are constructed with

$$\pi_k(\gamma|y) \propto p(y|\gamma)^{t_k}\pi(\gamma), \qquad k = 1, \ldots, m$$

where the parameters $0 < t_1 < t_2 < \cdots < t_m = 1$ are referred to as temperatures (with smaller t_j referring to higher temperatures). The density $\pi_m(\gamma|y)$ is the posterior density $p(\gamma|y)$ of interest. The distribution becomes flatter at higher temperatures. The sequence of posterior distribution can be sampled using Markov chain Monte Carlo methods by defining the joint target for $\gamma^\star = (\gamma_1^\star, \ldots, \gamma_m^\star)$, $\pi(\gamma^\star|y) = \prod_{k=1}^m \pi_k(\gamma_k^\star|y)$ where $\pi_k(\gamma_k^\star|y) \propto p(y|\gamma_k^\star)^{t_k}\pi(\gamma_k^\star)$. The Markov chain Monte Carlo algorithm uses two types of moves. Firstly, $\gamma_k^\star$ is updated for all values of k. Secondly, a Metropolis-Hastings update proposes to swap $\gamma_k^\star$ with $\gamma_{k+1}^\star$ for k chosen uniformly at random from $\{1, \ldots, m-1\}$. The temperature schedule is chosen adaptively using the method proposed by Miasojedow et al. (2013). As the temperature increases (as $t_k \to 0$) , the dependence becomes weaker between γ_i and γ_j under $\pi_k(\gamma_k^\star|y)$.

C Derivation of Rao-Blackwellised estimate of π_j

The Rao-Blackwellised estimate of $\pi_j = p(\gamma_j = 1|y)$ using posterior samples $\gamma^{(1)}, \ldots, \gamma^{(N)}$ is

$$\hat{\pi}_j = \frac{1}{N}\sum_{k=1}^{N} p\left(\gamma_j = 1 \left| \gamma_{-j}^{(k)}, y \right.\right).$$

We know that

$$p(\gamma|y) \propto p(y|\gamma)p(\gamma)$$

where

$$p(y|\gamma) = |Z_\gamma^T Z_\gamma + \Lambda_\gamma|^{-1/2} g^{-p_\gamma/2} \left(y^T y - y^T Z_\gamma (Z_\gamma^T Z_\gamma + \Lambda_\gamma)^{-1} Z_\gamma^T y\right)^{-n/2} \equiv m(\gamma)$$

and $\Lambda_\gamma = \text{diag}(0, \overbrace{g^{-1}, \ldots, g^{-1}}^{p_\gamma\text{-times}})$. Then, if h is fixed,

$$p(\gamma_j = 1|\gamma_{-j}, y) = \frac{hm(\gamma_j^\uparrow)}{(1-h)m(\gamma_j^\downarrow) + hm(\gamma_j^\uparrow)} = \frac{h\text{BF}_j(\gamma_{-j})}{1-h+h\text{BF}_j(\gamma_{-j})}$$

where $\gamma_j^\uparrow$ represents γ_{-j} and $\gamma_j = 1$, $\gamma_j^\downarrow$ represents γ_{-j} and $\gamma_j = 0$, and $\text{BF}_j(\gamma_{-j}) = m(\gamma_j^\uparrow)/m(\gamma_j^\downarrow)$ is the Bayes factor in favour of including the j-th variable given γ_{-j}. If $h \sim \text{Be}(a, b)$,

$$p(\gamma_j = 1|\gamma_{-j}, y) = \frac{h^\star m(\gamma_j^\uparrow)}{(1-h^\star)m(\gamma_j^\downarrow) + h^\star m(\gamma_j^\uparrow)} = \frac{h^\star\text{BF}_j(\gamma_{-j})}{1-h^\star+h^\star\text{BF}_j(\gamma_{-j})}$$

where $h^\star = \frac{\#\gamma_{-j}+1+a}{p+a+b}$.

Therefore, calculating the Rao-Blackwellised estimates reduces to calculating the Bayes factors $\text{BF}_j(\gamma_{-j})$ using values calculated in the MCMC chain. We distinguish the two cases, $\gamma_j = 0$ and $\gamma_j = 1$ in the current state of the chain.

If $\gamma_j = 0$, we define $Z_{\gamma_j^\uparrow} = [Z_\gamma\ x_j]$ and the Bayes factor is

$$\begin{aligned}\text{BF}_j(\gamma_{-j}) &= \frac{|Z_{\gamma_j^\uparrow}^T Z_{\gamma_j^\uparrow} + \Lambda_{\gamma_j^\uparrow}|^{-1/2}}{|Z_\gamma^T Z_\gamma + \Lambda_\gamma|^{-1/2}}\, g^{-1/2} \left(\frac{y^T y - y^T Z_{\gamma_j^\uparrow}(Z_{\gamma_j^\uparrow}^T Z_{\gamma_j^\uparrow} + \Lambda_{\gamma_j^\uparrow})^{-1} Z_{\gamma_j^\uparrow}^T y}{y^T y - y^T Z_\gamma (Z_\gamma^T Z_\gamma + \Lambda_\gamma)^{-1} Z_\gamma^T y}\right)^{-n/2} \\ &= \frac{|Z_{\gamma_j^\uparrow}^T Z_{\gamma_j^\uparrow} + \Lambda_{\gamma_j^\uparrow}|^{-1/2}}{|Z_\gamma^T Z_\gamma + \Lambda_\gamma|^{-1/2}}\, g^{-1/2} \left(\frac{y^T y - y^T Z_{\gamma_j^\uparrow} \tilde{F} Z_{\gamma_j^\uparrow}^T y}{y^T y - y^T Z_\gamma F Z_\gamma^T y}\right)^{-n/2}\end{aligned}$$

where $F = (Z_\gamma^T Z_\gamma + \Lambda_\gamma)^{-1}$ and $\tilde{F} = \left(Z_{\gamma_j^\uparrow}^T Z_{\gamma_j^\uparrow} + \Lambda_{\gamma_j^\uparrow}\right)^{-1}$. The standard formulae for the Schur complement show that

$$\left|Z_{\gamma_j^\uparrow}^T Z_{\gamma_j^\uparrow} + \Lambda_{\gamma_j^\uparrow}\right| = \left|Z_\gamma^T Z_\gamma + \Lambda_\gamma\right| d_j^\uparrow \tag{34}$$

and

$$\tilde{F} = \left(Z_{\gamma_j^\uparrow}^T Z_{\gamma_j^\uparrow} + \Lambda_{\gamma_j^\uparrow}\right)^{-1} = \begin{pmatrix} F + F(Z_\gamma^T x_j)(x_j^T Z_\gamma)F/d_j^\uparrow & -F(Z_\gamma^T x_j)/d_j^\uparrow \\ -x_j^T Z_\gamma F/d_j^\uparrow & 1/d_j^\uparrow \end{pmatrix} \tag{35}$$

where $d_j^\uparrow = x_j^T x_j + g^{-1} - (x_j^T Z_\gamma)F(Z_\gamma^T x_j)$. The result follows from the application of (C1) and (C2) to the expression for $\text{BF}_j(\gamma_{-j})$ and noticing that

$$y^T Z_{\gamma_j^\uparrow} \tilde{F} Z_{\gamma_j^\uparrow}^T y = y^T Z_\gamma F Z_\gamma^T y + \frac{1}{d_j^\uparrow}\left(y^T Z_\gamma F(Z_\gamma^T x_j)(x_j^T Z_\gamma) F Z_\gamma^T y - 2y^T Z_\gamma F Z_\gamma^T x_j x_j^T y + y^T x_j x_j^T y\right).$$

Similarly, if $\gamma_j = 1$, we can write $Z_\gamma = [Z_{\gamma_j^\downarrow}\ x_j]$

$$\mathrm{BF}_j(\gamma_{-j}) = \frac{|Z_\gamma^T Z_\gamma + \Lambda_\gamma|^{-1/2}}{|Z_{\gamma_j^\downarrow}^T Z_{\gamma_j^\downarrow} + \Lambda_{\gamma_j^\downarrow}|^{-1/2}}\, g^{-1/2} \left(\frac{y^T y - y^T Z_\gamma (Z_\gamma^T Z_\gamma + \Lambda_\gamma)^{-1} Z_\gamma^T y}{y^T y - y^T Z_{\gamma_j^\downarrow} (Z_{\gamma_j^\downarrow}^T Z_{\gamma_j^\downarrow} + \Lambda_{\gamma_j^\downarrow})^{-1} Z_{\gamma_j^\downarrow}^T y} \right)^{-n/2}.$$

Again, the standard formulae for the Schur complement show that

$$\left|Z_\gamma^T Z_\gamma + \Lambda_\gamma\right| = \left| Z_{\gamma_j^\downarrow}^T Z_{\gamma_j^\downarrow} + \Lambda_{\gamma_j^\downarrow} \right| d_j^\downarrow \tag{36}$$

and

$$F = \left(Z_\gamma^T Z_\gamma + \Lambda_\gamma\right)^{-1} = \begin{pmatrix} \tilde{F} + \tilde{F}(Z_{\gamma_j^\downarrow}^T x_j)(x_j^T Z_{\gamma_j^\downarrow})\tilde{F}/d_j^\downarrow & -\tilde{F}(Z_{\gamma_j^\downarrow}^T x_j)/d_j^\downarrow \\ -x_j^T Z_{\gamma_j^\downarrow}\tilde{F}/d_j^\downarrow & 1/d_j^\downarrow \end{pmatrix} \tag{37}$$

where $\tilde{F} = (Z_{\gamma_j^\downarrow}^T Z_{\gamma_j^\downarrow} + \Lambda_{\gamma_j^\downarrow})^{-1}$ and $d_j^\downarrow = x_j^T x_j + g^{-1} - (x_j^T Z_{\gamma_j^\downarrow})\tilde{F}(Z_{\gamma_j^\downarrow}^T x_j)$. We have that $\tilde{F} = F_{1:p_\gamma,1:p_\gamma} - F_{1:\gamma,p_\gamma+1} F_{p_\gamma+1,1:p_\gamma} \frac{1}{F_{p_\gamma+1,p_\gamma+1}}$ and $d_j^\downarrow = \frac{1}{F_{p_\gamma+1,p_\gamma+1}}$. Using

$$y^T Z_{\gamma_j^\downarrow} F_{1:p_\gamma,1:p_\gamma} Z_{\gamma_j^\downarrow} y = y^T Z_\gamma F Z_\gamma^T y - y^T x_j F_{p_\gamma+1,p_\gamma+1} x_j y - 2\, y^T Z_{\gamma_j^\downarrow} F_{1:p_\gamma,p_\gamma+1} x_j^T y$$

we can write

$$\begin{aligned} y^T Z_{\gamma_j^\downarrow} (Z_{\gamma_j^\downarrow}^T Z_{\gamma_j^\downarrow} + \Lambda_{\gamma_j^\downarrow})^{-1} Z_{\gamma_j^\downarrow}^T y &= y^T Z_{\gamma_j^\downarrow} \tilde{F} Z_{\gamma_j^\downarrow} y \\ &= y^T Z_\gamma F Z_\gamma^T y - y^T x_j \frac{1}{d_j^\downarrow} x_j y - 2\, y^T Z_{\gamma_j^\downarrow} F_{1:p_\gamma,p_\gamma+1} x_j^T y \\ &\quad - y^T Z_{\gamma_j^\downarrow} F_{1:p_\gamma,p_\gamma+1} F_{p_\gamma+1,1:p_\gamma} d_j^\downarrow Z_{\gamma_j^\downarrow} y^T \\ &= y^T Z_\gamma F Z_\gamma^T y - (a_j + b_j)^2 \end{aligned}$$

where

$$a_j = y^T x_j (d_j^\downarrow)^{-1/2}$$

$b_j = y^T Z_{\gamma_j^\downarrow} {d_j^\downarrow}^{1/2} F_{1:p_\gamma,p_\gamma+1} = {d_j^\downarrow}^{1/2} \left(y^T Z_\gamma F_{\cdot,p_\gamma+1} - y^T x_j F_{p_\gamma+1,p_\gamma+1}\right) = {d_j^\downarrow}^{1/2} y^T Z_\gamma F_{\cdot,p_\gamma+1} - y^T x_j (d_j^\downarrow)^{-1/2}$.
This sum simplifies to

$$a_j + b_j = y^T x_j (d_j^\downarrow)^{-1/2} + {d_j^\downarrow}^{1/2} y^T Z_\gamma F_{\cdot,p_\gamma+1} - y^T x_j (d_j^\downarrow)^{-1/2} = {d_j^\downarrow}^{1/2} y^T Z_\gamma F_{\cdot,p_\gamma+1}.$$

Applying this with (C3) to $\mathrm{BF}_j(\gamma_{-j})$ leads to the result.

D Additional graphical material for Section 2

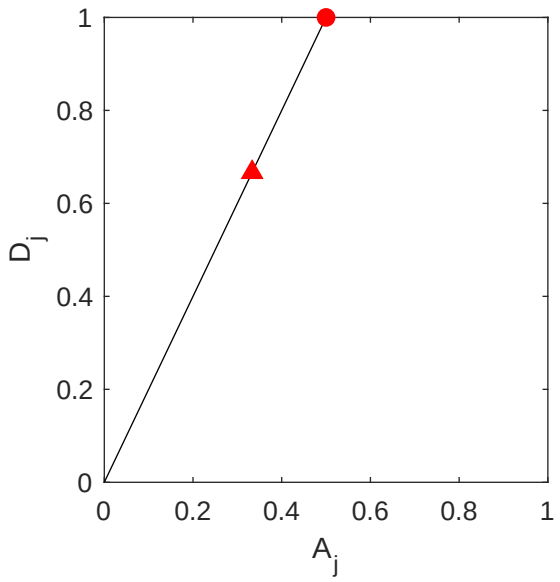
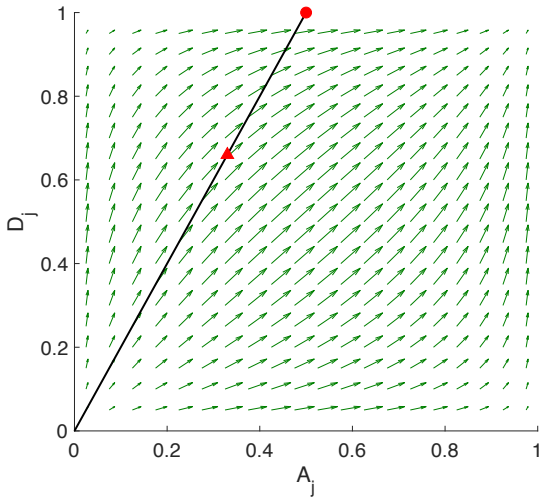
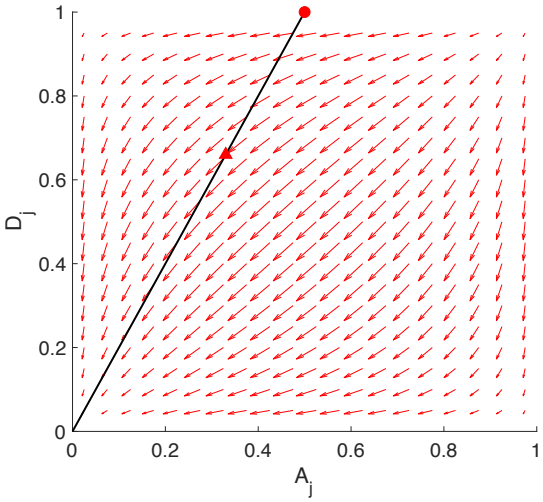


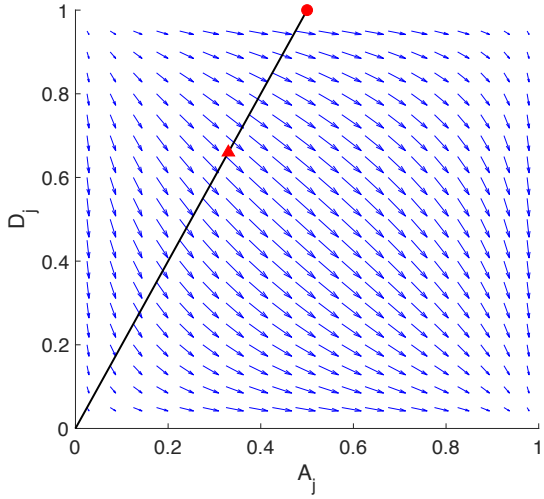
Figure 1: The solid black segment presents A_j's and D_j's corresponding to different choices of $\zeta_j \in [0,1]$ in (8) with $\pi_j = 1/3$. Any point in the segment results in acceptance probability $= 1$ for the idealized target (5). The iid sampling (i), marked with a triangle, is a shrunk version of the superefficient sampling (ii), marked with a bullet.



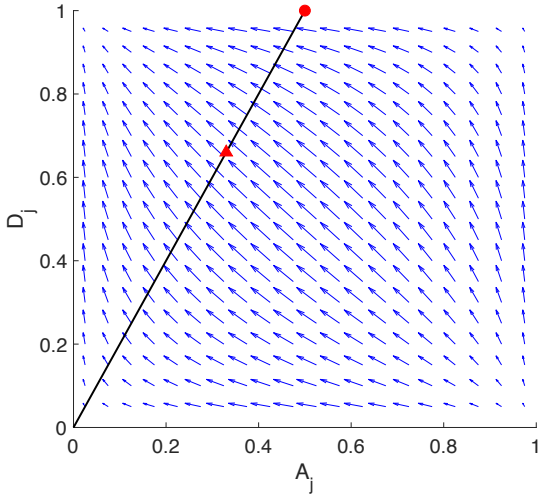
Gradient field of the *expansion step* for $\gamma^{A(i)}$ and $\gamma^{D(i)}$



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



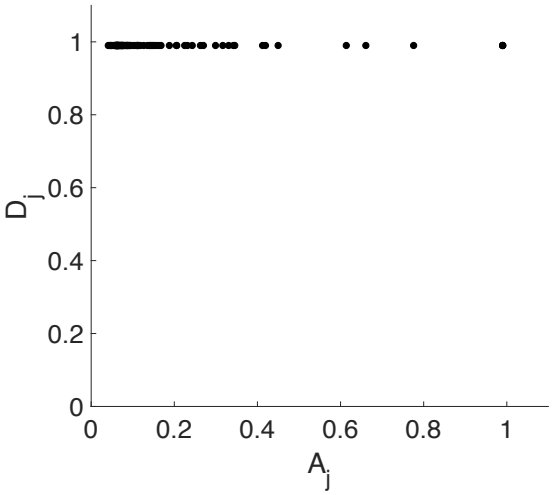
Gradient field of the *correction step* for $\gamma^{D(i)}$



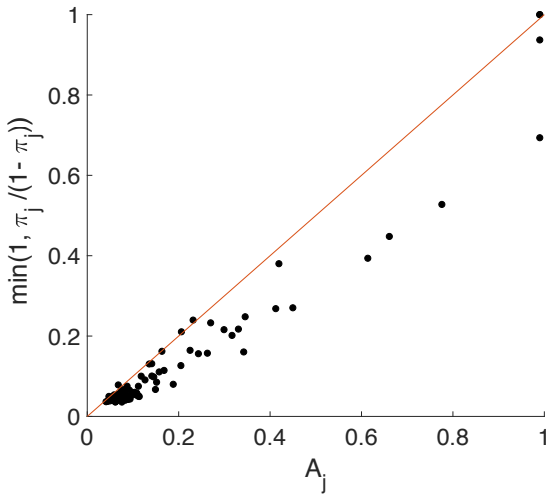
Gradient field of the *correction step* for $\gamma^{A(i)}$

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

E Graphical results for the examples with real data



(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, corresponding 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: Tecator data: the adaptive parameter $\eta = (A, D)$ for the exploratory individual adaptation algorithm.

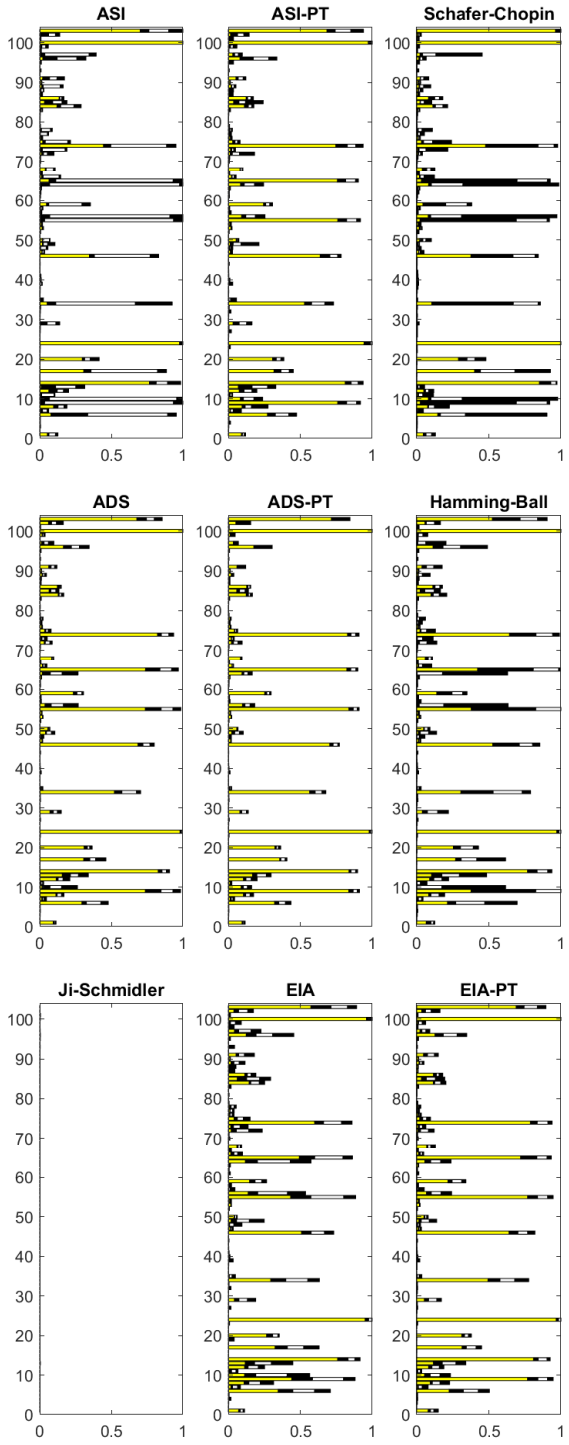


Figure 4: Boston housing data: Inclusion probabilities boxplots using adaptively scaled individual adaptation (ASI), exploratory individual adaptation (EIA), add-delete-swap (ADS) and the sequential Monte Carlo algorithm of Schäfer and Chopin (2013), with parallel tempering (PT) versions of the first three algorithms. We also consider the Hamming Ball and the Ji-Schmidler samplers

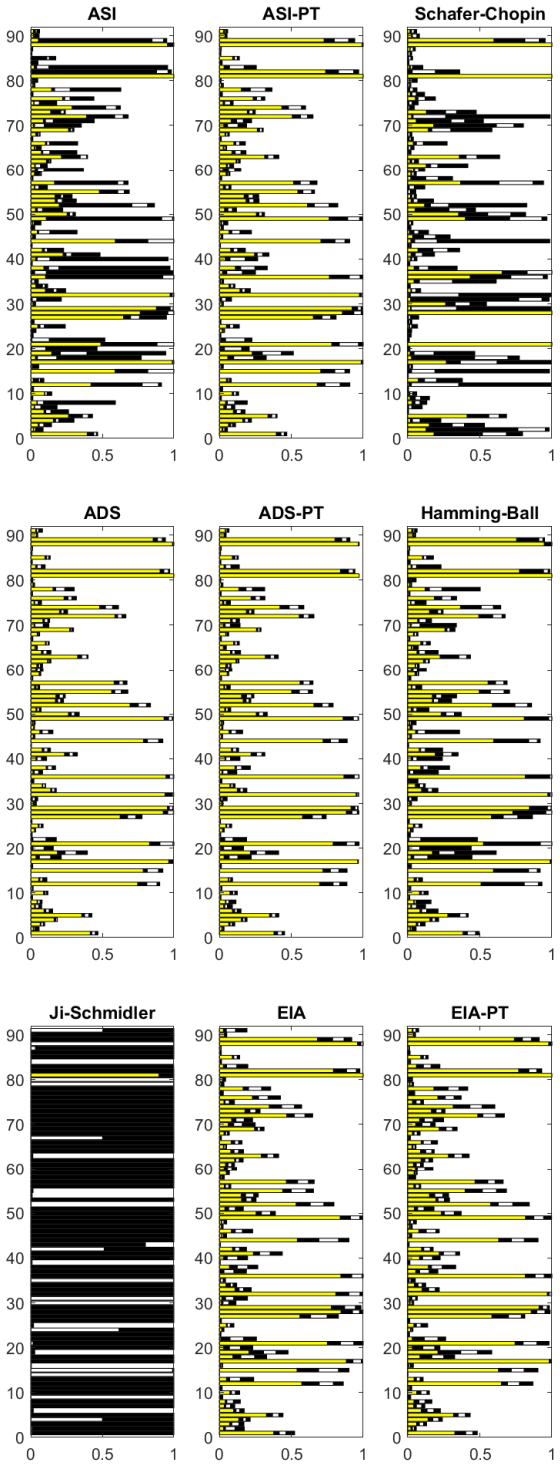


Figure 5: Concrete data: Inclusion probabilities boxplots using adaptively scaled individual adaptation (ASI), exploratory individual adaptation (EIA), add-delete-swap (ADS) and the sequential Monte Carlo algorithm of Schäfer and Chopin (2013), with parallel tempering (PT) versions of the first three algorithms. We also consider the Hamming Ball and the Ji-Schmidler samplers

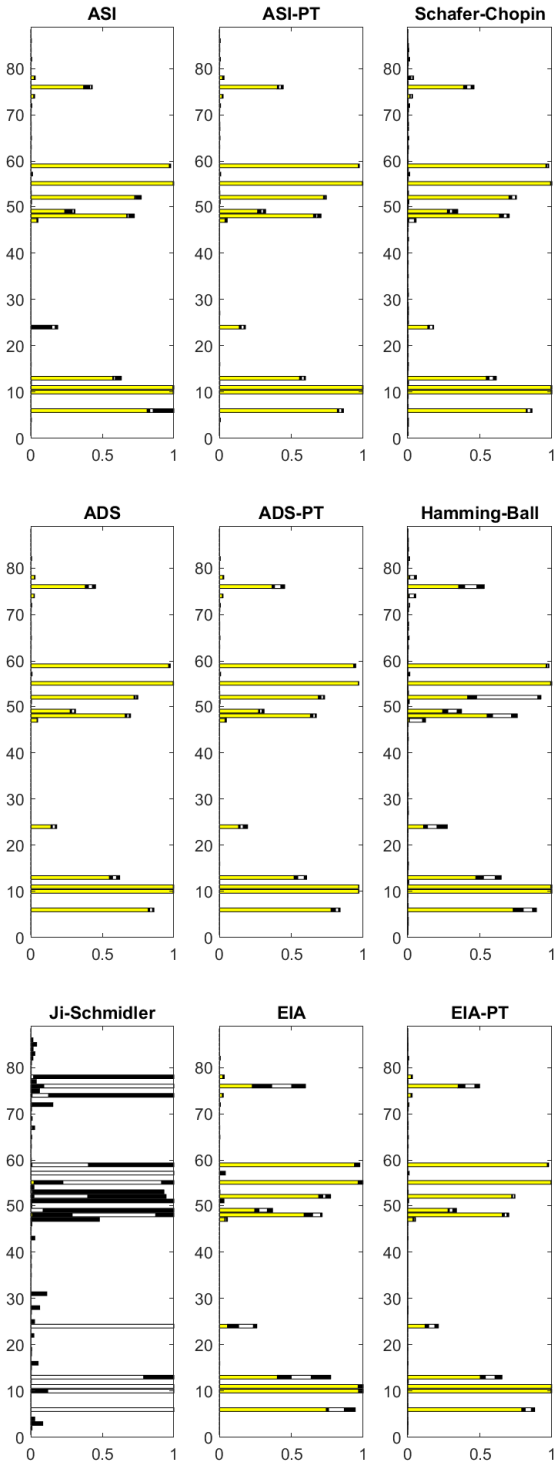


Figure 6: Protein data: Inclusion probabilities boxplots using adaptively scaled individual adaptation (ASI), exploratory individual adaptation (EIA), add-delete-swap (ADS) and the sequential Monte Carlo algorithm of Schäfer and Chopin (2013), with parallel tempering (PT) versions of the first three algorithms. We also consider the Hamming Ball and the Ji-Schmidler samplers

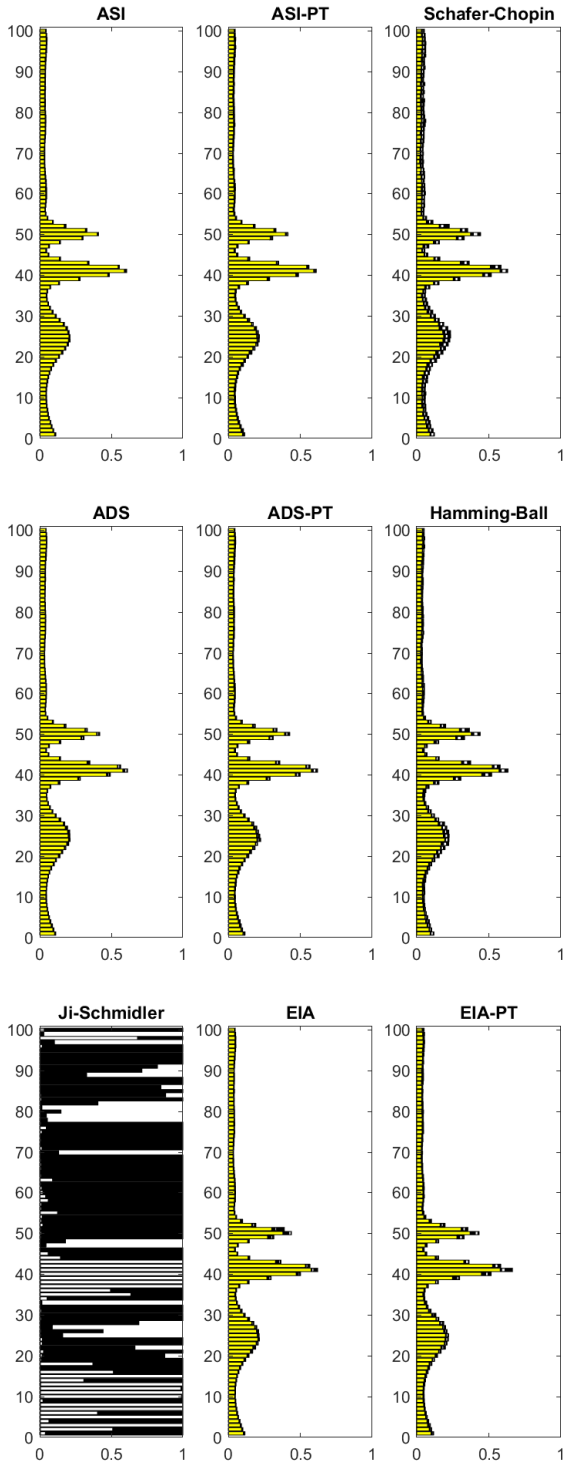


Figure 7:

Tecator data: Inclusion probabilities boxplots using adaptively scaled individual adaptation (ASI), exploratory individual adaptation (EIA), add-delete-swap (ADS) and the sequential Monte Carlo algorithm of Schäfer and Chopin (2013), with parallel tempering (PT) versions of the first three algorithms. We also consider the Hamming Ball and the Ji-Schmidler samplers

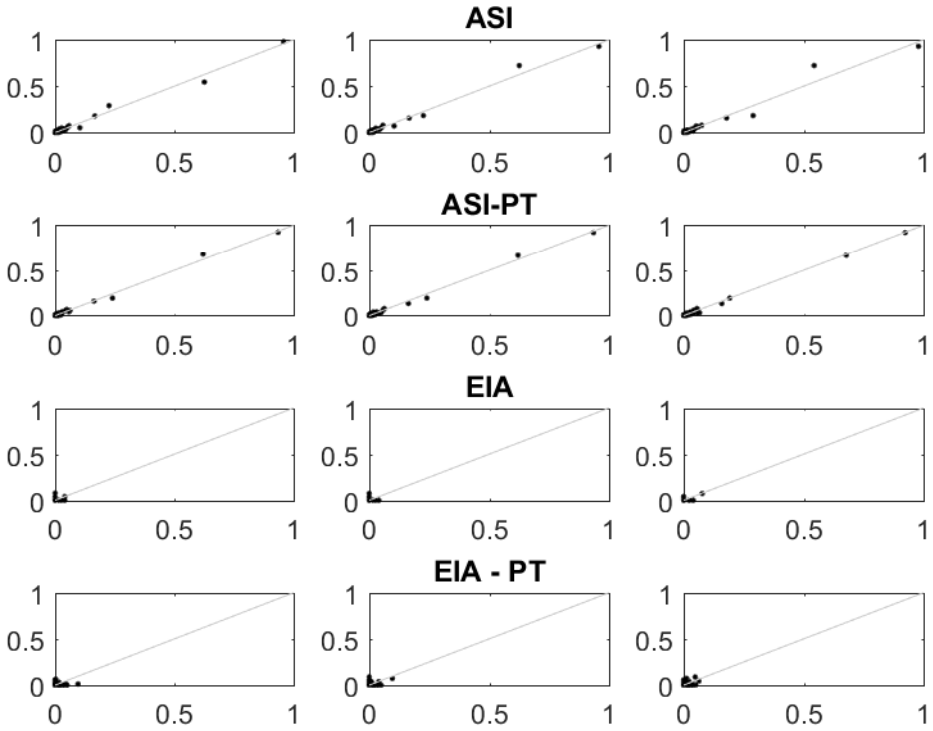


Figure 8: PCR1 data: comparisons of pairs of runs with random g and h using adaptively scaled individual adaptation (ASI), adaptively scaled individual adaptation with parallel tempering (ASI-PT), exploratory individual adaptation (EIA) and exploratory individual adaptation with parallel tempering (EIA-PT)

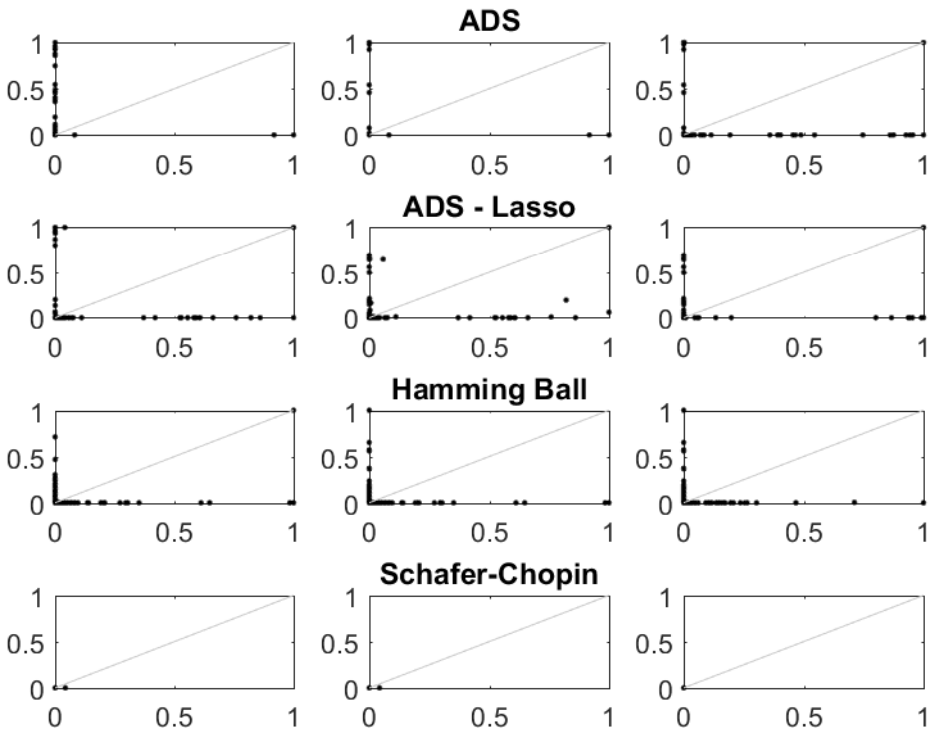


Figure 9: PCR1 data: comparisons of pairs of runs with random g and h using add-delete-swap (ADS), add-delete-swap with lasso start (ADS-Lasso), Hamming Ball and Schäfer-Chopin samplers

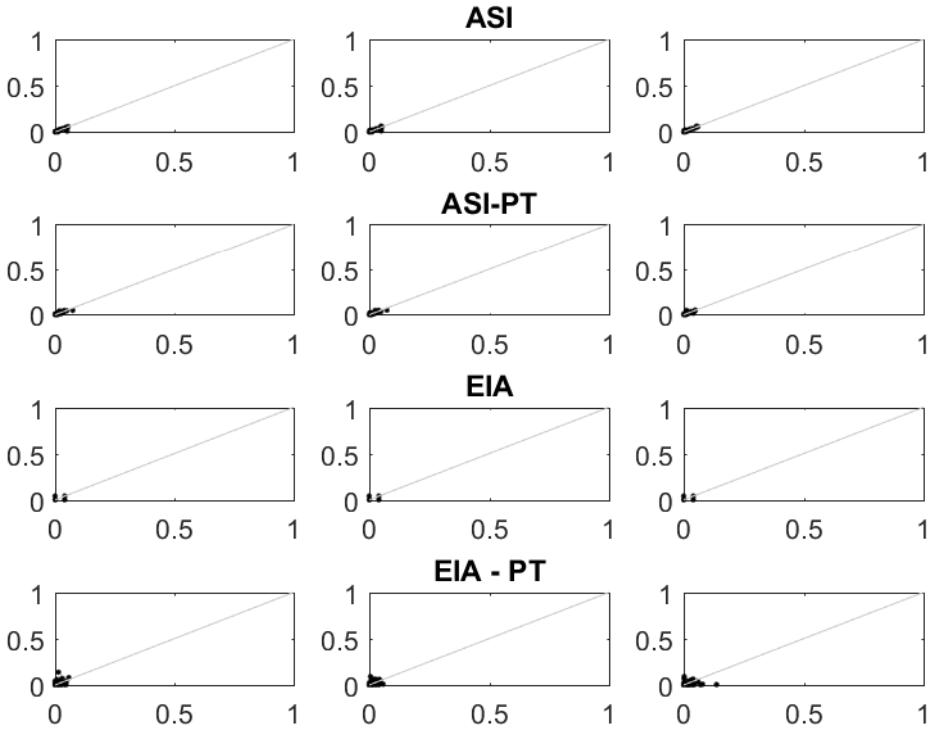


Figure 10: PCR2 data: comparisons of pairs of runs with random g and h using adaptively scaled individual adaptation (ASI), adaptively scaled individual adaptation with parallel tempering (ASI-PT), exploratory individual adaptation (EIA) and exploratory individual adaptation with parallel tempering (EIA-PT)

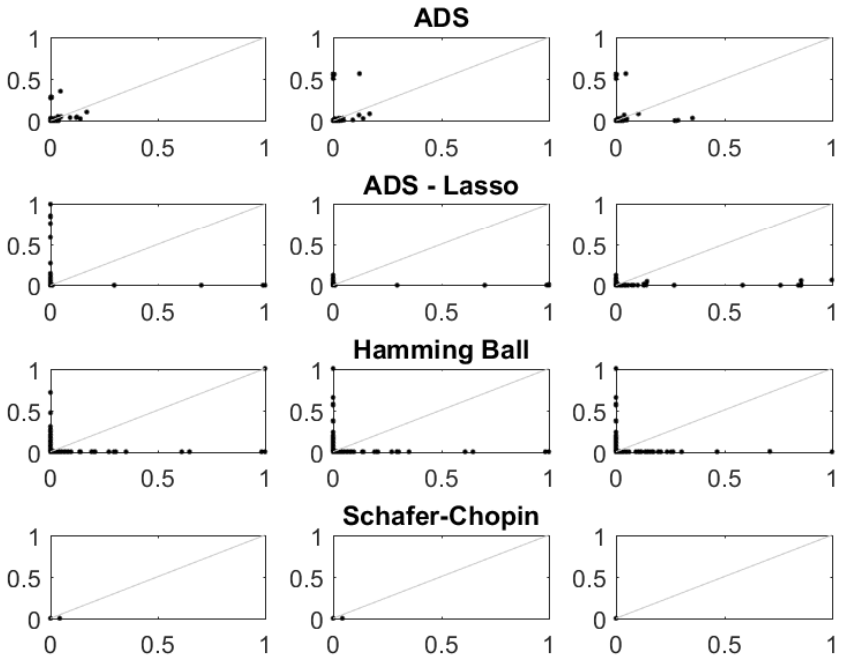


Figure 11: PCR2 data: comparisons of pairs of runs with random g and h using add-delete-swap (ADS), add-delete-swap with lasso start (ADS-Lasso), Hamming Ball and Schäfer-Chopin samplers

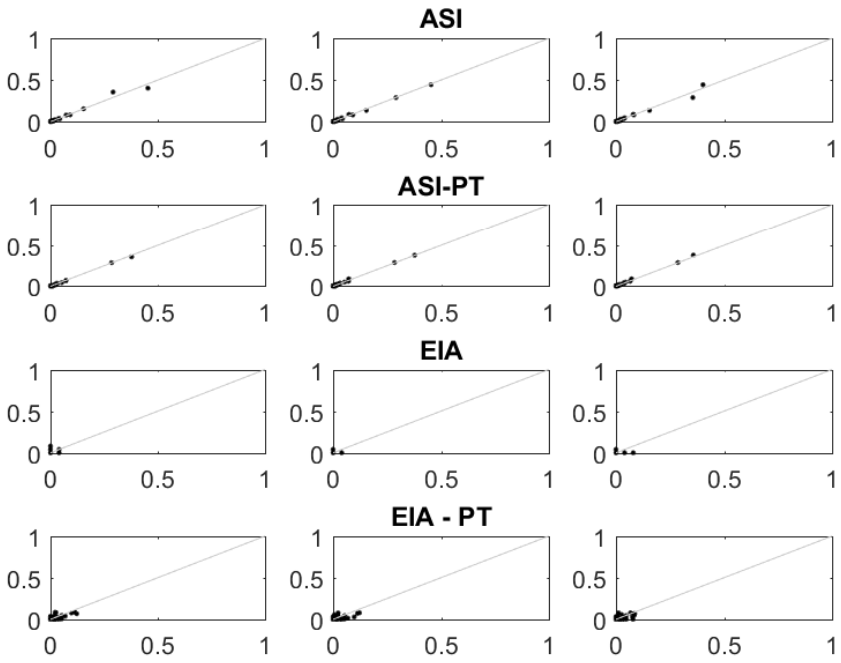


Figure 12: PCR3 data: comparisons of pairs of runs with random g and h using adaptively scaled individual adaptation (ASI), adaptively scaled individual adaptation with parallel tempering (ASI-PT), exploratory individual adaptation (EIA) and exploratory individual adaptation with parallel tempering (EIA-PT)

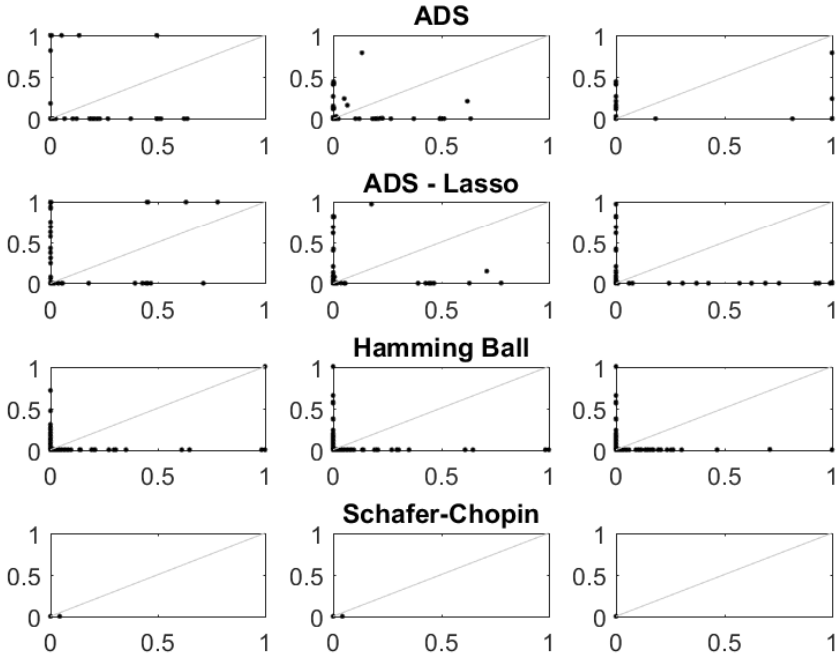


Figure 13: PCR3 data: comparisons of pairs of runs with random g and h using add-delete-swap (ADS), add-delete-swap with lasso start (ADS-Lasso), Hamming Ball and Schäfer-Chopin samplers

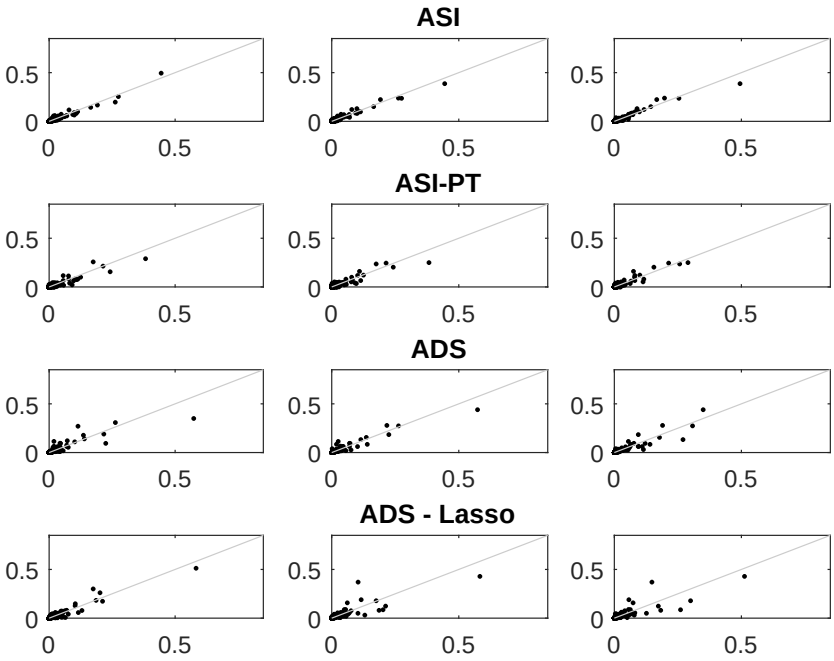


Figure 14: SNP data: comparisons of pairs of runs with random g and fixed h using adaptively scaled individual adaptation (ASI), adaptively scaled individual adaptation with parallel tempering (ASI-PT), add-delete-swap (ADS) and add-delete-swap with lasso start (ADS-L)