Statistics and Computing (2022) 32:84
https://doi.org/10.1007/s11222-022-10137-8

Adaptive random neighbourhood informed Markov chain Monte Carlo for high-dimensional Bayesian variable selection

Xitong Liang[1] · Samuel Livingstone[1] · Jim Griffin[1]

Received: 26 October 2021 / Accepted: 16 August 2022

Abstract
We introduce a framework for efficient Markov chain Monte Carlo algorithms targeting discrete-valued high-dimensional distributions, such as posterior distributions in Bayesian variable selection problems. We show that many recently introduced algorithms, such as the locally informed sampler of Zanella (J Am Stat Assoc 115(530):852–865, 2020), the locally informed with thresholded proposal of Zhou et al. (Dimension-free mixing for high-dimensional Bayesian variable selection, 2021) and the adaptively scaled individual adaptation sampler of Griffin et al. (Biometrika 108(1):53–69, 2021), can be viewed as particular cases within the framework. We then describe a novel algorithm, the *adaptive random neighbourhood informed* sampler, which combines ideas from these existing approaches. We show using several examples of both real and simulated data-sets that a computationally efficient point-wise implementation (PARNI) provides more reliable inferences on a range of variable selection problems, particularly in the very large p setting.

Keywords Bayesian computation · Variable selection · Spike-and-slab priors · Markov chain Monte Carlo · Random neighbourhood samplers · Locally informed Metropolis-Hastings schemes

1 Introduction

Despite their long history, linear regression models remain a key building block of many present-day statistical analyses. In the modern setting, practitioners not only show interest in making good predictions but also intend to investigate underlying low-dimensional structure based on the belief that only a small subset of predictors play a crucial role in predicting the response. These problems can be addressed by *variable selection*. A variable selection method is an automatic procedure that selects the best (small) subset of covariates that explains most of the variation in the response (Chipman et al. 2001). Frequentist approaches focus on model comparisons through information criteria or point estimates, using e.g. maximum penalised likelihood under sparsity assumptions (Hastie et al. 2015). Alternatively the Bayesian approach can be taken by imposing an appropriate prior on all possible models and computing the posterior.

We consider Bayesian variable selection (BVS) with spike-and-slab priors (Mitchell and Beauchamp 1988), which lead to natural uncertainty measures such as posterior model probabilities and marginal posterior variable inclusion probabilities. Suppose a linear regression model with p candidate covariates is given, we focus on a random variable $\gamma \in \Gamma = \{0, 1\}^p$ where $\gamma_j = 1$ indicates that the j-th covariate is included in the model. The exact posterior distribution of γ is challenging to compute, and when $p > 30$ Markov chain Monte Carlo (MCMC) algorithms are typically used to estimate posterior summaries of interest (George and McCulloch 1993; Chipman et al. 2001). Garcia-Donato and Martinez-Beneito (2013) discuss the use of the Gibbs sampler whereas Madigan et al. (1995) (MC^3) and Brown et al. (1998) (Add-Delete-Swap) propose random-walk Metropolis-Hastings algorithms. Yang et al. (2016) provide conditions on the the Add-Delete-Swap algorithm for *rapid mixing* in the sense that the mixing time grows at most polynomially in p under some mild conditions on the posterior distributions. These approaches can, however, suffer from an unexpectedly long mixing time and therefore slow convergence when p is

✉ Xitong Liang
xitong.liang.18@ucl.ac.uk

Samuel Livingstone
samuel.livingstone@ucl.ac.uk

Jim Griffin
j.griffin@ucl.ac.uk

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

Published online: 30 September 2022

large. For this reason, alternative *informed* MCMC schemes have gained popularity for problems with discrete parameter spaces (having already achieved prominence in the continuous setting). Informed MCMC schemes are those in which the Metropolis-Hastings proposal exploits some information about the target distribution. Intuitively, the success of informed proposals relies on avoiding models with low posterior model probabilities (Zhou et al. 2021). Titsias and Yau (2017) describe the Hamming ball sampler (HBS) in which models are proposed in proportion to their locally-truncated posterior probability within a Hamming ball neighbourhood. Zanella and Roberts (2019) consider a Tempered Gibbs sampler (TGS), which involves importance sampling and more frequently updates components with lower conditional distributions. A more general class of locally informed and balanced proposals is introduced by Zanella (2020). These locally balanced proposals can be obtained by weighting a base kernel using a balancing function, which is a function of the posterior distribution that satisfies a certain functional property. The base kernel is typically concentrated on a neighbourhood of the current state, resulting in a proposal that is informed and balanced using "local" information about the posterior. The author shows that a random walk proposal is asymptotically dominated by its locally balanced counterpart in the Peskun sense as dimensionality increases under mild conditions on the target distribution (Peskun 1973; Tierney 1998). Zhou et al. (2021) present a Locally Informed and Thresholded proposal (LIT) which replaces the balancing function by a thresholded weighting function (i.e. a thresholding function). The LIT scheme is closely connected to the locally balanced proposal because the thresholding function behaves like a flexible composition of globally and locally balanced functions. This novel scheme has been shown to have a dimension-free mixing time bound under similar conditions as in Yang et al. (2016). For other developments concerning locally informed proposals, see e.g. Livingstone and Zanella (2019); Gagnon (2021); Power and Goldman (2019).

Since the posterior distribution is discrete-valued, the above random-walk or informed MCMC schemes can be viewed as neighbourhood samplers. A *neighbourhood sampler* is an MCMC scheme which can be decomposed into two stages: (i) construct a *neighbourhood* that is a set of states (models) around the current state (model); (ii) propose a new state (model) within the neighbourhood constructed in stage (i). For example, the MC^3 and locally balanced schemes propose a new model on an identical neighbourhood which consists of models which only differ from the current model in 1 position (i.e. a Hamming neighbourhood), whereas their second stage is a random walk and an informed proposal respectively. The LIT algorithm of Zhou et al. (2021) is similar to the locally balanced scheme whereas it takes an identical neighbourhood generation mechanism to an Add-Delete-Swap scheme but its second stage uses a thresholding function. The design of the neighbourhoods is a crucial factor to the performance of MCMC schemes, especially in those informed schemes for two major reasons. The first reason is the "quality" of neighbourhood in the sense that we should generate neighbourhoods including many promising models. Encouraging better quality neighbourhood construction will improve the mixing of the chain and avoid it getting struck in some low probability models. The second reason is the size of the neighbourhood. Informed MCMC schemes often mix quickly and have good convergence properties, but the computation of each transition can be prohibitively expensive. For example, the number of models in the locally balanced proposal will be at least linear in p and will tend to include large numbers of unimportant variables under standard sparsity assumptions. Neighbourhoods have also been considered previously in the context of stochastic search. Hans et al. (2007) describe a novel *Shotgun Stochastic Search* (SSS) algorithm whilst Chen et al. (2016) consider a paired-move multiple-try stochastic search algorithm. Both schemes identify a subset of probable models and move to new models within the neighbourhood according to posterior model probabilities.

In this paper we propose a method which generates good neighbourhoods while controlling computational cost with large p by introducing a framework for constructing flexible and efficient MCMC algorithms based on *random* neighbourhoods. We refer to the scheme as a random neighbourhood sampler and show that if they are well-constructed such schemes can lead to Markov chains with good convergence properties and controlled computational cost per iteration. Our method uses an adaptive scheme to achieve a flexible neighbourhood generating mechanism. Adaptive MCMC is a sub-class of algorithms in which tuning parameters are automatically updated "on the fly" (e.g. Andrieu and Thoms 2008). Several adaptive methods have been developed in the context of BVS (Ji and Schmidler 2013; Lamnisos et al. 2009, 2013). We build on Griffin et al. (2021) who develop the *Adaptively-Scaled Individual Adaptation* sampler (ASI), which is able to adapt to the importance of each candidate covariate and propose multiple swaps per iteration in high-dimensional settings. We show that the ASI algorithm is a random neighbourhood sampler whose second stage is a random-walk proposal in this paper. Based on this discovery, we design a random neighbourhood informed sampler with the same neighbourhood generating mechanism as ASI but replace its second stage by an informed within-neighbourhood proposal. To illustrate the power of the framework, we develop a new MCMC algorithm for Bayesian variable selection in linear regression, namely the *Point-wise Adaptive Random Neighbourhood Informed* (PARNI) sampler. This combines the strengths of ASI for good neighbourhood generation and locally informed pro-

posals for avoiding random walk behaviour. An extensive set of empirical results on both real and simulated data-sets show that the PARNI sampler yields good estimates for posterior quantities of interest and performs particularly well for well-known large p examples such as the PCR ($p = 22,575$) and SNP ($p = 79,748$) data-sets.

The rest of this paper is structured as follows. In Sect. 2, we review BVS for the linear model along with prior specification. We also briefly describe both the ASI scheme of Griffin et al. (2021) and the locally informed methods of Zanella (2020) and Zhou et al. (2021). In Sect. 3, we characterise the construction of random neighbourhood proposals and illustrate that locally informed proposals and the ASI scheme fall within this framework. Section 4 presents the construction of adaptive random neighbourhood and informed samplers. Following this structure, we present the ARNI and PARNI samplers. In addition, we establish both the ergodicity and a strong law of large numbers for the PARNI algorithm. We implement the PARNI sampler in Sect. 5 on both simulated and real data. Comparisons between the PARNI samplers and other state-of-the-art MCMC algorithms are carried out to showcase their capacity and efficiency. In Sect. 6 we discuss limitations and possible future work. Detailed explanations and proofs are provided in the supplement.

2 Background

2.1 Bayesian variable selection for the linear regression model

Consider a data-set $\{(y_i, x_{i1}, ..., x_{ip})\}_{i=1}^n$, where the vector $y = (y_1, ..., y_n) \in \mathbb{R}^n$ is called the response variable and each $x_j = (x_{1j}, ..., x_{nj})$ is one of p predictor variables or covariates. The variable selection problem is concerned with finding the best $q \ll p$ covariates that are most associated with the response. Assuming that each regression includes an intercept, then there are 2^p possible models that can be formulated to predict the response. We refer to each model as M_γ where the models are indexed by the indicator variable $\gamma = (\gamma_1, \ldots, \gamma_p) \in \Gamma = \{0, 1\}^p$, where $\gamma_j = 1$ if the j-th variable is included in model M_γ and $\gamma_j = 0$ otherwise. We refer to Γ as model space and let $p_\gamma := \sum_j \gamma_j$. The model M_γ associated with γ is then

$$y = \alpha \mathbf{1}_n + X_\gamma \beta_\gamma + \epsilon \tag{1}$$

where $\epsilon \sim N_n(0, \sigma^2 I_n)$, y is an n-dimensional response vector, X_γ is an $(n \times p_\gamma)$ design matrix which consists of the "active" variables in γ (those for which $\gamma_j = 1$), α is an intercept term and $\beta_\gamma \in \mathbb{R}^{p_\gamma}$. In the Bayesian framework, we consider a commonly-used conjugate prior specification

$$p(\alpha) \propto 1, \quad \beta_\gamma | \gamma, \sigma^2 \sim N(0, g\sigma^2 V_\gamma), \quad p(\sigma^2) \propto \sigma^{-2},$$
$$p(\gamma) = h^{p_\gamma}(1-h)^{p-p_\gamma}.$$

For simplicity, we can remove the intercept term α by centering y and X_j for all j. Chipman et al. (2001) highlight that this method can be motivated from a formal Bayesian perspective by integrating out the coefficients corresponding to those fixed regressors with respect to an improper uniform prior. The covariance matrix V_γ is often chosen as $(X_\gamma^T X_\gamma)^{-1}$ (a g-prior) or identity matrix I_{p_γ} (an independent prior). In what follows, we will focus on the independence prior where $V_\gamma = I_{p_\gamma}$. For both of these choices, the marginal likelihood $p(y|\gamma)$ is analytically tractable. Suitable values for the global scale parameter g are suggested in Fernandez et al. (2001). It can also be driven by a hyperprior, yielding a fully Bayesian model (see Liang et al. (2008) for details). The hyperparameter $h \in (0, 1)$ is the prior probability that each variable is included in the model. Steel and Ley (2007) suggest against using fixed h unless strong information is given, and instead placing a hyperprior on it such as a Beta prior $h \sim \text{Beta}(a, b)$, leading to a Beta-binomial prior on the model size. The choices of g and h will be specified later for each set of data. In the following sections, we will develop efficient sampling schemes targeting the posterior distribution $\pi(\gamma) \propto p(y|\gamma)p(\gamma)$.

Remark 1 For a linear regression model with p candidate covariates, it has been shown that spike-and-slab priors often lead to *posterior consistency* in the sense that the posterior collapses to a Dirac measure on the true model as more observations are gathered (Fernandez et al. 2001; Liang et al. 2008; Yang et al. 2016), even in high-dimensional setting where p grows with n (Shang and Clayton 2011; Narisetty and He 2014). Another approach is to employ continuous shrinkage priors (e.g. Polson and Scott 2010; Griffin and Brown 2021), which only give posterior inference on regression coefficients but can result in a more computationally tractable posterior distribution.

2.2 Adaptively scaled individual adaptation algorithm

Griffin et al. (2021) introduce a scalable adaptive MCMC algorithm targeting high-dimensional BVS posterior distributions together with a method that automatically updates the tuning parameters. They consider the class of proposal kernels

$$q_\eta(\gamma, \gamma') = \prod_{j=1}^{p} q_{\eta,j}(\gamma_j, \gamma_j') \tag{2}$$

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 $q_{\eta,j}(\gamma_j = 1, \gamma_j' = 0) = D_j$, with

for $i = 1$ to $i = N$ **do**
  for $l = 1$ to $l = L$ **do**
    Sample $\gamma^{l,\prime} \sim q_{\zeta^{(i)}\eta^{(i)}}(\gamma^{l,(i)}, \cdot)$ as in (2) and $U \sim U(0,1)$;
    If $U < \alpha_{\zeta^{(i)}\eta^{(i)}}(\gamma^{l,(i)}, \gamma^{l,\prime})$ as in (3), then $\gamma^{l,(i+1)} = \gamma^{l,\prime}$, else $\gamma^{l,(i+1)} = \gamma^{l,(i)}$;
  end for
  Update $\hat{\pi}_j^{(i+1)}$ as in (5), set $\tilde{\pi}_j^{(i+1)} = \pi_0 + (1 - 2\pi_0)\hat{\pi}_j^{(i+1)}$ for $j = 1, \ldots, p$;
  Update $\zeta^{(i+1)}$ as in (7);
  Update $A_j^{(i+1)} = \min\left\{1, \tilde{\pi}_j^{(i+1)}/(1 - \tilde{\pi}_j^{(i+1)})\right\}$;
  Update $D_j^{(i+1)} = \min\left\{1, (1 - \tilde{\pi}_j^{(i+1)})/\tilde{\pi}_j^{(i+1)}\right\}$;
  Set $\eta^{(i+1)} = (A_j^{(i+1)}, D_j^{(i+1)})$
end for

Algorithm 1: Adaptively Scaled Individual Adaptation (ASI)

Metropolis-Hastings acceptance probability

$$\alpha_\eta(\gamma, \gamma') = \left\{1, \frac{\pi(\gamma')q_\eta(\gamma', \gamma)}{\pi(\gamma)q_\eta(\gamma, \gamma')}\right\}. \tag{3}$$

This proposal mainly benefits from two aspects. Firstly, the flexibility offered by $2p$ tuning parameters allows the proposal to be tailored to the data. Secondly, this form of proposal also allows multiple variables to be added or deleted from the model in a single iteration, which in turn allows the algorithm to make large jumps in model space.

Griffin et al. (2021) suggest an optimal choice of $\eta = (A, D)$ in Peskun sense while assuming that all variables are independent. If π_j denotes the posterior inclusion probability of the j-th regressor, the optimal choice of $\eta^{\text{opt}} = (A^{\text{opt}}, D^{\text{opt}})$ is given as

$$A_j^{\text{opt}} = \min\left\{1, \frac{\pi_j}{1 - \pi_j}\right\}, \quad D_j^{\text{opt}} = \min\left\{1, \frac{1 - \pi_j}{\pi_j}\right\}. \tag{4}$$

The independence assumption is usually violated due to the correlation between regressors and therefore a scaled proposal with parameters $\eta = \zeta\eta^{\text{opt}}$ for a scaling parameter $\zeta \in (0, 1)$ is suggested. This scaling parameter ζ controls the number of variables that differ between the current state γ and the proposed state γ'. Smaller values of η can be used to avoid overly ambitious moves with low probabilities of acceptance and so control the average acceptance rate. They also suggest multiple chain acceleration with common adaptive parameters since running multiple independent chains with shared adaptive parameters can facilitate the convergence of the adaptive parameters (Craiu et al. 2009). This phenomenon is demonstrated in their simulation studies where the schemes with 25 multiple chains outperform the schemes with only 5 multiple chains in terms of relative efficiency especially for large p data-sets. Suppose L chains are used and let $\gamma^{l,(i)}$ and $\gamma^{l,\prime}$ denote the current state and proposal for the l-th chain respectively. We also defined a vector $\gamma_{-j} = (\gamma_1, \ldots, \gamma_{j-1}, \gamma_{j+1}, \ldots, \gamma_p)$ to denote the vector of γ without γ_j. The tuning parameters of the proposal are updated on the fly using a Rao-Blackwellised estimate of the posterior inclusion probability of the j-th regressor which, at the N-th iteration, is

$$\hat{\pi}_j^{(N)} = \frac{1}{NL}\sum_{i=1}^{N}\sum_{l=1}^{L}\frac{\pi(\gamma_j = 1, \gamma_{-j}^{l,(i)}|y)}{\pi(\gamma_j = 1, \gamma_{-j}^{l,(i)}|y) + \pi(\gamma_j = 0, \gamma_{-j}^{l,(i)}|y)}. \tag{5}$$

The use of the Rao-Blackwellised estimates of the posterior inclusion probabilities can swiftly distinguish unimportant variables. Griffin et al. (2021) show how these Rao-Blackwellised estimates can be calculated in $\mathcal{O}(p)$ operations which leads to a scalable MCMC scheme in large p BVS problems. At the i-th iteration, the proposal parameters are $\eta = \zeta^{(i)} \times \eta^{(i)}$ where $\eta^{(i)} = (A^{(i)}, D^{(i)})$,

$$A_j^{(i)} = \min\left\{1, \frac{\hat{\pi}_j^{(i)}}{1 - \hat{\pi}_j^{(i)}}\right\}, \quad D_j^{(i)} = \min\left\{1, \frac{1 - \hat{\pi}_j^{(i)}}{\hat{\pi}_j^{(i)}}\right\} \tag{6}$$

and the scaling parameter $\zeta^{(i)}$ is tuned using the Robbins-Monro scheme

$$\text{logit}_\epsilon \zeta^{(i+1)} = \text{logit}_\epsilon \zeta^{(i)} + \frac{\phi_i}{L}\sum_{l=1}^{L}(\alpha_{\zeta^{(i)}\eta^{(i)}}(\gamma^{l,(i)}, \gamma^{l,\prime}) - \tau) \tag{7}$$

for a target rate of acceptance τ and the mapping $\text{logit}_\epsilon : (\epsilon, 1 - \epsilon) \to \mathbb{R}$ is a modified logistic function (or logit function) defined by

$$\text{logit}_\epsilon(x) = \log(x - \epsilon) - \log(1 - x - \epsilon) \tag{8}$$

for some small $\epsilon \in (0, 1/2)$. The full description of the sampler is given in Algorithm 1. The resulting algorithm is called *Adaptively Scaled Individual Adaptation* (ASI). Griffin et al. (2021) establish the π-ergodicity and a strong law of large numbers for the ASI sampler.

Remark 2 The performance of the ASI algorithm is crucially related to the choice of appropriate values of parameters and hyperparameters. The parameters $\eta^{(i)}$ and $\zeta^{(i)}$ are updated on the fly. The hyperparameters are chosen as follows: $\phi_i = i^{-0.7}$, $\tau = 0.234$, $\epsilon = 0.1/p$ and $\pi_0 = 0.001$. This hyperparameter specification is suggested by Griffin et al. (2021) and they shows that it works well in general based on the empirical performances. See their paper for the discussion on the choice of hyperparmeters.

2.3 Locally informed proposals for discrete-valued variables

In continuous sample space, MCMC algorithms often utilise gradients of the target distribution, e.g. the *Metropolis-adjusted Langevin algorithm* (Grenander and Miller 1994) and *Hamiltonian Monte Carlo* (Duane et al. 1987). These methods are defined on continuous spaces but Zanella (2020) develop a class of informed proposals as an analog for discrete spaces. The approach assumes that we can define a random walk Metropolis proposal kernel Q on a neighbourhood $N \subset \Gamma$ with mass function q. In this paper, we consider the following construction of informed proposals that are described by Zanella (2020) as follows

$$q_g(\gamma, \gamma') = \begin{cases} \frac{g\left(\frac{\pi(\gamma')}{\pi(\gamma)}\right) q(\gamma, \gamma')}{Z_g(\gamma)}, & \gamma \in N \\ 0, & \text{otherwise} \end{cases} \tag{9}$$

where $g : [0, \infty) \to [0, \infty)$ is a monotone continuous weighting function and $Z_g(\gamma)$ is a normalising constant such that

$$Z_g(\gamma) = \sum_{\gamma' \in N} g\left(\frac{\pi(\gamma')}{\pi(\gamma)}\right) q(\gamma, \gamma'). \tag{10}$$

The choice of the weighting function g is crucial for the performance of Q_g since it determines how the target distribution π drives the proposal. When g is the constant function $g(t) = 1$, the resulting informed proposal Q_g will coincide with the base kernel Q and this is referred to as a non-informed proposal. Zanella (2020) mainly discussed the *locally balanced proposals* which are formed by the *balancing functions* that satisfy $g(t) = tg(1/t)$ for all $t > 0$. The locally balanced proposals are approximately π-reversible if Q is restricted to local moves. The neighbourhoods are normally chosen to be $N = \mathcal{H}_m(\gamma) := \{\gamma' \in \Gamma | d_H(\gamma', \gamma) \leq m\}$ for which $d_H(\cdot, \cdot)$ denotes the measure of Hamming distance (i.e. $d_H(\gamma, \gamma') = \sum_{j=1}^{p} |\gamma_j - \gamma'_j|$) and the proposal kernel Q would be a uniform distribution on the neighbourhood N. When m is taken to be 1, the base kernel Q is identical to the MC^3 sampler. In addition, taking g as the identity function (i.e. $g(t) = t$) will lead to a globally balanced proposal Q_g where Q_g is π-reversible when the neighbourhood N is the whole sample space.

Theorem 5 of Zanella (2020) shows that using a uniform based kernel on neighbourhood $\mathcal{H}_m(\gamma)$ combined with a balancing function g as described above will be asymptotically optimal relative to the un-informed or globally balanced proposals, in terms of Peskun ordering, as the dimensionality goes to infinity under the condition that $\sup_{\gamma \in \Gamma, \gamma' \in N} Z_g(\gamma)/Z_g(\gamma') \to 1$ holds. However, for a Bayesian variable selection problem, Zhou et al. (2021) argue that the behavior of the function $\gamma \to Z_g(\gamma)$ is difficult to predict and the assumption may not hold. They therefore suggest a modified weighting function with upper- and lower-bounds

$$g(t) = \min\{\max\{p^l, t\}, p^L\} \tag{11}$$

where p is the total number of regressors and $-\infty < l < L < \infty$ are some constants. In what follows, the weighting function in (11) is referred to as the *thresholding function*. The thresholding function is flexible in the sense that it includes globally and locally balanced functions for specific values of l and L.

Their Locally informed with Thresholded (LIT) algorithm works on neighbourhoods derived from the Add-Delete-Swap scheme and allows the values of l and L to change with the type of move. Under the conditions that the posterior mass concentrations on a small set and the chain starts at a model that is not too far from the true data-generating model, they prove that the LIT algorithm can achieve a dimension-free mixing rate if the parameters of the LIT algorithm are properly selected.

3 Random neighbourhood samplers and the ASI algorithm

Let us recall the idea of a neighbourhood sampler from Sect. 1. In general, the neighbourhoods can be random and tailored to the target distribution π. This is referred to as a *random neighbourhood sampler*. In this section, we will properly present the random neighbourhood sampler in detail and show using Theorem 1 that the ASI sampler is a random neighbourhood sampler.

3.1 Random neighbourhood samplers

We consider a framework for constructing Metropolis-Hastings proposals to sample from $\pi(\gamma)$ in which a new state is proposed within a *random neighbourhood* around the current state. The random neighbourhoods are generated using an auxiliary variable k as a neighbourhood indicator. This auxiliary variable k is a discrete random variable defined on a countable set $\mathcal{K}$ such the probability of generating a neighbourhood $N = N(\gamma, k)$ is the same as the probability of generating k (i.e. $p(N|\gamma) = p(k|\gamma)$). Suppose γ is the current state and Q_k is a Metropolis-Hastings proposal kernel (conditioned on k) with mass function q_k. A new state γ' is drawn from kernel Q_k after a value of k has been generated. In updating k at each iteration, we usually consider proposing a new state k' conditional on the current state k through a deterministic bijection $\rho : \mathcal{K} \to \mathcal{K}$ such that $k' = \rho(k)$. The mapping ρ should be an involution which

is a self-inverse function which satisfies $\rho(\rho(k)) = k$. We call an MCMC algorithm that uses the above construction to generate Metropolis-Hastings proposals a *random neighbourhood sampler*. The followings are some examples of random neighbourhood samplers.

Example 1 *(Samplers with non-stochastic neighbourhoods)* In fact, samplers with non-stochastic neighbourhoods are also random neighbourhood samplers where the specific neighbourhoods are generated with constant probability of 1 at each state γ. In such cases, the choices of k and ρ can be arbitrary. For instance, the MC^3 sampler can be viewed as a random neighbourhood sampler for which the neighbourhood N consists of models that are 1-Hamming distance from γ. In particular, the locally balanced samplers of Zanella (2020) also belong to this class with neighbourhood N as defined in Sect. 2.3.

Example 2 *(Add-Delete-Swap sampler and LIT proposal)* In each iteration of an Add-Delete-Swap (ADS) sampler, a strategy from "addition", "deletion" and "swap" is uniformly chosen which implies that the auxiliary variable k is uniformed distributed over the sample space $\mathcal{K} =$ {"addition", "deletion", "swap"} and therefore construct a neighbourhood $N(\gamma, k)$ as in Yang et al. (2016). A new state γ' is uniformly proposed from $N(\gamma, k)$. The corresponding mapping ρ is then a function that sends the auxiliary variable to an opposite strategy, e.g. it sends "addition" to "deletion" and vice versa. Note that the opposite strategy of "swap" is itself. The Locally Informed and Thresholded (LIT) proposal by Zhou et al. (2021) has an identical neighbourhood construction to an ADS sampler but it proposes a new model using an informed weighted proposal that uses weighting functions bounded above and below.

Example 3 *(Hamming ball sampler)* A Hamming ball sampler with radius m is described by Titsias and Yau (2017). This algorithm selects a neighbourhood from $\mathcal{H}_m(\gamma) \subset \Gamma$, which is the set of states at most m-Hamming distance away from γ. The auxiliary variable k is equivalent to U in their design in which k is uniformly distributed over the set $\mathcal{K} = \mathcal{H}_m(\gamma)$ and a neighbourhood $N(\gamma, k) = \mathcal{H}_m(k)$ is used to draw a new state. The Hamming ball sampler proposes a new state according to the truncated posterior model probability in the neighbourhood $N(\gamma, k)$. In this scheme, the mapping ρ is the identity function, meaning the same auxiliary variable is used in reversed moves.

The full update of a random neighbourhood sampler uses the three stages below:

(i) (*Neighbourhood construction*) Sample a neighbourhood indicator k from $p(\cdot|\gamma)$, and construct the corresponding neighbourhood $N(\gamma, k)$;

(ii) (*Within-neighbourhood proposal*) Propose a new model γ' in $N(\gamma, k)$ according to $Q_k(\gamma, \cdot)$;

(iii) (*Accept/reject step*) Calculate the probability of the reverse move, $q_{\rho(k)}(\gamma', \gamma)$, by constructing the reverse neighbourhood $N(\gamma', \rho(k))$. Move to the new state γ' with probability $\alpha_k(\gamma, \gamma')$ where $\alpha_k(\gamma, \gamma')$ is the Metropolis-Hastings acceptance probability

$$\alpha_k(\gamma, \gamma') = \min\left\{1, \frac{\pi(\gamma')p(\rho(k)|\gamma')q_{\rho(k)}(\gamma', \gamma)}{\pi(\gamma)p(k|\gamma)q_k(\gamma, \gamma')}\right\}. \tag{12}$$

Throughout this article, we refer to the above three stages as *neighbourhood construction*, *within-neighbourhood proposal* and *accept/reject step* respectively. To preserve the reversibility of the chain, it is better to design a neighbourhood generation scheme where the law

$$\gamma' \in N(\gamma, k) \iff \gamma \in N(\gamma', \rho(k)) \tag{13}$$

holds for any γ, γ' and k. Upon this law, we assume that the condition

$$p(k|\gamma)q_k(\gamma, \gamma') > 0 \iff p(\rho(k)|\gamma)q_{\rho(k)}(\gamma', \gamma) > 0 \tag{14}$$

is satisfied. This assumption is a generalisation of the paired-move strategy in Chen et al. (2016) and it results in the correctness and reversibility of such a scheme through the following proposition.

Proposition 1 *A random neighbourhood sampler is π-reversible provided that condition (14) holds, $p(k|\gamma)$ is a valid probability measure on $\mathcal{K}$ and $q_k(\gamma, \gamma')$ is a valid probability measure on neighbourhood $N(\gamma, k)$ for all $\gamma \in \Gamma$ and $k \in \mathcal{K}$.*

Remark 3 To generalise the framework of random neighbourhood samplers, it is possible to use a continuous auxiliary variable k. In such a case, the acceptance probability in (12) should include the Jacobian term.

We show in the next part that the ASI sampler is also a random neighbourhood sampler. Unlike the locally balanced proposals, it focuses on constructing sophisticated random neighbourhoods which are more likely to contain promising models and employs a random walk within-neighbourhood proposal.

3.2 Another take on the ASI scheme

It is not straightforward to observe that the ASI sampler is a random neighbourhood sampler, however we show below that, in fact, it can be. To do so, we introduce a random

neighbourhood sampler, the *Adaptive Random Neighbourhood* (ARN) sampler, and prove that the ARN and ASI samplers are equivalent if they share some common adaptive parameters. The ARN sampler follows a random walk within neighbourhood but, compared to the locally informed approach, puts more efforts into neighbourhood construction.

We consider a random neighbourhood sampler with algorithmic tuning parameter $\theta = (\xi\eta^{\text{opt}}, \omega) \in (\epsilon, 1-\epsilon)^{2p+1} := \Delta_\epsilon^{2p+1}$ and a small $\epsilon \in (0, 1/2)$, where η^{opt} is given in (4), and the tuning parameters ξ and ω are used in the random neighbourhood construction and the within-neighbourhood proposal respectively. In the random neighbourhood construction, the neighbourhood indicator variable $k = (k_1, \ldots, k_p) \in \mathcal{K} = \{0,1\}^p$ is generated from the distribution

$$p^{\text{RN}}_{\xi\eta^{\text{opt}}}(k|\gamma) = \prod_{j=1}^{p} p^{\text{RN}}_{\xi\eta^{\text{opt}},j}(k_j|\gamma_j) \tag{15}$$

where $p^{\text{RN}}_{\xi\eta^{\text{opt}},j}(k_j = 1|\gamma_j = 0) = \xi A_j^{\text{opt}}$ and $p^{\text{RN}}_{\xi\eta^{\text{opt}},j}(k_j = 1|\gamma_j = 1) = \xi D_j^{\text{opt}}$. This is equivalent to the ASI proposal in (2) where $k_j = 1$ if and only if $\gamma_j \neq \gamma'_j$. A neighbourhood $N(\gamma, k)$ is obtained from γ and k for which γ is the "centre" of $N(\gamma, k)$ and k indicates the possible indices altered from γ. These tuning parameters ξ and η^{opt} are abortively updated on the fly. For any $\gamma^* \in N(\gamma, k)$, $k_j = 0$ implies that $\gamma^*_j = \gamma_j$. This identity can be used to state a formal definition of the neighbourhood $N(\gamma, k)$ as

$$N(\gamma, k) = \{\gamma^* \in \Gamma | \gamma_j = \gamma^*_j,\ \forall k_j = 0\}.$$

The neighbourhood contains 2^{p_k} models where p_k is the number of 1s in k (i.e. $p_k := \sum_{j=1}^{p} k_j$). The parameter ξ affects the p_k and therefore controls neighbourhood size. So we call ξ the neighbourhood scaling parameter.

The mapping ρ is chosen to be the identity function. The within-neighbourhood proposal in this *adaptive* random neighbourhood scheme is also based on the same proposal in (2) over the neighbourhood $N(\gamma, k)$. It can be characterised as choosing the variables to be added or deleted from the model by thinning from within the set $\{j|\ k_j = 1\}$ with the thinning probability set to be $\omega \in (0, 1)$. We refer to this parameter as the unique *within-neighbourhood proposal tuning parameter*. A larger value of ω increases the probability of proposing γ' further away from γ in Hamming distance. This can be written formally as the proposal in (2) with tuning parameter $\eta^{\text{THIN}} = (A^{\text{THIN}}, D^{\text{THIN}}) = (\omega k, \omega k)$, that is $A_j^{\text{THIN}} = D_j^{\text{THIN}} = \omega$ for $k_j = 1$ and $A_j^{\text{THIN}} = D_j^{\text{THIN}} = 0$ otherwise. The resulting proposal is termed as $q^{\text{THIN}}_{\omega,k}$ which is formulated as

Sample $k \sim p^{\text{RN}}_{\xi\eta^{\text{opt}}}(\cdot|\gamma)$ as in (15);
Sample $\gamma' \sim q^{\text{THIN}}_{\omega,k}(\gamma, \cdot)$ as in (16) and $U \sim U(0,1)$;
if $U < \alpha^{\text{ARN}}_{\theta,k}(\gamma, \gamma')$ as in (17), then accept γ'

Algorithm 2: Adaptive Random Neighbourhood sampler (ARN)

$$q^{\text{THIN}}_{\omega,k}(\gamma, \gamma') = \prod_{j=1}^{p} q^{\text{THIN}}_{\omega,k_j}(\gamma_j, \gamma'_j), \tag{16}$$

where $q^{\text{THIN}}_{\omega,1}(\gamma_j, 1-\gamma_j) = \omega$ and $q^{\text{THIN}}_{\omega,0}(\gamma_j, 1-\gamma_j) = 0$. The proposal $q^{\text{THIN}}_{\omega,k}$ is symmetric and only generates new states inside the neighbourhood $N(\gamma, k)$. This is because the probabilities of proposing flips on coordinates other than j such that $k_j = 1$ are 0. The scheme is completed by accepting or rejecting the proposal using a standard Metropolis-Hastings acceptance probability

$$\alpha^{\text{ARN}}_{\theta,k}(\gamma, \gamma') = \left\{1, \frac{\pi(\gamma')p^{\text{RN}}_{\xi\eta^{\text{opt}}}(k|\gamma')q^{\text{THIN}}_{\omega,k}(\gamma', \gamma)}{\pi(\gamma)p^{\text{RN}}_{\xi\eta^{\text{opt}}}(k|\gamma)q^{\text{THIN}}_{\omega,k}(\gamma, \gamma')}\right\}. \tag{17}$$

Remark 4 An alternative formulation to (16) in terms of Hamming distance between γ and γ' is

$$\begin{aligned} q^{\text{THIN}}_{\omega,k}(\gamma, \gamma') &= \omega^{d_H(\gamma,\gamma')}(1-\omega)^{p_k - d_H(\gamma,\gamma')}\mathbb{I}\{\gamma' \in N(\gamma, k)\} \\ &= \left(\frac{\omega}{1-\omega}\right)^{d_H(\gamma,\gamma')}(1-\omega)^{p_k}\mathbb{I}\{\gamma' \in N(\gamma, k)\} \end{aligned} \tag{18}$$

where $d_H(\gamma, \gamma')$ is the measure of Hamming distance between two models γ and γ'.

Remark 5 When ω is chosen to be $1/2$, the within-neighbourhood proposal $q^{\text{THIN}}_{\omega=1/2,k}$ is uniformly distributed over the local neighbourhood $N(\gamma, k)$.

Algorithm 2 describes how a new state γ' is proposed using the ARN scheme. We indicate the transition kernel by p^{ARN}_θ and the corresponding sub-transition kernel conditional on k by $p^{\text{ARN}}_{\theta,k}$. They obey the relationship

$$p^{\text{ARN}}_\theta(\gamma, \gamma') = \sum_{k \in \mathcal{K}} p^{\text{ARN}}_{\theta,k}(\gamma, \gamma').$$

The following proposition helps to show that the ARN sampler is π-reversible.

Proposition 2 *For any tuning parameter $\theta = (\eta, \omega) \in \Delta_\epsilon^{2p+1} = (\epsilon, 1-\epsilon)^{2p+1}$, the condition (14) holds, the conditional distribution of k, $p^{RN}_\eta(k|\gamma)$, within the ARN sampler is a valid distribution on $\mathcal{K} = \{0,1\}^p$. In addition, for any*

$\gamma \in \Gamma$ and $k \in \mathcal{K}$, the within-neighbourhood proposal of the ARN sampler $q^{THIN}_{\omega,k}(\gamma, \gamma')$ is also a valid probability distribution on $N(\gamma, k)$.

Proposition 1 together with Proposition 2 show that the ARN transition kernel is π-reversible and therefore generates samples that preserve the target distribution π. In fact ARN and ASI are mathematically equivalent provided that the tuning parameter choices are made in a prescribed manner. To see this suppose that the tuning parameters of both the ARN and ASI schemes are fixed and share the same tuning parameter η. The following theorem shows that their transition probabilities from γ to γ' are equal when $\zeta = \xi \times \omega$ holds.

Theorem 1 *Suppose that $\eta \in \Delta_\epsilon^{2p}$ and ζ, ξ, $\omega \in \Delta_\epsilon$ for small $\epsilon \in (0, 1/2)$, $p^{\text{ARN}}_{(\xi\eta,\omega)}$ and $p^{\text{ASI}}_{\zeta\eta}$ are transition kernels of the ARN and ASI schemes respectively. If $\zeta = \xi \times \omega$ and, then*

$$p^{\text{ARN}}_{(\xi\eta,\omega)}(\gamma, \gamma') = p^{\text{ASI}}_{\zeta\eta}(\gamma, \gamma') \tag{19}$$

holds for any γ and $\gamma' \in \Gamma$.

In addition we deduce the following corollary.

Corollary 1 *Setting $\xi_1 \times \omega_1 = \xi_2 \times \omega_2$ implies*

$$p^{\text{ARN}}_{(\xi_1\eta,\omega_1)}(\gamma, \gamma') = p^{\text{ARN}}_{(\xi_2\eta,\omega_2)}(\gamma, \gamma')$$

for any γ and $\gamma' \in \Gamma$.

Corollary 1 shows that two ARN kernels with different tuning parameters coincide in probability if the products of the neighbourhood scaling parameter ξ and proposal thinning parameter ω are equal. This corollary also suggests that magnitudes of ξ and ω can shift to each other without modifying the resulting proposal as long as their product preserves.

4 Adaptive random neighbourhood and informed samplers

It should be clear from the above discussion that both the locally informed proposals and ASI schemes can be viewed as random neighbourhood samplers, and that the former focuses on selecting good proposals within a neighbourhood, while the latter focuses on constructing neighbourhoods of models which are more likely to be accepted in the Metropolis-Hastings update. Our main methodological contribution is to design a random neighbourhood sampler for which both the neighbourhood construction and within-neighbourhood proposal are designed in an informed way. We therefore consider using an adaptive random neighbourhood approach to construct neighbourhoods, followed by a locally informed approach to select a proposal from this neighbourhood.

The advantages of combining the two schemes in this manner are worth highlighting. A key strength of ASI is that generating proposals is computationally cheap, but when components of the posterior distribution are highly correlated then the assumption of independence that is embedded into the proposal generation can lead to overly ambitious moves that will be rejected. To combat this, the scaling parameter must be used to control the acceptance rate, but in the presence of high correlation this can lead to small moves and slow mixing. The locally informed sampler can cope well with high levels of correlation in the posterior distribution, but the (un-informed) neighbourhood in high-dimensions often, either contain no sensible models, or be so large that the cost of computing all of the posterior probability of models within it becomes prohibitive. Combining the two schemes is therefore an attractive proposition, as an intelligent neighbourhood that is also not too large can be constructed using ASI, and then correlation can be controlled for at the second stage by choosing the within-neighbourhood proposal using the locally informed approach.

We give the details of this adaptive random neighbourhood and informed sampler below, which we call the *Adaptive Random Neighbourhood Informed* (ARNI) sampler. After this we define the point-wise ARNI (PARNI) scheme, which enjoys the benefits of ARNI but with much lower computational cost.

4.1 Adaptive random neighbourhood informed algorithm

We first describe a general construction of the *random neighbourhood informed proposals*. Suppose a random neighbourhood sampler is given with neighbourhood indicator variable $k \in \mathcal{K}$ and a update mapping ρ together with a within-neighbourhood proposal kernel Q_k. The variable k follows a conditional distribution $p(k|\gamma)$ whereas the proposal Q_k produces a new state γ' within neighbourhood $N(\gamma, k)$ in an uninformed manner. We consider a class of *random neighbourhood informed proposals* $Q_{g,k}$ with mass function

$$q_{g,k}(\gamma, \gamma') = \begin{cases} \dfrac{g\left(\frac{\pi(\gamma')p(\rho(k)|\gamma')q_{\rho(k)}(\gamma',\gamma)}{\pi(\gamma)p(k|\gamma)q_k(\gamma,\gamma')}\right)q_k(\gamma,\gamma')}{Z_{g,k}(\gamma)}, & \gamma \in N(\gamma, k) \\ 0, & \text{otherwise} \end{cases} \tag{20}$$

where $g : [0, \infty) \to [0, \infty)$ is a continuous monotone weighting function, and $Z_{g,k}(\gamma)$ is a normalising constant defined by

$$Z_{g,k}(\gamma) = \sum_{\gamma^* \in N(\gamma,k)} g\left(\frac{\pi(\gamma^*)p(\rho(k)|\gamma^*)q_{\rho(k)}(\gamma^*,\gamma)}{\pi(\gamma)p(k|\gamma)q_k(\gamma,\gamma^*)}\right) q_k(\gamma,\gamma^*). \tag{21}$$

The generated new state γ' is accepted using the Metropolis-Hastings rule

$$\alpha_{g,k}(\gamma,\gamma') = \min\left\{1, \frac{\pi(\gamma')p(\rho(k)|\gamma')q_{g,\rho(k)}(\gamma',\gamma)}{\pi(\gamma)p(k|\gamma)q_{g,k}(\gamma,\gamma')}\right\}. \tag{22}$$

The proposal collapses to the locally balanced proposal of Zanella (2020) when the neighbourhood is non-stochastic, the weighting function g is a balancing function that satisfies $g(t) = tg(1/t)$ and the within-neighbourhood proposal is symmetric. In what follows, we combine the above random neighbourhood informed proposal with the ARN scheme and develop an *Adaptive Random Neighbourhood Informed* (ARNI) proposal that uses an informed proposal at the within-neighbourhood proposal stage. In the ARNI scheme, the mapping ρ is chosen to be the identity function where $\rho(k) = k$ and the within-neighbourhood proposal in Algorithm 2 is replaced by

$$\begin{aligned} q^{\text{ARNI}}_{\theta,k}(\gamma,\gamma') &\propto g\left(\frac{\pi(\gamma')p^{\text{RN}}_{\xi\eta^{\text{opt}}}(k|\gamma')q^{\text{THIN}}_{\omega,k}(\gamma',\gamma)}{\pi(\gamma)p^{\text{RN}}_{\xi\eta^{\text{opt}}}(k|\gamma)q^{\text{THIN}}_{\omega,k}(\gamma,\gamma')}\right) q^{\text{THIN}}_{\omega,k}(\gamma,\gamma') \\ &= g\left(\frac{\pi(\gamma')p^{\text{RN}}_{\xi\eta^{\text{opt}}}(k|\gamma')}{\pi(\gamma)p^{\text{RN}}_{\xi\eta^{\text{opt}}}(k|\gamma)}\right) q^{\text{THIN}}_{\omega,k}(\gamma,\gamma'). \end{aligned} \tag{23}$$

for some weighting function g and some parameters $\theta = (\xi\eta^{\text{opt}}, \omega) \in \Delta_\epsilon^{2p+1} = (\epsilon, 1-\epsilon)^{2p+1}$. The last equation follows since the within-neighbourhood proposal $q^{\text{THIN}}_{\omega,k}$ is symmetric and therefore $q^{\text{THIN}}_{\omega,k}(\gamma',\gamma)/q^{\text{THIN}}_{\omega,k}(\gamma,\gamma') = 1$ for all $\gamma' \in N(\gamma,k)$. The Metropolis-Hastings acceptance probability is tailored to the new informed proposal as

$$\alpha^{\text{ARNI}}_{\theta,k}(\gamma,\gamma') = \min\left\{1, \frac{\pi(\gamma')p^{\text{RN}}_{\xi\eta^{\text{opt}}}(k|\gamma')q^{\text{ARNI}}_{\theta,k}(\gamma',\gamma)}{\pi(\gamma)p^{\text{RN}}_{\xi\eta^{\text{opt}}}(k|\gamma)q^{\text{ARNI}}_{\theta,k}(\gamma,\gamma')}\right\}. \tag{24}$$

The optimal choice of informed weighting function is unclear in the ARNI scheme. The thresholding function is not appropriate since the neighbourhoods generated by ARNI cannot be divided into the addition and deletion neighbourhoods as in the LIT scheme. We therefore recommend to use a balancing function which satisfies $g(t) = tg(1/t)$ and form an ARNI balanced proposal.

To boost the convergence of these adaptive tuning parameters, the same multiple chain strategy as ASI should be implemented. In addition to the notations used in ASI, $k^{l,(i)}$ denotes the neighbourhood indicator variable for the l-chain at time i. For L multiple chains, the tuning parameters η^{opt} are updated following the same scheme of ASI as in (6) and (7). Two scaling parameters ξ and ω can be updated using the Robbins-Monro schemes

$$\text{logit}_\epsilon \xi^{(i+1)} = \text{logit}_\epsilon \xi^{(i)} + \frac{\phi_i}{L}\sum_{l=1}^{L}(p_{k^{l,(i)}} - s) \tag{25}$$

$$\text{logit}_\epsilon \omega^{(i+1)} = \text{logit}_\epsilon \omega^{(i)} + \frac{\phi_i}{L}\sum_{l=1}^{L}(\alpha_i^l - \tau) \tag{26}$$

where p_k is the size of k as mentioned previously, s is the target size of k, α_i^l is the acceptance probability at the ith iteration for the l-th chain and τ is the target average acceptance rate.

Remark 6 For practical convenience, it is often useful to chose the diminishing sequence ϕ_i of the form $\phi_i = i^{-\lambda}$ for $\lambda \in (1/2, 1)$ since the condition $\phi_i = \mathcal{O}(i^{-\lambda})$ is not be violated by this choice of ϕ_i. Choosing $\lambda > 1$ would result in finite adaptation (Roberts and Rosenthal 2007) in which the adaptation stops after a finite stopping time, and using $\lambda < 1/2$ is uncommon because of finite sample stability concerns. We therefore recommend using $\phi_i = i^{-0.7}$ for both updating schemes. See Remark 3 in Griffin et al. (2021) for further discussion.

While the informed proposal is powerful in accelerating the convergence of the chains, it also introduces extra computational costs since the posterior probabilities of all models in a neighbourhood are required. Given a k of size p_k, the resulting neighbourhood $N(\gamma,k)$ consists of 2^{p_k} models. Although it is possible to speed up the posterior calculations using Gray codes as introduced in George and McCulloch (1997), evaluating 2^{p_k} models is still computationally expensive when p_k is very large and leads to an inefficient scheme. One way to address the issue is to tune the neighbourhood scaling parameter to generate neighbourhoods with a desired size, say let s be 5. In our experience, such control of the size of k comes at the cost of reduced exploration of the model space and the ARNI scheme fails to achieve better performance than ASI. This motivated us to develop a more efficient implementation of this approach that controls computational cost but maintains good exploration properties.

4.2 The PARNI sampler

We consider a point-wise implementation of the ARNI scheme (for short, the PARNI scheme). This approach is motivated by the block-wise implementation in Zanella (2020) and the block design strategy in Titsias and Yau (2017). The main idea is that a large neighbourhood is divided into a series of smaller blocks and the new model is proposed by sequentially adding or deleting variables in each block. The block design can lead to a significant reduction in the total number of models considered and so require less computational effort. vadjustFor instance, suppose that there are p_k non-zero neighbourhood indicator variables, which are divided into m equally sized blocks. The neighbourhoods generated by each block will have 2^m models. Working through each block to propose a new state requires evaluating $2^m\, p_k/m$ posteriorvadjust probabilities. As the computational cost is proportional to the total number of models considered, the computational cost is largest when $m = p_k$ where the only building block is the entire neighbourhood of $N(\gamma, k)$. In contrast, the smallest computational cost occurs when $m = 1$ where each block has one variable and therefore contains two models only. Throughout the section, we consider the latter block design when $m = 1$ and the resulting algorithm is the PARNI sampler.

4.2.1 Main algorithm

We now formally present the PARNI algorithm and show how a new model γ' is proposed from the current model γ. We use the same random neighbourhood construction as the ARNI scheme, in addition, the neighbourhood scaling parameter ξ is set to be fixed at 1 to indicate that neighbourhood sizes are not reduced at this stage. In other words the neighbourhoods are generated with the optimal values η^{opt} as in (4). After a neighbourhood $N(\gamma, k)$ is sampled, we sequentially propose new models with only 1-Hamming distance differences inside $N(\gamma, k)$. We define $K = \{K_1, \ldots, K_{p_k}\} = \{j | k_j = 1\}$ to be the set of variables for which $k_j = 1$ (the order of variables is random). We also define a sequence of models, $\gamma(1), \ldots, \gamma(p_k)$ and neighbourhoods, $N(1), \ldots, N(p_k)$ to sample the final proposal γ'. To introduce more flexibility, we allow different weighting functions for each sub-proposal so p_k weighting functions $g_1, \ldots, g_{p_k}$ are defined. Finally, let $e(1), \ldots, e(p)$ be the basis vector of a p-dimensional Cartesian space where $e(j)_j = 1$ and $e(j)_{j'} = 0$ whenever $j' \neq j$. We consider the neighbourhoods constructed according to $\gamma(j)$ and $e(K_r)$ for r from 1 to p_k. The first neighbourhood is $N(1) = N(\gamma, e(K_1))$ and propose a model $\gamma(1)$ from

$$q^{\text{PARNI}}_{\theta,K_1}(\gamma, \gamma(1)) \propto \begin{cases} g_1\left(\frac{\pi(\gamma(1))\, p^{\text{RN}}_{\eta^{\text{opt}}}(e(K_1)|\gamma(1))}{\pi(\gamma)\, p^{\text{RN}}_{\eta^{\text{opt}}}(e(K_1)|\gamma)}\right) q^{\text{THIN}}_{\omega,e(K_1)}(\gamma, \gamma(1)), & \text{if } \gamma(1) \in N(1) \\ 0, & \text{otherwise} \end{cases} \tag{27}$$

for some algorithmic parameters $\theta = (\eta^{\text{opt}}, \omega) \in \Delta_\epsilon^{2p+1}$. We repeat this process to construct the second neighbourhood $N(2) = N(\gamma(1), e(K_2))$ and propose the model $\gamma(2)$ from $N(2)$. In general, at time r, we defined $N(r) = N(\gamma(r-1), e(K_r))$ and propose a model $\gamma(r)$ from

$$q^{\text{PARNI}}_{\theta,K_r}(\gamma(r-1), \gamma(r)) \propto \begin{cases} g_r\left(\frac{\pi(\gamma(r))\, p^{\text{RN}}_{\eta^{\text{opt}}}(e(K_r)|\gamma(r))}{\pi(\gamma(r-1))\, p^{\text{RN}}_{\eta^{\text{opt}}}(e(K_r)|\gamma(r-1))}\right) q^{\text{THIN}}_{\omega,e(K_r)}(\gamma(r-1), \gamma(r)), & \text{if } \gamma(r) \in N(r) \\ 0, & \text{otherwise.} \end{cases} \tag{28}$$

Each sub-proposal above only allows the value in position K_r to change. Figure 1 provides a flowchart of the PARNI scheme which only involves enumerating at most $2p_k$ models rather than 2^{p_k} models in the ARNI proposal. The parameters of the proposal are $\theta = (\eta^{\text{opt}}, \omega)$.

To construct a π-reversible chain, the probability of the reverse moves is required. These reverse moves use $K' = \rho(K)$ as their auxiliary variables. The mapping ρ reverses the order of elements in K so that the variable K' contains the same elements in K but with reverse order. The typical benefit is that it leads to identical intermediate models of forward and reverse proposals and the posterior probabilities of p_k models are required instead of $2p_k$. Suppose that $\gamma'(r)$ for $r = 0, \ldots, p_k$ are consecutive intermediate models used in the reverse move and $N'(r)$ for $r = 0, \ldots, p_k$ are the neighbourhoods used in the reverse move. These models and neighbourhoods are identical to those ones used in the proposal move but with opposite order, in particular $\gamma'(r) = \gamma(p_k - r)$ for $r = 0, \ldots, p_k$ and $N'(r) = N(p_k - r + 1)$ for $r = 1, \ldots, p_k$. The second benefit is that the design leads

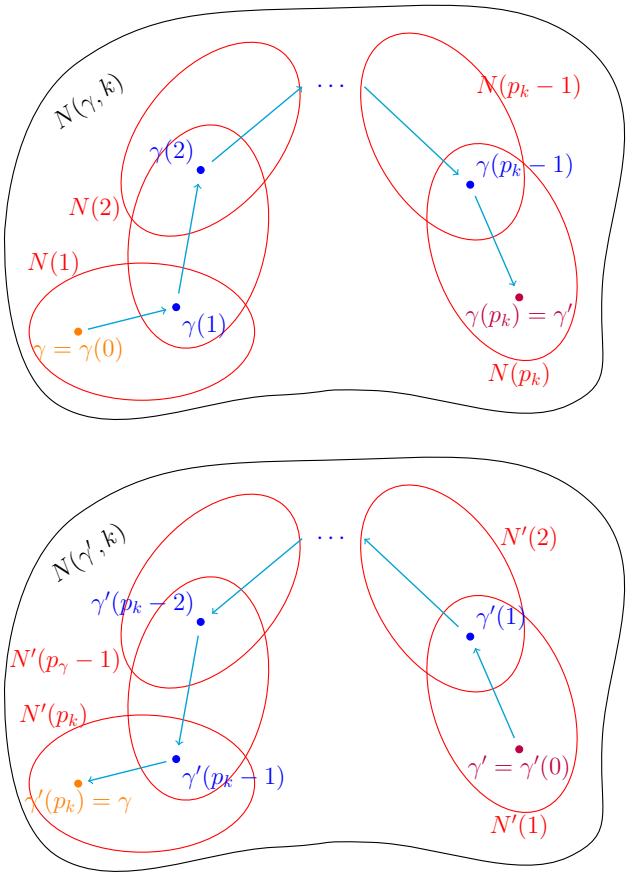

Fig. 1 Flowcharts of the pointwise implementation of adaptive random neighbourhood informed proposal in one iteration. Top panel: proposed direction. Bottom panel: reversed direction. The black neighbourhoods $N(\gamma, k)$ and $N(\gamma', k)$ are the original large neighbourhoods. The red neighbourhoods $N(r)$ and $N'(r)$ are subsequent small neighbourhoods used for each intermediate proposals. The orange model γ is the current state and the cerise model γ' is the final proposal. The blue models $\gamma(r)$ and $\gamma'(r)$ are intermediate models. The light blue arrows indicate the position-wise proposals. (Color figure online)

to a simpler form of the Metropolis-Hastings probability of acceptance. Let $Z(r)$ be the normalising constant of the r-th sub-proposal, $q^{\text{PARNI}}_{\theta,K_r}(\gamma(r-1), \gamma(r))$, and $Z'(j)$ denote the normalising constant of r-th sub-proposal in the reverse move, $q^{\text{PARNI}}_{\theta,K'_r}(\gamma'(r-1), \gamma'(r))$ with weighting functions g'_r. We have that

$$\begin{aligned} Z(r) &= \sum_{\gamma^* \in N(r)} g_r\left(\frac{\pi(\gamma^*) p^{\text{RN}}_{\eta^{\text{opt}}}(e(K_r)|\gamma^*)}{\pi(\gamma(r-1)) p^{\text{RN}}_{\eta^{\text{opt}}}(e(K_r)|\gamma(j-1))}\right) \\ &\quad q^{\text{THIN}}_{\omega,e(K_r)}(\gamma(j-1), \gamma^*) \\ Z'(r) &= \sum_{\gamma^* \in N'(r)} g'_r\left(\frac{\pi(\gamma^*) p^{\text{RN}}_{\eta^{\text{opt}}}(e(K'_r)|\gamma^*)}{\pi(\gamma'(r-1)) p^{\text{RN}}_{\eta^{\text{opt}}}(e(K'_r)|\gamma'(j-1))}\right) \\ &\quad q^{\text{THIN}}_{\omega,e(K'_r)}(\gamma'(j-1), \gamma^*). \end{aligned} \tag{29}$$

We let $q^{\text{PARNI}}_{\theta,k}(\gamma', \gamma)$ be the full proposal kernel that satisfies

$$q^{\text{PARNI}}_{\theta,k}(\gamma, \gamma') = \prod_{r=1}^{p_k} q^{\text{PARNI}}_{\theta,K_j}(\gamma(r-1), \gamma(r)) \tag{30}$$

where $\gamma(0)$ is current state γ and $\gamma(p_k)$ is the final proposal γ'. The Metropolis-Hastings acceptance probability of the PARNI proposal is given as

$$\alpha^{\text{PARNI}}_{\theta,k}(\gamma, \gamma') = \left\{1, \frac{\pi(\gamma') p^{\text{RN}}_{\eta^{\text{opt}}}(k|\gamma') q^{\text{PARNI}}_{\theta,k}(\gamma', \gamma)}{\pi(\gamma) p^{\text{RN}}_{\eta^{\text{opt}}}(k|\gamma) q^{\text{PARNI}}_{\theta,k}(\gamma, \gamma')}\right\}. \tag{31}$$

In specifying these weighting functions g_r for $r = 1, \ldots, p_k$, because each sub-proposal in the PARNI proposal can be treated as addition/deletion move, it is feasible to choose the thresholding function as LIT in Zhou et al. (2021) for different moves. We consider the following thresholded weighting function

$$g_r(t) = \begin{cases} \min\{\max\{p^{-1}, t\}, p\}, & \text{if } \gamma(r)_{K_r} = 0 \\ \min\{\max\{p^{-1}, t\}, 1\}, & \text{if } \gamma(r)_{K_r} = 1 \end{cases} \tag{32}$$

for $r = 1, \ldots, p_k$. The weighting functions in the reverse move are defined similarly. Alternatively, we can also use a balancing function g in PARNI. The choice of balancing function mainly focuses on three particular candidates: square root function $g_{\text{sq}}(t) = \sqrt{t}$, Hastings' choice $g_{\text{H}}(t) = \min\{1, t\}$ and Barker's choice $g_{\text{B}} = t/(1+t)$. The comparisons of these balancing functions in Supplement B.1.3 of Zanella (2020) illustrate two major findings. The Hastings' and Barker's choices only differ by at most a factor of 2 due to their similar asymptotic behaviors. The square root function mixes the worst outside the burn-in phase. Therefore, we consider the Hastings' choice throughout (i.e.

$$g_r(t) = \min\{1, t\} \tag{33}$$

for all $r = 1, \ldots, p_k$) in the rest of the paper. Similar results are also expected for the Barker's choice. Using the balancing function would lead to a simpler form of the Metropolis-Hastings acceptance probability and this is illustrated by the following proposition:

Proposition 3 *Suppose $\gamma, \gamma' \in \Gamma$ are fixed. For any $\theta = (\eta, \omega) \in \Delta_\epsilon^{2p+1}$ and k such that $\gamma' \in N(\gamma, k)$, if the weighting function g_r satisfies $g_r(t) = t g_r(1/t)$ for all r, the Metropolis-Hastings acceptance probability in (31) can be written*

$$\alpha^{PARNI}_{\theta,k}(\gamma, \gamma') = \min\left\{1, \prod_{r=1}^{p_k} \frac{Z(r)}{Z'(r)}\right\} \tag{34}$$

where $Z(r)$, $Z'(r)$ for $r = 1, \cdots, p_k$ are the normalising constants given in (29).

The PARNI proposal which uses thresholding function is referred to as PARNIT whereas one uses balancing function is referred to as PARNIB.

4.2.2 Adaptation schemes for algorithmic parameters

The last building block to complete the PARNI sampler is the adaptation mechanism of the tuning parameters. The posterior inclusion probabilities π_j are updated as in the ASI scheme in (5). The magnitude of the proposal thinning parameter ω is crucial in the mixing time and convergence rate of chains. Therefore, we consider two adaptation schemes for updating ω, the Robins-Monro adaptation scheme (RM) and the Kiefer–Wolfowitz adaptation scheme (KW). For the rest of section, we assume L multiple chains are used for the PARNI sampler.

The Robbins-Monro adaptation scheme is widely used in updating tuning parameters of adaptive MCMC rithms. Andrieu and Thoms (2008) review several adaptive MCMC algorithms using variants of the Robbins-Monro process. Given a specified probability of acceptance τ, the Robbins-Monro adaptation scheme automatically adjusts ω according to the comparison between the current probability of acceptance and τ. It is generally considered to be a robust adaption scheme. Given the acceptance probability of the l-th chain at the i-th iteration α_i^l, the tuning parameter ω is updated through the law

$$\text{logit}_\epsilon \omega^{(i+1)} = \text{logit}_\epsilon \omega^{(i)} + \frac{\phi_i}{L} \sum_{l=1}^{L} (\alpha_i^l - \tau). \tag{35}$$

for $\phi_i = O(i^{-\lambda})$ for some constant $1/2 < \lambda < 1$. The theoretical optimal value of τ may not exist for every candidate proposal kernel and choice of posterior distribution. We recommend using the diminishing sequence $\phi_i = i^{-0.7}$ and using a target acceptance rate of 0.65 based on a large number of experiments that will be illustrated in Sect. 5.

Apart from the above Robbins-Monro scheme, the Kiefer–Wolfowitz scheme is another possible adaption in tuning ω for the PARNI sampler. The Kiefer–Wolfowitz scheme is a stochastic approximation algorithm and modification of the Robbins-Monro scheme in which a finite difference approximation to the derivative is used. In this scheme the tuning parameter is updated to target the optimiser of an objective function of interest. According to the work of Pasarica and Gelman (2010), one can use the *expected squared jumping distance* as the objective function because the expected squared jumping distance is closely linked to the mixing and convergence properties of a Markov chain. The expected squared jumping distance can be estimated by the *average squared jumping distance*. An alternative objective function would be the generalised speed measure introduced in Titsias and Dellaportas (2019).

To estimate the finite difference approximation to the derivative of the average squared jumping distance, we exploit the multiple chain implementation of PARNI. The multiple independent chains naturally provide independent samples which fits the Kiefer–Wolfowitz approximation. Our implementation of the Kiefer–Wolfowitz adaption scheme proceeds as follows. We first evenly divide L multiple chains into two equally sized batches, L^+ and L^-. Let c_i be a diminishing sequence, new proposals are generated using $\omega^+ = \omega^{(i)} + c_i$ for chains in L^+ and $\omega^- = \omega^{(i)} - c_i$ for chains in L^-. The average squared jumping distances for these batches (i.e. $\text{ASJD}^{+,(i)}$ and $\text{ASJD}^{-,(i)}$) are estimated using the new proposals and their corresponding probabilities of acceptance. The tuning parameter ω is then updated according to the rule

$$\text{logit}_\epsilon \omega^{(i+1)} = \text{logit}_\epsilon \omega^{(i)} + a_i \left(\frac{\text{ASJD}^{+,(i)} - \text{ASJD}^{-,(i)}}{2c_i} \right). \tag{36}$$

We suggest using $a_i = i^{-1}$ and $c_i = i^{-0.5}$ in the Kiefer–Wolfowitz scheme. Further details of the Kiefer–Wolfowitz adaption scheme are given in A.1 of the supplementary material and a feasibility analysis of the Kiefer–Wolfowitz adaption scheme is carried out in C.2 of the supplementary material.

Remark 7 Blum (1954) show the Kiefer–Wolfowitz scheme converges if the diminishing sequences a_i and c_i satisfy $\sum_{i=0}^{\infty} a_i^2 c_i^{-2} = \infty$. According to Remark 6, the sequences a_i and c_i should have diminishing rate between -0.5 and -1. Therefore, the only possible pair would be $a_i = i^{-1}$ and $c_i = i^{-0.5}$.

Remark 8 Alternative to adapting the thinning parameter ω through the above adaptive schemes, one can set ω to a fixed value of $1/2$ for simplicity and the base kernel q^{THIN} becomes uniformly distributed. Note that fixing ω at $1/2$ does not necessarily lead to optimal mixing for the PARNI scheme.

Pseudocode of the PARNI samplers are given in Algorithm 3. The corresponding transition kernel is referred to as $p_\theta^{\text{PARNI}(*)-\bullet}$ for $* = \text{T}$ or B and $\bullet = \text{RM}$ or KW. In the next section we show that the PARNI sampler is π-ergodic and satisfy a strong law of large numbers.

4.2.3 Ergodicity and strong law of large numbers

The multiple chain acceleration can be thought of the realisation of L runs on a product space $\Gamma^{\otimes L}$ with joint variable

If $*$ is T, then use a thresholding function as in (32); else if $*$ is B, then use a balancing function as in (33);
for $i = 1$ to $i = N$ **do**
  If $\bullet$ is KW, then divide L-chains into L^+ and L^- and compute ω^+ and ω^-;
  for $l = 1$ to $l = L$ **do**
    Sample $k \sim p_{\eta^{(i)}}^{\mathrm{RN}}(\cdot|\gamma^{l,(i)})$ as in (15);
    Set $\gamma(0) = \gamma^{(i)}$, $p_k = \sum_{j=1}^{p} k_j$, $K = \{j|k_j = 1\}$;
    for $r = 1$ to $r = p_k$ **do**
      Sample $\gamma(r) \sim q_{\theta^{(i)},K_j}^{\mathrm{PARNI}}(\gamma(r-1), \cdot)$ as in (28);
      Calculate $Z(r)$ and $Z'(p_k - r + 1)$ as in (29);
    end for
    Set $\gamma' = \gamma(p_k)$ Sample $U \sim U(0, 1)$;
    If $U < \alpha_{\theta^{(i)},k}^{\mathrm{PARNI}}(\gamma^{(i)}, \gamma')$ as in (31), then $\gamma^{(i+1)} = \gamma'$, else $\gamma^{(i+1)} = \gamma^{(i)}$;
    Update $\hat{\pi}_j^{(i+1)}$ as in (5), set $\tilde{\pi}_j^{(i+1)} = \pi_0 + (1 - 2\pi_0)\hat{\pi}_j^{(i+1)}$ for $j = 1, \ldots, p$;
  end for
  Update $A_j^{(i+1)} = \min\left\{1, \tilde{\pi}_j^{(i+1)}/(1 - \tilde{\pi}_j^{(i+1)})\right\}$;
  Update $D_j^{(i+1)} = \min\left\{1, (1 - \tilde{\pi}_j^{(i+1)})/\tilde{\pi}_j^{(i+1)}\right\}$;
  If $\bullet$ is RM, then update $\omega^{(i+1)}$ as in (35);
else if $\bullet$ is KW, then update $\omega^{(i+1)}$ as in (36);
  Set $\eta^{(i+1)} = (A^{(i+1)}, D^{(i+1)})$ and $\theta^{(i+1)} = (\eta^{(i+1)}, \omega^{(i+1)})$
end for

Algorithm 3: Pointwise Adaptive Random Neighbourhood Sampler with Informed proposal (PARNI($*$)-$\bullet$)

$\gamma^{\otimes L} = (\gamma^1, \ldots, \gamma^L) \in \Gamma^{\otimes L}$. Without loss of generality, suppose further that $L \geq 1$ for the Robbins-Monro adaptation scheme and $L \geq 2$ for the Kiefer–Wolfowitz adaptation scheme. We consider a posterior distribution π on the space Γ which is of the form

$$\pi(\gamma) \propto p(y|\gamma)p(\gamma) \tag{37}$$

where both $p(y|\gamma)$ and $p(\gamma)$ are analytically available. In addition, the joint posterior distribution $\pi^{\otimes L}$ on the product set $\Gamma^{\otimes L}$ is given as

$$\pi^{\otimes L}(\gamma^{\otimes L}) = \prod_{l=1}^{L} \pi(\gamma^l). \tag{38}$$

In this section, the symbol $*$ denotes either T or B and the symbol $\bullet$ represents either KW or RM. The sub-proposal mass function of the PARNI($*$)$-\bullet$ sampler given neighbourhood indicator variable k and tuning parameter $\theta = (\eta, \omega)$ is defined by

$$\psi_{\theta,k}^{\mathrm{PARNI}(*)-\bullet}(\gamma, \gamma') = p_\eta^{\mathrm{RN}}(k|\gamma) q_{\theta,k}^{\mathrm{PARNI}(*)-\bullet}(\gamma, \gamma'). \tag{39}$$

The full transition kernel of the PARNI sampler is marginalised over all possible k

$$P_\theta^{\mathrm{PARNI}(*)-\bullet}(\gamma, S) = \sum_{k \in \mathcal{K}} P_{\theta,k}^{\mathrm{PARNI}(*)-\bullet}(\gamma, S) \tag{40}$$

where the sub-transition kernels given k are

$$\begin{aligned} P_{(\theta,k)}^{\mathrm{PARNI}(*)-\bullet}(\gamma, S) &= \sum_{\gamma' \in S} p_{\theta,k}^{\mathrm{PARNI}(*)-\bullet}(\gamma, \gamma') \\ &= \sum_{\gamma' \in S} \psi_{\theta,k}^{\mathrm{PARNI}(*)-\bullet}(\gamma, \gamma') \alpha_{\theta,k}^{\mathrm{PARNI}}(\gamma, \gamma') \\ &\quad + \mathbb{I}\{\gamma \in S\} \sum_{\gamma' \in \Gamma} \psi_{\theta,k}^{\mathrm{PARNI}(*)-\bullet}(\gamma, \gamma')(1 - \alpha_{\theta,k}^{\mathrm{PARNI}}(\gamma, \gamma')) \end{aligned} \tag{41}$$

and $\alpha_{\theta,k}^{\mathrm{PARNI}}$ are Metropolis-Hastings acceptance rates in (31). The Markov chain transition kernel that works on the product space $\Gamma^{\otimes L}$ is given as

$$P_{(\theta,k^{\otimes L})}^{\mathrm{PARNI}(*)-\bullet}(\gamma^{\otimes L}, S^{\otimes L}) = \prod_{l=1}^{L} P_{(\theta,k^l)}^{\mathrm{PARNI}(*)-\bullet}(\gamma^l, S^l). \tag{42}$$

To establish the ergodicity and a SLLN of the PARNI sampler and its multiple chain acceleration, we require the following assumptions:

(A.1) The weighting function $g : \mathbb{R}^+ \to \mathbb{R}^+$ is C_g-Lipschitz. That is to say for any $t_2 > t_1 > 0$, there exists a constant C_g such that the weighting function g satisfies

$$|g(t_2) - g(t_1)| \leq C_g|t_2 - t_1|. \tag{43}$$

The thresholded function of LIT clearly satisfies this assumption. This is also a common condition for the proper choice of balancing functions. For example, Hastings' choice $g_{\mathrm{H}}(t) = \min\{1, t\}$ follows (45) immediately for $C_g = 1$ and Barker's choice $g_{\mathrm{B}} = t/(1 + t)$ also follows (45) when $C_g = 1$ (i.e. the maximum derivative).

(A.1.a) Given a small positive real number c that is the universal minimum value of the ratio below

$$\frac{\pi(\gamma')p_\eta^{\mathrm{RN}}(k|\gamma')}{\pi(\gamma)p_\eta^{\mathrm{RN}}(k|\gamma)} \tag{44}$$

for all $\gamma, \gamma' \in \Gamma$ and $k \in \mathcal{K}$ and $\eta \in \Delta_\epsilon^p = (\epsilon, 1 - \epsilon)^p$, the weighting function $g : (c, \infty) \to (c, \infty)$ is C_g-Lipschitz. That is to say for any $t_2 > t_1 > c > 0$, there exists a constant C_g such that the weighting function g satisfies

$$|g(t_2) - g(t_1)| \leq C_g|t_2 - t_1|. \tag{45}$$

The square root function $g_{\mathrm{sq}}(t) = \sqrt{t}$ satisfy this condition by setting C_g to be $c^{-1/2}/2$.

(A.2) The posterior distribution π is everywhere positive and bounded, that is, there exists a positive $\Pi \in$

$(1, \infty)$ such that

$$\frac{1}{\Pi} \leq \frac{\pi(\gamma')}{\pi(\gamma)} \leq \Pi$$

for all $\gamma, \gamma' \in \Gamma$.

(A.3) Recall the interval $\Delta_\epsilon^{2p+1} = (\epsilon, 1-\epsilon)^{2p+1}$, the tuning parameters $\theta^{(i)} = (\eta^{(i)}, \omega^{(i)})$ are bounded away from 0 and 1, and lie in this interval

$$\theta^{(i)} \in \Delta_\epsilon^{2p+1} \tag{46}$$

for some small $\epsilon \in (0, 1/2)$.

The analysis of convergence and ergodicity often relies on the distribution of the Markov chain at time i along with its associated total variation distance $\|\cdot\|_{TV}$ at an arbitrary starting point. Given $\{\gamma^{l,(i)}\}_{i=0}^{\infty}$ these are defined as

$$\mathcal{L}^{l,(i)}[(\gamma^l, \theta), S] := \Pr\left[\gamma^{l,(i)} \in S | \gamma^{l,(0)} = \gamma^l, \theta^0 = \theta\right], \tag{47}$$

$$\lim_{i\to\infty} T^l(\gamma^l, \theta, i) := \|\mathcal{L}^{l,(i)}[(\gamma^l, \theta), \cdot] - \pi(\cdot)\|_{TV}. \tag{48}$$

We show here that the PARNI sampler is ergodic and satisfies a strong law of large numbers (SLLN). In mathematical terms for any starting point $\gamma^{\otimes L} \in \Gamma^{\otimes L}$ and $\theta \in \Delta_\epsilon^{2p+1}$ ergodicity means that

$$\lim_{i\to\infty} T^l(\gamma^l, \theta, i) \to 0, \tag{49}$$

for any $l = 1, \ldots, L$, while a strong law of large numbers (SLLN) implies that

$$\frac{1}{NL}\sum_{i=0}^{N-1}\sum_{l=1}^{L} f(\gamma^{l,(i)}) \to \pi(f) \tag{50}$$

almost surely, for any $f : \Gamma \to \mathbb{R}$. We first establish two technical results before presenting the main theorem of this section.

Lemma 1 (Simultaneous Uniform Ergodicity) *The MCMC transition kernel $P_\theta^{PARNI(*)-\bullet}$ in (40) with target distribution π in (37) is simultaneously uniformly ergodic for any choice of $\epsilon \in (0, 1/2)$ in (46). i.e. for any $\delta > 0$, there exists $N = N(\delta, \epsilon)$ such that*

$$\left\| \left(P_\theta^{PARNI(*)-\bullet}(\gamma^{\otimes L}, \cdot) \right)^N - \pi^{\otimes L}(\cdot) \right\|_{TV} \leq \delta$$

holds for any any starting point $\gamma^{\otimes L} \in \Gamma^{\otimes L}$ and any value $\theta \in \Delta_\epsilon^{2p+1}$.

Lemma 2 (Diminishing adaptation) *Let the constant of adaptation rate λ be in $(1/2, 1)$ for $\bullet = RM$ and be $1/2$ for $\bullet = KW$, for any $\epsilon \in (0, 1/2)$ and $\pi_0 \in (0, 1)$, the PARNI sampler satisfies diminishing adaptation, that is, its transition kernel satisfies*

$$\sup_{\gamma\in\Gamma} \left\| P_{\theta^{(i+1)}}^{PARNI(*)-\bullet}(\gamma, \cdot) - P_{\theta^{(i)}}^{PARNI(*)-\bullet}(\gamma, \cdot) \right\|_{TV} \leq C i^{-\lambda} \tag{51}$$

for some constant $C < \infty$.

Theorem 2 (Ergodicity and SLLN) *Consider a target distribution $\pi(\gamma)$ in (37), constant of adaptation rate $\lambda \in (1/2, 1)$ for $\bullet = RM$ and $\lambda = 1/2$ for $\bullet = KW$ and $\epsilon \in (0, 1/2)$ that lead to a adaptation rate $\mathcal{O}(i^{-\lambda})$, and the parameter $\pi_0 > 0$ in Algorithm 3. Then ergodicity (49) and a strong law of large numbers (50) hold for all the PARNI(T)-KW, PARNI(T)-RM, PARNI(B)-KW and PARNI(B)-RM samplers as described in Algorithm 3 and its corresponding multiple chain acceleration versions.*

5 Numerical studies

5.1 Simulated data

We consider the data generation model introduced by Yang et al. (2016), and replicated in simulation studies conducted by Griffin et al. (2021) and Zanella and Roberts (2019). Suppose a linear model with n observations and p covariates is needed, data are generated from the model specification

$$y = X^*\beta^* + \epsilon$$

where $\epsilon \sim N_n(0, \sigma^2 I_n)$ for pre-specified residual variance σ^2 and $\beta^* = \text{SNR} \times \tilde{\beta}\sqrt{(\sigma^2 \log p)/n}$ in which SNR represents the signal-to-noise ratios. Let $\tilde{\beta} = (2, -3, 2, 2, -3, 3, -2, 3, -2, 3, 0, \cdots, 0)$ and each row of the design matrix X_i^* follow a multivariate normal distribution with mean zero and covariance Σ with entries $\Sigma_{jj} = 1$ for all j and $\Sigma_{ij} = 0.6^{|i-j|}$ for $i \neq j$. We consider four choices of SNR, namely 0.5, 1, 2 and 3, two choices of n, namely 500 and 1,000 and three choices of p, namely 500, 5000 and 50,000.

We use the same prior parameter values $V_\gamma = I_{p_\gamma}$, $g = 9$ and $h = 10/p$ as specified in Griffin et al. (2021). In the same work, a detailed description of the resulting posterior distributions is given. In the presence of a low SNR (SNR = 0.5), there is too much noise to detect the true non-zero variables and the resulting posterior is rather flat, with no variables having posterior inclusion probabilities larger than 0.1. The posterior distributions are completely different when the SNR is large (SNR = 2 and SNR = 3). In

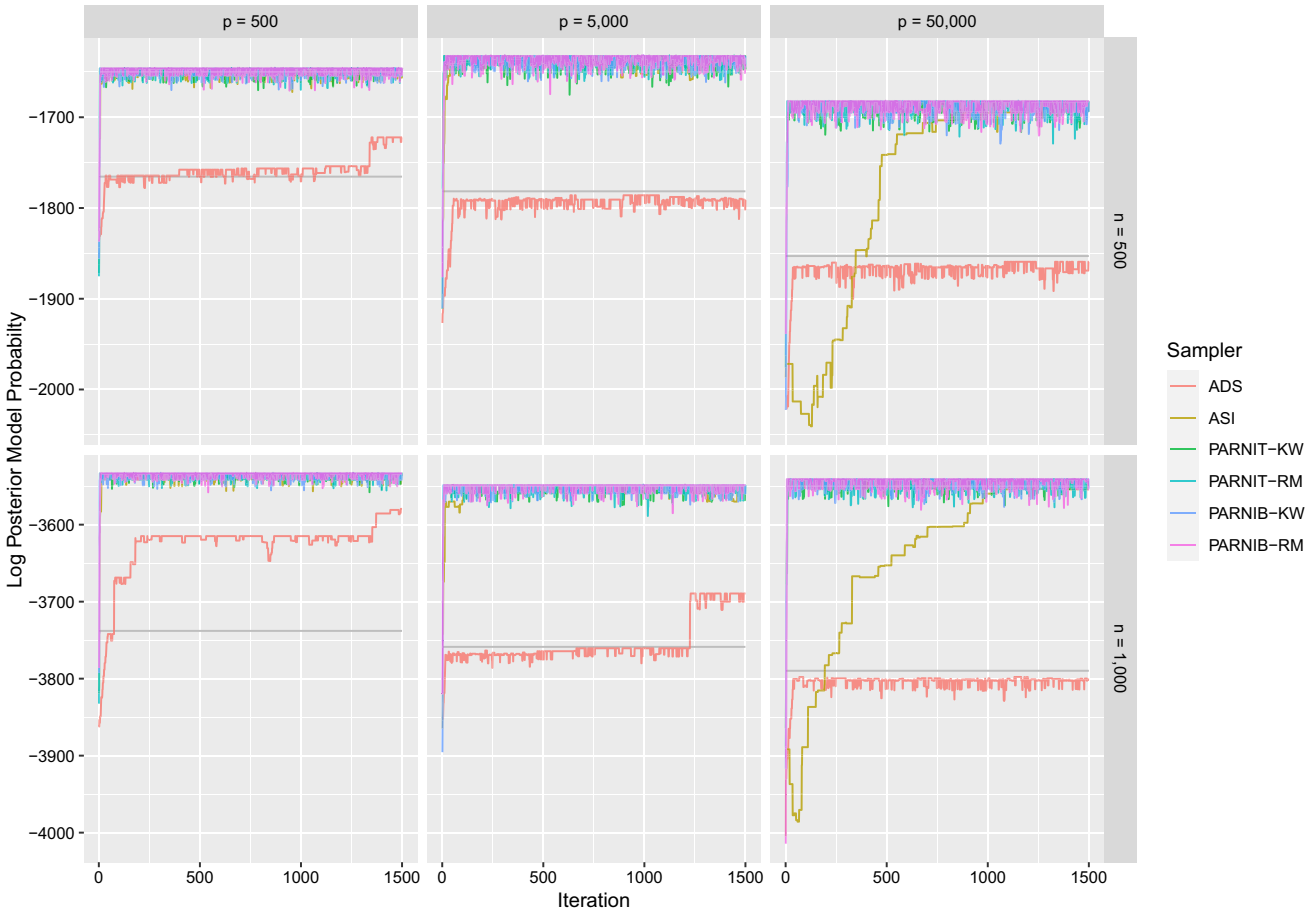

Fig. 2 Simulated data: trace plots of log posterior model probability from the Add-Delete-Swap (ADS), Adaptively Scaled Individual (ASI) adaptation, Pointwise implementation of Adaptive Random Neighbourhood Informed and Thresholded proposal with Kiefer–Wolfowitz update (PARNIT-KW), Pointwise implementation of Adaptive Random Neighbourhood Informed and Thresholded proposal with Robbins-Monro update (PARNIT-RM), Pointwise implementation of Adaptive Random Neighbourhood Informed and balanced proposal with Kiefer–Wolfowitz update (PARNIB-KW) and Pointwise implementation of Adaptive Random Neighbourhood Informed and balanced proposal with Robbins-Monro update (PARNIB-RM) samplers for the first 1500 iterations on simulated datasets with signal-to-noise ratio of 2

these cases all of the true non-zero variables have inclusion probabilities close to 1 as the posterior distributions are more concentrated. In the intermediate case SNR $= 1$ slightly less than half of the true non-zero variables have inclusion probabilities above 0.8. In general the problem of finding the true non-zero variables becomes more difficult in the cases with lower SNR, smaller n and larger p.

We are interested in comparing the performance of the ASI and PARNI schemes relative to an ADS sampler because the ASI scheme has been compared with several other state-of-the-art MCMC algorithms in Griffin et al. (2021). The adaptive algorithms are run with 25 multiple chains. The first third of the chain are identified as the period of burn-in. In addition, to reduce the computational budget, all the adaptations terminate after the period of burn-in.

Trace plots of chains are a straightforward way to visualise convergence. Figure 2 are the trace plots of posterior model probabilities from the ADS, ASI, PARNIT-KW, PARNIT-RM, PARNIB-KW and PARNIB-RM algorithms for the first 1500 iterations when the SNR = 2. The ADS scheme fails to converge for all choices of n and p and in particular becomes trapped at areas around the null model (i.e. the empty model) for a long period of time when $p = 50,000$. The ASI scheme converges reasonably quickly when p is 500 or 5000, but takes longer to reach high probability regions when $p = 50,000$. This suggests that ASI mixes worse and converges slower in high-dimensional data-sets. On the other hand, all the PARNI samplers mix rapidly in this setting for which they only take several moves to converge properly.

The trace plots are not truly a fair comparison as they do not take into account running time. To better address the issue of computational efficiency we ran all of the algorithms for 3 repetitions. Each individual chain was run for 15 min and we stored the estimates of posterior inclusion probabilities. We

Table 1 Simulated data: relative average mean squared errors for the Adaptively scaled individual (ASI), Pointwise implementation of Adaptive Random Neighbourhood Informed and Thresholded proposal with Kiefer–Wolfowitz update (PARNIT-KW), Pointwise implementation of Adaptive Random Neighbourhood Informed and Thresholded proposal with Robbins-Monro update (PARNIT-RM), Pointwise implementation of Adaptive Random Neighbourhood Informed and balanced proposal with Kiefer–Wolfowitz update (PARNIB-KW) and Pointwise implementation of Adaptive Random Neighbourhood Informed and balanced proposal with Robbins-Monro update (PARNIB-RM) schemes on estimating posterior inclusion probabilities over important and unimportant variables respectively against a standard Add-Delete-Swap algorithm

(n, p)	Samplers	SNR			
		0.5	1	2	3
(500, 500)	ASI	− 4.67(− **3.49**)	− 1.61(0.42)	− 1.53(− **0.79**)	− 1.25(− 1.01)
	PARNIT-KW	− 4.48 (− 3.34)	− 1.57 (− 0.69)	− 1.49 (− 0.71)	− 0.86(− 1.04)
	PARNIT-RM	− 4.58 (− 3.33)	− 1.55 (− **0.70**)	− **1.77** (− 0.67)	− **1.28** (− 0.97)
	PARNIB-KW	− **4.71**(− 3.35)	− **1.69**(− 0.69)	− 1.73(− 0.67)	− 0.97(− **1.02**)
	PARNIB-RM	− 4.49(− 3.40)	− 1.58(− **0.70**)	− 1.54(− 0.61)	− 0.74(− 0.94)
(1000, 500)	ASI	− 4.36(− 3.93)	− 0.39(− 0.15)	− 1.45(− **1.00**)	− 0.75(− **1.07**)
	PARNIT-KW	− 3.77 (− 3.96)	− 1.20 (− 1.86)	− 1.24 (− 0.95)	− **0.89** (− 1.02)
	PARNIT-RM	− 3.67 (− 3.88)	− 1.88 (− 1.83)	− 1.23 (− 0.93)	− **0.89** (− 0.95)
	PARNIB-KW	− 4.33(− **3.99**)	− **2.06**(− **1.80**)	− 1.19(− 0.96)	− 0.87(− 0.97)
	PARNIB-RM	− **4.45**(− 3.93)	− 1.92(− 1.48)	− **1.45**(− 0.94)	− 0.72(− 0.91)
(500, 5000)	ASI	− 1.97(− 0.12)	− 1.32(− 0.04)	− 1.20(0.11)	− 1.90(0.13)
	PARNIT-KW	− 2.29 (− **0.62**)	− 1.68 (− 0.12)	− 1.71 (− 0.28)	− 1.93 (− 0.20)
	PARNIT-RM	− 2.52 (− 0.59)	− 1.73 (− **0.53**)	− 1.61 (− 0.29)	− **2.07** (− **0.23**)
	PARNIB-KW	− 2.27(− 0.60)	− **1.98**(− 0.10)	− **1.77**(− 0.28)	− 1.86(− 0.15)
	PARNIB-RM	− **2.66**(− 0.58)	− 1.97(− 0.18)	− 1.66(− **0.33**)	− 1.78(− 0.19)
(1000, 5000)	ASI	− 3.76(− 2.05)	− 0.86(− 0.02)	− 0.66(− 0.01)	− 1.28(− 0.06)
	PARNIT-KW	− 4.01 (− 2.65)	− 0.98 (− 0.45)	− 0.05 (− 0.29)	− 1.79 (− 0.34)
	PARNIT-RM	− 4.05 (− **2.86**)	− 1.05 (− 0.54)	− 0.72 (− **0.45**)	− 2.37 (− 0.25)
	PARNIB-KW	− 4.09(− 2.67)	− **1.18**(− 0.44)	− **1.76**(− 0.38)	− **2.42**(− 0.29)
	PARNIB-RM	− **4.76**(− 2.69)	− 1.14(− **0.52**)	− 0.33(− 0.39)	− 1.80(− **0.35**)
(500, 50, 000)	ASI	− 1.37(0.42)	− 2.55(1.19)	− 3.95(− 0.54)	− 3.35(− 0.52)
	PARNIT-KW	− 2.78 (− 0.90)	− 4.25 (− 2.42)	− 5.15 (− 1.94)	− **5.13** (− 2.06)
	PARNIT-RM	− 2.51 (− **0.91**)	− 4.48 (− **2.44**)	− 4.96 (− 1.95)	− 4.90 (− 2.14)
	PARNIB-KW	− 2.32(− 0.88)	− **4.57**(− 2.41)	− **5.33**(− **1.97**)	− 4.91(− 2.10)
	PARNIB-RM	− **2.77**(− 0.89)	− 4.51(− 2.42)	− 5.19(− 1.93)	− 5.02(− **2.16**)
(1000, 50, 000)	ASI	− 1.60(− 0.42)	− 2.65(0.50)	− 3.94(− 0.87)	− 2.40(1.80)
	PARNIT-KW	− **3.31** (− 1.57)	− **4.13** (− 0.54)	− 4.40 (− 1.38)	− **4.88** (− 0.77)
	PARNIT-RM	− 2.42 (− 1.60)	− 3.24 (− 0.34)	− **4.83** (− 1.45)	− 4.03 (− 0.78)
	PARNIB-KW	− 2.17(− **1.63**)	− 3.72(− **0.61**)	− 4.77(− 1.41)	− 4.35(− 0.76)
	PARNIB-RM	− 2.68(− 1.62)	− 3.83(− 0.58)	− 4.60(− **1.52**)	− 4.57(− **0.80**)

The quantities outside the brackets are for important variables which have posterior inclusion probabilities greater than 0.01 whereas the quantities inside the brackets are for unimportant variables which have posterior inclusion probabilities less than 0.01. Values are presented in logarithm to base 10. Smaller values always indicate better estimates. Values in bold are those methods which have the best performance for each simulated dataset

calculated mean squared errors of these estimates compared to "gold standard" estimates taken from a weighted tempered Gibbs sampler that was run for roughly 12 h. We show results in the form of performance relative to the ADS scheme in Table 1. Smaller values always indicate better performance of the scheme. The value of −1 indicates the scheme yields 10 times smaller mean squared errors compared to those from the ADS scheme in this specific data-set. Generally speaking, the mean squared errors for important variables are greater than those for important variables for almost every data-set and scheme. The choice of n does not significantly affect the performance of the samplers. Concentrating on the results for important variables, the ASI scheme leads to an order of magnitude improvement in efficiency over the ADS sampler, which match the results in Griffin et al. (2021). The four PARNI algorithms with different weighting functions and adaptations lead to similar levels of accuracy and dominate both the ASI and ADS schemes in every case except

$p = 500$. In particular, the PARNI schemes result in roughly 10^5 times improvements over ADS and more than 10 times improvements over ASI when $p = 50,000$ and SNR$= 2$. On the other hand, the ADS scheme is quite adept at removing the unimportant variables when the true model size is small compared to the number of covariates. When $p = 50,000$ and SNR> 1 the ASI scheme struggles with unimportant variables and leads to worse estimates than ADS, but the PARNI algorithms produce better estimates even for these unimportant variables. Overall, the results suggest PARNI samplers are more computationally efficient than alternatives when p is large. More results from simulated data are provided in Section C.3 of the supplementary material.

5.2 Real data

We consider eight real data-sets implemented in Griffin et al. (2021), four of them with moderate p and four with larger p.

The first data-set is the Tecator data-set, which is previously analysed by Brown and Griffin (2010) in Bayesian linear regression and implemented by Lamnisos et al. (2013) and Griffin et al. (2021) in the context of Bayesian variable selection. It contains 172 observations and 100 explanatory variables. We also consider three small p data sets constructed by Schäfer and Chopin (2013) to illustrate the performance of sequential Monte Carlo algorithms on Bayesian variable selection problems, the Boston Housing data ($n = 506$, $p = 104$), the Concrete data ($n = 1030$, $p = 79$) and the Protein data ($n = 96$, $p = 88$). These data sets are extended by squared and interaction terms which lead to high dependencies and multicollinearity.

The last four data sets are high-dimensional problems with very large p. Three of them come from an experiment conducted by Lan et al. (2006) to examine the genetics of two inbred mouse populations. The experiment resulted in a set of data with 60 observations in total that were used to monitor the expression levels of 22, 575 genes of 31 female and 29 male mice. Bondell and Reich (2012) first considered this data-set in the context of variable selection. Three physiological phenotypes are also measured by quantitative real-time polymerase chain reaction (PCR), they are used as possible responses and are named PCRi for $i = 1, 2, 3$ respectively. For more details, see Lan et al. (2006); Bondell and Reich (2012). The last data-set concerns genome-wide mapping of a complex trait. The data are illustrated in Carbonetto et al. (2017). They are body and testis weight measurements recorded for 993 outbred mice, and genotypes at 79,748 single nucleotide polymorphisms (SNPs) for the same mice. The main purpose of the study is to identify genetic variants contributing to variation in testis weight. Thus, we consider the testis weight as response, the body weight as a regressor that is always included in the model and variable selection is performed on the 79,748 SNPs.

Before analysing the performance of MCMC algorithms on the above data-sets, it is worth discussing the selection of an optimal acceptance rate for the PARNI-RM sampler. The optimal scaling property of a Gaussian random walk proposal on some specific forms of target distribution is a well-studied problem. The most commonly used guideline is to seek an average acceptance rate of 0.234 (Gelman et al. 1997). The optimal acceptance rates for sophisticated *informed* proposals involving gradient information are typically larger, e.g. 0.57 for the *Metropolis-adjusted Langevin algorithm* (Grenander and Miller 1994; Roberts and Rosenthal 1998) and 0.65 for *Hamiltonian Monte Carlo* (Duane et al. 1987; Beskos et al. 2013). As our balanced random neighbourhood proposals can be viewed as a discrete analog to these gradient-based algorithms, it is natural to think that the PARNI samplers will have a larger optimal acceptance rate than a random walk Metropolis. To test this, we ran the PARNIT-RM and PARNIB-RM schemes targeting different rates of acceptance on the above data-sets. Figures 3 and 4 show the effect of the average acceptance rate on the expected squared jumping distance and average mean squared errors of these two schemes respectively. Both of the figures imply the same conclusions. Parts (a) and (b) of the figure illustrate the relation between the thinning parameter ω and the average acceptance rate. Bigger values of ω are synonymous with larger jumps and therefore can lead to a smaller average acceptance rate. Parts (c) and (d) of the figure suggest that the maximum average squared jumping distance occurs when the acceptance rate is around 0.65 for all data-sets. Parts (e) and (f) show that the average mean squared error is minimised when the average acceptance rate is around a similar region. Therefore, for problems we have looked at, targeting an average acceptance rate of 0.65 does not perform badly. Similar results for the simulated data-sets of 5.1 are presented in C.1 of the supplementary material. We stress that the PARNIT-KW and PARNIB-KW schemes does not require a target acceptance rate to be chosen, so users who are uncomfortable with having to choose this quantity for a particular data-set are recommended to use this version of the sampler.

We consider a total of ten different MCMC schemes for these sets of data. In addition to the six schemes used in the simulation study (ADS, ASI, PARNIT-KW, PARNIT-RM, PARNIB-KW and PARNIB-RM), we also implement four state-of-the-art algorithms, the Hamming ball sampler (HBS) with radius of 1 of Titsias and Yau (2017), both the tempered Gibbs sampler (TGS) and weighted tempered Gibbs sampler (WTGS) of Zanella and Roberts (2019), and also the Locally Informed and Thresholded (LIT) scheme by Zhou et al. (2021) (which uses same weighting function as the LIT-MH-1 scheme in their paper). All algorithms are run for the same amount of time and compared using average mean squared errors. Only the adaptive schemes are run with

Table 2 Table of prior specifications of 8 real dataset

Dataset	Prior on β_γ	Prior on γ
Tecator, Concrete, Boston Housing, Protein	$\beta_\gamma \sim N_{p_\gamma}(0, 100 I_{p_\gamma})$	$p(\gamma) = h^{p_\gamma}(1-h)^{p-p_\gamma}, h = 5/100$
PCR1, PCR2, PCR3	$\beta_\gamma \sim N_{p_\gamma}(0, 1/2 \times I_{p_\gamma})$	$p(\gamma) = h^{p_\gamma}(1-h)^{p-p_\gamma}, h \sim Be(1, (p-5)/5)$
SNP	$\beta_\gamma \sim N_{p_\gamma}(0, 1/4 \times I_{p_\gamma})$	$p(\gamma) = h^{p_\gamma}(1-h)^{p-p_\gamma}, h = 5/p$

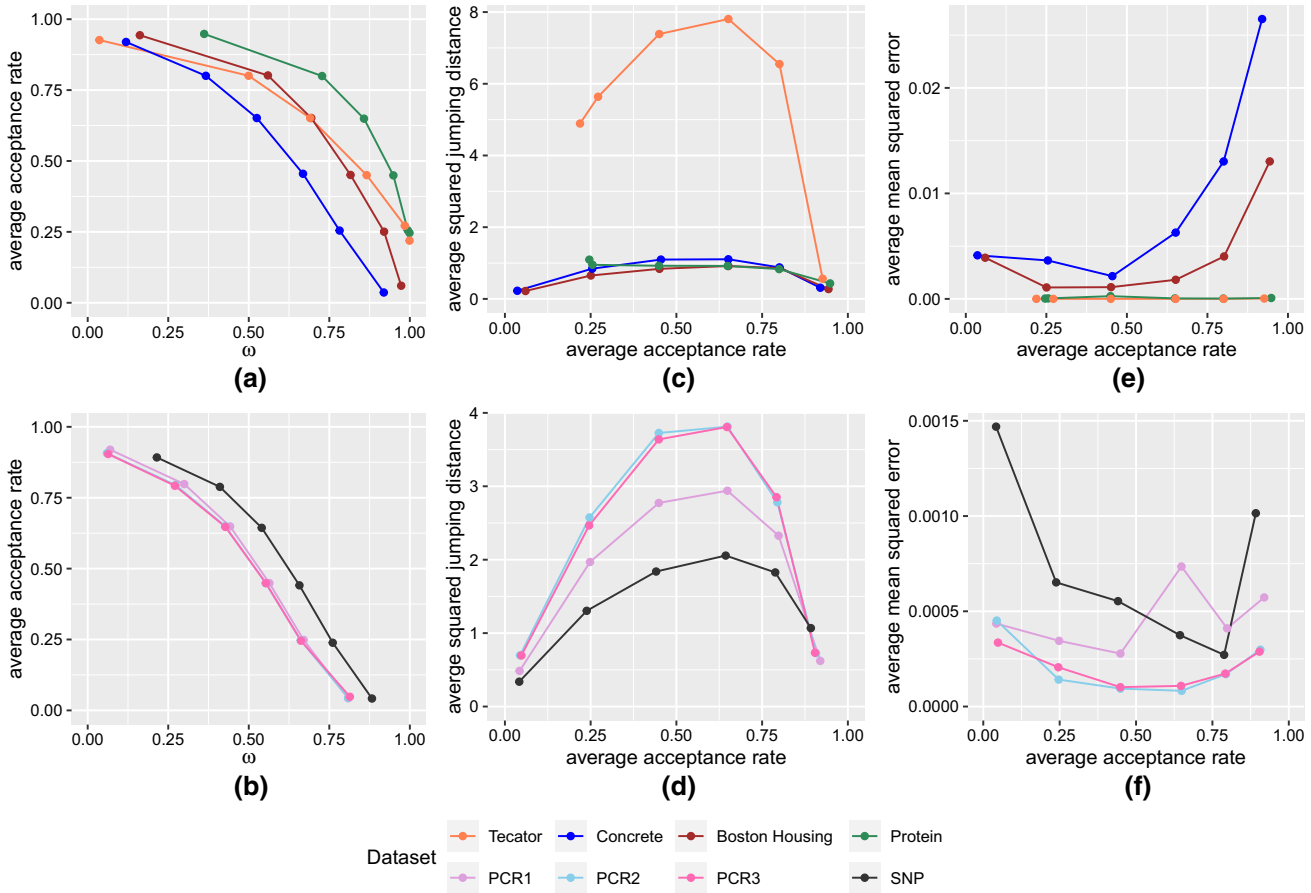

Fig. 3 Real data: plots of expected squared jumping distance and average mean square error again average acceptance rate and ω for the Pointwise implementation of Adaptive Random Neighbourhood Informed and Thresholded proposal with Robbins-Monro update (PARNIT-RM). **a** average acceptance rate against ω for 4 small-p real datasets; **b** average acceptance rate against ω for 4 large-p real datasets; **c** expected squared jumping distance against average acceptance rate for 4 small-p real datasets; **d** expected squared jumping distance against average acceptance rate for 4 large-p real datasets; **e** average mean squared error against average acceptance rate for 4 small-p real datasets; **f** average mean squared error against average acceptance rate for 4 large-p real datasets

25 multiple shorter chains while other schemes use a single longer chain. The prior specification for each data-set is given in Table 2 (Table table:realspsmse).

Figures 5 and 6 show trace plots of posterior model probabilities from the ADS, ASI, PARNIT-KW, PARNIT-RM, PARNIB-KW and PARNIB-RM algorithms for the first 1500 iterations in all eight real data-sets. Overall, the PARNI algorithms perform better than the ADS and ASI schemes in both convergence and mixing. It is clear that the ADS scheme does not mix well since it struggles to explore model space. All algorithms do reach high-probability regions for data-sets with moderate p in roughly the same number of iterations, however the PARNI schemes can reach these high-probability regions faster and accept more jumps inside the model space. In the large p data-sets, these algorithms lead to different behaviour. The ADS scheme gets trapped in the null model and only proposes models around it and the ASI algorithm does not converge properly for the first 1500 iterations either. The PARNI schemes, by contrast, accept almost every proposed states and mix very quickly. They are able

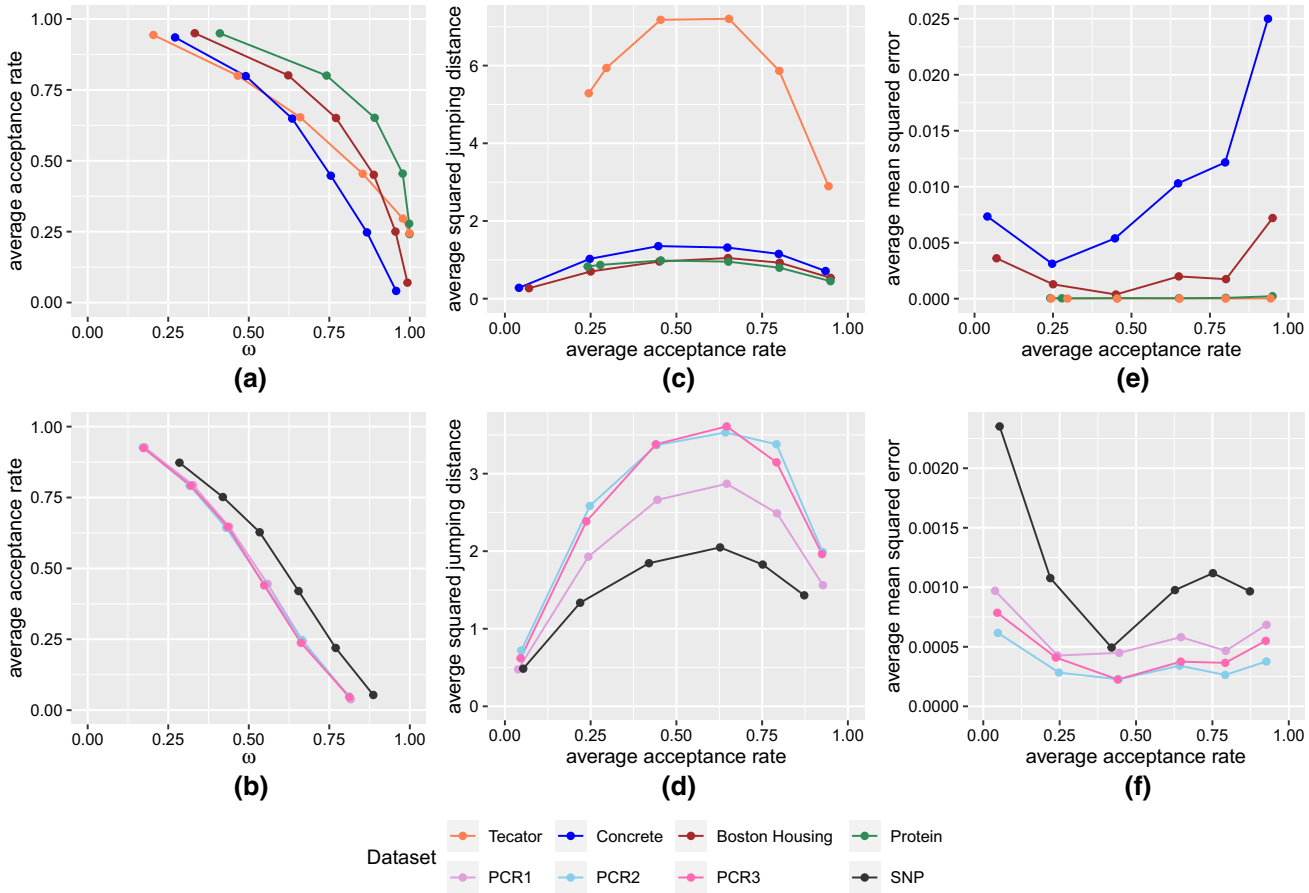

Fig. 4 Real data: plots of expected squared jumping distance and average mean square error again average acceptance rate and ω for the Pointwise implementation of Adaptive Random Neighbourhood Informed and Balanced proposal with Robbins-Monro update (PARNIB-RM). **a** average acceptance rate against ω for 4 small-p real datasets; **b** average acceptance rate against ω for 4 large-p real datasets; **c** expected squared jumping distance against average acceptance rate for 4 small-p real datasets; **d** expected squared jumping distance against average acceptance rate for 4 large-p real datasets; **e** average mean squared error against average acceptance rate for 4 small-p real datasets; **f** average mean squared error against average acceptance rate for 4 large-p real datasets

to propose and accept models with relatively low posterior probabilities and explore the sample space efficiently.

We next turn attention to the average mean squared errors on these eight real data-sets. These results are shown in Table 3. In moderate p data-sets, the PARNI samplers do not dominate other schemes, but they still lead to good results. However, PARNI performs worst for the Boston Housing and Concrete data-sets, which are multi-modal and contain intricately correlated covariates. This implies that the point-wise sub-proposals of PARNI can become trapped at isolated local modes. The ADS scheme performs well in terms of computational efficiency for the Tecator and Concrete data-sets due to a convenient computational implementation which has the cheapest computational costs among competing schemes. Due to the dimension-free mixing property, the LIT scheme outperforms ADS except on the Tecator data-set where all covariates carry non-negligible weights and all covariates are therefore in the potentially influential subset S of Γ (see Sect. 2.3 of Zhou et al. (2021) for more detail). For large p problems all the PARNI schemes significantly outperform other samplers. Surprisingly, the HBS and TGS schemes lead to worse estimates than ADS. This can be explained by the computational cost per iteration of the HBS, TGS and WTGS algorithms, which is linear in p. The combination of these large computational costs and the issue of rarely exploring important variables lead to low efficiencies for HBS and TGS. The WTGS algorithm still outperforms TGS, which coincides with the conclusions gathered in Zanella and Roberts (2019) where the WTGS algorithm is shown to have smaller relaxation time than TGS. The ASI algorithm gives competitive estimates to WTGS in high-dimension but is eventually dominated by the PARNI schemes. The LIT scheme leads to better results in the SNP data-set ($n = 993$) but not in PCR data-sets ($n = 60$) since the dimension-free mixing of LIT only holds when n is comparatively large. And it yields larger average mean squared errors than the PARNI samplers

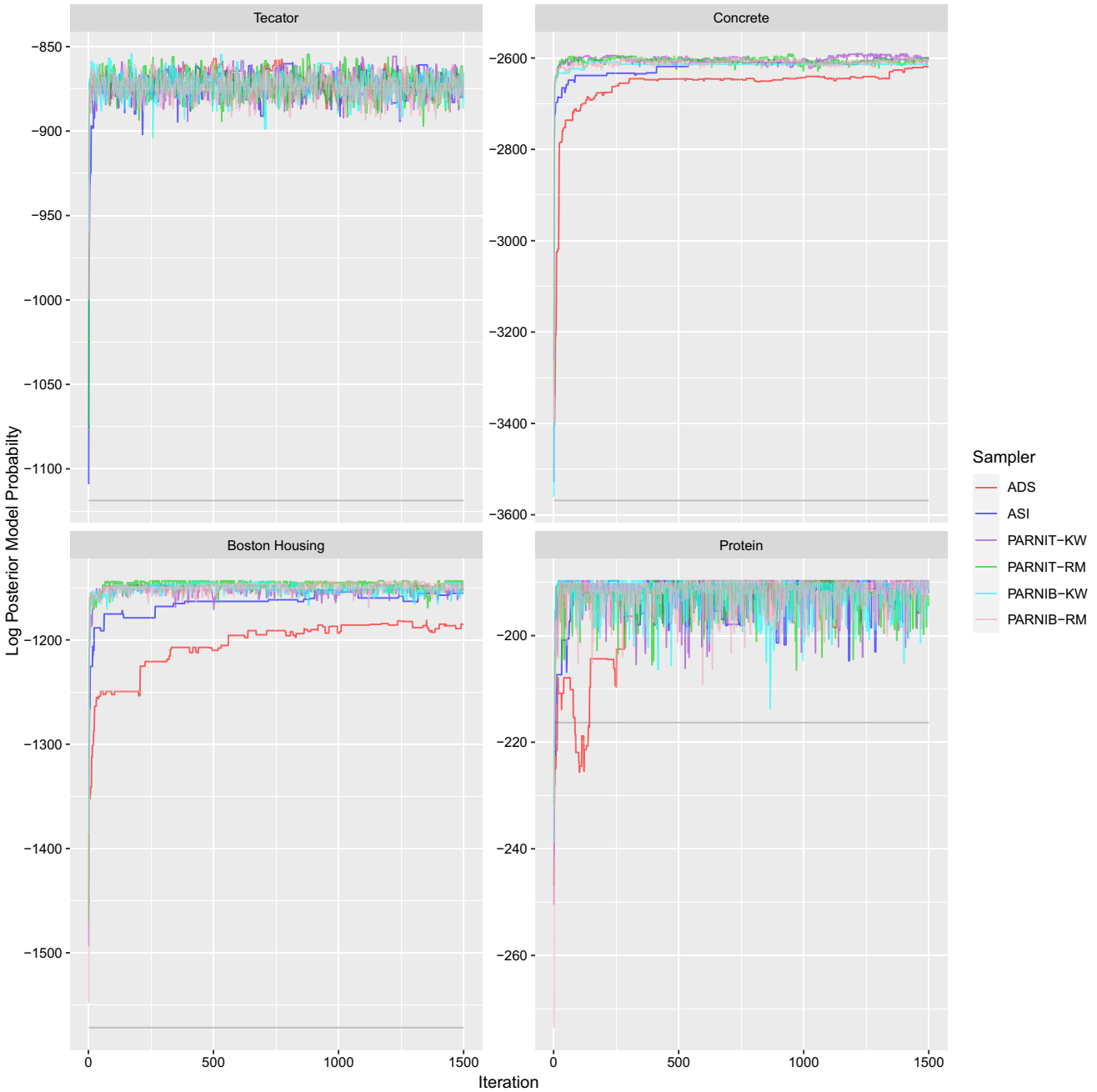

Fig. 5 Real data: trace plots of log posterior model probability from the Add-Delete-Swap (ADS), Adaptively Scaled Individual (ASI) adaptation, Pointwise implementation of Adaptive Random Neighbourhood Informed and Thresholded proposal with Kiefer–Wolfowitz update (PARNIT-KW), Pointwise implementation of Adaptive Random Neighbourhood Informed and Thresholded proposal with Robbins-Monro update (PARNIT-RM), Pointwise implementation of Adaptive Random Neighbourhood Informed and balanced proposal with Kiefer–Wolfowitz update (PARNIB-KW) and Pointwise implementation of Adaptive Random Neighbourhood Informed and balanced proposal with Robbins-Monro update (PARNIB-RM) samplers for the first 1500 iterations on 4 moderate-p datasets

in all large p data-sets because the ADS type neighbourhoods of the LIT scheme only contains models with at most 2 changes in Hamming distance and the jumping distance of LIT therefore is bounded by 2 whereas PARNI can potentially propose larger jumps. Among the PARNI schemes, the two weighting schemes (thresholding and balancing function) have a similar level of efficiency. Specifically, using the thresholding function estimates the posterior inclusion probabilities with lower relative average mean square errors than the balancing function in all four moderate p data-sets and the PCR1 data-set. The performance of all PARNI schemes is similar for the SNP data-set and outperforms that of competi-

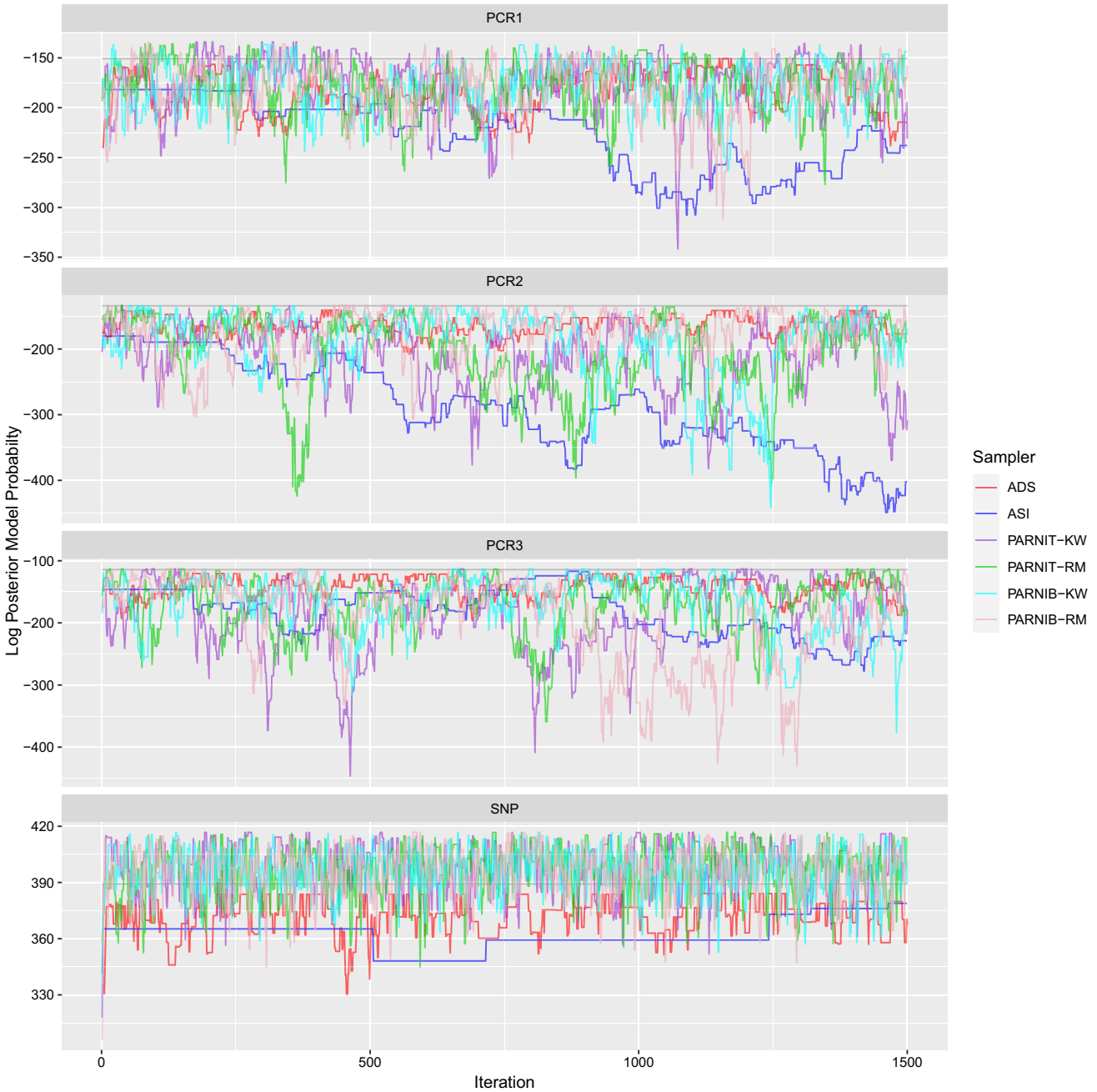

Fig. 6 Real data: trace plots of log posterior model probability from the Add-Delete-Swap (ADS), Adaptively Scaled Individual (ASI) adaptation, Pointwise implementation of Adaptive Random Neighbourhood Informed and Thresholded proposal with Kiefer–Wolfowitz update (PARNIT-KW), Pointwise implementation of Adaptive Random Neighbourhood Informed and Thresholded proposal with Robbins-Monro update (PARNIT-RM), Pointwise implementation of Adaptive Random Neighbourhood Informed and balanced proposal with Kiefer–Wolfowitz update (PARNIB-KW) and Pointwise implementation of Adaptive Random Neighbourhood Informed and balanced proposal with Robbins-Monro update (PARNIB-RM) samplers for the first 1500 iterations on 4 large-p real datasets

tors. In terms of the adaptation schemes, the PARNI sampler with Kiefer–Wolfowitz adaption generally performs better than the Robbins-Monro version, but only by a small margin. This is due to the fact that the optimal acceptance rates are problem-specific and not exactly 0.65 for every data-set.

6 Discussion and future work

In this paper we present a framework for neighbourhood based MCMC algorithms, and propose a new scheme as an *informed* counterpart to the ASI algorithm in Griffin et al. (2021), using elements from locally informed Metropolis-

Table 3 Real data: relative average mean squared errors for the Adaptively Scaled Individual (ASI), Hamming Ball Sampler (HBS), Tempered Gibbs Sampler (TGS), Weighted Tempered Gibbs Sampler (WTGS), Locally Informed and Thresholded (LIT), Pointwise implementation of Adaptive Random Neighbourhood Informed and Thresholded proposal with Kiefer–Wolfowitz update (PARNIT-KW), Pointwise implementation of Adaptive Random Neighbourhood Informed and Thresholded proposal with Robbins-Monro update (PARNIT-RM), Pointwise implementation of Adaptive Random Neighbourhood Informed and balanced proposal with Kiefer–Wolfowitz update (PARNIB-KW) and Pointwise implementation of Adaptive Random Neighbourhood Informed and balanced proposal with Robbins-Monro update (PARNIB-RM) schemes on estimating posterior inclusion probabilities over important variables against a standard Add-Delete-Swap algorithm

dataset	n	p	Samplers								
			ASI	HBS	TGS	WTGS	LIT	PARNIT-KW	PARNIT-RM	PARNIB-KW	PARNIB-RM
Tecator	172	100	**−0.50**	0.31	0.41	−0.25	0.55	0.11	0.06	0.40	0.08
Boston Housing	506	104	0.73	**−1.14**	−0.92	−0.72	−0.57	−0.57	−0.33	−0.38	−0.54
Concrete	1030	79	0.09	0.45	−0.12	**−0.17**	−0.15	0.33	0.40	0.78	0.66
Protein	96	88	−0.47	1.54	−0.51	**−1.05**	−0.31	−0.57	−0.58	−0.10	−0.16
PCR1	60	22,575	−0.91	0.52	0.75	−1.15	−0.20	**−1.42**	−1.24	−1.11	−1.03
PCR2	60	22,575	−0.30	1.12	1.21	−0.56	0.01	−0.68	−0.64	−0.79	**−0.87**
PCR3	60	22,575	−0.11	1.37	1.30	−0.42	0.19	−0.51	−0.43	**−0.58**	−0.50
SNP	993	79,748	−0.57	0.39	0.41	−0.83	−0.97	−1.46	−1.42	**−1.48**	−1.32

Important variables are those variables have posterior inclusion probabilities greater than 0.01. Values are presented in logarithm to base 10. Smaller values always indicate better estimates. Values in bold are those methods which have the best performance for each real dataset

Hastings introduced in Zanella (2020) and Zhou et al. (2021). To address the expensive computational costs introduced by the informed proposal, we introduce two less computationally costly algorithms, the PARNI schemes, which can lead to a dramatic improvement in computational efficiency. In addition, we offer two options of informed weighting functions, the thresholding function and balancing function. The PARNI schemes also allow two different adaptation schemes, the Kiefer–Wolfowitz and Robbins-Monro schemes. The numerical results from Sect. 5 support the power of the algorithmic structure of PARNI. The success of these new schemes is attributed to two aspects. Firstly the adaptation helps to explore the areas of interest (mainly with high posterior probabilities), and secondly the locally informed proposals are able to stabilise random walk behaviour in high-dimensions and lead to rapidly mixing samplers in practice. From the numerical studies on both simulated and real data-sets, we recommend using a PARNI sampler with the Kiefer–Wolfowitz scheme for tackling high-dimensional (or large p) Bayesian variable selection problems. We note that it can still be challenging for the PARNI samplers to move across low probabilistic regions, which could affect performance when the posterior has very isolated modes. This phenomenon is due to the fact that the PARNI samplers propose models sequentially where each sub-proposal can alter only 1 position at most. On the other hand, the original ARNI scheme can take larger jumps and is more able to explore well-separated modes, albeit with a substantial increase in computational costs. In summary, new schemes like PARNI show the potential of combining adaptive, random neighbourhood and informed proposals. We look forward to adding more theoretical support to the numerical evidence shown here in future work. In addition, the code to run the PARNI samplers and aforementioned numerical studies can be downloaded from https://github.com/XitongLiang/The-PARNI-scheme.git.

There are many directions for extensions and future work. Some recent work has shed light on the issues of extra computational costs that come with informed proposals. Grathwohl et al. (2021) develop an accelerated locally informed proposal that uses derivatives with respect to the log mass functions. It is possible to derive the gradient of the posterior mass function with respect to γ with minor modifications to representations of the posterior distribution $\pi(\gamma)$. To address the lack of mode jumping in the PARNI schemes, we can first try to construct larger blocks intelligently so that separated models are covered in one single block. This solution can be achieved by introducing basis vectors beyond the Cartesian case in the block construction. One can also use the sequential Monte Carlo methods of Schäfer and Chopin (2013) and Ma (2015), which are more able to handle multimodality. Combining them with PARNI yields the chance of producing efficient methods on highly multimodal posterior

distributions with well-separated modes. Another option in this direction is the JAMS algorithm of Pompe et al. (2020) that first locates each individual mode and then produces a mixture proposal that involves jumps within and between modes.

We also intend to study the performance of the PARNI schemes in generalised linear models as in Wan and Griffin (2021) or a more flexible Bayesian variable selection model such as that suggested by Rossell and Rubio (2018). In these cases, regression coefficients and residual variance are no longer integrated out analytically and the likelihood of γ is not available in closed form. Informed proposals for such models are computationally challenging because the proposals involve the evaluations of these likelihood but the required approximations and estimates of the marginal likelihood are computationally intensive. One possible approach is the data-augmentation method using the Pólya-gamma distribution as described in Polson et al. (2013). The design does however require some care to avoid inefficiency causing by introducing a large number of auxiliary variables in large n problems. We also believe that random neighbourhood samplers can be used beyond variable selection, and aim to consider applications to other discrete-valued sampling problems in future work.

Acknowledgements We would like to thank an Associated Editor and two referees for their helpful comments. XL thanks Dr. Krzysztof Łatuszyński for the discussion and help for the proof of Lemma 2 and related results. SL is supported by an Engineering and Physical Sciences Research Council New Investigator Award EP/V055380/1.

Declarations

Conflict of interest The authors have declared that no conflicts of interest exist.

A Additional materials

A.1 The Kiefer–Wolfowitz adaption scheme

The optimal scaling property of a Gaussian random walk proposal on some specific forms of target distribution is well-studied. The most commonly used way to achieve optimal mixing time is to tune scaling parameters, which leads to an average acceptance rate of 0.234. In practice, even in those cases where the posterior distribution does not strictly obey the assumptions, the average acceptance rate of 0.234 is often a suitable guide and results in good practical performance. For those proposals beyond random walks, the guidelines for optimal tuning are often unknown. Due to this fact, we develop an adaptation scheme which is able to adapt the tuning parameters in which the mixing time and convergence rate are optimised without knowing any theoretical results in advance.

We design a scheme that maximises the *Expected Squared Jumping Distance* (ESJD). The ESJD is an efficiency measure which accounts for the jumping distances between two consecutive states from a Markov chain which is highly related to first order autocorrelation (Pasarica and Gelman 2010). Suppose that $\xi \in \mathbb{R}$ is a continuous tuning parameter of a π-reversible transition kernel p_ξ. In the PARNI scheme, the scaling parameter ω lies in the interval $(0, 1)$, so we consider a transformation $\xi = \text{logit}_\epsilon \omega$ such that $\xi \in \mathbb{R}$. The definition of ESJD given parameter ξ is given as follows

$$\text{ESJD}(\xi) = \sum_{\gamma\in\Gamma}\sum_{\gamma'\in\Gamma}\left(\sum_{j=1}^{p}(\gamma_j - \gamma_j')^2\right)\pi(\gamma)p_\xi(\gamma,\gamma'). \tag{A.1}$$

If p_ξ is a Metropolis-Hastings transition kernel and can be decomposed into a product of a proposal kernel Q_ξ with mass function q_ξ and a term of the Metropolis-Hasting acceptance probability $\alpha_\xi(\gamma, \gamma')$, the definition of the ESJD above is equivalent to

$$\text{ESJD}(\xi) = \sum_{\gamma\in\Gamma}\sum_{\gamma'\in\Gamma}\left(\sum_{j=1}^{p}(\gamma_j - \gamma_j')^2\right)\pi(\gamma)q_\xi(\gamma,\gamma')\alpha_\xi(\gamma,\gamma'). \tag{A.2}$$

The ESJD is often infeasible to compute since it involves double sum over the sample space. To access the value of ESJD, we consider an estimator *Average Squared Jumping Distance* (ASJD) which depends on the past chain and ASJD

is defined as follows:

$$\text{ASJD}(\xi) = \frac{1}{N}\sum_{i=0}^{N-1}\left(\sum_{j=1}^{p}(\gamma_j^{(i)} - \gamma_j^{(i+1)})^2\right) \tag{A.3}$$

or alternatively

$$\text{ASJD}(\xi) = \frac{1}{N}\sum_{i=0}^{N-1}\left(\sum_{j=1}^{p}(\gamma_j^{(i)} - (\gamma^{(i)})'_j)^2\right)\alpha_\xi(\gamma^{(i)}, (\gamma^{(i)})') \tag{A.4}$$

where $(\gamma^{(i)})'$ is the proposal of $\gamma^{(i)}$ through Q_ξ. From above, the main advance of using ASJD is that ASJD can be easily estimated in each individual iteration.

The objective is to locate the value of ξ that leads to the largest ASJD. This is equivalent to solving the following optimisation problem of the tuning parameter

$$\xi^* := \arg\max_{\xi}\ \text{ASJD}(\xi). \tag{A.5}$$

If objective function ESJD(ξ) is unimodal and smooth, ξ^* can be found by solving the first order ordinary differential equation

$$\frac{d}{\mathrm{d}\xi}\text{ASJD}(\xi) = 0. \tag{A.6}$$

The Robbins-Monro scheme can be applied here to adaptively update the optimal θ when those derivatives exist analytically. In most cases, however, the derivatives are not available analytically, which makes the Robbins-Monro scheme impossible to use. The Kiefer–Wolfowitz scheme (Kiefer and Wolfowitz 1952), on the other hand, is an alternative to the Robbins-Monro algorithm where the derivatives are estimated using a finite difference method.

The following is how a Kiefer–Wolfowitz scheme proceeds. Let $M(\xi)$ be an objective function with a maximum θ^*. If $M(\xi)$ is assumed to be unknown but some random observations $\mathcal{M}(\xi)$ are given such that $M(\xi) = \mathbb{E}[\mathcal{M}(\xi)]$, ξ is updated following an iterative algorithm as follows

$$\xi_{i+1} = \xi_i + a_i\left(\frac{\mathcal{M}(\xi_i + c_i) - \mathcal{M}(\xi_i - c_i)}{2c_i}\right) \tag{A.7}$$

where a_i and c_i are two diminishing sequences of forward positive step sizes and finite difference widths used respectively. In each iteration, we need two independent observations, $\mathcal{M}(\xi_i + c_i)$ and $\mathcal{M}(\xi_i - c_i)$, with tuning parameters $\xi_i + c_i$ and $\xi_i - c_i$ respectively. If the objective function $M(\xi)$ satisfies certain regularity conditions, it can be shown that ξ_i will converge to the optimal value ξ^* as $n \to \infty$. Blum (1954) show that this convergence is almost sure provided that some other conditions hold, most importantly that the diminishing sequences a_i and c_i satisfy

1. $c_i \to 0$ as $i \to \infty$;
2. $\sum_{i=0}^{\infty} a_i = \infty$;
3. $\sum_{i=0}^{\infty} a_i c_i < \infty$;
4. $\sum_{i=0}^{\infty} a_i^2 c_i^{-2} = \infty$.

We design an adaptive MCMC sampler which combines the Kiefer–Wolfowitz scheme and parallel chain implementation. The parallel chain implementation involves a number of independent chains which only share the same tuning parameters and provides independent observations as the Kiefer–Wolfowitz scheme requires. We consider a sampler which involves L parallel chains. In the sampler a new state is proposed through the kernel Q_ξ, which is then accepted with probability α_ξ. An adaptation scheme with the Kiefer–Wolfowitz updates is given as follows: compute $a_i = i^{-\phi_a}$ and $c_i = i^{-\phi_c}$; calculate $\xi^+ = \xi_i + c_i$ and $\xi^- = \xi_i - c_n$; separate the N chains into two groups, L^- and L^+,; for each $l \in L^+$, propose new state $(\gamma^{(l,i)})'$ using $Q_{\xi^+}(\gamma^{(l,i)}, \cdot)$, accept $(\gamma^{(l,i)})'$ with probability $\alpha_{\xi^+}(\gamma^{(l,i)}, (\gamma^{(l,i)})')$; for each $l \in L^-$, propose new state $(\gamma^{(l,i)})'$ using $Q_{\xi^-}(\gamma^{(l,i)}, \cdot)$, accept $(\gamma^{(l,i)})'$ with probability $\alpha_{\xi^-}(\gamma^{(l,i)}, (\gamma^{(l,i)})')$; compute the ASJD for the current iteration by averaging over the chains in groups L^+ and L^- respectively as follows:

$$\begin{aligned}&\text{ASJD}^{\bullet,(i)}\\ &\approx \frac{1}{|L^\bullet|}\sum_{l\in L^\bullet}\left(\sum_{j=1}^{p}(\gamma_j^{(l,i)} - (\gamma^{(l,i)})'_j)^2\right)\alpha_{\theta^\bullet}(\gamma^{(l,i)}, (\gamma^{(l,i)})')\end{aligned} \tag{A.8}$$

where $\bullet$ is either $+$ or $-$; update the tuning parameter for the next iteration by

$$\xi^{(i+1)} = \xi^{(i)} + a_i\left(\frac{\text{ASJD}^{+,(i)} - \text{ASJD}^{-,(i)}}{2c_i}\right). \tag{A.9}$$

B Proofs

B.1 Proof of Proposition 1

The proof relies on Proposition 1 from Andrieu et al. (2020), which in turn relies on Theorem 3 in the same work. The proposition below is a concise summary of the two results sufficient for our needs.

Proposition 4 *Consider a Borel space $(E, \mathcal{E})$ such that any $\xi \in E$ can be written $\xi := (x, y)$ for $x \in \mathsf{X}$ and $y \in \mathsf{Y}$. Define the probability measure μ on E such that $\mu(d\xi) :=$*

$\pi(dx)v_x(dy)$. *Given an involution* $\phi : E \to E$, *then the deterministic Markov kernel*

$$\Pi(\xi, d\xi') := a(\xi)\delta_{\phi(\xi)}(d\xi') + (1 - a(\xi))\delta_{\xi}(d\xi')$$

is μ*-reversible, where* $a(\xi) := \min(1, \mu \circ \phi(d\xi)/\mu(d\xi))$ *if* $\xi \in S$ *and 0 otherwise, for a suitably defined* $S \subset E$ *(see Theorem 3(a) of* Andrieu et al. (2020) *for details). In addition, the marginal transition kernel*

$$P(x, \mathrm{d}x') := \int_{\mathsf{Y}} \Pi((x, y), (\mathrm{d}x', \mathsf{Y}))\nu_x(\mathrm{d}y)$$

is both Markov and π*-reversible.*

Proof of Proposition 1 Define the probability mass function $\mu(\gamma, \gamma', k) := \pi(\gamma)p(k|\gamma)q_k(\gamma, \gamma')$. Then the algorithm can be viewed as a Markov chain on the larger space $E := \Gamma \times \Gamma \times \mathcal{K}$, which alternates between the following steps:

1. Re-sample $k, \gamma'|\gamma$ from its conditional distribution with mass function $p(k|\gamma)q_k(\gamma, \gamma')$
2. Perform a Metropolis step with deterministic proposal

$$\phi\begin{pmatrix}\gamma\\ \gamma'\\ k\end{pmatrix} = \begin{pmatrix}\gamma'\\ \gamma\\ \rho(k)\end{pmatrix}$$

and acceptance probability $\min\left(1, \mu \circ \phi(\gamma, \gamma', k)/\mu(\gamma, \gamma', k)\right)$.

Note that $\phi \circ \phi(\gamma, \gamma', k) = (\gamma, \gamma', k)$, meaning ϕ is an involution. Therefore setting $\mathsf{X} := \Gamma$ and $\mathsf{Y} := \Gamma \times \mathcal{K}$, step 2 can be identified with the deterministic kernel Π in Proposition 4 above, and steps 1 and 2 combined can be identified with the marginal kernel P on the state space Γ, which is therefore π-reversible. □

B.2 Proof of Proposition 2

Proof of Proposition 2 Recall the interval $\Delta_\epsilon^{2p} = (\epsilon, 1-\epsilon)^{2p}$, Since $\eta \in \Delta_\epsilon^{2p}$, it is clear that $p_\eta^{\mathrm{RN}}(k|\gamma) > \epsilon^p$ for all $\gamma \in \Gamma$ and $k \in \mathcal{K}$. Recall that $q_{\omega,k}^{\mathrm{THIN}}$ is symmetric, so we have $q_{\omega,k}^{\mathrm{THIN}}(\gamma, \gamma') = q_{\omega,k}^{\mathrm{THIN}}(\gamma', \gamma)$. The condition

$$\begin{aligned} &p_\eta^{\mathrm{RN}}(k|\gamma)q_{\omega,k}^{\mathrm{THIN}}(\gamma, \gamma') > 0 \\ &\quad\iff p_\eta^{\mathrm{RN}}(k|\gamma')q_{\omega,k}^{\mathrm{THIN}}(\gamma', \gamma) > 0 \end{aligned} \tag{B.1}$$

follows immediately.

To show that $p_\eta^{\mathrm{RN}}(\cdot|\gamma)$ is a probability measure on $\mathcal{K}$ for any $\gamma \in \Gamma$, we also need to show that

$$\sum_{k\in\mathcal{K}} p_\eta^{\mathrm{RN}}(k|\gamma) = 1. \tag{B.2}$$

Starting from the right hand side gives

$$\begin{aligned}\sum_{k\in\mathcal{K}} p_\eta^{\mathrm{RN}}(k|\gamma) &= \sum_{k\in\mathcal{K}}\prod_{j=1}^{p} p_{\eta,j}^{\mathrm{RN}}(k_j|\gamma_j) \\ &= \prod_{j=1}^{p}\left(p_{\eta,j}^{\mathrm{RN}}(k_j = 0|\gamma_j) + p_{\eta,j}^{\mathrm{RN}}(k_j = 1|\gamma_j)\right). \\ &= 1\end{aligned}$$

as required.

We then show $q_{\omega,k}(\gamma, \cdot)$ is a probability measure on set $N(\gamma, k)$ for any $\gamma \in \Gamma$ and $k \in \mathcal{K}$. Let J be a projection of k to a set $J(k)$ consisting of the indices j for which $k_j = 1$ (i.e. $J(k) = \{j|k_j = 1\}$). Starting from Eq. (18) of *Remark* 4 together with the identity used above $\int_\Gamma p(d\gamma) = \prod_j \int_{\Gamma_j} p_j(\mathrm{d}\gamma_j)$, we obtain

$$\begin{aligned}&\sum_{\gamma'\in N(\gamma,k)} q_{\omega,k}^{\mathrm{THIN}}(\gamma, \gamma') \\ &= \prod_{j\in J(k)}\sum_{\gamma'_j\in\{0,1\}}\left(\frac{\omega}{1-\omega}\right)^{d_H(\gamma_j,\gamma'_j)}(1-\omega) \\ &= \prod_{j\in J(k)}\sum_{\gamma'_j\in\{\gamma_j,1-\gamma_j\}}\left(\frac{\omega}{1-\omega}\right)^{|\gamma_j-\gamma'_j|}(1-\omega) \\ &= \prod_{j\in J(k)}((1-\omega)+\omega) \\ &= 1\end{aligned}$$

as required. □

B.3 Proof of Theorem 1

Before proving Theorem 1, we first draw some conclusions on acceptance probability of the ASI and ARN proposals.

Proposition 5 *Suppose* γ *is the current state,* γ' *is the proposed state and* $\eta \in \Delta_\epsilon^{2p} = (\epsilon, 1-\epsilon)^{2p}$ *is fixed parameter. For any* k *that constructs neighbourhood containing* γ *and* γ' *and any choices of* $\xi \in \Delta_\epsilon = (\epsilon, 1-\epsilon)$ *and* $\omega \in \Delta_\epsilon$, *the Metropolis-Hastings acceptance probability of the ARN proposal,* $\alpha_{(\xi\eta,\omega),k}^{ARN}$, *as in* (17) *is fixed. In addition, this term is also the acceptance probability of the ASI proposal, i.e.*

$$\alpha_{(\xi\eta,\omega),k}^{ARN}(\gamma, \gamma') = \alpha_{\zeta\eta}^{ASI}(\gamma, \gamma').$$

for any $\zeta \in \Delta_\epsilon$.

Proof of Proposition 5 Suppose that $\gamma, \gamma' \in \Gamma$, $\eta \in \Delta_\epsilon^{2p}$ are given and fixed. We consider all the $k \in \mathcal{K}$ such that $\gamma' \in N(\gamma, k)$. We are going to show that for any ξ and

$\omega \in \Delta_\epsilon$, the acceptance probability $\alpha^{\text{ARN}}_{\theta,k}(\gamma,\gamma')$ is free from the choice of k, ξ and ω.

To locate the different positions between γ and γ', we define the set $J(\gamma,\gamma') := \{j|\gamma_j \neq \gamma'_j\} \subseteq J(k)$. From (17), we have

$$\begin{aligned}\alpha^{\text{ARN}}_{\theta,k}(\gamma,\gamma') &= \min\left\{1, \frac{\pi(\gamma')p^{\text{RN}}_{\xi\eta^{\text{opt}}}(k|\gamma')q^{\text{THIN}}_{\omega,k}(\gamma',\gamma)}{\pi(\gamma)p^{\text{RN}}_{\xi\eta^{\text{opt}}}(k|\gamma)q^{\text{THIN}}_{\omega,k}(\gamma,\gamma')}\right\}\\ &= \min\left\{1, \frac{\pi(\gamma')p^{\text{RN}}_{\xi\eta^{\text{opt}}}(k|\gamma')}{\pi(\gamma)p^{\text{RN}}_{\xi\eta^{\text{opt}}}(k|\gamma)}\right\}\end{aligned}$$

where the last equality follows since $q^{\text{THIN}}_{\omega,k}$ is symmetric. Substituting $p^{\text{RN}}_{\xi\eta}$ yields

$$\begin{aligned}&\alpha^{\text{ARN}}_{\theta,k}(\gamma,\gamma')\\ &= \min\left\{1, \frac{\pi(\gamma')}{\pi(\gamma)}\prod_{j=1}^{p}\frac{(\xi A_j)^{(1-\gamma'_j)k_j}(1-\xi A_j)^{(1-\gamma'_j)(1-k_j)}(\xi D_j)^{\gamma'_j k_j}(1-\xi D_j)^{\gamma'_j(1-k_j)}}{(\xi A_j)^{(1-\gamma_j)k_j}(1-\xi A_j)^{(1-\gamma_j)(1-k_j)}(\xi D_j)^{\gamma_j k_j}(1-\xi D_j)^{\gamma_j(1-k_j)}}\right\}\\ &= \min\left\{1, \frac{\pi(\gamma')}{\pi(\gamma)}\prod_{j\in J(\gamma,\gamma')}\frac{(A_j)^{(1-\gamma'_j)}(D_j)^{\gamma'_j}}{(A_j)^{(1-\gamma_j)}(D_j)^{\gamma_j}}\right\}.\end{aligned}$$

The value of $\alpha^{\text{ARN}}_{\theta,k}(\gamma,\gamma')$ only depends on the choice of η and it does not involve the terms ξ, ω and k. Therefore, the proposition is proved. In addition, this is also the Metropolis-Hastings acceptance probability for ASI of proposing γ' to γ following the definition of the ASI sampler. □

Now we formally prove Theorem 1.

Proof of Theorem 1 Rewriting the transition kernel of $p^{\text{ARN}}_{(\xi\eta,\omega)}$ gives

$$\begin{aligned}p^{\text{ARN}}_{(\xi\eta,\omega)}(\gamma,\gamma') &= \sum_{k\in\mathcal{K}} p^{\text{RN}}_{\xi\eta}(k|\gamma)q^{\text{THIN}}_{\omega,k}(\gamma,\gamma')\alpha^{\text{ARN}}_{(\xi\eta,\omega),k}(\gamma\gamma')\\ &= \alpha^{\text{ASI}}_{\zeta\eta}(\gamma,\gamma')\sum_{k\in\mathcal{K}} p^{\text{RN}}_{\xi\eta}(k|\gamma)q^{\text{THIN}}_{\omega,k}(\gamma,\gamma')\end{aligned} \tag{B.3}$$

for which the last line follows from Proposition 5 and the fact that $\zeta = \xi \times \omega$.

Recall the definitions of the conditional distribution of k in (15) and the within-neighbourhood proposal in (18). Together with $\eta = (A, D)$, we have

$$\begin{aligned}p^{\text{RN}}_{\xi\eta}(k|\gamma) &= \prod_{j=1}^{p} p^{\text{RN}}_{\xi\eta,j}(k_j|\gamma_j)\\ &= \prod_{j=1}^{p}(\xi A_j)^{(1-\gamma_j)k_j}(1-\xi A_j)^{(1-\gamma_j)(1-k_j)}\\ &\quad (\xi D_j)^{\gamma_j k_j}(1-\xi D_j)^{\gamma_j(1-k_j)}\end{aligned}$$

and

$$\begin{aligned}q^{\text{THIN}}_{\omega,k}(\gamma,\gamma') &= \left(\frac{\omega}{1-\omega}\right)^{d_H(\gamma,\gamma')}(1-\omega)^{\sum_{j=1}^{p}k_j}\\ &= \left(\frac{\omega}{1-\omega}\right)^{d_H(\gamma,\gamma')}\prod_{j=1}^{p}(1-\omega)^{k_j}\end{aligned}$$

where the last line follows since $(1-\omega)^{k_j} = 1$ if $k_j = 0$. Substituting the above into (B.3) yields

$$\begin{aligned}&p^{\text{ARN}}_{(\xi\eta,\omega)}(\gamma,\gamma') = \alpha^{\text{ASI}}_{\zeta\eta}(\gamma,\gamma')\\ &\sum_{k\in\mathcal{K}}\left[\prod_{j=1}^{p}(\xi A_j)^{(1-\gamma_j)k_j}(1-\xi A_j)^{(1-\gamma_j)(1-k_j)}\right.\\ &\left.(\xi D_j)^{\gamma_j k_j}(1-\xi D_j)^{\gamma_j(1-k_j)}\left(\frac{\omega}{1-\omega}\right)^{d_H(\gamma,\gamma')}(1-\omega)^{k_j}\right].\end{aligned} \tag{B.4}$$

Let $J(\gamma,\gamma')$ be a set which consists of the indices j for which $\gamma_j \neq \gamma'_j$ and $J(\gamma,\gamma')^c$ be the complement to $J(\gamma,\gamma')$. By definition k_j must be 1 when $j \in J(\gamma,\gamma')$ and k_j can take values either 0 or 1 when $j \in J(\gamma,\gamma')^c$. Continuing from (B.4), we divide $j = 1$ to p into two groups, $J(\gamma,\gamma')$ and $J(\gamma,\gamma')^c$, and obtain

$$\begin{aligned}&p^{\text{ARN}}_{(\xi\eta,\omega)}(\gamma,\gamma')\\ &= \alpha^{\text{ASI}}_{\zeta\eta}(\gamma,\gamma')\left(\frac{\omega}{1-\omega}\right)^{|J(\gamma,\gamma')|}\\ &\quad\prod_{j\in J(\gamma,\gamma')}\left((\xi A_j)^{(1-\gamma_j)}(\xi D_j)^{\gamma_j}(1-\omega)\right)\\ &\quad\times\prod_{j\in J(\gamma,\gamma')^c}\left[\sum_{k_j\in\{0,1\}}(\xi A_j)^{(1-\gamma_j)k_j}(1-\xi A_j)^{(1-\gamma_j)(1-k_j)}\right.\\ &\quad\left.(\xi D_j)^{\gamma_j k_j}(1-\xi D_j)^{\gamma_j(1-k_j)}(1-\omega)^{k_j}\right]\end{aligned}$$

$$= \alpha^{\mathrm{ASI}}_{\zeta\eta}(\gamma,\gamma') \prod_{j\in J(\gamma,\gamma')} \left((\xi A_j)^{(1-\gamma_j)}(\xi D_j)^{\gamma_j}\cdot\omega\right) \prod_{j\in J(\gamma,\gamma')^c} \times \underbrace{\left((1-\xi A_j)^{1-\gamma_j}(1-\xi D_j)^{\gamma_j} + (\xi A_j)^{1-\gamma_j}(\xi D_j)^{\gamma_j}(1-\omega)\right)}_{I_j}$$

We am going to further investigate the terms I_j for $j \in J(\gamma,\gamma')^c$. Clearly that, if $\gamma_j = 1$, then

$$\begin{aligned} I_j &= (1-\xi D_j) + \xi D_j(1-\omega) \\ &= 1-\omega\xi D_j, \end{aligned}$$

similarly that when $\gamma_j = 0$, we have

$$\begin{aligned} I_j &= (1-\xi A_j) + \xi A_j(1-\omega) \\ &= 1-\omega\xi A_j. \end{aligned}$$

Putting everything back to $p^{\mathrm{ARN}}_{\xi\eta,\omega}$, and reconstructing the product in j from 1 to p gives

$$\begin{aligned} p^{\mathrm{ARN}}_{\xi\eta,\omega}(\gamma,\gamma') &= \alpha^{\mathrm{ASI}}_{\zeta\eta}(\gamma,\gamma') \times \prod_{j\in J(\gamma,\gamma')} (\xi\omega A_j)^{(1-\gamma_j)}(\xi\omega D_j)^{\gamma_j} \\ &\quad \times \prod_{j\in J(\gamma,\gamma')^c} (1-\xi\omega A_j)^{(1-\gamma_j)}(1-\xi\omega D_j)^{\gamma_j} \\ &= \alpha^{\mathrm{ASI}}_{\zeta\eta}(\gamma,\gamma') \times \prod_{j=1}^{p} \Big[(\xi\omega A_j)^{(1-\gamma_j)\gamma'_j}(\xi\omega D_j)^{\gamma_j(1-\gamma'_j)} \\ &\quad (1-\xi\omega A_j)^{(1-\gamma_j)(1-\gamma'_j)}(1-\xi\omega D_j)^{\gamma_j\gamma'_j} \Big]. \end{aligned}$$

Rewriting the above in terms of $\zeta = \xi \times \omega$ yields

$$\begin{aligned} p^{\mathrm{ARN}}_{(\xi\eta,\omega)}(\gamma,\gamma') &= \alpha^{\mathrm{ASI}}_{\zeta\eta}(\gamma,\gamma') \\ &\quad \times \prod_{j=1}^{p} \Big[(\zeta A_j)^{(1-\gamma_j)\gamma'_j}(\zeta D_j)^{\gamma_j(1-\gamma'_j)} \\ &\quad \times (1-\zeta A_j)^{(1-\gamma_j)(1-\gamma'_j)}(1-\zeta D_j)^{\gamma_j\gamma'_j} \Big] \\ &= p^{\mathrm{ASI}}_{\zeta\times\eta}(\gamma,\gamma') \end{aligned}$$

as required. □

B.4 Proof of Corollary 1

Proof of Corollary 1 It is clear that the argument holds from the Theorem 1 and its proof. □

B.5 Proof of Proposition 3

Proof We define $J(k)$ to be a set that consists of the positions that $k_j = 1$ (i.e. $J(k) = \{j|k_j = 1\}$) and $J(\gamma,\gamma')$ to be a set which consists of the different variables (i.e. $J(\gamma,\gamma') = \{j|\gamma_j \neq \gamma'_j\}$). They obey the relationship $J(\gamma,\gamma') \subseteq J(k)$ if $\gamma' \in N(\gamma,k)$. The j-th conditional distribution of k_j, $p^{\mathrm{RN}}_{\eta,j}$, satisfies

$$p^{\mathrm{RN}}_{\eta,j}(k_j|\gamma_j) = p^{\mathrm{RN}}_{\eta,j}(k_j|\gamma'_j)$$

if j is outside of $J(\gamma,\gamma')$. This is because $\gamma_j = \gamma'_j$ for $j \in J(\gamma,\gamma')$.

Start with simplifying the ratio $p^{\mathrm{RN}}_{\eta}(k|\gamma')/p^{\mathrm{RN}}_{\eta}(k|\gamma)$. Following the similar steps in the proof of Proposition 5 and suppose $\gamma_j = 1-\gamma'_j$ for all $j \in J(\gamma,\gamma')$ and $\eta = (A,D)$, we can show the following

$$\begin{aligned} \frac{p^{\mathrm{RN}}_{\eta}(k|\gamma')}{p^{\mathrm{RN}}_{\eta}(k|\gamma)} &= \prod_{j\in J(\gamma,\gamma')} \frac{A_j^{1-\gamma_j}D_j^{\gamma_j}}{A_j^{\gamma_j}D_j^{1-\gamma_j}} \\ &= \prod_{j\in J(\gamma,\gamma')} \left(\frac{A_j}{D_j}\right)^{1-2\gamma_j} = \prod_{j\in J(\gamma,\gamma')} \left(\frac{A_j}{D_j}\right)^{2\gamma'_j-1}. \end{aligned} \tag{B.5}$$

Next step is simplifying the second ratio $q^{\mathrm{PARNI}}_{\theta,k}(\gamma',\gamma)/q^{\mathrm{PARNI}}_{\theta,k}(\gamma,\gamma')$ and showing that this term can be decomposed into 3 parts. Since the sampling process is sequential, the model $\gamma(r)$ is proposed from $\gamma(r-1)$ at time r. For reversed move, the model $\gamma'(r)$ is proposed from $\gamma'(r-1)$ at time r. Moreover, $\gamma(0)$ and $\gamma'(p_k)$ are the current state γ and $\gamma'(p_k)$ and $\gamma'(0)$ are the final proposal γ'. We correlate r and $r' = p_k - r + 1$ since $\gamma(r) = \gamma'(r'-1)$ and $K_r = K'_{r'}$. We consider the ratio $q^{\mathrm{PARNI}}_{\theta,K'_{r'}}(\gamma'(r'-1),\gamma'(r'))/q^{\mathrm{PARNI}}_{\theta,K_r}(\gamma(r-1),\gamma(r))$. From (28) and therefore have

$$\begin{aligned} &\frac{q^{\mathrm{PARNI}}_{\theta,K_{r'}}(\gamma'(r'-1),\gamma'(r'))}{q^{\mathrm{PARNI}}_{\theta,K_r}(\gamma(r-1),\gamma(r))} \\ &= \frac{g\left(\frac{\pi(\gamma'(r'))p^{\mathrm{RN}}_{\eta}(e(K_{r'})|\gamma'(r'))}{\pi(\gamma'(r'-1))p^{\mathrm{RN}}_{\eta}(e(K_{r'})|\gamma'(r'-1))}\right)/Z'(r')}{g\left(\frac{\pi(\gamma(r))p^{\mathrm{RN}}_{\eta}(e(K_r)|\gamma(r))}{\pi(\gamma(r-1))p^{\mathrm{RN}}_{\eta}(e(K_r)|\gamma(r-1))}\right)/Z(r)}. \end{aligned}$$

Since g is a balancing function and satisfies $g(t) = tg(1/t)$ for any positive t and $\gamma(r) = \gamma'(r'+1)$, we have

$$\begin{aligned} &\frac{q^{\mathrm{PARNI}}_{\theta,K_{r'}}(\gamma'(r'-1),\gamma'(r'))}{q^{\mathrm{PARNI}}_{\theta,K_r}(\gamma(r-1),\gamma(r))} \\ &= \frac{\pi(\gamma(r-1))p^{\mathrm{RN}}_{\eta}(e(K_r)|\gamma(r-1))}{\pi(\gamma(r))p^{\mathrm{RN}}_{\eta}(e(K_r)|\gamma(r))} \cdot \frac{Z(r)}{Z'(r')}. \end{aligned}$$

The product of the above ratio from $r = 1$ to p_k yields the term $q^{\mathrm{PARNI}}_{\theta,k}(\gamma',\gamma)/q^{\mathrm{PARNI}}_{\theta,k}(\gamma,\gamma')$ as follows

$$\frac{q^{\mathrm{PARNI}}_{\theta,k}(\gamma',\gamma)}{q^{\mathrm{PARNI}}_{\theta,k}(\gamma,\gamma')}$$

$$= \frac{\prod_{r'=1}^{p_k} q_{\theta,K_{r'}}^{\text{PARNI}}(\gamma'(r'-1),\gamma'(r'))}{\prod_{r=1}^{p_k} q_{\theta,K_r}^{\text{PARNI}}(\gamma(r-1),\gamma(r))}$$
$$= \prod_{r=1}^{p_k} \frac{\pi(\gamma(r-1))p_\eta^{\text{RN}}(e(K_r)|\gamma(r-1))}{\pi(\gamma(r))p_\eta^{\text{RN}}(e(K_r)|\gamma(r))} \cdot \frac{Z(r)}{Z'(r')}$$
$$= \underbrace{\left(\prod_{r=1}^{p_k} \frac{\pi(\gamma(r-1))}{\pi(\gamma(r))}\right)}_{\text{I}} \cdot \underbrace{\left(\prod_{r=1}^{p_k} \frac{p_\eta^{\text{RN}}(e(K_r)|\gamma(r-1))}{p_\eta^{\text{RN}}(e(K_r)|\gamma(r))}\right)}_{\text{II}} \cdot \underbrace{\left(\prod_{r=1}^{p_k} \frac{Z(r)}{Z'(r)}\right)}_{\text{III}}$$

since $r' = p_k - r$.

The first term I is equal to

$$\text{I} = \frac{\pi(\gamma)}{\pi(\gamma(1))} \frac{\pi(\gamma(1))}{\pi(\gamma(2))} \cdots \frac{\pi(\gamma(p_k-2))}{\pi(\gamma(p_k-1))} \frac{\pi(\gamma(p_k-1))}{\pi(\gamma')}.$$

Most terms can be cancelled out and this leaves the first numerator and the last denominator

$$\text{I} = \frac{\pi(\gamma)}{\pi(\gamma')}.$$

We now deal with the second term II. Substituting the values gives

$$\frac{p_\eta^{\text{RN}}(e(K_r)|\gamma(r-1))}{p_\eta^{\text{RN}}(e(K_r)|\gamma(r)))} = \left(\frac{A_{K_r}}{D_{K_r}}\right)^{1-2\gamma(r)_{K_r}}.$$

We know that the positions $K_1, \ldots, K_{p_k}$ are distinct and the vectors $e(K_1), \ldots, e(K_{p_k})$ can recover the auxiliary variable k. Therefore, we obtain

$$\text{II} = \prod_{j \in J(\gamma,\gamma')} \left(\frac{A_j}{D_j}\right)^{1-2\gamma_j}.$$

Following the above arguments, the product of sequence $j = 1, \ldots, p_k$ can be simplified

$$\frac{q_{\theta,k}^{\text{PARNI}}(\gamma',\gamma)}{q_{\theta,k}^{\text{PARNI}}(\gamma,\gamma')} = \underbrace{\frac{\pi(\gamma)}{\pi(\gamma')}}_{\text{I}} \cdot \underbrace{\prod_{j \in J(\gamma,\gamma')} \left(\frac{A_j}{D_j}\right)^{1-2\gamma_j}}_{\text{II}} \cdot \underbrace{\prod_{r=1}^{p_k} \frac{Z(r)}{Z'(r)}}_{\text{III}} \tag{B.6}$$

The Metropolis-Hastings acceptance probability in (31) is

$$\alpha_{\theta,k}^{\text{PARNI}}(\gamma,\gamma') = \left\{1, \frac{\pi(\gamma')p_\eta^{\text{RN}}(k|\gamma')q_{\theta,k}^{\text{PARNI}}(\gamma',\gamma)}{\pi(\gamma)p_\eta^{\text{RN}}(k|\gamma)q_{\theta,k}^{\text{PARNI}}(\gamma,\gamma')}\right\}$$
$$= \min\left\{1, \left(\frac{\pi(\gamma')}{\pi(\gamma)} \cdot \text{I}\right) \cdot \left(\frac{p_\eta^{\text{RN}}(k|\gamma')}{p_\eta^{\text{RN}}(k|\gamma)} \cdot \text{II}\right) \cdot \text{III}\right\}$$

From (B.5) and (B.6), we have

$$\frac{\pi(\gamma')}{\pi(\gamma)} \cdot \text{I} = 1$$
$$\frac{p_\eta^{\text{RN}}(k|\gamma')}{p_\eta^{\text{RN}}(k|\gamma)} \cdot \text{II} = 1$$

and we therefore obtain

$$\alpha_{\theta,k}^{\text{PARNI}}(\gamma,\gamma') = \min\{1, \text{III}\} = \min\left\{1, \prod_{j=1}^{p_k} \frac{Z(j)}{Z'(j)}\right\}$$

as required. □

B.6 Proof of Lemma 1

The proof of Lemma 1 is structured as proof of Lemma 1 in Griffin et al. (2021). We first recall a tailored version of a general result that is well-known in the literature, see e.g. Theorem 8 of Roberts and Rosenthal (2004), Theorem 1 of Roberts and Rosenthal (2007).

Proposition 6 *Consider a family of π-invariant Markov kernels $\{P_\theta\}_{\theta\in\Theta}$ defined on some countable state space Γ. If there exists $\epsilon > 0$ such that for every $(\gamma,\gamma') \in \Gamma \times \Gamma$ it holds that*

$$\inf_\theta P_\theta(\gamma,\gamma') \geq \epsilon\pi(\gamma'), \tag{B.7}$$

then the family $\{P_\theta\}_{\theta\in\Theta}$ satisfies a simultaneous uniform ergodicity condition. Namely, for every $\delta > 0$ there exists a finite $N := N(\delta)$ such that

$$\|P_\theta^N(\gamma,\cdot) - \pi(\cdot)\|_{TV} < \delta$$

for all $\gamma \in \Gamma$ and all $\theta \in \Theta$.

Proof of Lemma 1 We first introduce some preliminary work. Since $\theta = (\eta,\omega) \in \Delta_\epsilon^{2p+1} = (\epsilon, 1-\epsilon)^{2p+1}$, we have

$$\epsilon^p \leq p_\eta^{\text{RN}}(k|\gamma) \leq (1-\epsilon)^p \tag{B.8}$$

for any k and γ. Similar arguments implies that

$$\left(\frac{\epsilon}{1-\epsilon}\right)^p \leq \frac{p_\eta^{\text{RN}}(k|\gamma')}{p_\eta^{\text{RN}}(k|\gamma)} \leq \left(\frac{1-\epsilon}{\epsilon}\right)^p. \tag{B.9}$$

From assumption (A.2), we know that there exists a constant Π such that

$$\frac{1}{\Pi} \leq \frac{\pi(\gamma')}{\pi(\gamma)} \leq \Pi. \tag{B.10}$$

Let $t_{\gamma,\gamma',k}$ be

$$t_{\gamma,\gamma',k} = \frac{\pi(\gamma')}{\pi(\gamma)} \cdot \frac{p_\eta^{\mathrm{RN}}(k|\gamma')}{p_\eta^{\mathrm{RN}}(k|\gamma)}. \tag{B.11}$$

Using (B.9) and (B.10) leads to

$$\frac{1}{\Pi} \cdot \left(\frac{\epsilon}{1-\epsilon}\right)^p \leq t_{\gamma,\gamma',k} \leq \Pi \cdot \left(\frac{1-\epsilon}{\epsilon}\right)^p, \tag{B.12}$$

for any $\gamma, \gamma' \in \Gamma$ and $k \in \mathcal{K}$, thus,

$$g\left(\frac{1}{\Pi} \cdot \left(\frac{\epsilon}{1-\epsilon}\right)^p\right) \leq g(t_{\gamma,\gamma',k}) \leq g\left(\Pi \cdot \left(\frac{1-\epsilon}{\epsilon}\right)^p\right)$$

since g is a non-decreasing function. We define the following qualities

$$g^\uparrow := g\left(\Pi \cdot \left(\frac{1-\epsilon}{\epsilon}\right)^p\right)$$
$$g^\downarrow := g\left(\frac{1}{\Pi} \cdot \left(\frac{\epsilon}{1-\epsilon}\right)^p\right)$$

Therefore, for all normalising constants $Z(r)$, it is bounded between $Z^\downarrow$ and $Z^\uparrow$ where

$$Z^\downarrow := 2\epsilon g^\downarrow \tag{B.13}$$
$$Z^\uparrow := 2(1-\epsilon) g^\uparrow \tag{B.14}$$

due to the fact that $\omega \in (\epsilon, 1-\epsilon)$.

Suppose k and corresponding K are given, we now bound each individual kernel $q_{\theta,K_r}^{\mathrm{PARNI}}(\gamma,\gamma')$ for any γ and r from 1 to p_k. Staring from the definition in (28),

$$\begin{aligned} q_{\theta,K_r}^{\mathrm{PARNI}}(\gamma,\gamma') =& g\left(\frac{\pi(\gamma')p_\eta^{\mathrm{RN}}(e(K_r)|\gamma')}{\pi(\gamma)p_\eta^{\mathrm{RN}}(e(K_r)|\gamma)}\right) \\ & q_{\omega,e(K_r)}^{\mathrm{THIN}}(\gamma,\gamma')/Z(r) \\ \geq& \frac{\epsilon g^\downarrow}{Z(r)} \\ \geq& \frac{\epsilon g^\downarrow}{2(1-\epsilon)g^\uparrow} \end{aligned} \tag{B.15}$$

where the last is followed by (B.14). We next consider the full update kernel of PARNI

$$q_{\theta,k}^{\mathrm{PARNI}}(\gamma,\gamma') = \prod_{r=1}^{p_k} q_{\theta,K_r}^{\mathrm{PARNI}}(\gamma(r-1),\gamma(r)) \tag{B.16}$$

where $\gamma(0) = \gamma$ and $\gamma(p_k) = \gamma'$. From (B.15), the full update kernel is bounded as follows

$$\begin{aligned} q_{\theta,k}^{\mathrm{PARNI}}(\gamma,\gamma') &\geq \left(\frac{\epsilon g^\downarrow}{2(1-\epsilon)g^\uparrow}\right)^{p_k} \\ &\geq \left(\frac{\epsilon g^\downarrow}{2(1-\epsilon)g^\uparrow}\right)^{p} \end{aligned} \tag{B.17}$$

since $p_k \leq p$ for all k. We also bound the Metropolis-Hastings acceptance probability in (31) from below

$$\begin{aligned} \alpha_{\theta,k}^{\mathrm{PARNI}}(\gamma,\gamma') &= \min\left\{1, \frac{\pi(\gamma')p_\eta^{\mathrm{RN}}(k|\gamma')q_{\theta,k}^{\mathrm{PARNI}}(\gamma',\gamma)}{\pi(\gamma)p_\eta^{\mathrm{RN}}(k|\gamma)q_{\theta,k}^{\mathrm{PARNI}}(\gamma,\gamma')}\right\} \\ &\geq \pi(\gamma')p_\eta^{\mathrm{RN}}(k|\gamma')q_{\theta,k}^{\mathrm{PARNI}}(\gamma',\gamma) \\ &\geq \pi_m\left(\frac{\epsilon g^\downarrow}{2(1-\epsilon)g^\uparrow}\right)^p \epsilon^p \end{aligned}$$

where $\pi_m = \min_\gamma \pi(\gamma)$.

Finally, we can chose b such that

$$b = \pi_m\left(\frac{\epsilon^2 g^\downarrow}{2(1-\epsilon)g^\uparrow}\right)^{2p}$$

and therefore

$$\begin{aligned} p_\theta^{\mathrm{PARNI}}(\gamma,\gamma') &= \sum_{k\in\mathcal{K}} p_{\theta,k}^{\mathrm{PARNI}}(\gamma,\gamma') \\ &= \sum_{k\in\mathcal{K}} q_{\theta,k}^{\mathrm{PARNI}}(\gamma,\gamma')\alpha_{\theta,k}^{\mathrm{PARNI}}(\gamma,\gamma') \\ &\geq \sum_{k\in\mathcal{K}} b \geq b. \end{aligned}$$

Since $\pi(\gamma') < 1$ for any $\gamma' \in \Gamma$, it follows that Eq. (B.7) is satisfied for $\{p_\theta^{\mathrm{PARNI}}\}_{\theta\in\Theta}$, and therefore by Proposition 6 the family of kernels is simultaneously uniformly ergodic.

For L multiple chains, the arguments are similar, but instead, the target distribution is now $\pi^{\otimes L}(\gamma^{\otimes L})$ for $\gamma^{\otimes L} \in \Gamma^{\otimes L}$ and $\Gamma^{\otimes L}$ is $(1, b^{\otimes L}, \pi^{\otimes L}(\cdot))$-small for which

$$b^{\otimes L} = \pi_m^L\left(\frac{\epsilon^2 g^\downarrow}{2(1-\epsilon)g^\uparrow}\right)^{2pL}.$$

□

B.7 Proof of Lemma 2

Before proving the lemma, we require the following inequalities and its generalised version

Lemma 3

$$\left|\prod_{j=1}^p a_j - \prod_{j=1}^p b_j\right| \leq \sum_{j=1}^p |a_j - b_j| \tag{B.18}$$

for all $a_j, b_j \in [0, 1]$.

Proof of Lemma 3 Let $D(p)$ is the LHS of (B.18)

$$D(p) = \prod_{j=1}^{p} a_j - \prod_{j=1}^{p} b_j. \tag{B.19}$$

In addition, we define a telescopic sum $S(k)$ such that

$$S(p) = \sum_{i=1}^{p} a_1 \times \cdots \times a_{i-1}(a_i - b_i)b_{i+1} \times \cdots \times b_p. \tag{B.20}$$

The proof of Lemma 3 is structured as follows: (1) show that $D(p) = S(p)$ for $p \geq 2$ by induction; (2) using triangular inequality and condition that all a_j and b_j are bounded between 0 and 1 to prove the LHS of (B.18) is equal to its RHS.

Step (1), we prove

$$D(p) = S(p) \tag{B.21}$$

for $p \geq 2$ by induction.

Base case: When $p = 2$, rearranging the RHS of (B.21)

$$\begin{aligned} S(2) &= (a_1 - b_1)b_2 + a_1(a_2 - b_2) \\ &= a_1b_2 - b_1b_2 + a_1a_2 - a_1b_2 \\ &= a_1a_2 - b_1b_2 = D(2) \end{aligned}$$

which equals to the RHS. We therefore proved that (B.21) holds when $p = 2$.

Induction step: Let $k \leq 2$ be given and suppose (B.21) is true for $p = k$. Then

$$\begin{aligned} D(k+1) &= \prod_{j=1}^{k+1} a_j - \prod_{j=1}^{k+1} b_j \\ &= \left(\prod_{j=1}^{k} a_j\right) a_{k+1} - \left(\prod_{j=1}^{k} b_j\right) b_{k+1}. \end{aligned}$$

By applying $(a_1 - b_1)b_2 + a_1(a_2 - b_2) = a_1a_2 - b_1b_2$ which we just showed, we obtain

$$\begin{aligned} D(k+1) &= \left(\prod_{j=1}^{k} a_j - \prod_{j=1}^{k} b_j\right) b_{k+1} \\ &\quad + \left(\prod_{j=1}^{k} a_j\right)(a_{k+1} - b_{k+1}) \\ &= S(k)b_{k+1} + \left(\prod_{j=1}^{k} a_j\right)(a_{k+1} - b_{k+1}) \\ &= S(k+1). \end{aligned}$$

Therefore, (B.21) holds for $p = k + 1$.

Conclusion: By the principle of induction, (B.21) holds for $p \geq 2$.

Step (2), we start with the RHS of (B.18) and use the above statement

$$\begin{aligned} \left|\prod_{j=1}^{p} a_j - \prod_{j=1}^{p} b_j\right| &= \left|D(p)\right| = \left|S(p)\right| \\ &= \left|\sum_{i=1}^{p} a_1 \times \cdots \times a_{i-1}(a_i - b_i)b_{i+1} \times \cdots \times b_p\right|. \end{aligned}$$

Applying the triangular inequality gives

$$\begin{aligned} &\left|\prod_{j=1}^{p} a_j - \prod_{j=1}^{p} b_j\right| \\ &\leq \sum_{i=1}^{p} \left|a_1 \times \cdots \times a_{i-1}(a_i - b_i)b_{i+1} \times \cdots \times b_p\right| \\ &\quad \times \sum_{i=1}^{p} a_1 \times \cdots \times a_{i-1}\left|a_i - b_i\right|\left|b_{i+1} \times \cdots \times b_p\right| \end{aligned}$$

where the last term follows since all a_j and b_j are non-negative. Because a_j and b_j are bounded between 0 and 1, we then have

$$\left|\prod_{j=1}^{p} a_j - \prod_{j=1}^{p} b_j\right| \sum_{i=1}^{p} \left|a_i - b_i\right|$$

which is the inequality (B.18). Therefore, we completed the proof. □

We can generalise the above lemma and obtain the following.

Lemma 4 *If* $a_j, b_j \leq C$ *for some constant* $C > 0$, *then*

$$\left|\prod_{j=1}^{p} a_j - \prod_{j=1}^{p} b_j\right| \leq C_1 \sum_{j=1}^{p} |a_j - b_j| \tag{B.22}$$

for some constant C_1. C_1 *can be chosen to be* C^{p-1}.

Proof

$$\left|\prod_{j=1}^{p} a_j - \prod_{j=1}^{p} b_j\right| = C^p \left|\prod_{j=1}^{p} \frac{a_j}{C} - \prod_{j=1}^{p} \frac{b_j}{C}\right| := A$$

As $a_j, b_j \leq C$, a_j/C and b_j/C are smaller than 1.

$$A \leq C^p \sum_{j=1}^{p} \left|\frac{a_j}{C} - \frac{b_j}{C}\right|$$

$$= C^{p-1} \sum_{j=1}^{p} |a_j - b_j|$$

as required. □

The following lemma shows the diminishing rate of the proposal thinning parameter ω for both schemes (the PARNI-KW and PARNI-RM proposals).

Lemma 5 *The diminishing rate of adaptive parameter ω in both (35) and (36) between two consecutive iterations satisfies*

$$|\omega^{(i+1)} - \omega^{(i)}| = \mathcal{O}(i^{-\lambda}) \tag{B.23}$$

for some $\lambda > 0$. In particular, setting $\phi_i = i^{-\lambda}$ for $\lambda \in (1/2, 1)$ in (35), $a_i = i^{-1}$ and $c_i = i^{-0.5}$ in (36) as suggested, (B.23) holds for $\lambda \in (1/2, 1)$ and $\lambda = 0.5$ respectively.

Proof of Lemma 5 The update rule of (35) immediately leads to

$$|\omega^{(i+1)} - \omega^{(i)}| = \mathcal{O}(i^{-\lambda})$$

for $\lambda \in (1/2, 1)$.

For the Kiefer–Wolfowitz updating law in (36), the values of tuning parameters ω adopted involve diminishing sequence c_i. We are therefore interested in

$$\left| |\omega^{(i+1)} \pm c_{i+1}| - |\omega^{(i)} \pm c_i| \right|. \tag{B.24}$$

By applying the reverse triangle inequality, we obtain

$$\begin{aligned}
&\left| |\omega^{(i+1)} \pm c_{i+1}| - |\omega^{(i)} \pm c_i| \right| \\
&\leq \left| (\omega^{(i+1)} \pm c_{i+1}) - (\omega^{(i)} \pm c_i) \right| \\
&\leq \left| \omega^{(i+1)} + c_{i+1} - \omega^{(i)} + c_i \right| \\
&\leq \left| \omega^{(i+1)} - \omega^{(i)} \right| + \left| c_{i+1} + c_i \right| := S_1 + S_2,
\end{aligned}$$

Starting from the first term S_1 and rearranging (A.9), we obtain

$$\begin{aligned}
S_1 \leq & \left| a_i \left(\frac{\text{ASJD}^{+,(i)} - \text{ASJD}^{-,(i)}}{2c_i} \right) \right| \\
\leq & p \left| \frac{a_i}{c_i} \right| \\
= & \mathcal{O}(i^{-(\phi_a - \phi_c)}) = \mathcal{O}(i^{-0.5})
\end{aligned}$$

where the second line follows from the fact that the expected jumping distances are bounded above by p.

Substituting the definition of c_i into S_2 yields

$$\begin{aligned}
S_2 \leq & \left| (i+1)^{-\phi_c} + (i)^{-\phi_c} \right| \\
\leq & \left| 2i^{-\phi_c} \right| \\
= & \mathcal{O}(i^{-\phi_c}) = \mathcal{O}(i^{-0.5}).
\end{aligned}$$

Since both terms S_1 and S_2 are of the same order of $\mathcal{O}(i^{-0.5})$, the Eq. (B.24) is also $\mathcal{O}(i^{-0.5})$, which completes the proof. □

We also require the following lemma to bound transition kernels by proposal kernels. The lemma and its proof is inspired by Lemma 4.21 in Łatuszyński et al. (2013).

Lemma 6 *The sub-proposal kernel in (39) and sub-transition kernel in (41) obey the following relationship:*

$$\begin{aligned}
&\sup_{\gamma \in \Gamma} \sup_{\gamma' \in \Gamma} \sup_{k \in \mathcal{K}} \left| p^{PARNI}_{\theta^{(i+1)},k}(\gamma, \gamma') - p^{PARNI}_{\theta^{(i)},k}(\gamma, \gamma') \right| \\
&\leq C \sup_{\gamma \in \Gamma} \sup_{\gamma' \in \Gamma} \sup_{k \in \mathcal{K}} \left| \psi^{PARNI}_{\theta^{(i+1)},k}(\gamma, \gamma') - \psi^{PARNI}_{\theta^{(i)},k}(\gamma, \gamma') \right|
\end{aligned} \tag{B.25}$$

for some constant $C < \infty$.

Proof of Lemma 6 Let the left-hand side and right-hand side of (B.25) be L and R respectively, namely

$$L = \sup_{\gamma \in \Gamma} \sup_{\gamma' \in \Gamma} \sup_{k \in \mathcal{K}} \left| p^{\text{PARNI}}_{\theta^{(i+1)},k}(\gamma, \gamma') - p^{\text{PARNI}}_{\theta^{(i)},k}(\gamma, \gamma') \right| \tag{B.26}$$

$$R = \sup_{\gamma \in \Gamma} \sup_{\gamma' \in \Gamma} \sup_{k \in \mathcal{K}} \left| \psi^{\text{PARNI}}_{\theta^{(i+1)},k}(\gamma, \gamma') - \psi^{\text{PARNI}}_{\theta^{(i)},k}(\gamma, \gamma') \right|. \tag{B.27}$$

Starting from the definition of sub-proposal kernel

$$\begin{aligned}
p^{\text{PARNI}}_{\theta,k}(\gamma, \gamma') = & \psi^{\text{PARNI}}_{\theta,k}(\gamma, \gamma') \alpha^{\text{PARNI}}_{\theta,k}(\gamma, \gamma') + \mathbb{I}\{\gamma = \gamma'\} \\
& \sum_{\gamma' \in \Gamma} \psi^{\text{PARNI}}_{\theta,k}(\gamma, \gamma')(1 - \alpha^{\text{PARNI}}_{\theta,k}(\gamma, \gamma'))
\end{aligned}$$

and substituting it into the left-hand side of (B.25), we obtain

$$\begin{aligned}
&\left| p^{\text{PARNI}}_{\theta^{(i+1)},k}(\gamma, \gamma') - p^{\text{PARNI}}_{\theta^{(i)},k}(\gamma, \gamma') \right| \\
&\leq \left| \psi^{\text{PARNI}}_{\theta^{(i+1)},k}(\gamma, \gamma') \alpha^{\text{PARNI}}_{\theta^{(i+1)},k}(\gamma, \gamma') \right. \\
&\quad \left. - \psi^{\text{PARNI}}_{\theta^{(i)},k}(\gamma, \gamma') \alpha^{\text{PARNI}}_{\theta^{(i)},k}(\gamma, \gamma') \right|
\end{aligned}$$

$$+\mathbb{I}\{\gamma=\gamma'\}\sum_{\gamma'\in\Gamma}\Big|\psi^{\text{PARNI}}_{\theta^{(i+1)},k}(\gamma,\gamma')(1-\alpha^{\text{PARNI}}_{\theta^{(i+1)},k}(\gamma,\gamma'))$$
$$-\psi^{\text{PARNI}}_{\theta^{(i)},k}(\gamma,\gamma')(1-\alpha^{\text{PARNI}}_{\theta^{(i)},k}(\gamma,\gamma'))\Big| := \text{I}+\text{II}$$

Starting with the term I and substituting in the definition of Metropolis-Hastings acceptance probability in (31) gives

$$\text{I}=\Big|\min\{\psi^{\text{PARNI}}_{\theta^{(i+1)},k}(\gamma,\gamma'),\frac{\pi(\gamma')}{\pi(\gamma)}\psi^{\text{PARNI}}_{\theta^{(i+1)},k}(\gamma',\gamma)\}$$
$$-\min\{\psi^{\text{PARNI}}_{\theta^{(i)},k}(\gamma,\gamma'),\frac{\pi(\gamma')}{\pi(\gamma)}\psi^{\text{PARNI}}_{\theta^{(i)},k}(\gamma',\gamma)\}\Big| \quad \text{(B.28)}$$

Using $|\min\{a,b\}-\min\{c,d\}|\le|a-c|+|b-d|$ to further split I

$$\begin{aligned}\text{I}&\le|\psi^{\text{PARNI}}_{\theta^{(i+1)},k}(\gamma,\gamma')-\psi^{\text{PARNI}}_{\theta^{(i)},k}(\gamma,\gamma')|\\&\quad+\frac{\pi(\gamma')}{\pi(\gamma)}|\psi^{\text{PARNI}}_{\theta^{(i+1)},k}(\gamma',\gamma)-\psi^{\text{PARNI}}_{\theta^{(i)},k}(\gamma',\gamma)|\\&\le(1+\Pi)R\end{aligned} \quad \text{(B.29)}$$

where the last line follows from the assumption (A.2). An analogous calculation gives

$$\text{II}\le(2+K)R, \quad \text{(B.30)}$$

which together with (B.29) implies that

$$\Big|p^{\text{PARNI}}_{\theta^{(i+1)},k}(\gamma,\gamma')-p^{\text{PARNI}}_{\theta^{(i)},k}(\gamma,\gamma')\Big|\le(3+2\Pi)R$$

for any values of γ,γ',k and O, hence, it is enough to prove

$$L\le C\cdot R \quad \text{(B.31)}$$

for $C=(3+2\Pi)$ as required. □

The proof of Lemma 2 is structured similarly to the proof of Lemma 2 in Griffin et al. (2021).

Proof of Lemma 2 The proof is structured as follows: firstly, we re-write the problem as a sum of sub-transition kernels and bound the sub-transition kernels by sub-proposal kernels; secondly, we break sub-proposal kernels into various parts; lastly, we bound each part individually and hence bound the proposal kernels.

Starting from the total variation norm in (51) and substituting in the definition of $P^{\text{PARNI}}_{\theta}$ in (40), we have

$$\begin{aligned}&\sup_{\gamma\in\Gamma}\|P^{\text{PARNI}}_{\theta^{(i+1)}}(\gamma,\cdot)-P^{\text{PARNI}}_{\theta^{(i)}}(\gamma,\cdot)\|_{TV}\\&=\sup_{\gamma\in\Gamma}\sum_{\gamma'\in\Gamma}\sum_{k\in\mathcal{K}}\Big|p^{\text{PARNI}}_{\theta^{(i+1)},k}(\gamma,\gamma')-p^{\text{PARNI}}_{\theta^{(i)},k}(\gamma,\gamma')\Big|\\&\le C_1\sup_{\gamma\in\Gamma}\sup_{\gamma'\in\Gamma}\sup_{k\in\mathcal{K}}\Big|p^{\text{PARNI}}_{\theta^{(i+1)},k}(\gamma,\gamma')-p^{\text{PARNI}}_{\theta^{(i)},k}(\gamma,\gamma')\Big|\\&:=\text{I}\end{aligned}$$

for some constant $C_1<\infty$. Using Lemma 6, the problem is reduced to bounding the largest variations in two consecutive proposal kernels

$$\text{I}\le C_2\sup_{\gamma\in\Gamma}\sup_{\gamma'\in\Gamma}\sup_{k\in\mathcal{K}}\Big|\psi^{\text{PARNI}}_{\theta^{(i+1)},k}(\gamma,\gamma')-\psi^{\text{PARNI}}_{\theta^{(i)},k}(\gamma,\gamma')\Big| \quad \text{(B.32)}$$

for some constant $C_2<\infty$. Plugging in the definition of $\psi^{\text{PARNI}}_{\theta,k}$ into (39), therefore,

$$\text{I}\le C_3\sup_{\gamma\in\Gamma}\sup_{\gamma'\in\Gamma}\sup_{k\in\mathcal{K}}\Big|p^{\text{RN}}_{\eta^{(i+1)}}(k|\gamma)q^{\text{PARNI}}_{\theta^{(i+1)},k}(\gamma,\gamma')-p^{\text{RN}}_{\eta^{(i)}}(k|\gamma)q^{\text{PARNI}}_{\theta^{(i)},k}(\gamma,\gamma')\Big| \quad \text{(B.33)}$$

for some constant $C_3<\infty$. Applying Lemma 3 to (B.33), we obtain

$$\text{I}\le C_3\sup_{\gamma\in\Gamma}\sup_{\gamma'\in\Gamma}\sup_{k\in\mathcal{K}}\Big(\underbrace{\Big|p^{\text{RN}}_{\eta^{(i+1)}}(k|\gamma)-p^{\text{RN}}_{\eta^{(i)}}(k|\gamma)\Big|}_{\text{II}}$$
$$+\underbrace{\Big|q^{\text{PARNI}}_{\theta^{(i+1)},k}(\gamma,\gamma')-q^{\text{PARNI}}_{\theta^{(i)},k}(\gamma,\gamma')\Big|}_{\text{III}}\Big).$$

Now we move our attention to the next part of the proof where we are going to bound terms II and III respectively.

Starting with II, recall the definition of p^{RN}_{η} in (15) and $\eta=(A,D)$, we have

$$\begin{aligned}p_\eta(k|\gamma)&=\prod_{j=1}^{p}p_{\eta,j}(k_j|\gamma_j)\\&=\prod_{j=1}^{p}\left(A_j\right)^{(1-\gamma_j)k_j}\left(1-A_j\right)^{(1-\gamma_j)(1-k_j)}\\&\quad\left(D_j\right)^{\gamma_j(1-k_j)}\left(1-D_j\right)^{\gamma_jk_j}.\end{aligned}$$

Following similar arguments to the proof of Lemma 2 of Griffin et al. (2021), we obtain

$$\begin{aligned}\text{II}&\le\sum_{j=1}^{p}\max\left\{|A_j^{(i+1)}-A_j^{(i)}|,|D_j^{(i+1)}-D_j^{(i)}|\right\}\\&\le p\max\left\{\max_j\left\{|A_j^{(i+1)}-A_j^{(i)}|\right\},\max_j\left\{|D_j^{(i+1)}-D_j^{(i)}|\right\}\right\}.\end{aligned}$$

From the definitions of A_j and D_j, we have

$$|A_j^{(i+1)} - A_j^{(i)}| = \left| \min\left\{1, \frac{\tilde{\pi}_j^{(i+1)}}{(1-\tilde{\pi}_j^{(i+1)})}\right\} - \min\left\{1, \frac{\tilde{\pi}_j^{(i)}}{(1-\tilde{\pi}_j^{(i)})}\right\} \right| \tag{B.34}$$

$$|D_j^{(i+1)} - D_j^{(i)}| = \left| \min\left\{1, \frac{(1-\tilde{\pi}_j^{(i+1)})}{\tilde{\pi}_j^{(i+1)}}\right\} - \min\left\{1, \frac{(1-\tilde{\pi}_j^{(i)})}{\tilde{\pi}_j^{(i)}}\right\} \right|. \tag{B.35}$$

The pseudo-code of the PARNI sampler in (3) states that $\tilde{\pi}_j^{(i)} = \pi_0 + (1-2\pi_0)\hat{\pi}_j^{(i)}$ and $\hat{\pi}_j^{(i)}$ are the Rao-Blackwellised estimates of posterior inclusion probabilities π_j. Recall the definition of γ_{-j} which is vector of γ without γ_j (i.e. $\gamma_{-j} = (\gamma_1, \ldots, \gamma_{j-1}, \gamma_{j+1}, \ldots, \gamma_p)$). Note that

$$\hat{\pi}_j^{(i)} = \frac{1}{i}\sum_{\tau=1}^{i} \Pr(\gamma_j = 1 | y, \gamma_{-j}^{(\tau)}) \tag{B.36}$$

where

$$\Pr(\gamma_j = 1 | y, \gamma_{-j}^{(\tau)}) = \frac{\pi(\gamma_j = 1, \gamma_{-j}^{(\tau)} | y)}{\pi(\gamma_j = 1, \gamma_{-j}^{(\tau)} | y) + \pi(\gamma_j = 0, \gamma_{-j}^{(\tau)} | y)} \tag{B.37}$$

for all j from 1 to p, and therefore

$$\begin{aligned} &|\hat{\pi}_j^{(i+1)} - \hat{\pi}_j^{(i)}| \\ &= \left| \frac{i}{i+1}\hat{\pi}_j^{(i)} + \frac{1}{i+1}\Pr(\gamma_j = 1 | y, \gamma_{-j}^{(i+1)}) - \hat{\pi}_j^{(i)} \right| \\ &\le \left| \frac{i}{i+1}\hat{\pi}_j^{(i)} - \hat{\pi}_j^{(i)} \right| + \frac{1}{i+1}\Pr(\gamma_j = 1 | y, \gamma_{-j}^{(i+1)}) \\ &\le \frac{2}{i+1} \end{aligned}$$

Note that $f_{\pi_0}(x) = \min\{1, (\pi_0 + (1-2\pi_0 x))/(\pi_0 + (1-2\pi_0(1-x))\}$ is Lipshitz with constant $1/\pi_0$, we obtain

$$\begin{aligned} &\left| \min\left\{1, \frac{\tilde{\pi}_j^{(i+1)}}{(1-\tilde{\pi}_j^{(i+1)})}\right\} - \min\left\{1, \frac{\tilde{\pi}_j^{(i)}}{(1-\tilde{\pi}_j^{(i)})}\right\} \right| \\ &\le \frac{1}{\pi_0}|\hat{\pi}_j^{(i+1)} - \hat{\pi}_j^{(i)}| \\ &\le \frac{1}{\pi_0} \cdot \frac{2}{i+1}. \end{aligned}$$

A similar conclusion can be drawn for each D_j, meaning

$$\text{II} \le C_4 i^{-1} \tag{B.38}$$

for some constant $C_4 < \infty$.

The second term, III, is more complicated. We start by substituting sub proposal kernels in (28) to III yielding

$$\text{III} = \left| \prod_{r=1}^{p_k} q^{\text{PARNI}}_{\theta^{(i+1)}, K_r}(\gamma(r-1), \gamma(r)) - \prod_{r=1}^{p_k} q^{\text{PARNI}}_{\theta^{(i)}, K_r}(\gamma(r-1), \gamma(r)) \right| \tag{B.39}$$

where $\gamma(0) = \gamma$ and $\gamma(p_k) = \gamma'$. Applying Lemma 3 to (B.39), we have

$$\begin{aligned} \text{III} &\le \sum_{r=1}^{p_k} \left| q^{\text{PARNI}}_{\theta^{(i+1)}, K_r}(\gamma(r-1), \gamma(r)) - q^{\text{PARNI}}_{\theta^{(i)}, K_r}(\gamma(r-1), \gamma(r)) \right| \\ &\le p \max_r \left| q^{\text{PARNI}}_{\theta^{(i+1)}, K_r}(\gamma(r-1), \gamma(r)) - q^{\text{PARNI}}_{\theta^{(i)}, K_r}(\gamma(r-1), \gamma(r)) \right| \\ &:= \text{IV}. \end{aligned}$$

The sub-proposal kernels typically contain two moves, either flipping position K_r or keeping it. Term IV is smaller than the maximum probability of these two moves. Let V be the absolute difference in flipping and VI be the absolute difference in keeping, then

$$\text{IV} \le q \max_r\{\max\{\text{V}, \text{VI}\}\}.$$

We next consider terms V and VI. Starting with V and substituting (28) in V gives

$$\text{V} = \left| \frac{\omega^{(i+1)} g\left(\frac{\pi(\gamma(r))}{\pi(\gamma(r-1))}\left(\frac{A_{K_r}^{(i+1)}}{D_{K_r}^{(i+1)}}\right)^{\gamma(r)_{K_r} - \gamma(r-1)_{K_r}}\right)}{Z^{(i+1)}(r)} - \frac{\omega^{(i)} g\left(\frac{\pi(\gamma(r))}{\pi(\gamma(r-1))}\left(\frac{A_{K_r}^{(i)}}{D_{K_r}^{(i)}}\right)^{\gamma(r)_{K_r} - \gamma(r-1)_{K_r}}\right)}{Z^{(i)}(r)} \right|.$$

Reduce the fractions to a common denominator to yield

$$\mathrm{V} = \left| \frac{\omega^{(i+1)} g\left(\frac{\pi(\gamma(r))}{\pi(\gamma(r-1))}\left(\frac{A_{K_r}^{(i+1)}}{D_{K_r}^{(i+1)}}\right)^{\gamma^d(r)}\right) Z^{(i)}(r) - \omega^{(i)} g\left(\frac{\pi(\gamma(r))}{\pi(\gamma(r-1))}\left(\frac{A_{K_r}^{(i)}}{D_{K_r}^{(i)}}\right)^{\gamma^d(r)}\right) Z^{(i+1)}(r)}{Z^{(i+1)}(r) Z^{(i)}(r)} \right|$$

$$\leq \frac{1}{(Z^{\downarrow})^2} \left| \omega^{(i+1)} g\left(\frac{\pi(\gamma(r))}{\pi(\gamma(r-1))}\left(\frac{A_{K_r}^{(i+1)}}{D_{K_r}^{(i+1)}}\right)^{\gamma^d(r)}\right) Z^{(i)}(r) - \omega^{(i)} g\left(\frac{\pi(\gamma(r))}{\pi(\gamma(r-1))}\left(\frac{A_{K_r}^{(i)}}{D_{K_r}^{(i)}}\right)^{\gamma^d(r)}\right) Z^{(i+1)}(r) \right|$$

where $\gamma^d(r) = \gamma(r)_{K_r} - \gamma(r-1)_{K_r}$ and the last line follows from (B.13) in the proof of Lemma 1 where all normalising constants can be bounded above and below. Using Lemma 4, we obtain

$$\begin{aligned}
\mathrm{V} \leq\ & C_5 \underbrace{\left|\omega^{(i+1)} - \omega^{(i)}\right|}_{:=\mathrm{VII}} \\
& + C_6 \underbrace{\left| g\left(\frac{\pi(\gamma(r))}{\pi(\gamma(r-1))}\left(\frac{A_{K_r}^{(i+1)}}{D_{K_r}^{(i+1)}}\right)^{\gamma^d(r)}\right) - g\left(\frac{\pi(\gamma(r))}{\pi(\gamma(r-1))}\left(\frac{A_{K_r}^{(i)}}{D_{K_r}^{(i)}}\right)^{\gamma^d(r)}\right) \right|}_{:=\mathrm{VIII}} \\
& + C_7 \underbrace{\left|Z^{(n+1)}(r) - Z^{(n)}(r)\right|}_{:=\mathrm{IX}}
\end{aligned}$$

for some constants $C_5, C_6, C_7 < \infty$. We can apply similar arguments to VI giving

$$\begin{aligned}
\mathrm{VI} &= \left| \frac{(1-\omega^{(i+1)})g(1)}{Z^{(i+1)}(r)} - \frac{(1-\omega^{(i)})g(1)}{Z^{(i)}(r)} \right| \\
&\leq C_8 \underbrace{\left|\omega^{(i+1)} - \omega^{(i)}\right|}_{\mathrm{VII}} + C_9 \underbrace{\left|Z^{(i+1)}(r) - Z^{(i)}(r)\right|}_{\mathrm{IX}}
\end{aligned}$$

for some constants $C_8, C_9 < \infty$. Starting with IX, by substituting in the definitions in (29), we have

$$\begin{aligned}
\mathrm{IX} &= \left|Z^{(i+1)}(r) - Z^{(i)}(r)\right| \\
&= \left| \left(\omega^{(i+1)} g\left(\frac{\pi(\gamma(r))}{\pi(\gamma(r-1))}\left(\frac{A_{K_r}^{(i+1)}}{D_{K_r}^{(i+1)}}\right)^{\gamma^d(r)}\right) \right.\right. \\
&\quad \left.\left. + (1-\omega^{(i+1)})g(1) \right) \right. \\
&\quad \left. - \left(\omega^{(i)} g\left(\frac{\pi(\gamma(r))}{\pi(\gamma(r-1))}\left(\frac{A_{K_r}^{(i)}}{D_{K_r}^{(i)}}\right)^{\gamma^d(r)}\right) \right.\right. \\
&\quad \left.\left. + (1-\omega^{(i)})g(1) \right) \right|,
\end{aligned}$$

and applying the triangle inequality yields

$$\begin{aligned}
\mathrm{IX} \leq & \left| \omega^{(i+1)} g\left(\frac{\pi(\gamma(r))}{\pi(\gamma(r-1))}\left(\frac{A_{K_r}^{(i+1)}}{D_{K_r}^{(i+1)}}\right)^{\gamma^d(r)}\right) \right. \\
& \left. - \omega^{(i)} g\left(\frac{\pi(\gamma(r))}{\pi(\gamma(r-1))}\left(\frac{A_{K_r}^{(i)}}{D_{K_r}^{(i)}}\right)^{\gamma^d(r)}\right) \right| \\
& + g(1)\left|\omega^{(i+1)} - \omega^{(i)}\right| \\
\leq & C_{10} \underbrace{\left|\omega^{(i+1)} - \omega^{(i)}\right|}_{\mathrm{VII}}
\end{aligned}$$

$$+ C_{11} \underbrace{\left| g\left(\frac{\pi(\gamma(r))}{\pi(\gamma(r-1))} \left(\frac{A_{K_r}^{(i+1)}}{D_{K_r}^{(i+1)}} \right)^{\gamma^d(r)} \right) - g\left(\frac{\pi(\gamma(r))}{\pi(\gamma(r-1))} \left(\frac{A_{K_r}^{(i)}}{D_{K_r}^{(i)}} \right)^{\gamma^d(r)} \right) \right|}_{\text{VIII}}$$

for some constants $C_{10}, C_{11} < \infty$. The last line follows after applying Lemma 3.

The diminishing adaptation of tuning parameter ω is shown in Lemma 5 where

$$\text{VII} \leq C_{12} i^{-\lambda} \tag{B.40}$$

for some constant $C_{12} < \infty$, $\lambda \in (1/2, 1)$ for the Robins-Monro adaptation scheme and $\lambda = 0.5$ for the Kiefer–Wolfowitz adaptation scheme.

We now consider term VIII. From assumption (A.1), we have

$$g(t_2) - g(t_1) \leq C_g(t_2 - t_1)$$

and therefore

$$|g(t_2) - g(t_1)| \leq C_g |t_2 - t_1| \tag{B.41}$$

for any $t_1, t_2 > 0$. We then have

$$\begin{aligned}
\text{VIII} &\leq C_g \left| \frac{\pi(\gamma(r))}{\pi(\gamma(r-1))} \left(\frac{A_{K_r}^{(i+1)}}{D_{K_r}^{(i+1)}} \right)^{\gamma^d(r)} \right. \\
&\quad \left. - \frac{\pi(\gamma(r))}{\pi(\gamma(r-1))} \left(\frac{A_{K_r}^{(i)}}{D_{K_r}^{(i)}} \right)^{\gamma^d(r)} \right| \\
&\leq \Pi C_g \left| \left(\frac{A_{K_r}^{(i+1)}}{D_{K_r}^{(i+1)}} \right)^{\gamma^d(r)} - \left(\frac{A_{K_r}^{(i)}}{D_{K_r}^{(i)}} \right)^{\gamma^d(r)} \right|,
\end{aligned}$$

where the last line follows from Assumption (A.2). Considering two possible values of $\gamma^d(r)$, namely 1 and -1, we show that VIII is bounded by the maximum of those values

$$\text{VIII} \leq \Pi C_g \max \left\{ \left| \frac{A_{K_r}^{(i+1)}}{D_{K_r}^{(i+1)}} - \frac{A_{K_r}^{(i)}}{D_{K_r}^{(i)}} \right|, \left| \frac{D_{K_r}^{(i+1)}}{A_{K_r}^{(i+1)}} - \frac{D_{K_r}^{(i)}}{A_{K_r}^{(i)}} \right| \right\}.$$

Next, multiplying the common denominator yields

$$\begin{aligned}
\text{VIII} &\leq \Pi C_g \max \left\{ \left| \frac{A_{K_r}^{(i+1)} D_{K_r}^{(i)} - A_{K_r}^{(i)} D_{K_r}^{(i+1)}}{D_{K_r}^{(i+1)} D_{K_r}^{(i)}} \right|, \right. \\
&\quad \left. \left| \frac{D_{K_r}^{(i+1)} A_{K_r}^{(i)} - D_{K_r}^{(i)} A_{K_r}^{(i+1)}}{A_{K_r}^{(i+1)} A_{K_r}^{(i)}} \right| \right\} \\
&\leq \frac{\Pi C_g}{\pi_0^2} \max \left\{ \left| A_{K_r}^{(i+1)} D_{K_r}^{(i)} - A_{K_r}^{(i)} D_{K_r}^{(i+1)} \right|, \right. \\
&\quad \left. \left| D_{K_r}^{(i+1)} A_{K_r}^{(i)} - D_{K_r}^{(i)} A_{K_r}^{(i+1)} \right| \right\},
\end{aligned}$$

which holds because $\pi_0 \leq A_j, D_j \leq 1$ from the proof of Lemma 1. Then, by applying Lemma 3, we have

$$\begin{aligned}
\text{VIII} &\leq \frac{\Pi C_g}{\pi_0^2} \max \left\{ \left| A_{K_r}^{(i+1)} - A_{K_r}^{(i)} \right| + \left| D_{K_r}^{(i+1)} - D_{K_r}^{(i)} \right|, \right. \\
&\quad \left. \left| A_{K_r}^{(i+1)} - A_{K_r}^{(i)} \right| + \left| D_{K_r}^{(i+1)} - D_{K_r}^{(i)} \right| \right\} \\
&\leq \frac{\Pi C_g}{\pi_0^2} \left(\left| A_{K_r}^{(i+1)} - A_{K_r}^{(i)} \right| + \left| D_{K_r}^{(i+1)} - D_{K_r}^{(i)} \right| \right) \\
&\leq C_{13} \left(\max_j \left\{ \left| A_j^{(i+1)} - A_j^{(i)} \right| \right\} \right. \\
&\quad \left. + \max_j \left\{ \left| D_j^{(i+1)} - D_j^{(i)} \right| \right\} \right)
\end{aligned}$$

for some constant $C_{13} < \infty$.

As we have previously showed that

$$\max_j \left\{ \left| A_j^{(i+1)} - A_j^{(i)} \right| \right\} \leq \frac{1}{\pi_0} \cdot \frac{2}{i+1} \tag{B.42}$$

$$\max_j \left\{ \left| D_j^{(i+1)} - D_j^{(i)} \right| \right\} \leq \frac{1}{\pi_0} \cdot \frac{2}{i+1}, \tag{B.43}$$

this leads to

$$\begin{aligned}
\text{VIII} &\leq C_{14} \frac{1}{\pi_0} \cdot \frac{2}{i+1} \\
&\leq C_{15} i^{-1}
\end{aligned}$$

for some constants $C_{14}, C_{15} < \infty$.

We complete the proof by stating that IV $\leq C_{16} i^{-\lambda}$ for some constant $C_{16} < \infty$, and hence I $\leq C_{17} i^{-\lambda}$ for some constant $C_{17} < \infty$, $\lambda \in (1/2, 1)$ for the Robins-Monro adaptation scheme and $\lambda = 0.5$ for the Kiefer–Wolfowitz adaptation scheme, which shows that the diminishing adaptation for the PARNI sampler has established. □

B.8 Proof of Theorem 2

Proof *Simultaneous uniform ergodicity* together with *diminishing adaption* are enough to show that π is stationary for

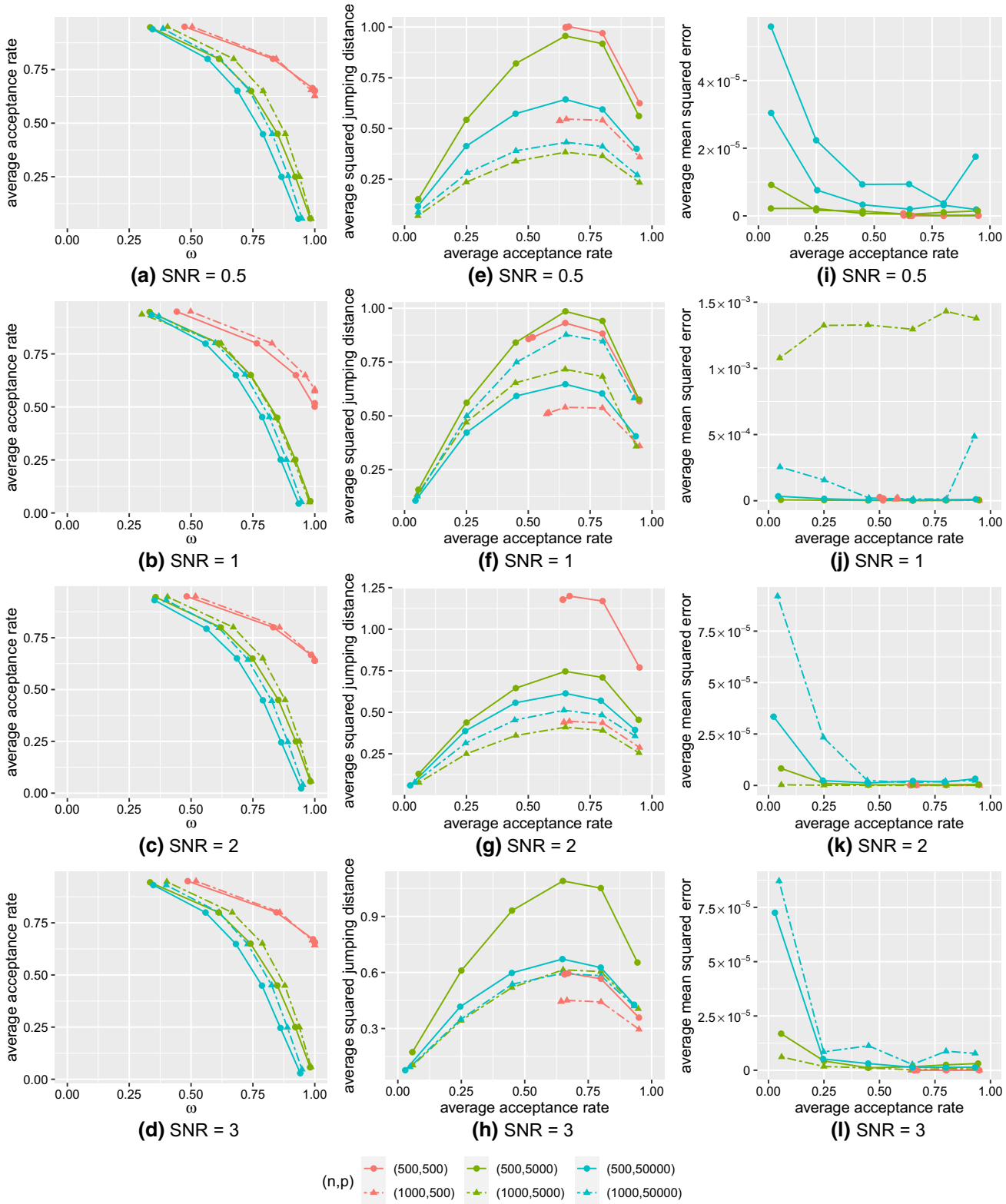

Fig. 7 Simulated data: plots of average squared jumping distance and average mean square error against average acceptance rate and ω for the Pointwise implementation of Adaptive Random Neighbourhood Informed and Thresholded proposal with Robbins-Monro update (PARNIT-RM). **a**–**d** average acceptance rate against ω for simulated data-sets with signal-to-noise ratio of 0.5, 1, 2 and 3; **e**–**h** average squared jumping distance against average acceptance rate for simulated data-sets with signal-to-noise ratio of 0.5, 1, 2, 3; **i**–**j** average mean squared error against average acceptance rate for simulated data-sets with signal-to-noise ratio of 0.5, 1, 2 and 3

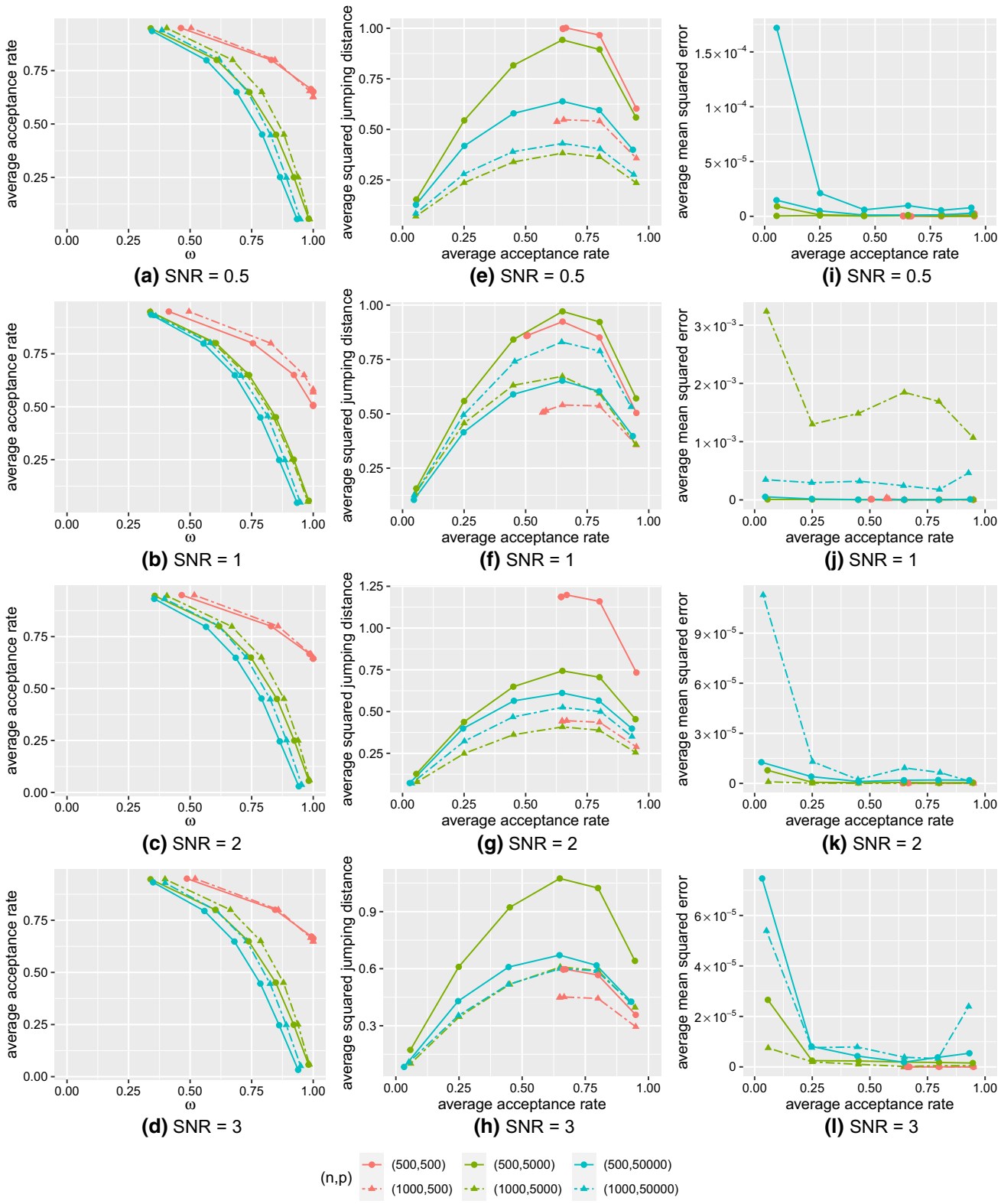

Fig. 8 Simulated data: plots of average squared jumping distance and average mean square error against average acceptance rate and ω for the Pointwise implementation of Adaptive Random Neighbourhood Informed and Balanced proposal with Robbins-Monro update (PARNIB-RM). **a**–**d** average acceptance rate against ω for simulated data-sets with signal-to-noise ratio of 0.5, 1, 2 and 3; **e**–**h** average squared jumping distance against average acceptance rate for simulated data-sets with signal-to-noise ratio of 0.5, 1, 2, 3; **i**–**j** average mean squared error against average acceptance rate for simulated data-sets with signal-to-noise ratio of 0.5, 1, 2 and 3

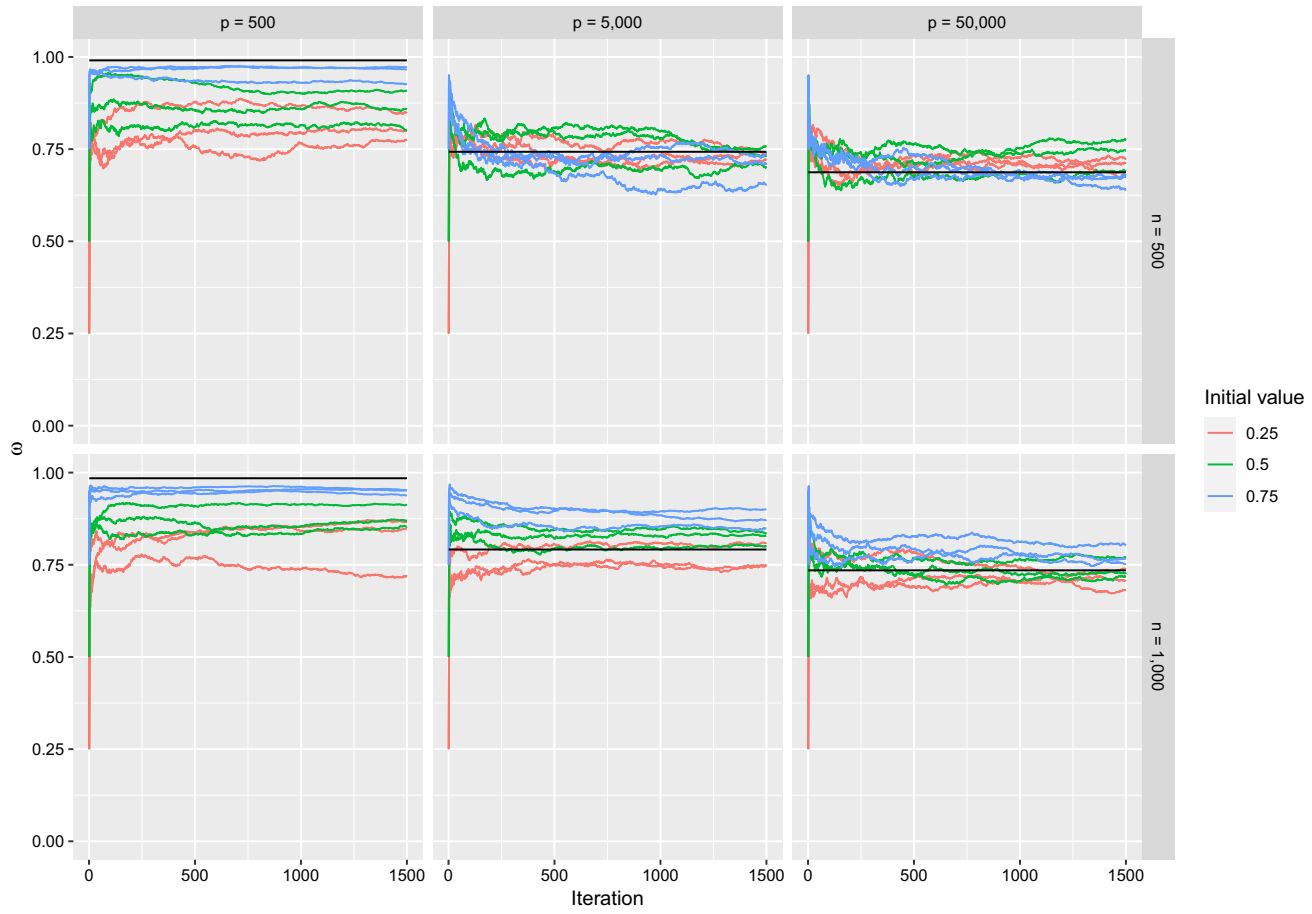

Fig. 9 Simulated data: trace plots of ω from the point-wise implementation of the Adaptive Random Neighbourhood Informed and Thresholded proposal with Kiefer–Wolfowitz update (PARNIT-KW) for the first 1500 iterations on simulated data-sets with signal-to-noise ratio of 0.5 and three choices of initial values (0.25, 0.5 and 0.75). The black line indicates the optimal values of ω for each data-set. (Color figure online)

each kernel P_θ^{PARNI} and the adaptive algorithm is ergodic from Theorem 1 in Roberts and Rosenthal (2007). Its multiple chain version is also ergodic with respect to $\pi^{\otimes L}$.

The proof of the Strong Law of Large Numbers (SLLN) contains two steps. Firstly, we show that each individual chain satsifies a SLLN, that is

$$\frac{1}{N}\sum_{i=0}^{N-1} f(\gamma^{l,(i)}) \to \pi(f) \quad \text{almost surely as } N \to \infty. \tag{A.44}$$

Then by averaging over L parallel chains, we can show that the SLLN in (50) is satisfied for the multiple chain version.

We use Theorem 2.7 in Fort et al. (2011) to show that (A.44) holds. To do so, three conditions are required.

(Con.1) The measurable function V can be chosen to be the constant function $V \equiv 1$, where V-variation distance norm reduces to the total variation distance. It is obvious that the below is met if with $\lambda_\theta = 1/2$, $b_\theta = 1$, the measure ν_θ is uniformly distributed on the sample space Γ, that is

$$\nu_\theta(\gamma) = \frac{1}{2^p},$$

with $\delta_\theta = b$ (the lower bound for a single chain in Lemma 1), then

$$\begin{aligned} P_\theta^{\text{PARNI}} V &\le \frac{1}{2}V + 1 \\ P_\theta^{\text{PARNI}}(\gamma, \cdot) &\ge b\nu_\theta(\cdot) I\{V \le c_\theta\}(\gamma), \\ c_\theta &= 2b_\theta(1-\lambda_\theta)^{-1} - 1 = 3 \end{aligned}$$

is satisfied.

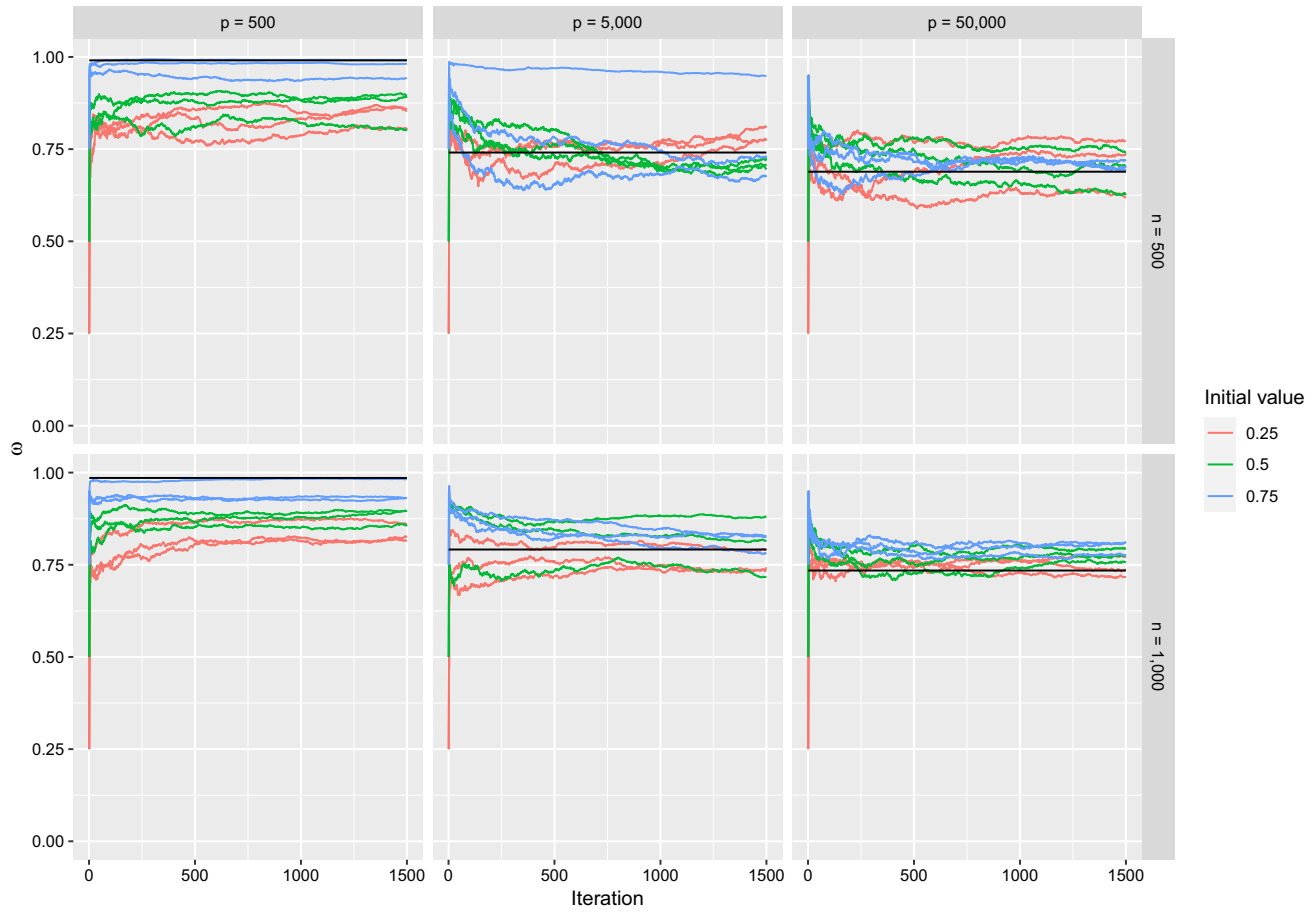

Fig. 10 Simulated data: trace plots of ω from the point-wise implementation of the Adaptive Random Neighbourhood Informed and Balanced proposal with Kiefer–Wolfowitz update (PARNIB-KW) for the first 1500 iterations on simulated data-sets with signal-to-noise ratio of 0.5 and three choices of initial values (0.25, 0.5 and 0.75). The black line indicates the optimal values of ω for each data-set. (Color figure online)

(Con.2) Using the same parameters specified in (Con.1), the problem is reduced to

$$\sum_{i=1}^{\infty} i^{-1} \sup_{\gamma \in \Gamma} \| P^{\text{PARNI}}_{\theta^{(i+1)}}(\gamma, \cdot) - P^{\text{PARNI}}_{\theta^{(i)}}(\gamma, \cdot) \| < +\infty.$$

This is satisfied since the PARNI sampler satisfies diminishing adaption by Lemma 2.

(Con.3) Condition A5 in *Fort et al.* Fort et al. (2011) is trivially satisfied with the parameters chosen in (Con.1).

We have established (Con.1), (Con.2), and (Con.3), therefore by Corollary 2.8 in Fort et al. (2011), (A.44) holds, and so does (50). □

C Additional numerical results

This section will provide more numerical results in addition to Sect. 5.

C.1 Sensitivity analysis on thinning parameter ϖ for simulated data-sets

We repeat the experiment which studies the optimal value of ω and optimal average acceptance rate in Sect. 5.2 on simulated data-sets used in Sect. 5.1. Figures 7 and 8 show the effect of manipulating the target average acceptance rate on average squared jumping distance and average mean squared errors for the PARNIT and PARNIB proposals respectively. We split the plots into 4 sets where each set of graphs corresponds to a level of signal-to-noise ratio. In both figures, panels (a)–(d) show the negative relationship of ω against average acceptance rate. Panels (e)–(h) plot the average squared jumping distance against the average acceptance rate. Finally, panels (i)–(l) show the average acceptance rate and the average mean squared errors. These plots suggest the similar conclusion in Sect. 5.2 for which the average acceptance rate of 0.65 yields the largest average squared jumping distance. The smallest average mean squared errors are also located around the region of 0.65 average acceptance rate.

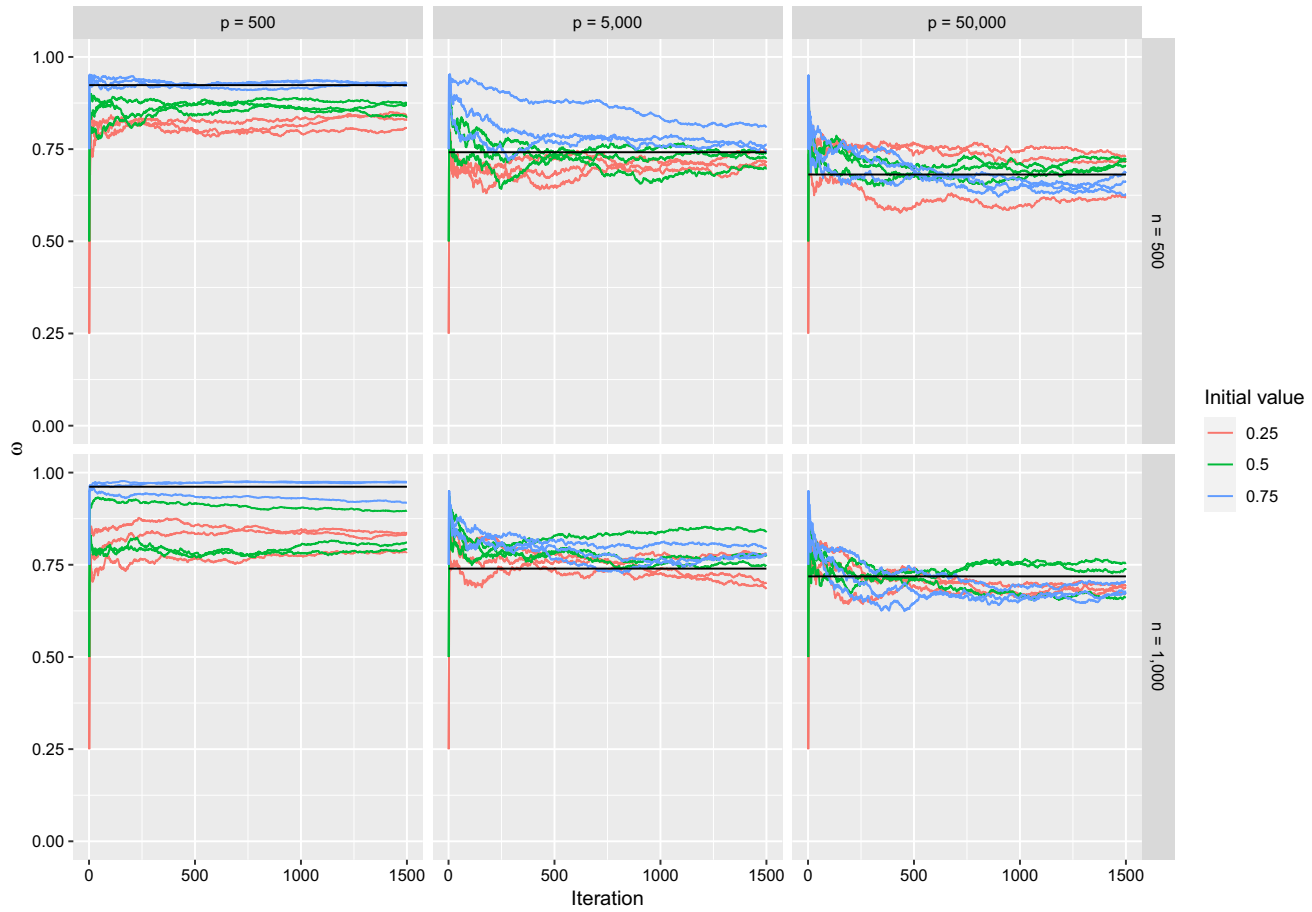

Fig. 11 Simulated data: trace plots of ω from the point-wise implementation of the Adaptive Random Neighbourhood and Thresholded Informed proposal with Kiefer–Wolfowitz update (PARNIT-KW) for the first 1500 iterations on simulated data-sets with signal-to-noise ratio of 1 and three choices of initial values (0.25, 0.5 and 0.75). The black line indicates the optimal values of ω for each data-set. (Color figure online)

C.2 Additional results from Kiefer–Wolfowitz adaptation scheme

This section is to examine whether applying the Kiefer–Wolfowitz adaptation scheme to the PARNI sampler would lead to the optimal scaling property of the chains. We ran the PARNIT-KW and PARNIB-KW samplers for 1500 iterations with 3 different initial values on the 24 simulated data-sets of Sect. 5.1 and the 8 real data-sets of Sect. 5.2 and recorded the values of ω. The trace plots of these ω values are given in Figs. 9, 10, 11, 12, 13, 14, 15 , and 16, for the simulated data-sets and Figs. 17 and 18 for the real data-sets. The black horizontal lines in these plots indicate the empirical optimal values of ω gathered for each data-set from Figs. 3, 4, 7 and 8. The optimal values decrease along with the dimensionality of p and they are also influenced by the correlation structure for which more complicated correlation structures imply smaller values of ω. It appears that the values of ω are moving towards the black lines and converging to them regardless of initial values.

There is a significant trend that the ω values are approaching the region around the optimal values fairly quickly. Surprisingly, the parameter ω converges even faster in high-dimensional problems, for example, when $p = 50,000$ in simulated data-sets and the SNP data-set. But there is still a rare chance that the Kiefer–Wolfowitz scheme does not lead to the optimal choice. Some of ω values become trapped on the value of 1. This issue is mainly caused by two reasons. Firstly, the ASJD estimators are often too noisy to capture the true signal in the expected squared jumping distances. These estimators can be viewed as simple Monte Carlo with only a few samples and therefore we may obtain estimates with extremely large estimation errors. Large errors introduce uncertainty into the true direction that should be updated and make ω take longer to converge or converge to a suboptimal value. Secondly, the use of the logistic transformation makes ω difficult to be updated in the boundary areas and therefore ω easily get trapped in values of 0 or 1.

Overall, the Kiefer–Wolfowitz adaptation scheme is relatively robust in tuning ω for the PARNI sampler, and we

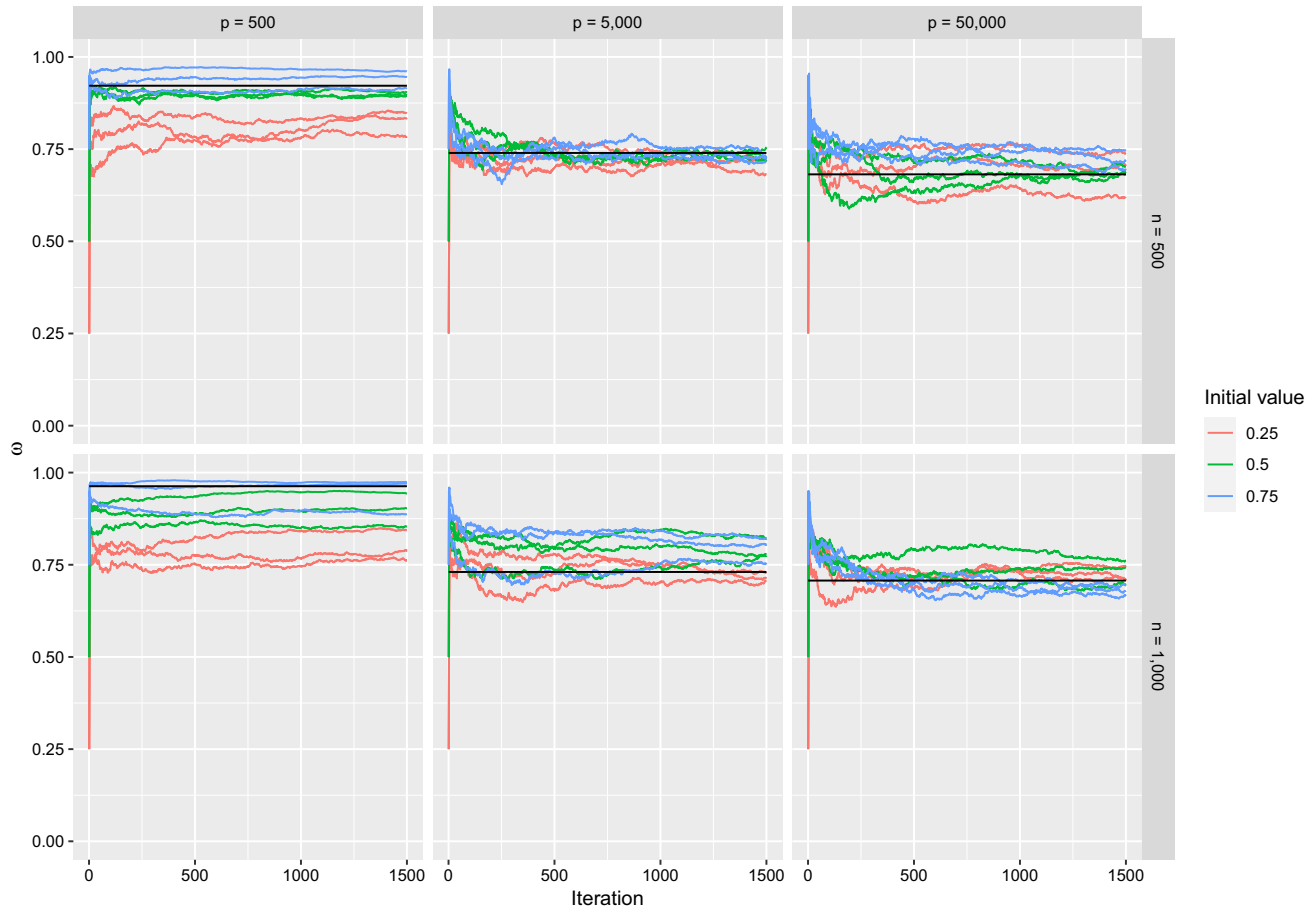

Fig. 12 Simulated data: trace plots of ω from the point-wise implementation of the Adaptive Random Neighbourhood Informed and Balanced proposal with Kiefer–Wolfowitz update (PARNIB-KW) for the first 1500 iterations on simulated data-sets with signal-to-noise ratio of 1 and three choices of initial values (0.25, 0.5 and 0.75). The black line indicates the optimal values of ω for each data-set. (Color figure online)

believe it can also be applied to other adaptive MCMC schemes in tuning the scaling parameters.

C.3 More results from simulated data-sets

In addition to Figs. 2, 19, 20 and 21 are trace plots of log posterior model probabilities from the ADS, ASI, PARNIT-RM, PARNIT-KW, PARNIB-RM and PARNIB-KW schemes on the simulated data-sets of Sect. 5.1 when SNR $= 0.5$, 1 and 3. Generally speaking, all the PARNI algorithms mix better than the ADS and ASI schemes on all data-sets. Except for the data-sets for which the posterior distributions do not concentrate in a few models (when SNR $= 0.5$), the ADS scheme always get struck on the empty model and struggles to include important variables and reach the high probability region within the first 1500 iterations. The ASI algorithm mixes quite well when p is relative small, but this algorithm is taking longer to converge and it is inefficient to jump between different models when p reaches $50,000$. On the other hand, the PARNIT-RM, PARNIT-KW, PARNIB-RM and PARNIB-KW samplers only take dozens of iterations to converge properly in all settings. In conclusion, the plots suggest that all the PARNI schemes outperform ADS and ASI in terms of the mixing time and convergence rate on the simulated data-sets. They always propose models with high probability of being accepted and therefore sufficiently explore the sample space.

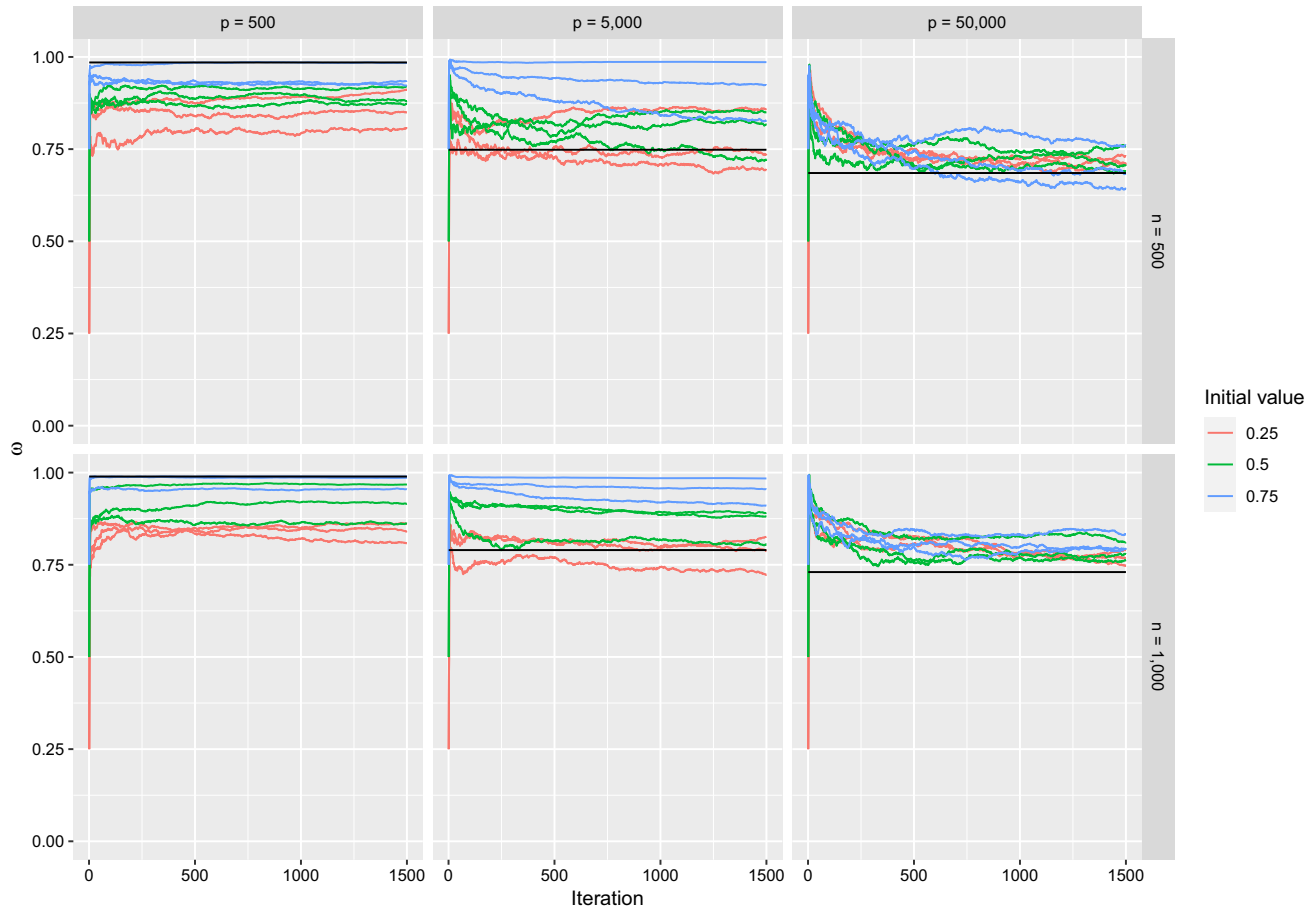

Fig. 13 Simulated data: trace plots of ω from the point-wise implementation of the Adaptive Random Neighbourhood Informed and Thresholded proposal with Kiefer–Wolfowitz update (PARNIT-KW) for the first 1500 iterations on simulated data-sets with signal-to-noise ratio of 2 and three choices of initial values (0.25, 0.5 and 0.75). The black line indicates the optimal values of ω for each data-set. (Color figure online)

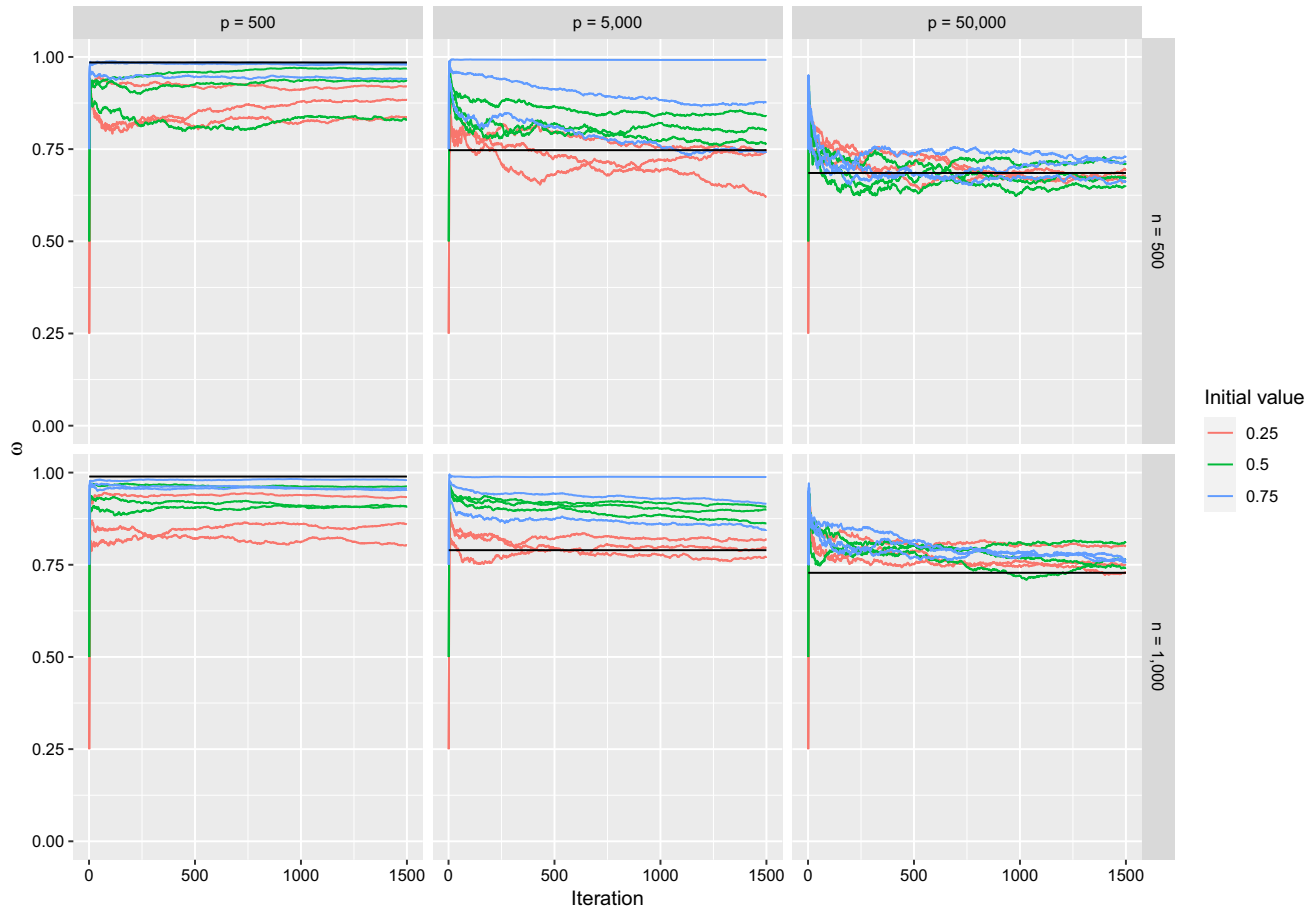

Fig. 14 Simulated data: trace plots of ω from the point-wise implementation of the Adaptive Random Neighbourhood Informed and Balanced proposal with Kiefer–Wolfowitz update (PARNIB-KW) for the first 1500 iterations on simulated data-sets with signal-to-noise ratio of 2 and three choices of initial values (0.25, 0.5 and 0.75). The black line indicates the optimal values of ω for each data-set. (Color figure online)

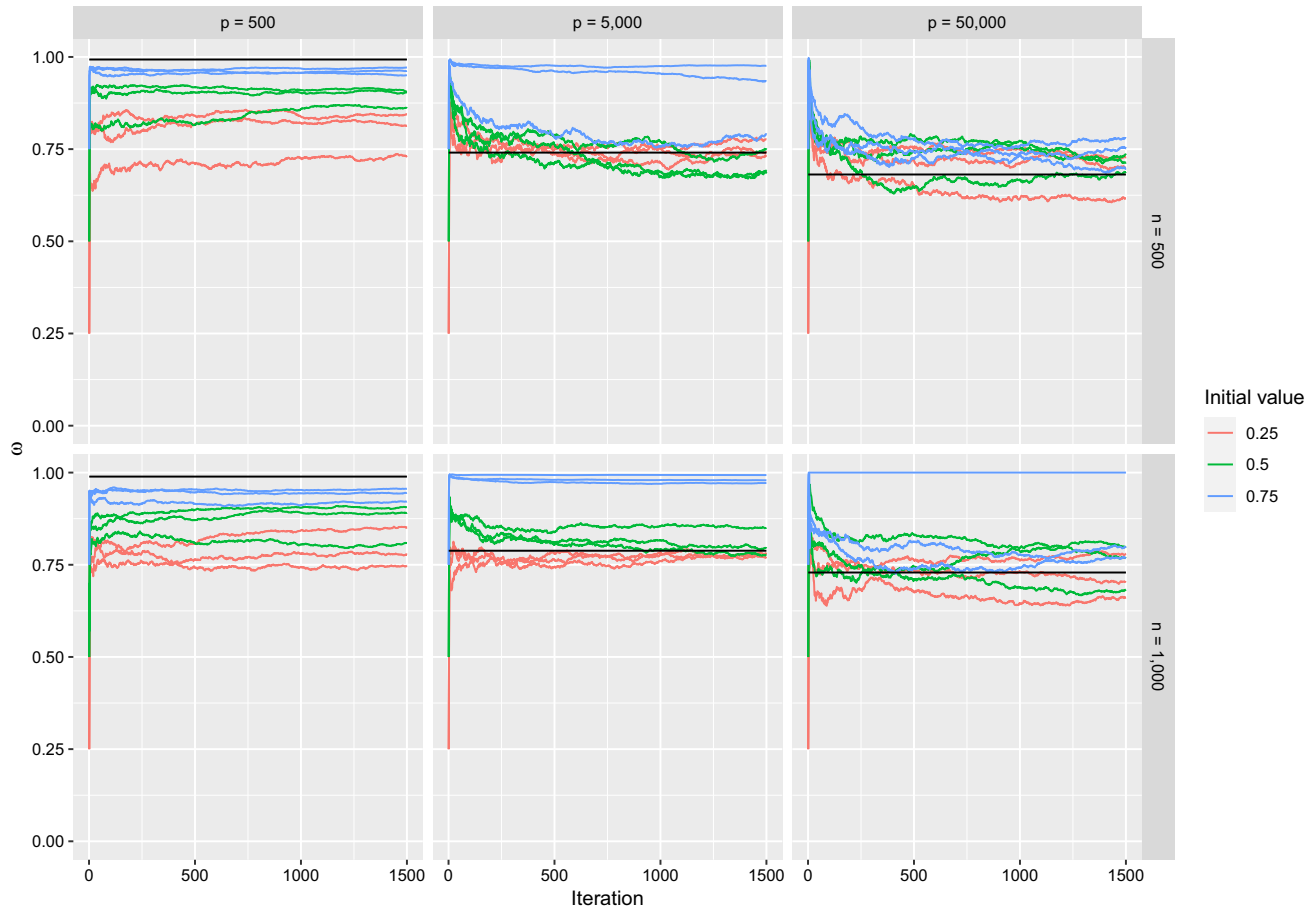

Fig. 15 Simulated data: trace plots of ω from the point-wise implementation of the Adaptive Random Neighbourhood Informed and Thresholded proposal with Kiefer–Wolfowitz update (PARNIT-KW) for the first 1500 iterations on simulated data-sets with signal-to-noise ratio of 3 and three choices of initial values (0.25, 0.5 and 0.75). The black line indicates the optimal values of ω for each data-set. (Color figure online)

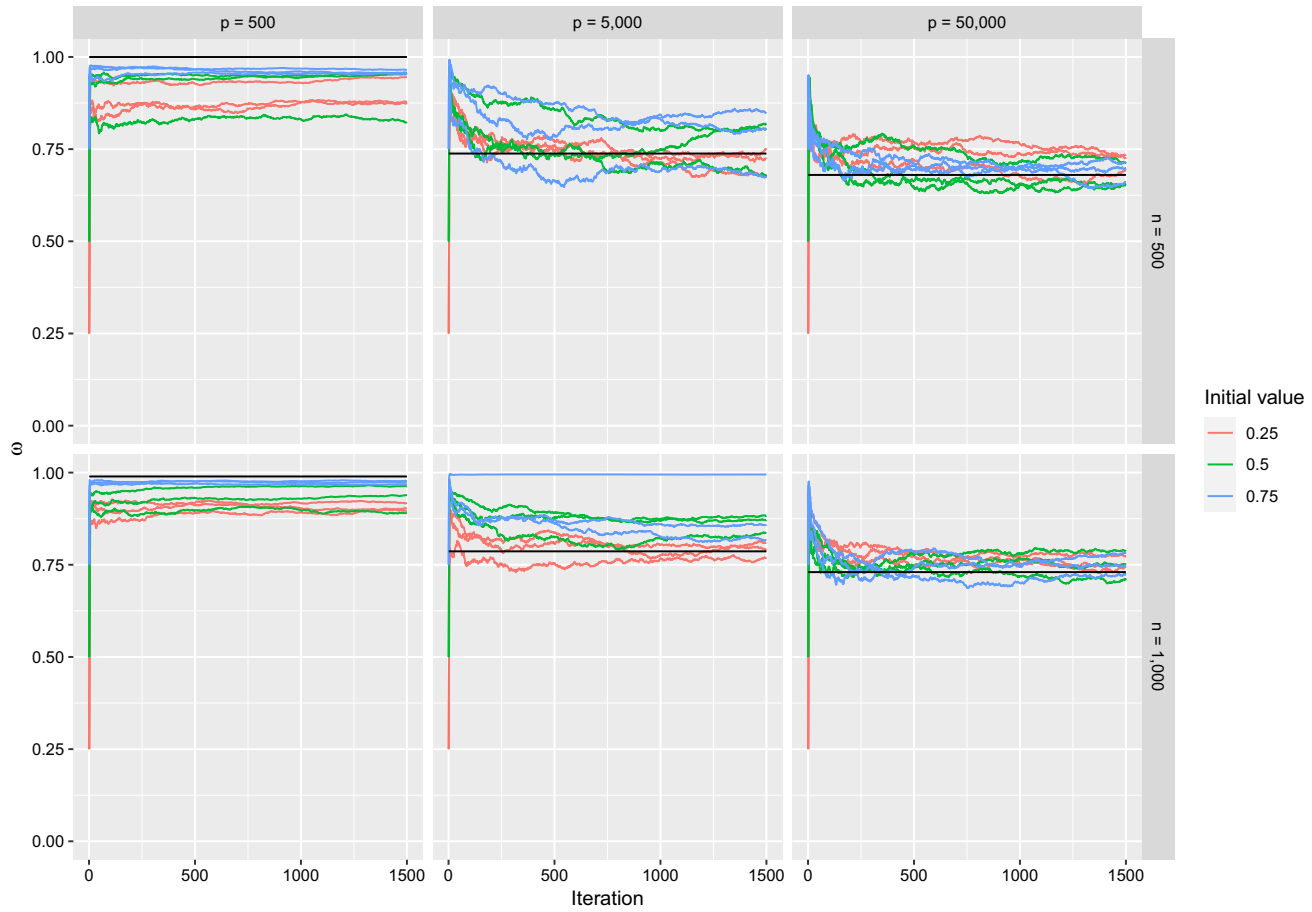

Fig. 16 Simulated data: trace plots of ω from the point-wise implementation of the Adaptive Random Neighbourhood Informed and Balanced proposal with Kiefer–Wolfowitz update (PARNIB-KW) for the first 1500 iterations on simulated data-sets with signal-to-noise ratio of 3 and three choices of initial values (0.25, 0.5 and 0.75). The black line indicates the optimal values of ω for each data-set. (Color figure online)

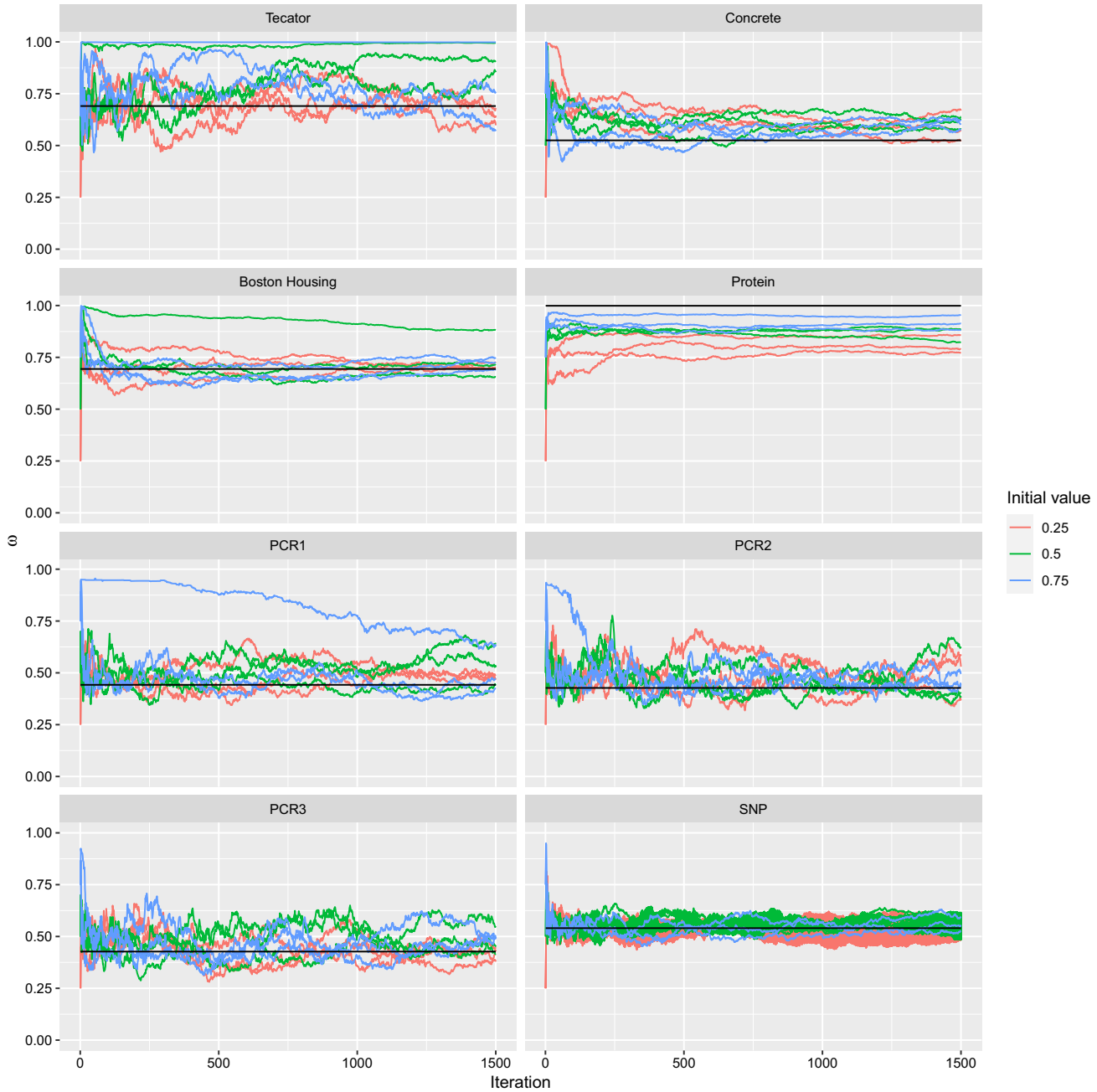

Fig. 17 Real data: trace plots of ω from the point-wise implementation of the Adaptive Random Neighbourhood Informed and Thresholded proposal with Kiefer–Wolfowitz update (PARNIT-KW) for the first 1500 iterations on real data-sets with three choices of initial values (0.25, 0.5 and 0.75). The black line indicates the optimal values of ω for each data-set. (Color figure online)

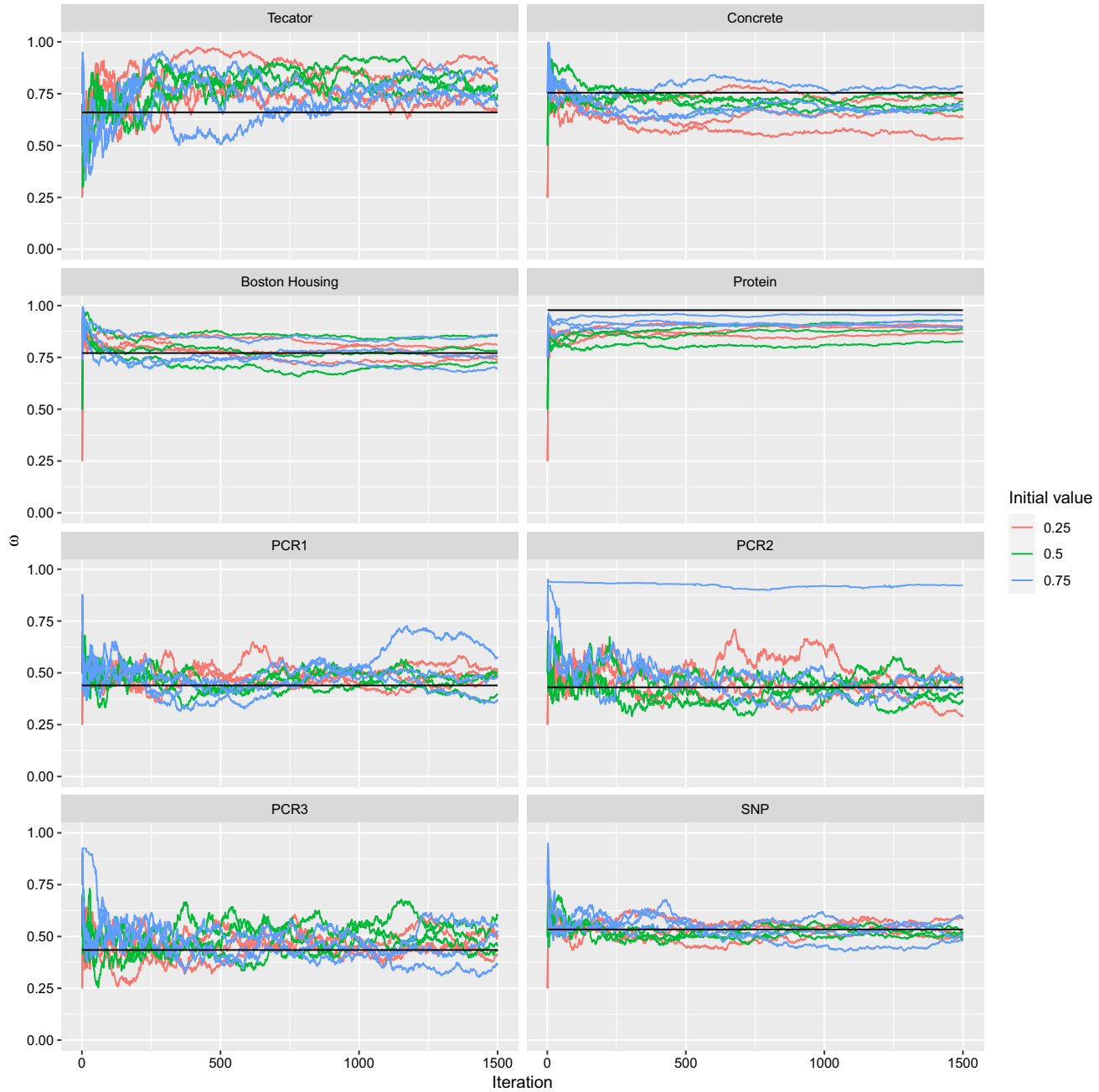

Fig. 18 Real data: trace plots of ω from the point-wise implementation of the Adaptive Random Neighbourhood Informed and Balanced proposal with Kiefer–Wolfowitz (PARNIB-KW) update for the first 1500 iterations on real data-sets with three choices of initial values (0.25, 0.5 and 0.75). The black line indicates the optimal values of ω for each data-set. (Color figure online)

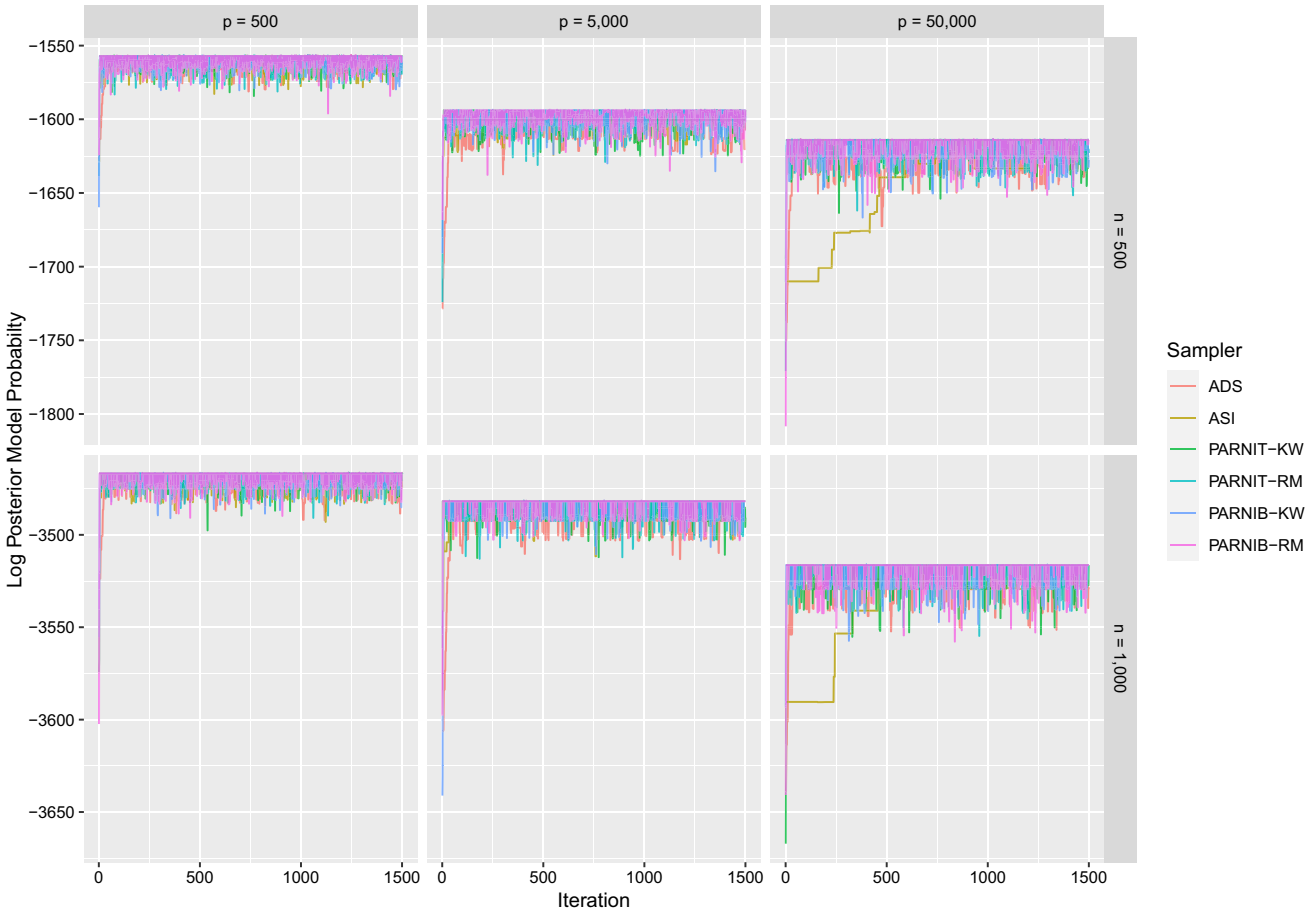

Fig. 19 Simulated data: trace plots of log posterior model probability from the Add-Delete-Swap (ADS), Adaptively Scaled Individual (ASI) adaptation, Pointwise implementation of Adaptive Random Neighbourhood Informed and Thresholded proposal with Kiefer–Wolfowitz update (PARNIT-KW), Pointwise implementation of Adaptive Random Neighbourhood Informed and Thresholded proposal with Robbins-Monro update (PARNIT-RM), Pointwise implementation of Adaptive Random Neighbourhood Informed and balanced proposal with Kiefer–Wolfowitz update (PARNIB-KW) and Pointwise implementation of Adaptive Random Neighbourhood Informed and balanced proposal with Robbins-Monro update (PARNIB-RM) samplers for the first 1500 iterations on simulated data-sets with signal-to-noise ratio of 0.5

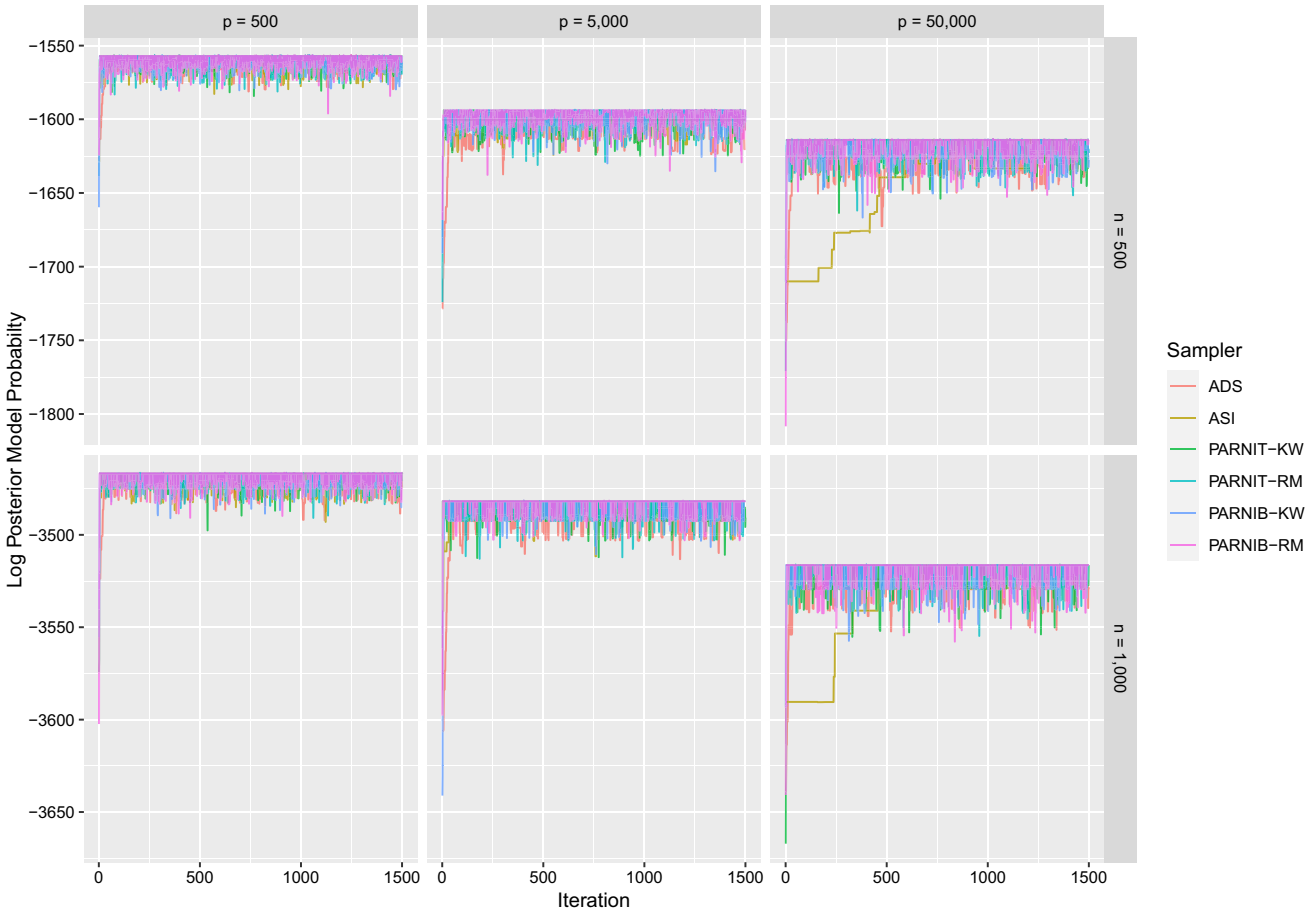

Fig. 20 Simulated data: trace plots of log posterior model probability from the Add-Delete-Swap (ADS), Adaptively Scaled Individual (ASI) adaptation, Pointwise implementation of Adaptive Random Neighbourhood Informed and Thresholded proposal with Kiefer–Wolfowitz update (PARNIT-KW), Pointwise implementation of Adaptive Random Neighbourhood Informed and Thresholded proposal with Robbins-Monro update (PARNIT-RM), Pointwise implementation of Adaptive Random Neighbourhood Informed and balanced proposal with Kiefer–Wolfowitz update (PARNIB-KW) and Pointwise implementation of Adaptive Random Neighbourhood Informed and balanced proposal with Robbins-Monro update (PARNIB-RM) samplers for the first 1500 iterations on simulated data-sets with signal-to-noise ratio of 1

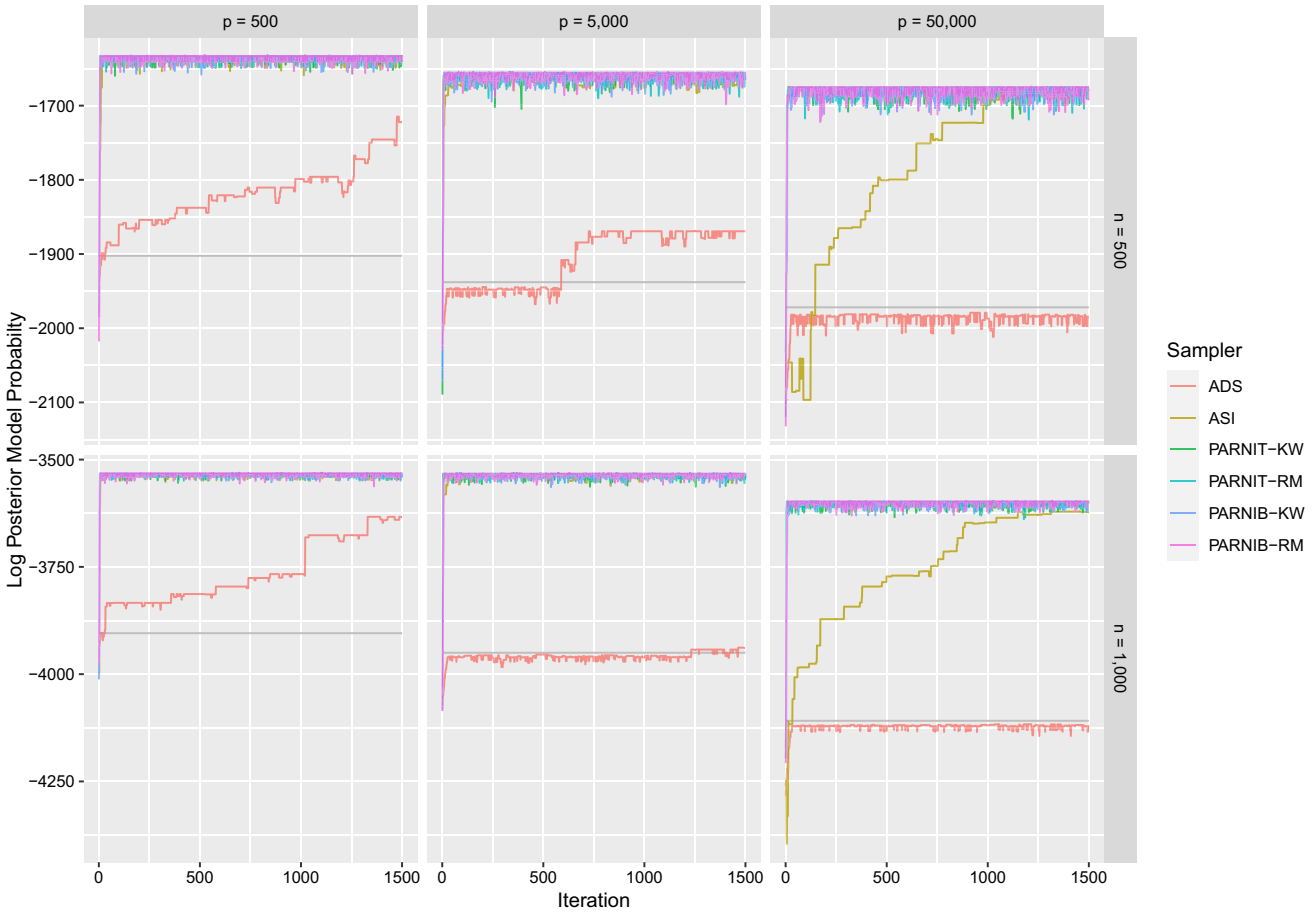

Fig. 21 Simulated data: trace plots of log posterior model probability from the Add-Delete-Swap (ADS), Adaptively Scaled Individual (ASI) adaptation, Pointwise implementation of Adaptive Random Neighbourhood Informed and Thresholded proposal with Kiefer–Wolfowitz update (PARNIT-KW), Pointwise implementation of Adaptive Random Neighbourhood Informed and Thresholded proposal with Robbins-Monro update (PARNIT-RM), Pointwise implementation of Adaptive Random Neighbourhood Informed and balanced proposal with Kiefer–Wolfowitz update (PARNIB-KW) and Pointwise implementation of Adaptive Random Neighbourhood Informed and balanced proposal with Robbins-Monro update (PARNIB-RM) samplers for the first 1500 iterations on simulated data-sets with signal-to-noise ratio of 3

References

Andrieu, C., Lee, A., Livingstone, S.: A general perspective on the Metropolis–Hastings kernel. arXiv:2012.14881 (2020)

Andrieu, C., Thoms, J.: A tutorial on adaptive MCMC. Stat. Comput. **18**(4), 343–373 (2008)

Beskos, A., Pillai, N., Roberts, G., Sanz-Serna, J.-M., Stuart, A.: Optimal tuning of the hybrid Monte Carlo algorithm. Bernoulli **19**(5A), 1501–1534 (2013)

Blum, J.R., et al.: Approximation methods which converge with probability one. Ann. Math. Stat. **25**(2), 382–386 (1954)

Bondell, H.D., Reich, B.J.: Consistent high-dimensional Bayesian variable selection via penalized credible regions. J. Am. Stat. Assoc. **107**(500), 1610–1624 (2012)

Brown, P.J., Griffin, J.E.: Inference with normal-gamma prior distributions in regression problems. Bayesian Anal. **5**(1), 171–188 (2010)

Brown, P.J., Vannucci, M., Fearn, T.: Bayesian wavelength selection in multicomponent analysis. J. Chemom. J. Chemom. Soc. **12**(3), 173–182 (1998)

Carbonetto, P., Zhou, X., Stephens, M.: varbvs: fast variable selection for large-scale regression (2017). arXiv:1709.06597

Chen, X., Qamar, S., Tokdar, S. T.: Paired-move multiple-try stochastic search for Bayesian variable selection (2016). arXiv:1611.09790

Chipman, H., George, E.I., McCulloch, R.E., Clyde, M., Foster, D.P., Stine, R.A.: The practical implementation of Bayesian model selection. Lecture Notes-Monograph Series pp. 65–134 (2001)

Craiu, R.V., Rosenthal, J., Yang, C.: Learn from thy neighbor: parallel-chain and regional adaptive MCMC. J. Am. Stat. Assoc. **104**(488), 1454–1466 (2009)

Duane, S., Kennedy, A.D., Pendleton, B.J., Roweth, D.: Hybrid Monte Carlo. Phys. Lett. B **195**(2), 216–222 (1987)

Fernandez, C., Ley, E., Steel, M.F.J.: Benchmark priors for Bayesian model averaging. J. Econom. **100**(2), 381–427 (2001)

Fort, G., Moulines, E., Priouret, P., et al.: Convergence of adaptive and interacting Markov chain Monte Carlo algorithms. Ann. Stat. **39**(6), 3262–3289 (2011)

Gagnon, P.: Informed reversible jump algorithms. Electron. J. Stat. **15**(2), 3951–3995 (2021)

Garcia-Donato, G., Martinez-Beneito, M.A.: On sampling strategies in Bayesian variable selection problems with large model spaces. J. Am. Stat. Assoc. **108**(501), 340–352 (2013)

Gelman, A., Gilks, W.R., Roberts, G.O.: Weak convergence and optimal scaling of random walk Metropolis algorithms. Ann. Appl. Probab. **7**(1), 110–120 (1997)

George, E.I., McCulloch, R.: Variable selection via Gibbs sampling. J. Am. Stat. Assoc. **88**(423), 881–889 (1993)

George, E.I., McCulloch, R.E.: Approaches for Bayesian variable selection. Stat. Sin. **7**, 339–373 (1997)

Grathwohl, W., Swersky, K., Hashemi, M., Duvenaud, D., Maddison, C.J.: Oops I took a gradient: scalable sampling for discrete distributions (2021). arXiv:2102.04509

Grenander, U., Miller, M.I.: Representations of knowledge in complex systems. J. R. Stat. Soc. Ser. B (Methodol.) **56**(4), 549–581 (1994)

Griffin, J.E., Brown, P.J.: Bayesian global-local shrinkage methods for regularisation in the high dimension linear model. Chemom. Intell. Lab. Syst. **210**, 104255 (2021)

Griffin, J., Łatuszyński, K., Steel, M.: In search of lost mixing time: adaptive Markov chain Monte Carlo schemes for Bayesian variable selection with very large p. Biometrika **108**(1), 53–69 (2021)

Hans, C., Dobra, A., West, M.: Shotgun stochastic search for large-p regression. J. Am. Stat. Assoc. **102**(478), 507–516 (2007)

Hastie, T., Tibshirani, R., Wainwright, M.: Statistical learning with sparsity: the lasso and generalizations. CRC Press, Boca Raton (2015)

Ji, C., Schmidler, S.C.: Adaptive Markov chain Monte Carlo for Bayesian variable selection. J. Comput. Graph. Stat. **22**(3), 708–728 (2013)

Kiefer, J., Wolfowitz, J.: Stochastic estimation of the maximum of a regression function. Ann. Math. Stat. **23**, 462–466 (1952)

Lamnisos, D., Griffin, J.E., Steel, M.F.J.: Transdimensional sampling algorithms for Bayesian variable selection in classification problems with many more variables than observations. J. Comput. Graph. Stat. **18**(3), 592–612 (2009)

Lamnisos, D., Griffin, J.E., Steel, M.F.J.: Adaptive MC^3 and Gibbs algorithms for Bayesian model averaging in linear regression models (2013). arXiv:1306.6028

Lan, H., Chen, M., Flowers, J.B., Yandell, B.S., Stapleton, D.S., Mata, C.M., Mui, E.T.-K., Flowers, M.T., Schueler, K.L., Manly, K.F., et al.: Combined expression trait correlations and expression quantitative trait locus mapping. PLoS Genet. **2**(1), e6 (2006)

Łatuszyński, K., Roberts, G.O., Rosenthal, J.S.: Adaptive Gibbs samplers and related MCMC methods. Ann. Appl. Probab. **23**(1), 66–98 (2013)

Liang, F., Paulo, R., Molina, G., Clyde, M.A., Berger, J.O.: Mixtures of g priors for Bayesian variable selection. J. Am. Stat. Assoc. **103**(481), 410–423 (2008)

Livingstone, S., Zanella, G.: The Barker proposal: combining robustness and efficiency in gradient-based MCMC (2019). arXiv:1908.11812

Ma, L.: Scalable Bayesian model averaging through local information propagation. J. Am. Stat. Assoc. **110**(510), 795–809 (2015)

Madigan, D., York, J., Allard, D.: Bayesian graphical models for discrete data. Int. Stat. Rev. **63**, 215–232 (1995)

Mitchell, T.J., Beauchamp, J.J.: Bayesian variable selection in linear regression. J. Am. Stat. Assoc. **83**(404), 1023–1032 (1988)

Narisetty, N.N., He, X.: Bayesian variable selection with shrinking and diffusing priors. Ann. Stat. **42**(2), 789–817 (2014)

Pasarica, C., Gelman, A.: Adaptively scaling the Metropolis algorithm using expected squared jumped distance. Stat. Sin. **20**, 343–364 (2010)

Peskun, P.H.: Optimum Monte-Carlo sampling using Markov chains. Biometrika **60**(3), 607–612 (1973)

Polson, N.G., Scott, J.G.: Shrink globally, act locally: sparse Bayesian regularization and prediction. Bayesian Stat. **9**(501–538), 105 (2010)

Polson, N.G., Scott, J.G., Windle, J.: Bayesian inference for logistic models using Pólya-Gamma latent variables. J. Am. Stat. Assoc. **108**(504), 1339–1349 (2013)

Pompe, E., Holmes, C., Łatuszyński, K., et al.: A framework for adaptive MCMC targeting multimodal distributions. Ann. Stat. **48**(5), 2930–2952 (2020)

Power, S., Goldman, J. V.: Accelerated sampling on discrete spaces with non-reversible Markov processes (2019). arXiv:1912.04681

Roberts, G.O., Rosenthal, J.S.: Optimal scaling of discrete approximations to Langevin diffusions. J. R. Stat. Soc. Ser. B (Stat. Methodol.) **60**(1), 255–268 (1998)

Roberts, G.O., Rosenthal, J.S.: Coupling and ergodicity of adaptive Markov chain Monte Carlo algorithms. J. Appl. Probab. **44**(2), 458–475 (2007)

Roberts, G.O., Rosenthal, J.S., et al.: General state space Markov chains and MCMC algorithms. Probab. Surv. **1**, 20–71 (2004)

Rossell, D., Rubio, F.J.: Tractable Bayesian variable selection: beyond normality. J. Am. Stat. Assoc. **113**(524), 1742–1758 (2018)

Schäfer, C., Chopin, N.: Sequential Monte Carlo on large binary sampling spaces. Stat. Comput. **23**(2), 163–184 (2013)

Shang, Z., Clayton, M.K.: Consistency of Bayesian linear model selection with a growing number of parameters. J. Stat. Plann. Inference **141**(11), 3463–3474 (2011)

Steel, M.F.J., Ley, E.: On the Effect of Prior Assumptions in Bayesian Model Averaging with Applications to Growth Regression. The World Bank, Washington (2007)
Tierney, L.: A note on Metropolis-Hastings kernels for general state spaces. Anna. Appl. Probab. **8**, 1–9 (1998)
Titsias, M., Dellaportas, P.: Gradient-based adaptive Markov chain Monte Carlo. Adv. Neural. Inf. Process. Syst. **32**, 15730–15739 (2019)
Titsias, M.K., Yau, C.: The Hamming ball sampler. J. Am. Stat. Assoc. **112**(520), 1598–1611 (2017)
Wan, K.Y.Y., Griffin, J.E.: An adaptive MCMC method for Bayesian variable selection in logistic and accelerated failure time regression models. Stat. Comput. **31**(1), 1–11 (2021)
Yang, Y., Wainwright, M.J., Jordan, M.I., et al.: On the computational complexity of high-dimensional Bayesian variable selection. Ann. Stat. **44**(6), 2497–2532 (2016)
Zanella, G.: Informed proposals for local MCMC in discrete spaces. J. Am. Stat. Assoc. **115**(530), 852–865 (2020)
Zanella, G., Roberts, G.: Scalable importance tempering and Bayesian variable selection. J. R. Stat. Soc. B **81**(3), 489–517 (2019)
Zhou, Q., Yang, J., Vats, D., Roberts, G.O., Rosenthal, J.S.: Dimension-free mixing for high-dimensional Bayesian variable selection (2021). arXiv:2105.05719