ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Statistical Methodology
Series B

J. R. Statist. Soc. B (2019)

Scalable importance tempering and Bayesian variable selection

Giacomo Zanella

Bocconi University, Milan, Italy

and Gareth Roberts

University of Warwick, Coventry, UK

[Received May 2018. Revised February 2019]

Summary. We propose a Monte Carlo algorithm to sample from high dimensional probability distributions that combines Markov chain Monte Carlo and importance sampling. We provide a careful theoretical analysis, including guarantees on robustness to high dimensionality, explicit comparison with standard Markov chain Monte Carlo methods and illustrations of the potential improvements in efficiency. Simple and concrete intuition is provided for when the novel scheme is expected to outperform standard schemes. When applied to Bayesian variable-selection problems, the novel algorithm is orders of magnitude more efficient than available alternative sampling schemes and enables fast and reliable fully Bayesian inferences with tens of thousand regressors.

Keywords: Bayesian variable selection; Computational complexity; Gibbs sampling; Importance sampling; Markov chain Monte Carlo sampling; Point mass priors

1. Introduction

Sampling from high dimensional probability distributions is a common task arising in many scientific areas, such as Bayesian statistics, machine learning and statistical physics. In this paper we propose and analyse a novel Monte Carlo scheme for generic, high dimensional target distributions that combines importance sampling and Markov chain Monte Carlo (MCMC) sampling.

There have been many attempts to embed importance sampling within Monte Carlo schemes for Bayesian analysis; see for example Smith and Gelfand (1992), Gramacy *et al.* (2010) and beyond. However, except where sequential Monte Carlo approaches can be adopted, pure Markov-chain-based schemes (i.e. schemes which simulate from precisely the right target distribution with no need for subsequent importance sampling correction) have been far more successful. This is because MCMC methods are usually much more scalable to high dimensional situations (see for example Frieze *et al.* (1994), Belloni and Chernozhukov (2009), Yang *et al.* (2016) and Roberts and Rosenthal (2016)), whereas importance sampling weight variances tend to grow (often exponentially) with dimension. In this paper we propose a natural way to combine the best of MCMC and importance sampling in a way that is robust in high dimensional contexts and ameliorates the slow mixing which plagues many Markov-chain-based schemes. The scheme proposed, which we call tempered Gibbs sampling (TGS), involves componentwise updating rather like Gibbs sampling (GS), with improved mixing properties and associated importance

Address for correspondence: Giacomo Zanella, Department of Decision Sciences, Bocconi University, Via Roentgen 1 Milano, Milan 20136, Italy.
E-mail: zanella.gcm@gmail.com

 1369–7412/19/81000

weights which remain stable as the dimension increases. Through an appropriately designed tempering mechanism, TGS circumvents the main limitations of standard GS, such as the slow mixing that is induced by strong posterior correlations. It also avoids the requirement to visit all co-ordinates sequentially, instead iteratively making state-informed decisions about which co-ordinate should be next updated.

Our scheme differentiates from classical simulated and parallel tempering (Marinari and Parisi, 1992; Geyer and Thompson, 1995) in that it tempers only the co-ordinate that is currently being updated, and compensates for the overdispersion that is induced by the tempered update by choosing to update components which are in the tail of their conditional distributions more frequently. The resulting dynamics can dramatically speed up convergence of the standard GS, both during the transient and the stationary phase of the algorithm. Moreover, TGS does not require multiple temperature levels (as in simulated and parallel tempering) and thus avoids the tuning issues that are related to choosing the number of levels and collection of temperatures, as well as the heavy computational burden that is induced by introducing multiple copies of the original state space.

We apply the novel sampling scheme to Bayesian variable-selection (BVS) problems, observing multiple orders of magnitude improvements compared with alternative Monte Carlo schemes. For example, TGS enables us to perform reliable, fully Bayesian inference for spike-and-slab models with over 10000 regressors in less than 2 min by using a simple R implementation and a single desktop computer.

The paper is structured as follows. The TGS scheme is introduced in Section 2. There we provide basic validity results and intuition on the potential improvement that is given by the novel scheme, together with an illustrative example. In Section 3 we develop a careful analysis of the scheme. First we show that, unlike common tempering schemes, TGS is robust to high dimensionality of the target as the co-ordinatewise tempering mechanism that is employed is actually improved rather than damaged by high dimensionality. Secondly we show that TGS cannot perform worse than standard GS by more than a constant factor that can be chosen by the user (in our simulations we set it to 2), while being able to perform orders of magnitude better. Finally we provide concrete insight regarding the type of correlation structures where TGS will perform much better than GS and the structures where GS and TGS will perform similarly. In Section 4 we provide a detailed application to BVS problems, including computational complexity results. Section 5 contains simulation studies. We review our findings in Section 6. Short proofs are directly reported in the paper, whereas longer proofs can be found in the on-line supplementary material.

2. The tempered Gibbs sampling scheme

Let $f(\mathbf{x})$ be a probability distribution with $\mathbf{x} = (x_1, \ldots, x_d) \in \mathcal{X}_1 \times \ldots \times \mathcal{X}_d = \mathcal{X}$. Each iteration of the classical random-scan GS scheme proceeds by picking i from $\{1, \ldots, d\}$ uniformly at random and then sampling $x_i \sim f(x_i|\mathbf{x}_{-i})$. We consider the following tempered version of the Gibbs sampler, which depends on a collection of modified full conditionals denoted by $\{g(x_i|\mathbf{x}_{-i})\}_{i,\mathbf{x}_{-i}}$ with $i \in \{1, \ldots, d\}$ and $\mathbf{x}_{-i} \in \mathcal{X}_{-i}$. The only requirement on $g(x_i|\mathbf{x}_{-i})$ is that, for all $\mathbf{x}_{-i}$, it is a probability density function on $\mathcal{X}_i$ absolutely continuous with respect to $f(x_i|\mathbf{x}_{-i})$, with no need to be the actual full conditional of some global distribution $g(\mathbf{x})$. The following functions play a crucial role in the definition of the TGS algorithm:

$$p_i(\mathbf{x}) = \frac{g(x_i|\mathbf{x}_{-i})}{f(x_i|\mathbf{x}_{-i})} \qquad \text{for } i = 1, \ldots, d; \qquad Z(\mathbf{x}) = \frac{1}{d} \sum_{i=1}^{d} p_i(\mathbf{x}). \tag{1}$$

TGS algorithm: at each iteration of the Markov chain,

(a) (*co-ordinate selection*) sample i from $\{1,\dots,d\}$ proportionally to $p_i(\mathbf{x})$;
(b) (*tempered update*) sample $x_i \sim g(x_i|\mathbf{x}_{-i})$;
(c) (*importance weighting*) assign to the new state $\mathbf{x}$ a weight $w(\mathbf{x}) = Z(\mathbf{x})^{-1}$.

The Markov chain $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots$ that is induced by steps (a) and (b) of the TGS algorithm is reversible with respect to fZ, which is a probability density function on $\mathcal{X}$ defined as $(fZ)(\mathbf{x}) = f(\mathbf{x})Z(\mathbf{x})$. We shall assume the following condition on Z which is stronger than necessary, but which holds naturally for our purposes later on:

$$Z(\mathbf{x}) \text{ is bounded away from 0, and bounded above on compact sets.} \tag{2}$$

Throughout the paper Z and w are the inverse of each other, i.e. $w(\mathbf{x}) = Z(\mathbf{x})^{-1}$ for all $\mathbf{x} \in \mathcal{X}$. As usual, we denote the space of f-integrable functions from $\mathcal{X}$ to $\mathbb{R}$ by $L^1(\mathcal{X}, f)$ and we write $\mathbb{E}_f[h] = \int_{\mathcal{X}} h(\mathbf{x}) f(\mathbf{x}) \mathrm{d}\mathbf{x}$ for every $h \in L^1(\mathcal{X}, f)$.

Proposition 1. fZ is a probability density function on $\mathcal{X}$ and the Markov chain $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots$ induced by steps (a) and (b) of the TGS algorithm is reversible with respect to fZ. Assuming that condition (2) holds and that TGS is fZ irreducible, then

$$\hat{h}_n^{\mathrm{TGS}} = \frac{\sum_{t=1}^{n} w(\mathbf{x}^{(t)}) h(\mathbf{x}^{(t)})}{\sum_{t=1}^{n} w(\mathbf{x}^{(t)})} \to \mathbb{E}_f[h], \qquad \text{as } n \to \infty, \tag{3}$$

almost surely for every $h \in L^1(\mathcal{X}, f)$.

Proof. Reversibility with respect to $f(\mathbf{x})Z(\mathbf{x})$ can be checked as in the proof of proposition 6 in section A.4 of the on-line supplement. Representing $f(\mathbf{x})Z(\mathbf{x})$ as a mixture of d probability densities on $\mathcal{X}$ we have

$$\int_{\mathcal{X}} f(\mathbf{x}) Z(\mathbf{x}) \mathrm{d}\mathbf{x} = \int_{\mathcal{X}} \frac{1}{d} \sum_{i=1}^{d} f(\mathbf{x}) \frac{g(x_i|\mathbf{x}_{-i})}{f(x_i|\mathbf{x}_{-i})} \mathrm{d}\mathbf{x} = \frac{1}{d} \sum_{i=1}^{d} \int_{\mathcal{X}} f(\mathbf{x}_{-i}) g(x_i|\mathbf{x}_{-i}) \mathrm{d}\mathbf{x} = 1.$$

The functions h and hw have identical support from condition (2). Moreover it is clear that $h \in L^1(\mathcal{X}, f)$ if and only if $hw \in L^1(\mathcal{X}, fZ)$ and that in fact

$$\mathbb{E}_f[h] = \int h(\mathbf{x}) f(\mathbf{x}) \mathrm{d}\mathbf{x} = \int h(\mathbf{x}) w(\mathbf{x}) f(\mathbf{x}) Z(\mathbf{x}) \mathrm{d}\mathbf{x} = \mathbb{E}_{fZ}[hw] \,.$$

Therefore from theorem 17.0.1 of Meyn and Tweedie (1993) applied to both numerator and denominator, result (3) holds since by hypothesis TGS is fZ irreducible so $(\mathbf{x}^{(t)})_{t=1}^{\infty}$ is ergodic. □

We note that fZ-irreducibility of TGS can be established in specific examples by using standard techniques; see for example Roberts and Smith (1994). Moreover under condition (2) conditions from Roberts and Smith (1994) which imply that f-irreducibility of the standard Gibbs sampler readily extend to demonstrating that TGS is fZ irreducible.

The implementation of TGS requires the user to specify a collection of densities $\{g(x_i|\mathbf{x}_{-i})\}_{i,\mathbf{x}_{-i}}$. Possible choices of these include tempered conditionals of the form

$$g(x_i|\mathbf{x}_{-i}) = f^{(\beta)}(x_i|\mathbf{x}_{-i}) = \frac{f(x_i|\mathbf{x}_{-i})^{\beta}}{\int_{\mathcal{X}_i} f(y_i|\mathbf{x}_{-i})^{\beta} \mathrm{d}y_i}, \tag{4}$$

where β is a fixed value in $(0, 1)$, and mixed conditionals of the form

$$g(x_i|\mathbf{x}_{-i}) = \tfrac{1}{2} f(x_i|\mathbf{x}_{-i}) + \tfrac{1}{2} f^{(\beta)}(x_i|\mathbf{x}_{-i}), \tag{5}$$

with $\beta \in (0, 1)$ and $f^{(\beta)}$ defined as in equation (4). Note that $g(x_i|\mathbf{x}_{-i})$ in equation (5) are not the full conditionals of $\frac{1}{2} f(\mathbf{x}) + \frac{1}{2} f^{(\beta)}(\mathbf{x})$ as the latter would have mixing weights depending on $\mathbf{x}$. Indeed $g(x_i|\mathbf{x}_{-i})$ in equation (5) are unlikely to be the full conditionals of any distribution.

The theory that is developed in Section 3 will provide insight into which choice for $g(x_i|\mathbf{x}_{-i})$ leads to effective Monte Carlo methods. Moreover, we shall see that building $g(x_i|\mathbf{x}_{-i})$ as a mixture of $f(x_i|\mathbf{x}_{-i})$ and a flattened version of $f(x_i|\mathbf{x}_{-i})$, as in equation (5), is typically a robust and efficient choice.

The modified conditionals need to be tractable, as we need to sample from them and to evaluate their density. In many cases, if the original full conditionals $f(x_i|\mathbf{x}_{-i})$ are tractable (e.g. Bernoulli, normal, beta or gamma distributions), then also the densities of the form $f^{(\beta)}(x_i|\mathbf{x}_{-i})$ are. More generally, one can use any flattened version of $f(x_i|\mathbf{x}_{-i})$ instead of $f^{(\beta)}(x_i|\mathbf{x}_{-i})$. For example in Section 3.5 we provide an illustration using a t-distribution for $g(x_i|\mathbf{x}_{-i})$ when $f(x_i|\mathbf{x}_{-i})$ is normal.

TGS has various potential advantages over GS. First it makes an 'informed choice' on which variable to update, choosing with higher probability co-ordinates whose value is currently in the tail of their conditional distribution. Secondly it induces potentially longer jumps by sampling x_i from a tempered distribution $g(x_i|\mathbf{x}_{-i})$. Finally, as we shall see in the next sections, the invariant distribution fZ has potentially much less correlation between variables compared with the original distribution f.

2.1. Illustrative example

Consider the following illustrative example, where the target is a bivariate Gaussian distribution with correlation $\rho = 0.999$. Posterior distributions with such strong correlations naturally arise in Bayesian modelling, e.g. in the context of hierarchical linear models with a large number of observations. Fig. 1(a) displays the first 200 iterations of GS. As expected, the strong correlation slows down the sampler dramatically and the chain hardly moves away from the starting point, in this case $(3, 3)$. Figs 1(b) and 1(c) display the first 200 iterations of TGS with modified conditionals given by equations (4) and (5) respectively, and $\beta = 1 - \rho^2$. See Section 3 for some discussion on the choice of β in practice. Now the tempered conditional distributions of TGS allow the chain to move freely around the state space despite correlation. However, the version of TGS that uses tempered conditionals as in equation (4), which we refer to as 'TGS vanilla' here, spends the majority of its time outside the region of high probability under the target. This results in high variability of the importance weights $w(\mathbf{x}^{(t)})$ (represented by the size of the black dots in Fig. 1), which deteriorates the efficiency of the estimators $\hat{h}_t^{\mathrm{TGS}}$ defined in expression (3). In contrast, the TGS scheme that uses tempered conditionals as in equation (5), which we refer to as TGS mixed here, achieves both fast mixing of the Markov chain $\mathbf{x}^{(t)}$ and low variance of the importance weights $w(\mathbf{x}^{(t)})$. For example, for the simulations of Fig. 1, the estimated variances of the importance weights for the TGS vanilla and TGS mixed methods are 16.2 and 0.88 respectively. In Section 3 we provide theoretical analysis, as well as intuition, to explain the behaviour of TGS schemes.

Remark 1. The TGS algorithm inherits the robustness and tuning-free properties of GS, such as invariance to co-ordinate rescalings or translations. More precisely, the MCMC algorithms that are obtained by applying TGS to the original target $f(\mathbf{x})$ or to the target that is obtained by

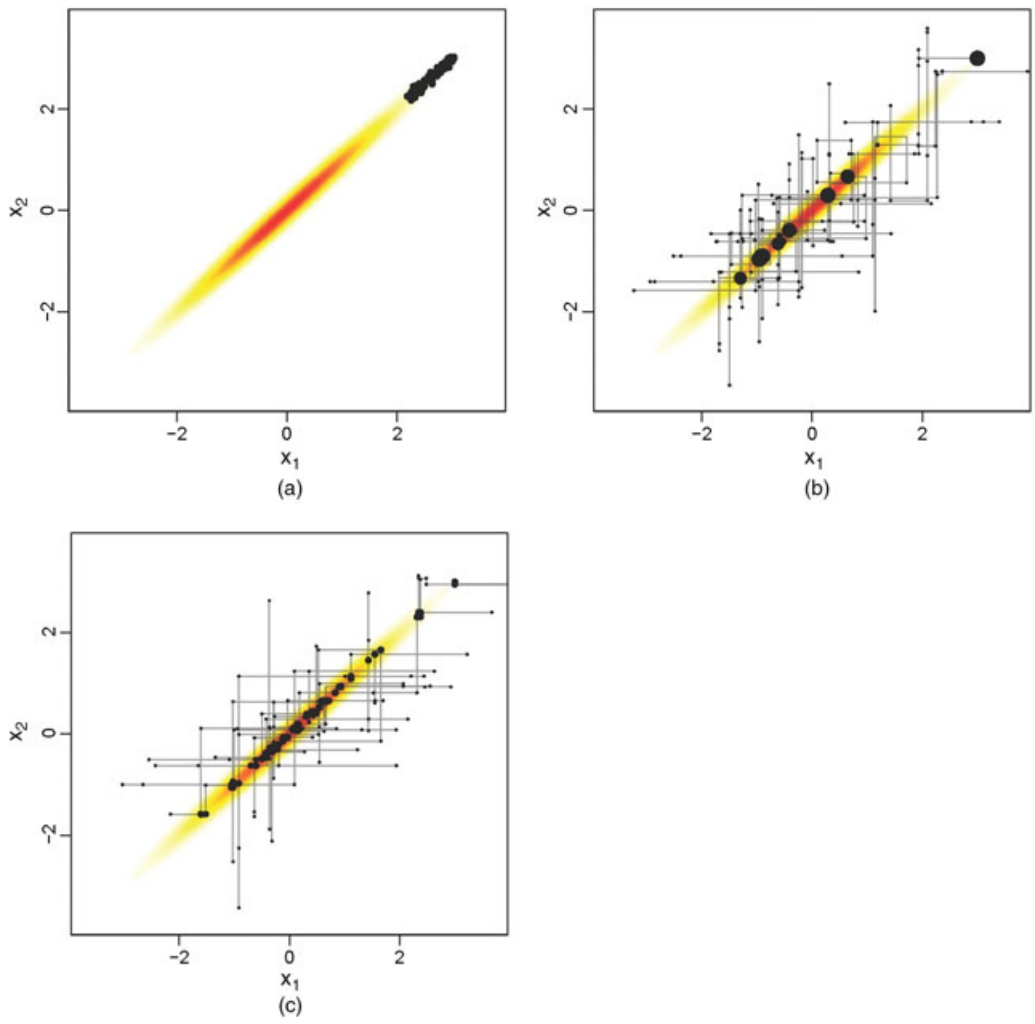

Fig. 1. Comparison of (a) GS with two versions of TGS, (b) 'vanilla' ($\widehat{\text{var}}(W) = 16.2$) and (c) 'mixed' ($\widehat{\text{var}}(W) = 0.88$) for $n = 200$ iterations on a strongly correlated bivariate distribution: the sizes of the black dots are proportional to the importance weights $(w(\mathbf{x}^{(t)}))_{t=1}^{n}$; $\widehat{\text{var}}(W)$ refers to the estimated normalized variance of the importance weights, defined as $\widehat{\text{var}}(W) = (1/n)\Sigma_{t=1}^{n} \bar{w}_t^2 - 1$, where $\bar{w}_t = w(\mathbf{x}^{(t)})/\{(1/n)\Sigma_{s=1}^{n} w(\mathbf{x}^{(s)})\}$

applying any bijective transformation to a co-ordinate x_i are equivalent, provided that $g(x_i|\mathbf{x}_{-1})$ are also transformed accordingly. A practical implication is that the TGS implementation does not require careful tuning of the scale of the proposal distribution such as typical Metropolis–Hasting algorithms do. It is also trivial to see that TGS is invariant to permutations of the order of co-ordinates.

Remark 2 (extended target interpretation). The TGS scheme has a simple alternative construction that will be useful in what follows. Consider the extended state space $\mathcal{X} \times \{1, \ldots, d\}$ with augmented target

$$\tilde{f}(\mathbf{x}, i) = \frac{1}{d} f(\mathbf{x}_{-i}) g(x_i|\mathbf{x}_{-i}) \qquad (\mathbf{x}, i) \in \mathcal{X} \times \{1, \ldots, d\}.$$

The integer i represents which co-ordinate of $\mathbf{x}$ is being tempered, and $g(x_i|\mathbf{x}_{-i})$ is the tempered version of $f(x_i|\mathbf{x}_{-i})$. The extended target $\tilde{f}$ is a probability density function over $\mathcal{X} \times \{1, \ldots, d\}$ with marginals over i and $\mathbf{x}$ given by

$$\tilde{f}(i) = \int \tilde{f}(\mathbf{x}, i) \mathrm{d}\mathbf{x} = \frac{1}{d},$$

$$\tilde{f}(\mathbf{x}) = \sum_{i=1}^{n} \tilde{f}(\mathbf{x}, i) = \frac{1}{d} \sum_{i=1}^{d} f(\mathbf{x}_{-i}) g(x_i|\mathbf{x}_{-i}) = f(\mathbf{x}) Z(\mathbf{x}).$$

TGS can be seen as a scheme that targets $\tilde{f}$ by alternating sampling from $\tilde{f}(i|\mathbf{x})$ and $\tilde{f}(x_i|i, \mathbf{x}_{-i})$, and then corrects for the difference between $\tilde{f}$ and f with $Z(\mathbf{x})^{-1}$. A direct consequence of this extended target interpretation is that the marginal distribution of i is uniform, meaning that each co-ordinate becomes updated every $1/d$ iterations on average.

3. Analysis of the algorithm

In this section we provide a careful theoretical and empirical analysis of the TGS algorithm. The first aim is providing theoretical guarantees on the robustness of TGS, both in terms of variance of the importance sampling weights in high dimensions and mixing of the resulting Markov chain compared with the GS mixing. The second aim is to provide understanding about which situations will be favourable to TGS and which will not. The main message is that the performances of TGS are never significantly worse than the GS performance and, depending on the situation, can be much better.

A key quantity in the discussion of TGS robustness is the following ratio between the original and the modified conditionals:

$$b = \sup_{i, \mathbf{x}} \frac{f(x_i|\mathbf{x}_{-i})}{g(x_i|\mathbf{x}_{-i})}. \tag{6}$$

To ensure robustness of TGS, we want the constant b to be finite and not too large. This can be easily achieved in practice. For example setting $g(x_i|\mathbf{x}_{-i})$ as in equation (5) we are guaranteed to have $b \leqslant 2$. More generally, choosing

$$g(x_i|\mathbf{x}_{-i}) = \frac{1}{1+\epsilon} f(x_i|\mathbf{x}_{-i}) + \frac{\epsilon}{1+\epsilon} f^{(\beta)}(x_i|\mathbf{x}_{-i})$$

we obtain $b \leqslant 1 + \epsilon$. The important aspect to note here is that equation (6) involves only ratios of one-dimensional densities rather than d-dimensional ratios (more precisely densities over $\mathcal{X}_i$ rather than over $\mathcal{X}$).

Throughout the paper, we measure the efficiency of Monte Carlo algorithms through their *asymptotic variances*. The smaller the asymptotic variance, the more efficient the algorithm. For any $h \in L^2(\mathcal{X}, f)$, the asymptotic variance that is associated with TGS is defined as $\mathrm{var}(h, \mathrm{TGS}) = \lim_{n \to \infty} n \, \mathrm{var}(\hat{h}_n^{\mathrm{TGS}})$, where $\hat{h}_n^{\mathrm{TGS}}$ is the TGS estimator defined in expression (3). The following lemma provides a useful representation of $\mathrm{var}(h, \mathrm{TGS})$.

Lemma 1. Let $h \in L^1(\mathcal{X}, f)$ and $\bar{h}(\mathbf{x}) = h(\mathbf{x}) - \mathbb{E}_f[h]$. If $\mathrm{var}(h, \mathrm{TGS}) < \infty$ then

$$\mathrm{var}(h, \mathrm{TGS}) = \mathbb{E}_f[\bar{h}^2 w] \left(1 + 2 \sum_{t=1}^{\infty} \rho_t \right), \tag{7}$$

where ρ_t is the lag t auto-correlation of $(w(\mathbf{x}^{(i)})\bar{h}(\mathbf{x}^{(i)}))_{i=1}^{\infty}$ and $(\mathbf{x}^{(i)})_{i=1}^{\infty}$ is the discrete time chain that is induced by TGS started in stationarity.

The term $\mathbb{E}_f[\bar{h}^2 w]$ in equation (7) equals the asymptotic variance of the hypothetical importance sampler that uses fZ as a proposal. More formally, for any $h \in L^1(\mathcal{X}, f)$ define the self-normalized importance sampling (SIS) estimator as

$$\hat{h}_n^{\mathrm{SIS}} = \frac{\sum_{i=1}^{n} w(\mathbf{y}^{(i)}) h(\mathbf{y}^{(i)})}{\sum_{i=1}^{n} w(\mathbf{y}^{(i)})},$$

where $(\mathbf{y}^{(i)})_{i=1}^{\infty}$ is a sequence of independent and identically distributed (IID) random variables with distribution fZ. Standard important sampling theory (see for example Deligiannidis and Lee (2018), section 3.2) tells us that $\mathbb{E}_f[\bar{h}^2 w] = \mathrm{var}(h, \mathrm{SIS})$, where $\mathrm{var}(h, \mathrm{SIS}) = \lim_{n\to\infty} n\mathrm{var}(\hat{h}_n^{\mathrm{SIS}})$. Therefore the two terms on the right-hand side of equation (7), $\mathbb{E}_f[\bar{h}^2 w]$ and $1 + 2\Sigma_{t=1}^{\infty}\rho_t$, can be interpreted as respectively the importance sampling and the MCMC contributions to $\mathrm{var}(h, \mathrm{TGS})$.

3.1. Robustness to high dimensionality

A major concern with classical importance tempering schemes is that they often collapse in high dimensional scenarios (see for example Owen (2013), section 9.1). The reason is that the 'overlap' between the target distribution f and a tempered version, such as $g = f^{(\beta)}$ with $\beta \in (0, 1)$, can be extremely low if f is a high dimensional distribution. In contrast, the importance sampling procedure that is associated with TGS is robust to high dimensional scenarios. This can be quantified by looking at the asymptotic variances $\mathrm{var}(h, \mathrm{SIS}) = \mathbb{E}_f[\bar{h}^2 w]$, or at the variance of the importance weights $W = w(\mathbf{X})$ for $\mathbf{X} \sim fZ$.

Proposition 2. Given $\mathbf{X} \sim fZ$ and $W = w(\mathbf{X})$, we have

$$\mathrm{var}(W) \leqslant b - 1$$

and

$$\mathrm{var}(h, \mathrm{SIS}) \leqslant b\,\mathrm{var}_f(h),$$

with b defined in equation (6) and $\mathrm{var}_f(h) = \mathbb{E}_f[h^2] - \mathbb{E}_f[h]^2$.

Proof. Equation (6) implies that $p_i(\mathbf{x}) \geqslant b^{-1}$ and thus $w(\mathbf{x}) = Z(\mathbf{x})^{-1} \leqslant b$ for every $\mathbf{x} \in \mathcal{X}$. Combining the latter inequality with $\mathrm{var}(W) = \mathbb{E}_{fZ}[w^2] - \mathbb{E}_{fZ}[w]^2 = \mathbb{E}_f[w] - 1$, we obtain $\mathrm{var}(W) = \mathbb{E}_f[w] - 1 \leqslant b - 1$. Again from $w(\mathbf{x}) \leqslant b$, we have $\mathrm{var}(h, \mathrm{SIS}) = \mathbb{E}_f[\bar{h}^2 w] \leqslant b\,\mathbb{E}_f[\bar{h}^2] = b\,\mathrm{var}_f(h)$.

Proposition 2 implies that, regardless of the dimensionality of the state space, the asymptotic variance $\mathrm{var}(h, \mathrm{SIS})$ is at most b times $\mathrm{var}_f(h)$. Therefore, by equation (7), setting b to a low value is sufficient to ensure that the importance sampling contribution to $\mathrm{var}(h, \mathrm{TGS})$ is well behaved. For example, if $g(x_i|\mathbf{x}_{-i})$ are chosen to be the mixed conditionals in equation (5) we are guaranteed to have $\mathrm{var}(W) \leqslant 1$ and $\mathrm{var}(h, \mathrm{SIS}) \leqslant 2\,\mathrm{var}_f(h)$. Note that the theoretical bound $\mathrm{var}(W) \leqslant 1$ is coherent with the estimated variance of the importance weights of the TGS mixed algorithm in Fig. 1.

An even stronger property of TGS than the bounds in proposition 2 is that, under appropriate assumptions, $\mathrm{var}(W)$ converges to 0 as $d \to \infty$. The underlying reason is that the weight function $w(\mathbf{x})$ depends on an average of d terms, namely $(1/d)\Sigma_{i=1}^{d} p_i(\mathbf{x})$, and the increase of dimensionality has a stabilizing effect on the latter. If, for example, the target has indepen-

dent components with common distribution f_0, $f(\mathbf{x}) = \Pi_{i=1}^{d} f_0(x_i)$, one can show that $\mathrm{var}(W)$ converges to 0 as $d \to \infty$.

Proposition 3. Suppose that $f(\mathbf{x}) = \Pi_{i=1}^{d} f_0(x_i)$ and $g(x_i|\mathbf{x}_{-i}) = g_0(x_i)$ where f_0 and g_0 are univariate probability density functions that are independent of i. If $\sup_{x_i} f_0(x_i)/g_0(x_i) \leqslant b < \infty$, then

$$\mathrm{var}(W) \to 0 \qquad \text{as } d \to \infty. \tag{8}$$

Proof. By assumption we have

$$w(\mathbf{x})^{-1} = \frac{1}{d} \sum_{i=1}^{d} \frac{g_0(x_i)}{f_0(x_i)}.$$

Thus, given $\mathbf{x} \sim f$, $w(\mathbf{x})^{-1}$ is the average of IID random variables with mean 1 and converges almost surely to 1 by the strong law of large numbers. It follows that $w(\mathbf{x}) \to 1$ almost surely as $d \to \infty$. Also, $\sup_{x_i} f_0(x_i)/g_0(x_i) \leqslant b$ implies that

$$w(\mathbf{x}) = \left\{ \frac{1}{d} \sum_{i=1}^{d} \frac{g_0(x_i)}{f_0(x_i)} \right\}^{-1} \leqslant b.$$

Thus by the bounded convergence theorem $\mathbb{E}_f[w] \to 1$ as $d \to \infty$. It follows that $\mathrm{var}(W) = (\mathbb{E}_f[w] - 1) \to 0$. □

By contrast, recall that the importance weights that are associated with classical tempering (e.g. setting $g = f^{(\beta)}$ as importance distribution) in an IID context such as proposition 3 would have a variance growing exponentially with d (see examples 9.1–9.3 of Owen (2013) for a more detailed discussion).

Proposition 3 makes the assumption of IID components for simplicity and illustration. In fact, inspecting the proof of proposition 3, we can see that result (8) holds whenever $b < \infty$ and $\lim_{d\to\infty} (1/d)\Sigma_{i=1}^{d} p_i(\mathbf{x}) = 1$ in probability for $\mathbf{x} \sim f$. Therefore, one could extend proposition 3 to any scenario where the law of large numbers for $\{p_i(\mathbf{x})\}_i$ holds. These include, for example, the case where f has independent but non-identical components such that the variance of $p_i(\mathbf{x})$ is bounded, i.e. $f(\mathbf{x}) = \Pi_{i=1}^{d} f_i(x_i)$, $g(x_i|\mathbf{x}_{-i}) = g_i(x_i)$ and

$$\int_{\mathcal{X}_i} \frac{g_i(x_i)}{f_i(x_i)} g_i(x_i) \mathrm{d}x_i$$

bounded over i. More generally, one could exploit laws of large numbers for dependent random variables in cases where the d components of $\mathbf{x} \sim f$ enjoy some appropriate local dependence structure which is sufficient to have $(1/d)\Sigma_{i=1}^{d} p_i(\mathbf{x})$ converging to a constant as $d \to \infty$.

3.2. Explicit comparison with standard Gibbs sampling

We now compare the efficiency of the Monte Carlo estimators that are produced by TGS with the estimaters that are produced by classical GS. For any function $h \in L^1(\mathcal{X}, f)$ define the GS estimator of $\mathbb{E}_f[h]$ as $\hat{h}_n^{\mathrm{GS}} = (1/n)\Sigma_{t=1}^{n} h(\mathbf{y}^{(t)})$, where $\mathbf{y}^{(1)}, \mathbf{y}^{(2)}, \ldots$ is the $\mathcal{X}$-valued Markov chain generated by GS, and denote the corresponding asymptotic variance by $\mathrm{var}(h, \mathrm{GS}) = \lim_{n\to\infty} n\, \mathrm{var}(\hat{h}_n^{\mathrm{GS}})$. The following theorem shows that the efficiency of TGS estimators can never be worse than the efficiency of GS estimators by a factor larger than b^2.

Theorem 1. For every $h \in L^2(\mathcal{X}, f)$ we have

$$\mathrm{var}(h, \mathrm{TGS}) \leqslant b^2 \mathrm{var}(h, \mathrm{GS}) + b^2 \mathrm{var}_f(h). \tag{9}$$

Remark 3. In most non-trivial scenarios, $\mathrm{var}_f(h)$ will be small in comparison with $\mathrm{var}(h, \mathrm{GS})$, because the asymptotic variance that is obtained by GS is typically much larger than that of an IID sampler. In such cases we can interpret inequality (9) as saying that the asymptotic variance of TGS is at most b^2 times those of GS plus a smaller order term. More generally, since the Markov kernel that is associated with GS is a positive operator, we have $\mathrm{var}(h, \mathrm{GS}) \geqslant \mathrm{var}_f(h)$ and thus, by inequality (9),

$$\mathrm{var}(h, \mathrm{TGS}) \leqslant 2b^2 \mathrm{var}(h, \mathrm{GS}) \qquad \text{for all } h \in L^2(\mathcal{X}, f). \tag{10}$$

Remark 4. Assuming that $b < \infty$, theorem 1 implies that whenever $\mathrm{var}(h, \mathrm{GS})$ is finite then also $\mathrm{var}(h, \mathrm{TGS})$ is finite. In general it is possible for $\mathrm{var}(h, \mathrm{TGS})$ to be finite when $\mathrm{var}(h, \mathrm{GS})$ is not. The simplest example can be obtained setting $d = 1$, in which case GS and TGS boil down to respectively IID sampling and importance sampling. In that case, any function h such that $\int_{\mathcal{X}} h(x)^2 f(x) \mathrm{d}x = \infty$ but $\int_{\mathcal{X}} h(\mathbf{x})^2 w(\mathbf{x}) f(\mathbf{x}) \mathrm{d}\mathbf{x} < \infty$ will satisfy $\mathrm{var}(h, \mathrm{GS}) = \infty$ and $\mathrm{var}(h, \mathrm{TGS}) < \infty$.

As discussed after equation (6), it is easy to set b to a desired value in practice, for example using a mixture structure as in equation (5), which leads to the following corollary.

Corollary 1. Let $\epsilon, \beta > 0$. If

$$g(x_i|\mathbf{x}_{-i}) = \frac{1}{1+\epsilon} f(x_i|\mathbf{x}_{-i}) + \frac{\epsilon}{1+\epsilon} f^{(\beta)}(x_i|\mathbf{x}_{-i})$$

then

$$\mathrm{var}(h, \mathrm{TGS}) \leqslant (1+\epsilon)^2 \mathrm{var}(h, \mathrm{GS}) + (1+\epsilon)^2 \mathrm{var}_f(h).$$

By choosing ϵ to be sufficiently small, we have theoretical guarantees that GS is not doing more than $(1+\epsilon)^2$ times better than TGS. Choosing ϵ too small, however, will reduce the potential benefit that is obtained with TGS, with TGS collapsing to GS for $\epsilon = 0$, so optimizing involves a compromise between these extremes. The optimal choice involves a trade-off between small variance of the importance sampling weights and fast mixing of the resulting Markov chain. In our examples we used $\epsilon = 1$, leading to equation (5), which is a safe and robust choice both in terms of importance sampling variance and of Markov chain mixing.

3.3. Tempered Gibbs sampling and correlation structure

Theorem 1 implies that, under suitable choices of $g(x_i|\mathbf{x}_{-i})$, TGS never provides significantly worse (i.e. worse by more than a controllable constant factor) efficiency than GS. In contrast, TGS performances can be much better than standard GS. The underlying reason is that the tempering mechanism can dramatically speed up the convergence of the TGS Markov chain $\mathbf{x}^{(t)}$ to its stationary distribution fZ by reducing correlations in the target. In fact, the covariance structure of fZ is substantially different from the structure of the original target f and this can prevent the sampler from becoming stuck in situations where GS would. Fig. 2 displays the original target f and the modified target fZ for a bivariate Gaussian distribution with increasing correlation. Here the modified conditionals are defined as in equation (4) with $\beta = 1 - \rho^2$. It can be seen that, even if the correlation of f goes to 1, the importance distribution fZ does not collapse on the diagonal (note that fZ is not Gaussian here). As we show in the next section, this

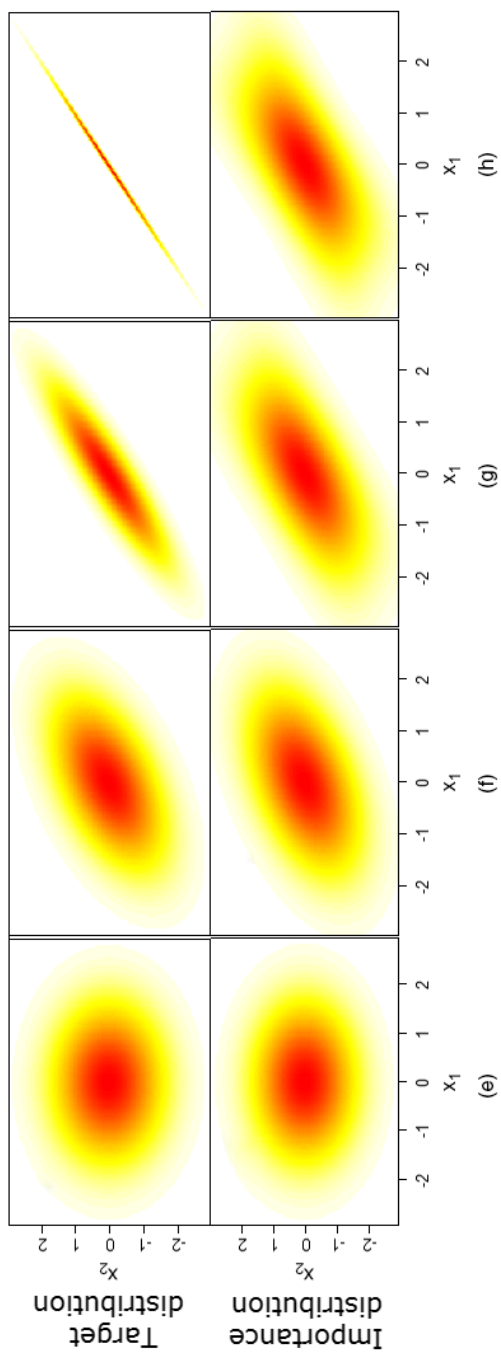

Fig. 2. Comparison between (a)–(d) f and (e)–(h) fZ, for increasing correlation (here f is a symmetric bivariate normal distribution with correlation ρ and $g = f^{(\beta)}$ with $\beta = 1 - \rho^2$): (a), (e) $\rho = 0$; (b), (f) $\rho = 0.5$; (c), (g) $\rho = 0.9$; (d), (h) $\rho = 0.999$

allows TGS to have a mixing time that is uniformly bounded over ρ. Clearly, the same property does not hold for GS, whose mixing time deteriorates as $\rho \to 1$.

A classical tempering approach would not help the Gibbs sampler in this context. In fact, a Gibbs sampler targeting $f^{(\beta)}$ with $\beta < 1$ may be as slow to converge as a sampler targeting f. For example, in the Gaussian case the covariance matrix of $f^{(\beta)}$ is simply β times the matrix of f and thus, using the results of Roberts and Sahu (1997), a Gibbs sampler targeting $f^{(\beta)}$ has exactly the same rate of convergence as a sampler targeting f. In the next section we provide some more rigorous understanding of the convergence behaviour of TGS to show the potential mixing improvements compared with GS.

3.4. Convergence analysis in the bivariate case

In general, the TGS Markov chain $\mathbf{x}^{(t)}$ evolves according to highly complex dynamics and providing generic results on its rate of convergence of fZ is extremely challenging. Nonetheless, we now show that, using the notion of deinitializing chains from Roberts and Rosenthal (2001) we can obtain quite an explicit understanding of the convergence behaviour of $\mathbf{x}^{(t)}$ in the bivariate case. The results suggest that, for appropriate choices of modified conditionals, the mixing time of $\mathbf{x}^{(t)}$ is uniformly bounded regardless of the correlation structure of the target. This must be contrasted with the chain that is induced by GS, whose mixing time diverges to ∞ as the target's correlation goes to 1.

Our analysis proceeds as follows. First we consider the augmented Markov chain $(\mathbf{x}^{(t)}, i^{(t)})_{t=0}^{\infty}$ on $\mathcal{X} \times \{1, \ldots, d\}$ obtained by including the index i, as in remark 2. The transition from $(\mathbf{x}^{(t)}, i^{(t)})$ to $(\mathbf{x}^{(t+1)}, i^{(t+1)})$ is given by the following two steps:

(a) sample $i^{(t+1)}$ from $\{1, \ldots, d\}$ proportionally to $(p_1(\mathbf{x}^{(t)}), \ldots, p_d(\mathbf{x}^{(t)}))$;
(b) sample $x_{i^{(t+1)}}^{(t+1)} \sim g(x_{i^{(t+1)}} | \mathbf{x}_{-i^{(t+1)}} = \mathbf{x}_{-i^{(t+1)}}^{(t)})$ and set $\mathbf{x}_{-i^{(t+1)}}^{(t+1)} = \mathbf{x}_{-i^{(t+1)}}^{(t)}$.

Once we augment the space with $i^{(t)}$, we can ignore the component $x_{i^{(t)}}^{(t)}$, whose distribution is fully determined by $\mathbf{x}_{-i^{(t)}}^{(t+1)}$ and $i^{(t)}$. More precisely, consider the stochastic process $(\mathbf{z}^{(t)}, i^{(t)})_{t=0}^{\infty}$ that is obtained by taking

$$\mathbf{z}^{(t)} = \mathbf{x}_{-i^{(t)}}^{(t)}, \qquad t \geqslant 0,$$

where $\mathbf{x}_{-i^{(t)}}^{(t)}$ denotes the vector $\mathbf{x}^{(t)}$ without the $i^{(t)}$th component. The following proposition shows that the process $(\mathbf{z}^{(t)}, i^{(t)})_{t=0}^{\infty}$ is Markovian and contains all the information that is needed to characterize the convergence to stationarity of $\mathbf{x}^{(t)}$.

Proposition 4. The process $(\mathbf{z}^{(t)}, i^{(t)})_{t=0}^{\infty}$ is a Markov chain and is deinitializing for $(\mathbf{x}^{(t)}, i^{(t)})_{t=0}^{\infty}$, meaning that

$$\mathcal{L}(\mathbf{x}^{(t)}, i^{(t)} | \mathbf{x}^{(0)}, i^{(0)}, \mathbf{z}^{(t)}, i^{(t)}) = \mathcal{L}(\mathbf{x}^{(t)}, i^{(t)} | \mathbf{z}^{(t)}, i^{(t)}), \qquad t \geqslant 1, \tag{11}$$

where $\mathcal{L}(\cdot|\cdot)$ denotes conditional distributions. It follows that for any starting state $\mathbf{x}_* \in \mathcal{X}$

$$\| \mathcal{L}(\mathbf{x}^{(t)} | \mathbf{x}^{(0)} = \mathbf{x}_*) - fZ \|_{\mathrm{TV}} = \| \mathcal{L}(\mathbf{z}^{(t)}, i^{(t)} | \mathbf{x}^{(0)} = \mathbf{x}_*) - \pi \|_{\mathrm{TV}}, \tag{12}$$

where '$\| \cdot \|_{\mathrm{TV}}$' denotes total variation distance and π is the stationary distribution of $(\mathbf{z}^{(t)}, i^{(t)})$.

Note that the conditioning on $\mathbf{x}^{(0)}$ in equation (12) is equivalent to conditioning on $(\mathbf{x}^{(0)}, i^{(0)})$, because the distribution of $(\mathbf{x}^{(t)}, i^{(t)})$ for $t > 1$ is independent of $i^{(0)}$.

Proposition 4 implies that the convergence to stationarity of $\mathbf{x}^{(t)}$ is fully determined by that of $(\mathbf{z}^{(t)}, i^{(t)})$. In some situations, by looking at the chain $(\mathbf{z}^{(t)}, i^{(t)})$ rather than $\mathbf{x}^{(t)}$, we can obtain a better understanding of the convergence properties of TGS. Consider for example the bivariate

case, with $\mathcal{X} = \mathbb{R}^2$ and target $f(x_1, x_2)$. In this context $(z^{(t)})_{t=0}^{\infty}$ is an $\mathbb{R}$-valued process, with stationary distribution $\frac{1}{2} f_1(z) + \frac{1}{2} f_2(z)$, where $f_1(z) = \int_{\mathbb{R}} f(z, x_2) \mathrm{d}x_2$ and $f_2(z) = \int_{\mathbb{R}} f(x_2, z) \mathrm{d}x_2$ are the target marginals. To keep the notation light and to have results that are easier to interpret, here we further assume exchangeability, i.e. $f(x_1, x_2) = f(x_2, x_1)$, whereas lemma 5 in the on-line supplementary material considers the generic case. The simplification that is given by exchangeability is that it suffices to consider the Markov chain $(z^{(t)})_{t=0}^{\infty}$ rather than $(z^{(t)}, i^{(t)})_{t=0}^{\infty}$.

Proposition 5. Let $\mathcal{X} = \mathbb{R}^2$ and f be a target distribution with $f(x_1, x_2) = f(x_2, x_1)$, and marginal on x_1 denoted by f_1. For any starting state $\mathbf{x}_* = (z_*, z_*) \in \mathbb{R}^2$ we have

$$\| \mathcal{L}(\mathbf{x}^{(t)} | \mathbf{x}^{(0)} = \mathbf{x}_*) - fZ \|_{\mathrm{TV}} = \| \mathcal{L}(z^{(t)} | z^{(0)} = z_*) - f_1 \|_{\mathrm{TV}},$$

where $z^{(t)}$ is an $\mathbb{R}$-valued Markov chain with stationary distribution $f_1(z)$ and transition kernel

$$P(z'|z) = r(z) \delta_{(z)}(z') + q(z'|z) \alpha_b(z'|z), \tag{13}$$

where

$$r(z) = 1 - \int_{\mathbb{R}} \alpha_b(z'|z) q(z'|z) \mathrm{d}z',$$

$$\alpha_b(z'|z) = \frac{f_1(z') q(z|z')}{f_1(z) q(z'|z) + f_1(z') q(z|z')}$$

and $q(z'|z) = g(x_i = z' | x_{-i} = z)$.

The transition kernel (13) coincides with that of an accept–reject algorithm with proposal distribution $q(z'|z) = g(x_i = z'|x_{-i} = z)$ and acceptance given by the Barker rule, i.e. accept with probability $\alpha_b(z'|z)$. The intuition behind the appearance of an accept–reject step is that updating the same co-ordinate x_i in consequent iterations of TGS coincides with not moving the chain $(z^{(t)})$ and thus having a rejected transition. Proposition 5 implies that, given the modified conditionals $g(x_i|x_{-i})$, the evolution of $(z^{(t)})_{t=0}^{\infty}$ depends on f only through the marginal distributions, f_1 or f_2, rather than on the joint distribution $f(x_1, x_2)$.

Proposition 5 provides quite a complete understanding of TGS convergence behaviour for bivariate exchangeable distributions. Consider for example a bivariate Gaussian target with correlation ρ, as in Section 2.1. From remark 1, we can assume without loss of generality that f has standard normal marginals and thus is exchangeable. In this case $(z^{(t)})_{t=0}^{\infty}$ is a Markov chain with stationary distribution $f_1 = N(0, 1)$ and proposal $q(z'|z) = g(x_i = z'|x_{-i} = z)$. For example, choosing modified conditionals as in equation (4) with $\beta = 1 - \rho^2$ we obtain $q(\cdot|z) = N(\rho z, 1)$. The worst-case scenario for such a chain is $\rho = 1$, where $q(\cdot|z) = N(z, 1)$. Nonetheless, even in this case the mixing of $(z^{(t)})_{t=0}^{\infty}$, and thus of $(\mathbf{x}^{(t)})_{t=0}^{\infty}$, does not collapse. By contrast, the convergence of GS in this context deteriorates as $\rho \to 1$ as it is closely related to the convergence of the auto-regressive process $z^{(t+1)}|z^{(t)} \sim N(\rho z, 1 - \rho^2)$. The discussion about the bivariate Gaussian case provides theoretical insight for the behaviour that was heuristically observed in Section 2.1. Proposition 5 is not limited to the Gaussian context and thus we would expect that the qualitative behaviour just described holds much more generally.

3.5. When does tempered Gibbs sampling work and when does it not?

The previous two sections showed that in the bivariate case TGS can induce much faster mixing compared with GS. A natural question is how much this extends to the case $d > 2$. In this section we provide insight into when TGS substantially outperforms GS and when instead they

are comparable (we know by theorem 1 that TGS cannot converge substantially slower than GS). Whether or not TGS substantially outperforms GS depends on the correlation structure of the target with intuition as follows. When sampling from a d-dimensional target $(x_1, \ldots, x_d)$, the tempering mechanism of TGS enables us to overcome strong pairwise correlations between any pair of variables x_i and x_j as well as strong k-wise *negative* correlations, i.e. negative correlations between blocks of k variables. In contrast TGS does not help significantly in overcoming strong k-wise *positive* correlations. We illustrate this behaviour with a simulation study considering multivariate Gaussian targets with increasing degree of correlations (controlled by a parameter $\rho \in [0, 1]$) under three scenarios. Given the scale and translation invariance properties of the algorithms under consideration, we can assume without loss of generality that the d-dimensional target has zero mean and covariance matrix Σ satisfying $\Sigma_{ii} = 1$ for $i = 1, \ldots, n$ in all scenarios. The first scenario considers pairwise correlation, with d being a multiple of 2 and $\Sigma_{2i-1,2i} = \rho$ for $i = 1, \ldots, d/2$ and $\Sigma_{ij} = 0$ otherwise; the second exchangeable, positively correlated distributions with $\Sigma_{ij} = \rho$ for all $i \neq j$; the third exchangeable, negatively correlated distributions with $\Sigma_{ij} = -\rho/(n-1)$ for all $i \neq j$. In all scenarios, as $\rho \to 1$ the target distribution collapses to some singular distribution and the GS convergence properties deteriorate (see Roberts and Sahu (1997) for related results).

Fig. 3 reports the (estimated) asymptotic variance of the estimators of the co-ordinates mean (i.e. $h(\mathbf{x}) = x_i$; the value of i is irrelevant) for $d = 10$. We compare GS with two versions of TGS. The first has mixed conditionals as in equation (5), with $\beta = 1 - \rho^2$. By choosing a value of β that depends on ρ we are exploiting explicit global knowledge on Σ in a potentially unrealistic way, matching the inflated conditional variance with the marginal variance. Thus we also consider a more realistic situation where we ignore global knowledge on Σ and set $g(x_i|\mathbf{x}_{-i})$ to be a t-distribution centred at $\mathbb{E}[x_i|\mathbf{x}_{-i}]$, with scale $\sqrt{\text{var}(x_i|\mathbf{x}_{-i})}$ and shape $\nu = 0.2$. As expected, the asymptotic variance of the estimators that are obtained with GS deteriorate in all cases. In contrast, TGS performances do not deteriorate or deteriorate very mildly as $\rho \to 1$ for scenarios 1 and 3. For scenario 2, TGS has very similar performances compared with GS. In all cases, the two versions of TGS perform quite similarly, with the first of the two being slightly more efficient. The qualitative conclusions of these simulations are not sensitive to various set-up details, such as the value of d, the order of variables (especially in scenario 1) or the degree of symmetry. Also, it is worth noting that TGS does not require prior knowledge of the global correlation structure or of which variables are strongly correlated to be implemented.

The reason for the presence or lack of improvements given by TGS lies in the different geometrical structure that is induced by positive and negative correlations. Intuitively, we conjecture that, if the limiting singular distribution for $\rho \to 1$ can be navigated with pairwise updates (i.e. moving on (x_i, x_j) 'planes' rather than (x_i) 'lines' as for GS), then TGS should perform well (i.e. uniformly good mixing over ρ for a good choice of β); otherwise it will not.

The intuition that is developed here will be useful in the BVS application of Section 4; see for example the discussion in Section 4.5.

3.6. Controlling the frequency of co-ordinate updating

In the extended target interpretation that was discussed in remark 2 we have shown that the marginal distribution of i under the extended target $\tilde{f}$ is uniform over $\{1, \ldots, d\}$. This implies that, for every $i, j \in \{1, \ldots, d\}$, the TGS scheme will update x_i and x_j the same number of times on average. In the absence of prior information on the structure of the problem under consideration, updating each co-ordinate the same number of times on average is a desirable robustness property as it prevents the algorithm for updating some co-ordinates too often and ignoring others. However, in some contexts, we may want to invest more computational effort

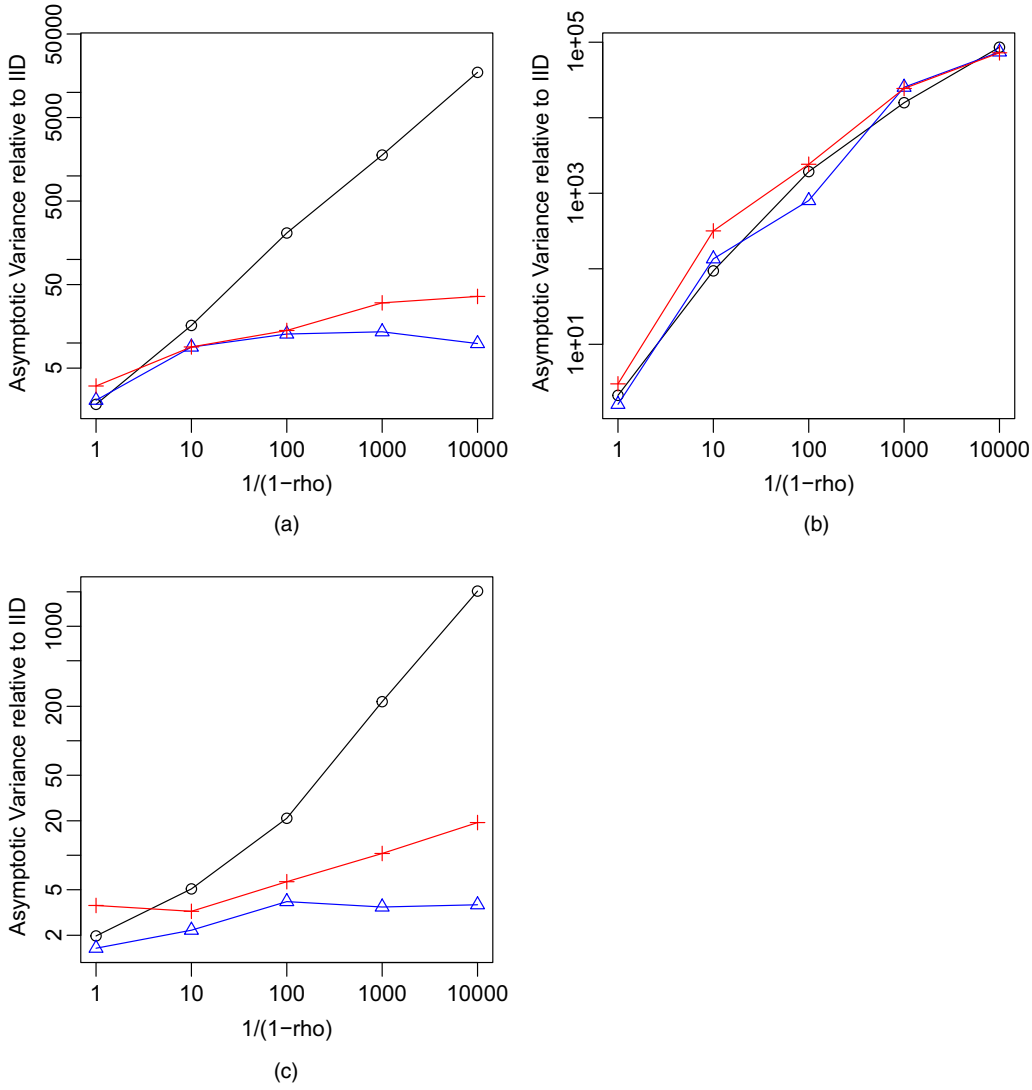

Fig. 3. Log–log-plots of estimated asymptotic variances for GS(○) compared with two versions of TGS (△, β; +, Student t) on Gaussian targets with different covariance structures: (a) pairwise correlation; (b) k-wise positive correlation; (c) k-wise negative correlation

in updating some co-ordinates rather than others (see for example the BVS problems that are discussed below). This can be done by multiplying the selection probability $p_i(\mathbf{x})$ for some weight function $\eta_i(\mathbf{x}_{-i})$, while leaving the rest of the algorithm unchanged. We call the resulting algorithm weighted tempered Gibbs sampling (WTGS).

WTGS algorithm: at each iteration of the Markov chain do

(a) sample i from $\{1,\dots,d\}$ proportionally to

$$p_i(\mathbf{x}) = \eta_i(\mathbf{x}_{-i})\frac{g(x_i|\mathbf{x}_{-i})}{f(x_i|\mathbf{x}_{-i})};$$

(b) sample $x_i \sim g(x_i|\mathbf{x}_{-i})$;
(c) weight the new state $\mathbf{x}$ with a weight $Z(\mathbf{x})^{-1}$ where $Z(\mathbf{x}) = \zeta^{-1}\Sigma_{i=1}^{d} p_i(\mathbf{x})$ and $\zeta = \Sigma_{i=1}^{d}\mathbb{E}_{x\sim f}[\eta_i(\mathbf{x}_{-i})]$.

The normalizing constant ζ in the definition of $Z(\mathbf{x})$ is designed so that $\mathbb{E}_f[Z] = 1$ as for TGS. When implementing WTGS, one needs to compute the weights $Z(\mathbf{x})^{-1}$ only up to proportionality and thus ζ need not be computed explicitly. TGS is a special case of WTGS obtained when $\eta_i(\mathbf{x}_{-i}) = 1$, in which case $\zeta = d$.

As shown by the following proposition, the introduction of the weight functions $\eta_i(\mathbf{x}_{-i})$ does not impact the validity of the algorithm and it results in having a marginal distribution over the updated component i proportional to $\mathbb{E}[\eta_i(\mathbf{x}_{-i})]$, where $\mathbf{x} \sim f$.

Proposition 6. The Markov chain $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots$ that is induced by steps (a) and (b) of the WTGS algorithm is reversible with respect to fZ. The frequency of updating of the ith co-ordinate equals $\zeta^{-1}\mathbb{E}_{x\sim f}[\eta_i(\mathbf{x}_{-i})]$.

By choosing $\eta_i(\mathbf{x}_{-i})$ appropriately, we can control the frequency with which we update each co-ordinate. In Section 4.3 we show an application of WTGS to BVS problems.

4. Application to Bayesian variable selection

We shall illustrate the theoretical and methodological conclusions of Section 3 in an important class of statistical models where Bayesian computational issues are known to be particularly challenging. Binary inclusion variables in BVS models typically have the kind of pairwise and/or negative dependence structures that have been conjectured to be conducive to successful application of TGS in Section 3.5 (see Section 4.5 for a more detailed discussion). Therefore, in this section we provide a detailed application of TGS to sampling from the posterior distribution of Gaussian BVS models. This is a widely used class of models where posterior inferences are computationally challenging because of high dimensional discrete parameters. In this context, the Gibbs sampler is the standard choice of algorithm to draw samples from the posterior distribution (see section B.6 in the on-line supplement for more details).

4.1. Model specification

BVS models provide a natural and coherent framework to select a subset of explanatory variables in linear regression contexts (Chipman *et al.*, 2001). In standard linear regression, an $n \times 1$ response vector Y is modelled as $Y|\beta, \sigma^2 \sim N(X\beta, \sigma^2)$, where X is an $n \times p$ design matrix and β an $n \times 1$ vector of coefficients. In BVS models a vector of binary variables $\gamma = (\gamma_1, \ldots, \gamma_p) \in \{0, 1\}^p$ is introduced to indicate which regressor is included in the model and which is not ($\gamma_i = 1$ indicates that the ith regressor is included in the model and $\gamma_i = 0$ that it is excluded). The resulting model can be written as

$$Y|\beta_\gamma, \gamma, \sigma^2 \sim N(X_\gamma\beta_\gamma, \sigma^2\mathbb{I}_n),$$
$$\beta_\gamma|\gamma, \sigma^2 \sim N(0, \sigma^2\Sigma_\gamma),$$
$$p(\sigma^2) \propto 1/\sigma^2,$$

where X_γ is the $n \times |\gamma|$ matrix containing only the included columns of the $n \times p$ design matrix X, β_γ is the $|\gamma| \times 1$ vector containing only the coefficients corresponding to the selected regressors and Σ_γ is the $|\gamma| \times |\gamma|$ prior covariance matrix for the $|\gamma|$ selected regressors. Here $|\gamma| = \Sigma_{i=1}^{p}\gamma_i$ denotes the number of 'active' regressors. The covariance Σ_γ is typically chosen to be equal to

a positive multiple of $(X_\gamma^{\mathrm{T}} X_\gamma)^{-1}$ or the identity matrix, i.e. $\Sigma_\gamma = c(X_\gamma^{\mathrm{T}} X_\gamma)^{-1}$ or $\Sigma_\gamma = c\mathbb{I}_{|\gamma|}$ for fixed $c > 0$. The binary vector γ is given a prior distribution $p(\gamma)$ on $\{0,1\}^p$, e.g. assuming that

$$\gamma_i | h \overset{\text{IID}}{\sim} \text{Bern}(h) \qquad i = 1, \ldots, p,$$

where h is a prior inclusion probability, which can either be set to some fixed value in $(0,1)$ or be given a prior distribution (e.g. a distribution belonging to the beta family).

Remark 5. One can also add an intercept to the linear model obtaining $Y|\beta_\gamma, \gamma, \sigma^2, \alpha \sim N(\alpha + X_\gamma \beta_\gamma, \sigma^2)$. Assigning a flat prior, $p(\alpha) \propto 1$, to such an intercept is equivalent to centring $Y, X_1, \ldots, X_p$ to have zero mean (Chipman *et al.* (2001), section 3).

Under this model set-up, the continuous hyperparameters β and σ can be analytically integrated and we are left with an explicit expression for $p(\gamma|Y)$. Sampling from such a $\{0,1\}^p$-valued distribution enables us to perform full posterior inferences for the BVS models that were specified above since $p(\beta_\gamma, \gamma, \sigma^2|Y) = p(\beta_\gamma, \sigma^2|\gamma, Y)\, p(\gamma|Y)$ and $p(\beta_\gamma, \sigma^2|\gamma, Y)$ is analytically tractable. The standard way to draw samples from $p(\gamma|Y)$ is by performing GS on the p components $(\gamma_1, \ldots, \gamma_p)$, repeatedly choosing $i \in \{1, \ldots, p\}$ either in a random or a deterministic scan fashion and then updating $\gamma_i \sim p(\gamma_i | Y, \gamma_{-i})$.

4.2. Tempered Gibbs sampling for Bayesian variable selection

We apply TGS to the problem of sampling from $\gamma \sim p(\gamma|Y)$. Under the notation of Section 2, this corresponds to $d = p$, $\mathcal{X} = \{0,1\}^p$ and $f(\gamma) = p(\gamma|Y)$. For every value of i and γ_{-i}, we set the tempered conditional distribution $g_i(\gamma_i|\gamma_{-i})$ to be the uniform distribution over $\{0,1\}$. It is easy to check that the supremum b that is defined in expression (6) is upper bounded by 2 and thus we have theoretical guarantees on the robustness of TGS from proposition 2 and theorem 1.

Since the target state space is discrete, it is more efficient to replace the Gibbs step of updating γ_i conditionally on i and γ_{-i}, with its Metropolized version (see for example Liu (1996)). The resulting specific instance of TGS is the following algorithm.

TGS for BVS algorithm: at each iteration of the Markov chain do

(a) sample i from $\{1, \ldots, p\}$ proportionally to $p_i(\gamma) = 1/\{2p(\gamma_i|\gamma_{-i}, Y)\}$;
(b) switch γ_i to $1 - \gamma_i$;
(c) weight the new state γ with a weight $Z(\gamma)^{-1}$ where $Z(\gamma) = (1/p)\Sigma_{i=1}^p p_i(\gamma)$.

In step (a), $p(\gamma_i|\gamma_{-i}, Y)$ denotes the probability that γ_i takes its current value conditionally on the current value of γ_{-i} and on the observed data Y. In the remainder of Section 4, the expression TGS will refer to this specific implementation of the generic scheme that was described in Section 2, and P_{TGS} to the Markov transition kernel of the resulting discrete time chain $(\gamma^{(t)})_{t=1}^\infty$.

4.3. Weighted tempered Gibbs sampling for Bayesian variable selection

As discussed in Section 6, TGS updates each co-ordinate with the same frequency. In a BVS context, however, this may be inefficient as the resulting sampler would spend most iterations updating variables that have low or negligible posterior inclusion probability, especially when p becomes large. A better solution would be to update more often components with a larger inclusion probability, thus having a more focused computational effort. In the WTGS framework of Section 3.6, this can be obtained by using non-uniform weight functions $\eta_i(\gamma_{-i})$. For example, proposition 6 implies that choosing $\eta_i(\gamma_{-i}) = p(\gamma_i = 1|\gamma_{-i}, Y)$ leads to a frequency of updating of the ith component equal to $\zeta^{-1}\mathbb{E}[\eta_i(\gamma_{-i})] = s^{-1} p(\gamma_i = 1|Y)$, where $s = \Sigma_{j=1}^p p(\gamma_j = 1|Y)$ is the expected number of active variables *a posteriori*. Here $p(\gamma_i = 1|Y)$ denotes the (marginal)

posterior probability that $\gamma_i = 1$, whereas $p(\gamma_i = 1|\gamma_{-i}, Y)$ denotes the probability of the same event conditional on both the observed data Y and the current value of γ_{-i}. With WTGS we can obtain a frequency of updating of the ith component proportional to $p(\gamma_i = 1|Y)$ without knowing the actual value of $p(\gamma_i = 1|Y)$, but rather using only the conditional expressions $p(\gamma_i = 1|\gamma_{-i}, Y)$.

The optimal choice of frequency of updating is related to an exploration *versus* exploitation trade-off. For example, choosing a uniform frequency of updating favours exploration, as it forces the sampler to explore new regions of the space by flipping variables with low conditional inclusion probability. In contrast, choosing a frequency of updating that focuses on variables with high conditional inclusion probability favours exploitation, as it allows the sampler to focus on the most important region of the state space. For this reason, we use a compromise between the choice of $\eta_i(\gamma_{-i})$ that was described above and the uniform TGS, obtained by setting $\eta_i(\gamma_{-i}) = p(\gamma_i = 1|\gamma_{-i}, Y) + k/p$ with k being a fixed parameter (in our simulations we used $k = 5$). Such a choice leads to frequencies of updating given by a mixture of the uniform distribution over $\{1, \ldots, p\}$ and the distribution proportional to $p(\gamma_i = 1|Y)$. More precisely we have

$$\zeta^{-1}\mathbb{E}[\eta_i(\gamma_{-i})] = \alpha\frac{p(\gamma_i = 1|Y)}{s} + (1-\alpha)\frac{1}{p},$$

where $\alpha = s/(k+s)$. The resulting scheme is as follows (see above for the definition of $p(\gamma_i = 1|\gamma_{-i}, Y)$).

WTGS algorithm for BVS: at each iteration of the Markov chain do

(a) sample i from $\{1, \ldots, p\}$ proportionally to

$$p_i(\gamma) = \frac{p(\gamma_i = 1|\gamma_{-i}, Y) + k/p}{2p(\gamma_i|\gamma_{-i}, Y)};$$

(b) switch γ_i to $1 - \gamma_i$;
(c) weight the new state γ with a weight $Z(\gamma)^{-1}$ where $Z(\gamma) \propto \Sigma_{i=1}^{p} p_i(\gamma)$.

In the remainder of Section 4, the expression WTGS will refer to this specific implementation of the generic scheme that was described in Section 3.6, and P_{WTGS} to the Markov transition kernel of the resulting discrete time Markov chain $(\gamma^{(t)})_{t=1}^{\infty}$.

4.4. Efficient implementation and Rao–Blackwellization

Compared with GS, TGS and WTGS provide substantially improved convergence properties at the price of an increased computational cost per iteration. The additional cost is computing $\{p(\gamma_i|Y, \gamma_{-i})\}_{i=1}^{p}$ given $\gamma \in \{0, 1\}^p$, which can be done efficiently through vectorized operations as described in section B.1 of the on-line supplement. Such efficient implementation is crucial to the successful application of these TGS schemes. The resulting cost per iteration of TGS and WTGS is of order $\mathcal{O}(np + |\gamma|p)$. For comparison, the cost per iteration of GS is $\mathcal{O}(n|\gamma| + |\gamma|^2)$. If $X^{\mathrm{T}}X$ has been precomputed before running the MCMC algorithm then the costs per iteration become $\mathcal{O}(|\gamma|p)$ for TGS and $\mathcal{O}(|\gamma|^2)$ for GS. In both cases, the relative additional cost of TGS over GS is $\mathcal{O}(p/|\gamma|)$. See section B.2 of the supplement for derivations of these expressions.

Interestingly, $\{p(\gamma_i|Y, \gamma_{-i})\}_{i=1}^{p}$ are the same quantity that is needed to compute Rao–Blackwellized estimators of the marginal posterior inclusion probabilities (PIPs) $\{p(\gamma_i = 1|Y)\}_{i=1}^{p}$. Therefore, using TGS enables us to implement Rao–Blackwellized estimators of PIPs (for all $i \in \{1, \ldots, p\}$ at each flip) without extra cost. See section B.3 of the on-line supplement for more details.

4.5. Computational complexity results for simple Bayesian variable-selection scenarios

In this section we provide quantitative results on the computational complexity of GS, TGS and WTGS in some simple BVS scenarios. In particular, we consider two extreme cases: one where all regressors in the design matrix X are orthogonal to each other (Section 4.5.2), and one where some of the regressors are perfectly collinear (Section 4.5.3). In the first case the posterior distribution $p(\gamma|Y)$ features independent components and thus it is the ideal case for GS, whereas the second case features some maximally correlated components and thus it is a worst-case scenario for GS. Our results show that the computational complexity of TGS and WTGS is not impacted by the change in correlation structure between the two scenarios. This is coherent with the conjecture of Section 3.5 that the convergence of TGS and WTGS is not slowed down by pairwise and/or negative correlation. In fact, a block of collinear regressors in the design matrix X induces a corresponding block of *negatively* correlated inclusion variables in $p(\gamma|Y)$. See section B.4 of the on-line supplement for a quantitative example. More generally, strong correlation between regressors induces strong *negative* correlation between the corresponding inclusion variables in $p(\gamma|Y)$. Intuitively, strongly correlated regressors provide the same type of information regarding Y. Thus, conditionally on the ith regressor being included in the model, the regressors that are strongly correlated with the ith regressor are not required to explain the data further and thus have a low probability of being included. Such a behaviour holds regardless of whether the original correlation between regressors is positive or negative.

As a preliminary step for the results in Sections 4.5.2 and 4.5.3, we now discuss the definition of computational complexity that we shall use.

4.5.1. Computational complexity for Markov chain Monte Carlo sampling and importance tempering

In classical contexts, we can define the computational complexity of an MCMC algorithm as the product between the cost per iteration and the number of iterations that are required to obtain Monte Carlo estimators with effective sample size of order 1. One way to define such a number of iterations is the so-called relaxation time, which is defined as the inverse of the spectral gap that is associated with the Markov kernel under consideration (for instance the second-largest eigenvalue in the case where the Markov kernel has a purely discrete spectrum). Such a definition is motivated by the fact that the asymptotic variances that are associated with an f-reversible Markov kernel P satisfy

$$\mathrm{var}(h, P) \leqslant \frac{2\,\mathrm{var}_f(h)}{\mathrm{gap}(P)} \qquad h \in L^2(\mathcal{X}, f), \tag{14}$$

where $\mathrm{gap}(P)$ is the spectral gap of P (Rosenthal (2003), proposition 1). Note that here $\mathrm{gap}(P)$ refers to the spectral gap of P and not the *absolute* spectral gap; see Rosenthal (2003) for more discussion. In what follows we denote the relaxation time of GS as $t_{\mathrm{GS}} = \mathrm{gap}(P_{\mathrm{GS}})^{-1}$. By condition (14), we can interpret $2t_{\mathrm{GS}}$ as the number of GS iterations that are required to have effective sample size equal to 1.

In contrast, TGS asymptotic variances include also an importance sampling contribution; see equation (7). Thus the direct analogue of condition (14), i.e. $\mathrm{var}(h, \mathrm{TGS}) \leqslant 2\mathrm{gap}(P_{\mathrm{TGS}})^{-1}\mathrm{var}_f(h)$, does not hold anymore and defining the TGS relaxation time as $\mathrm{gap}(P_{\mathrm{TGS}})^{-1}$ would be inappropriate. As shown by the following lemma, the problem can be circumvented by using the spectral gap of a continuous time version of TGS. To simplify the lemma's proof and notation, we assume that $|\mathcal{X}| < \infty$, which always holds in the BVS context. We expect an analogous result to hold in the context of general state spaces $\mathcal{X}$.

Lemma 2. Let $|\mathcal{X}| < \infty$. Define the jump matrix Q_{TGS} on $\mathcal{X}$ as $Q_{\mathrm{TGS}}(\gamma, \gamma') = Z(\gamma) P_{\mathrm{TGS}}(\gamma, \gamma')$ for all $\gamma' \neq \gamma$ and $Q_{\mathrm{TGS}}(\gamma, \gamma) = -\Sigma_{\gamma' \neq \gamma} Q_{\mathrm{TGS}}(\gamma, \gamma')$. Then

$$\mathrm{var}(h, \mathrm{TGS}) \leqslant \frac{2\mathrm{var}_f(h)}{\mathrm{gap}(Q_{\mathrm{TGS}})} \qquad h : \mathcal{X} \to \mathbb{R}, \tag{15}$$

where $\mathrm{gap}(Q_{\mathrm{TGS}})$ is the smallest non-zero eigenvalue of $-Q_{\mathrm{TGS}}$.

Lemma 2 implies that $\mathrm{gap}(Q_{\mathrm{TGS}})$ implicitly incorporates both the importance sampling and the auto-correlation terms in $\mathrm{var}(h, \mathrm{TGS})$. Motivated by condition (15), we define the relaxation time of TGS as $t_{\mathrm{TGS}} = \mathrm{gap}(Q_{\mathrm{TGS}})^{-1}$. By lemma 2, we can still interpret $2t_{\mathrm{TGS}}$ as the number of TGS iterations that are required to have effective sample size equal to 1. Similarly, we define the relaxation time of WTGS as the inverse spectral gap of its continuous time version (see section A.5 in the on-line supplement).

It can be shown that in cases where the importance tempering procedure coincides with classical MCMC sampling (i.e. when $Z(\gamma) = 1$) the two definitions of relaxation times discussed above coincide.

4.5.2. *Diagonal $X^{\mathrm{T}}X$*

Consider the case where all regressors are orthogonal to each other, i.e. $X^{\mathrm{T}}X$ is diagonal. This requires that $n \geqslant p$. The resulting posterior distribution for the inclusion variables $\gamma = (\gamma_1, \ldots, \gamma_p)$ is a collection of independent Bernoulli random variables. Denoting by q_i the PIP of the ith regressor, the posterior distribution of interest $f(\gamma) = p(\gamma|Y)$ has the form

$$f(\gamma) = \prod_{i=1}^{p} q_i^{\gamma_i} (1 - q_i)^{1-\gamma_i}. \tag{16}$$

Sampling from a target with independent components as in distribution (16) is the ideal scenario for GS, and we are interested in understanding how suboptimal TGS and WTGS are compared with GS in this context. The following theorem provides expressions for the relaxation times of GS, TGS and WTGS.

Theorem 2. Under distribution (16), the relaxation times of GS, TGS and WTGS satisfy

$$\left.\begin{array}{c} t_{\mathrm{GS}} = \alpha_1 p, \\ t_{\mathrm{TGS}} = \alpha_2 p, \\ t_{\mathrm{WTGS}} = s(1 - q_{\min}), \end{array}\right\} \tag{17}$$

where $\alpha_1 = \max\{q_{\max}, 1 - q_{\min}\}$, $\alpha_2 = \max_{i \in \{1,\ldots,p\}} q_i(1 - q_i)$, $q_{\max} = \max_{i \in \{1,\ldots,p\}} q_i$, $q_{\min} = \min_{i \in \{1,\ldots,p\}} q_i$ and $s = \Sigma_{i=1}^{p} q_i$.

Theorem 2 implies that t_{GS} and t_{TGS} are proportional to the total number of variables p, whereas t_{WTGS} depends only on the expected number of active variables $s = \Sigma_{i=1}^{p} q_i$, which is often much smaller than p. Assuming that α_2 and $q_{\min}$ are bounded away from respectively 0 and 1 as $p \to \infty$, the results in expression (17) imply that both GS and WTGS have $\mathcal{O}(pns)$ computational complexity, whereas the complexity of TGS is $\mathcal{O}(p^2 n)$. If $X^{\mathrm{T}}X$ is precomputed before the MCMC run (see Section 4.4), the complexities are reduced to $\mathcal{O}(ps^2)$ for GS and WTGS and to $\mathcal{O}(p^2 s)$ for TGS. It follows that, even in the case of independent components, WTGS has the same theoretical cost of GS. In contrast, TGS is suboptimal by an $\mathcal{O}(p/s)$ factor.

Remark 6. The analysis above ignores Rao–Blackwellization, which can be favourable to TGS and WTGS. In fact, when $X^{\mathrm{T}}X$ is diagonal the Rao–Blackwellized PIP estimators of TGS

and WTGS are deterministic and return the p exact PIPs in one iteration with cost $\mathcal{O}(np)$. By contrast, GS has an $\mathcal{O}(nps)$ cost for each IID sample.

4.5.3. *Fully collinear case*

We now consider the other extreme case, where there are maximally correlated regressors. In particular, suppose that m out of the p available regressors are perfectly collinear among themselves and with the data vector (i.e. each regressor fully explains the data), whereas the other $p-m$ regressors are orthogonal to the first m regressors. For simplicity, assume that $\Sigma_\gamma = c(X_\gamma^{\mathrm{T}} X_\gamma)^{-1}$ and $h \in (0,1)$ fixed. The $X^{\mathrm{T}}X$-matrix resulting from the scenario described above is not full rank. In such contexts, the standard definition of g-priors, $\Sigma_\gamma = c(X_\gamma^{\mathrm{T}} X_\gamma)^{-1}$, is not directly applicable and needs to be replaced by the more general definition involving generalized inverses (details are in section B.4 of the on-line supplement).

The posterior distribution of interest, $f(\gamma) = p(\gamma|Y)$, has the structure

$$f(\gamma) = f_0(\gamma_1, \ldots, \gamma_m) \prod_{i=m+1}^{p} q_i^{\gamma_i} (1-q_i)^{1-\gamma_i}, \tag{18}$$

where $q_i \in (0,1)$ is the posterior inclusion probability of the ith variable for $i = m+1, \ldots, p$ and f_0 denotes the joint distribution of the first m variables. By construction, the distribution f_0 is symmetric, meaning that $f_0(\gamma_1, \ldots, \gamma_m) = q\, \Sigma_{i=1}^m \gamma_i$ for some $q : \{0, \ldots, m\} \to [0,1]$. See section B.4 of the on-line supplement for the specific form of q. Under mild assumptions, we have $q(s)/q(1) \to 0$ as $p \to \infty$ for all $s \neq 1$, meaning that the distribution f_0 concentrates on the configurations having one and only one active regressor as p increases.

We study the asymptotic regime where m is fixed and $p \to \infty$. This corresponds to the commonly encountered scenario of having a small number of 'true' variables and a large number of noise variables. This asymptotic regime has been the focus of much of the recent BVS literature (Johnson and Rossell, 2012) and is motivated, for example, by applications to genomics (see the examples in Section 5.3). In our analysis, the number of data points n, as well as the hyperparameters c and h, can depend on p in an arbitrary manner, provided that the following technical assumption is satisfied.

Assumption 1. $\lim_{p\to\infty} h(1+c)^{-1/2} = 0$, $\limsup_{p\to\infty} h < 1$ and $\liminf_{p\to\infty} h^2(1+c)^{(n-2)/2} > 0$.

Assumption 1 is weak and satisfied in nearly any realistic scenario. For example, it is satisfied whenever $c \geqslant 1$ and $h \to 0$ at a slower than exponential rate in p. Note that the assumptions on $X^{\mathrm{T}}X$ impose the constraint $n \geqslant p - m + 1$.

The following theorem characterizes the behaviour of the relaxation times of GS, TGS and WTGS as p increases.

Theorem 3. As $p \to \infty$, the relaxation times of GS, TGS and WTGS satisfy

$$\left.\begin{array}{c} t_{\mathrm{GS}} \geqslant \mathcal{O}(c^{1/2} h^{-1} p), \\ t_{\mathrm{TGS}} \geqslant \mathcal{O}(p), \\ t_{\mathrm{WTGS}} = \mathcal{O}(s), \end{array}\right\} \tag{19}$$

where $s = \mathbb{E}_f[|\gamma|]$ is the expected number of active variables *a posteriori*.

Theorem 3 implies that WTGS has $\mathcal{O}(pns)$ computational complexity, whereas TGS has complexity at least $\mathcal{O}(p^2 n)$. We conjecture that $t_{\mathrm{TGS}} \leqslant \mathcal{O}(p)$ and we discuss a proof strategy in remark 7 of the on-line supplementary material. If such a conjecture is correct, then TGS

has complexity exactly $\mathcal{O}(p^2 n)$. In contrast expression (19) implies that the computational complexity of GS is at least $\mathcal{O}(pnsc^{1/2}h^{-1})$, whose asymptotic behaviour depends on the choices of c and h. In general, WTGS provides an improvement over GS of at least $\mathcal{O}(c^{1/2}h^{-1})$. If $h = \mathcal{O}(p^{-1})$ and $c = n$ such an improvement is at least $\mathcal{O}(pn^{1/2})$, whereas if $c = p^2$ it is at least $\mathcal{O}(p^2)$.

Theorems 2 and 3 suggest that the relaxation times of TGS and WTGS are not significantly impacted by changes in correlation structure between equations (16) and (18). As discussed in Section 4.5.1, this supports the conjectures of Section 3.5.

5. Simulation studies

In this section we provide simulation studies illustrating the performances of GS, TGS and WTGS in the BVS context that was described in Section 4.

5.1. Illustrative example

The differences between GS, TGS and WTGS can be well illustrated considering a scenario where two regressors with good explanatory power are strongly correlated.

In such a situation, models including one of the two variables will have high posterior probability, whereas models including both variables or none of the two will have a low posterior probability. As a result, the Gibbs sampler will become stuck in one of the two local modes corresponding to one variable being active and the other inactive.

Fig. 4 considers simulated data with $n = 100$ and $p = 100$, where the two correlated variables are number 1 and 2. The detailed simulation set-up is described in Section 5.2 (namely scenario 1 with signal-to-noise ratio SNR $= 3$). All chains were started from the empty model ($\gamma_i = 0$ for every i). TGS and WTGS, which have a roughly equivalent cost per iteration, were run for 30000 iterations, after a burn-in of 5000 iterations. GS was run for the same central processor unit time, performing multiple moves per iteration so that the cost per iteration matched the cost of TGS and WTGS. Figs 4(a) and 4(b) display the trace plots of the estimates for the PIP of variables 1 and 2 for GS, TGS and WTGS. The true PIP values are indicated with grey horizontal lines. Such values are accurate approximations to the exact PIP obtained by running an extremely long run of WTGS. For this illustration, it is reasonable to treat these values as exact as the associated Monte Carlo error is orders of magnitude smaller than the other Monte Carlo errors that are involved in the simulation. In the run displayed, GS became stuck in the mode corresponding to $(\gamma_1, \gamma_2) = (1, 0)$ and never flipped variable 1 or 2. In contrast, both TGS and WTGS manage

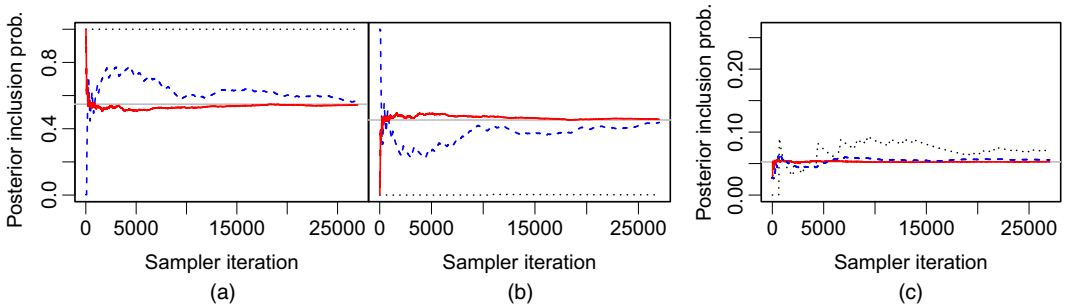

Fig. 4. Running estimates of PIPs for variables (a) 1, (b) 2 and (c) 3 produced by GS (· · · · · ·), TGS (– – –) and WTGS (———): here $p = n = 100$; thinning is used so that all schemes have the same cost per iteration (———, accurate approximations to the true values of the PIPs)

to move frequently between the two modes and indeed the resulting estimates of PIPs for both variables appear to converge to the correct value, with WTGS converging significantly faster. It is also interesting to compare the schemes' efficiency in estimating PIP for variables with lower but still non-negligible inclusion probability. For example variable 3 in these simulated data has a PIP of roughly 0.05. In this case the variable is rarely included in the model and the frequency-based estimators have a high variability, whereas the Rao–Blackwellized estimators produce nearly instantaneous good estimates; see Fig. 4(c).

Consider then an analogous simulated data set with $p = 1000$ and $n = 500$. In this case the larger number of regressors induces a more significant difference between TGS and WTGS as the latter focuses the computational effort on more important variables. In fact, as shown in Fig. 5, both TGS and WTGS manage to move across the $(\gamma_1, \gamma_2) = (0, 1)$ and $(\gamma_1, \gamma_2) = (1, 0)$ modes but WTGS does it much more often and produces estimates converging dramatically faster to the correct values. This is well explained by proposition 6, which implies that TGS flips each variable every $1/p$ iterations on average, whereas WTGS has frequency of flipping equal to $\zeta^{-1}\mathbb{E}[\eta_i(\gamma_{-i})]$ defined in Section 4.3, which is a function of $p(\gamma_j = 1|Y)$. The faster mixing of WTGS for the most influential variables accelerates also the estimation of lower but non-negligible PIPs, such as co-ordinates 3 and 600 in Figs 4 and 5 respectively.

To summarize, the main improvements of TGS and WTGS are due to

(a) tempering reducing correlation and helping to move across modes (see Figs 4(a) and 4(b)),
(b) Rao–Blackwellization producing more stable estimators (see Figs 4(c) and 5(c)) and
(c) the weighting mechanism of WTGS allowing us to focus computation on relevant variables (see Figs 5(a) and 5(b)).

The qualitative conclusions of this illustrative example would not change if we consider a scenario involving m strongly correlated variables, with $m > 2$.

5.2. Simulated data

In this section we provide a quantitative comparison between GS, TGS and WTGS under various simulated scenarios. Data are generated as $Y \sim N(X\beta^*, \sigma^2)$ with $\sigma^2 = 1$,

$$\beta^* = \mathrm{SNR}\sqrt{\left\{\frac{\sigma^2 \log(p)}{n}\right\}}\beta_0^*,$$

and each row $(X_{i1}, \ldots, X_{ip})$ of the design matrix X independently simulated from a multivariate normal distribution with zero mean and covariance $\Sigma^{(X)}$ having $\Sigma^{(X)}_{jj} = 1$ for all j. We set the prior probability h to $5/p$, corresponding to a prior expected number of active regressors equal

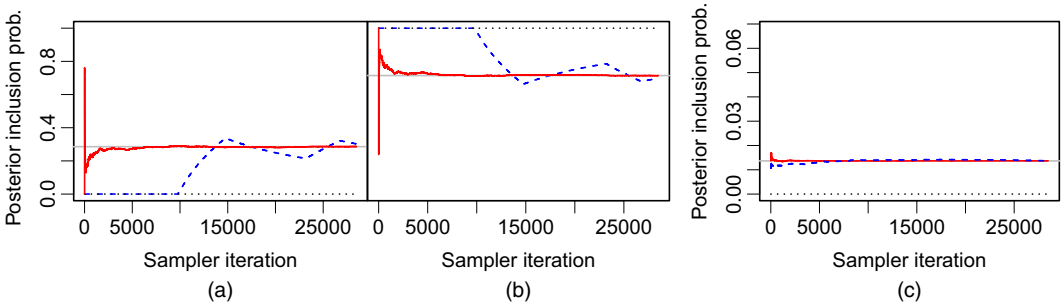

Fig. 5. Analogous to Fig. 4 with $p = 1000$ and $n = 500$: (a) variable 1; (b) variable 2; (c) variable 600

to 5. The values of β_0^* and $\Sigma_{ij}^{(X)}$ for $i \neq j$ vary depending on the scenario considered. In particular, we consider the following situations:

(a) *two strongly correlated variables*, $\beta_0^* = (1, 0, \ldots, 0)$, $\Sigma_{12}^{(X)} = \Sigma_{21}^{(X)} = 0.99$ and $\Sigma_{ij}^{(X)} = 0$ otherwise;
(b) *batches of correlated variables*, $\beta_0^* = (3, 3, -2, 3, 3, -2, 0, \ldots, 0)$, $\Sigma_{ij}^{(X)} = 0.9$ if $i, j \in \{1, 2, 3\}$ or $i, j \in \{4, 5, 6\}$ and $\Sigma_{ij}^{(X)} = 0$ otherwise;
(c) *uncorrelated variables*, $\beta_0^* = (2, -3, 2, 2, -3, 3, -2, 3, -2, 3, 0, \ldots, 0)$ and $\Sigma_{ij}^{(X)} = 0$ for all $i \neq j$.

Scenarios analogous to these have been previously considered in the literature. For example, Titsias and Yau (2017), section 3.2.3, considered a scenario similar to scenario (a), Wang *et al.* (2011), example 4, and Huang *et al.* (2016), section 4.2, a scenario similar to scenario (b) and Yang *et al.* (2016) a scenario analogous to scenario (c). We compare GS, TGS and WTGS on all three scenarios for a variety of values of n, p and SNR. To have a fair comparison, we implement the Metropolized version of GS, like we did for TGS and WTGS. To provide a quantitative comparison we consider a standard measure of relative efficiency: the ratio of the estimators' effective sample sizes over computational times. More precisely, we define the relative efficiency of TGS over GS as

$$\frac{\mathrm{Eff}_{\mathrm{TGS}}}{\mathrm{Eff}_{\mathrm{GS}}} = \frac{\mathrm{ess}_{\mathrm{TGS}}/T_{\mathrm{TGS}}}{\mathrm{ess}_{\mathrm{GS}}/T_{\mathrm{GS}}} = \frac{\sigma_{\mathrm{GS}}^2 T_{\mathrm{GS}}}{\sigma_{\mathrm{TGS}}^2 T_{\mathrm{TGS}}}, \tag{20}$$

where σ_{GS}^2 and σ_{TGS}^2 are the variances of the Monte Carlo estimators produced by GS and TGS respectively, whereas T_{GS} and T_{TGS} are the central processor unit time that is required to produce such estimators. An analogous measure is used for the relative efficiency of WTGS over GS. For each simulated data set, we computed the relative efficiency defined by equation (20) for each PIP estimator, thus obtaining p-values: one for each variable. Table 1 reports the median of such p-values for each data set under consideration. The variances in equation (20), such as σ_{GS}^2 and σ_{TGS}^2, were estimated with the sample variances of the PIP estimates obtained with 50 runs of each algorithm. See section B.5 of the on-line supplement for more details.

From Table 1 it can be seen that both TGS and WTGS provide orders of magnitude improvement in efficiency compared with GS, with median improvement of TGS over GS ranging from 1.7×10^3 to 2.1×10^7 and of WTGS over GS ranging from 8.0×10^3 to 1.1×10^8. Such a huge improvement, however, needs to be interpreted carefully. In fact, in all the simulated data sets the fraction of variables having non-negligible PIP is small (as is typical in large p BVS applications) and thus the median improvement refers to the efficiency in estimating a variable with very small PIP, e.g. below 0.001. When estimating such small probabilities, standard Monte Carlo estimators perform poorly compared with Rao–Blackwellized versions (see Figs 4(c) and 5(c)) and this explains such a huge improvement of TGS and WTGS over GS. In many practical scenarios, however, we are not interested in estimating the actual value of such a small PIP. Thus a more informative comparison can be obtained by restricting our attention to variables with moderately large PIP. Table 2 reports the mean relative efficiency for variables whose PIP is estimated to be larger than 0.05 by at least one of the algorithms under consideration. Empty values correspond to cells where either no PIP was estimated above 0.05 or where GS never flipped such a variable and thus we had no natural (finite) estimate of the variance in equation (20). In both such cases we expect the improvement in relative efficiency over GS to be extremely large (either corresponding to the values in Table 1, first case, or currently estimated at ∞, second case) and thus excluding those values from Table 2 is conservative and plays in favour of

Table 1. Median improvement over variables of TGS and WTGS relative to GS for simulated data†

(p,n)	*TGS versus GS for the following values of SNR:*				*WTGS versus GS for the following values of SNR:*			
	0.5	*1*	*2*	*3*	*0.5*	*1*	*2*	*3*
Scenario (a)								
(100,50)	4.0×10^5	2.4×10^4	2.0×10^4	6.6×10^4	2.1×10^6	2.6×10^5	3.4×10^5	1.9×10^5
(200,200)	1.0×10^6	4.2×10^6	4.9×10^5	2.1×10^6	1.6×10^7	5.3×10^7	1.0×10^7	2.4×10^7
(1000,500)	1.3×10^6	1.2×10^6	1.1×10^6	2.2×10^6	7.8×10^7	9.3×10^7	6.5×10^7	1.1×10^8
Scenario (b)								
(100,50)	1.0×10^4	2.9×10^3	1.7×10^3	3.9×10^4	1.5×10^5	4.1×10^4	9.3×10^3	1.6×10^5
(200,200)	1.1×10^5	1.0×10^5	8.2×10^3	1.4×10^7	1.8×10^6	2.8×10^6	1.5×10^5	3.2×10^6
(1000,500)	4.6×10^5	9.2×10^4	6.7×10^5	2.1×10^6	3.3×10^7	1.1×10^7	1.1×10^7	1.5×10^7
Scenario (c)								
(100,50)	2.5×10^3	4.2×10^3	7.7×10^3	7.4×10^4	2.9×10^4	3.9×10^4	8.0×10^3	1.5×10^4
(200,200)	9.1×10^4	4.3×10^4	2.8×10^7	3.5×10^6	1.0×10^6	3.1×10^5	2.9×10^6	8.0×10^5
(1000,500)	9.8×10^4	5.9×10^5	1.1×10^7	2.1×10^7	7.0×10^6	4.4×10^6	7.6×10^6	1.0×10^7

†Scenarios (a)–(c) are described in Section 5.2.

GS. The mean improvements that are reported in Table 2 are significantly smaller than that in Table 1 but still potentially very large, with ranges of improvement being $(1.4, 2.5\times10^6)$ for TGS and $(1.8\times10^1, 1.9\times10^4)$ for WTGS. There is no value below 1, meaning that in these simulations TGS or WTGS is always more efficient than GS, and that WTGS is more efficient than TGS in most scenarios. Also, especially for WTGS, the improvement over GS grows larger as p increases.

The value of c in the prior covariance matrix has a large effect on the concentration of the posterior distribution and thus on the resulting difficulty of the computational task. Different suggestions for the choice of c have been proposed in the literature, such as $c=n$ (Zellner, 1986), $c=\max\{n, p^2\}$ (Fernandez *et al.*, 2001) or a fixed value between 10 and 10^4 (Smith and Kohn, 1996). For the simulations that are reported in Tables 1 and 2 we set $c=10^3$, which provided results that are fairly representative in terms of relative efficiency of the algorithms that were considered. In Section 5.3 we shall consider both $c=n$ and $c=\max\{n, p^2\}$.

5.3. Real data

In this section we consider three real data sets with increasing number of covariates. We compare WTGS with GS and the Hamming ball sampler, which is a recently proposed sampling scheme designed for posterior distributions over discrete spaces, including BVS models (Titsias and Yau, 2017). We refer to the data sets as DLD data, TGFB172 data and TGFB data. The DLD data come from a genomic study by Yuan *et al.* (2016) based on ribonucleic acid sequencing and have a moderate number of regressors, $p=57$ and $n=192$. The version of the data set that we used is freely available from the supplementary material of Rossell and Rubio (2018). See section 6.5 therein for a short description of the data set and the inferential questions of interest. The second and third data sets are human microarray gene expression data in colon cancer patients from Calon *et al.* (2012). The TGFB172 data, which have $p=172$ and $n=262$, are obtained as a subset of the TGFB data, for which $p=10172$ and $n=262$. These two data

Table 2. Mean improvement of TGS and WTGS relative to GS over variables with PIP > 0.05†

(p,n)	*TGS versus GS for the following values of SNR:*				*WTGS versus GS for the following values of SNR:*			
	0.5	*1*	*2*	*3*	*0.5*	*1*	*2*	*3*
Scenario (a)								
(100,50)		7.2×10	1.8×10	2.8×10^2		5.8×10^2	4.2×10^2	3.1×10^3
(200,200)	4.9×10^3		6.6×10	1.9×10^2	1.1×10^4		1.8×10^3	1.6×10^4
(1000,500)	2.7×10^2	6.3×10^2	1.4	8.1×10	8.8×10^3	2.5×10^4	5.8×10^2	1.9×10^4
Scenario (b)								
(100,50)	4.8	1.4×10	3.3	2.0×10	1.3×10^2	2.4×10^2	1.8×10	1.4×10^2
(200,200)	8.6×10	4.7×10	3.4	2.5×10^6	2.3×10^3	2.1×10^3	6.0×10	4.1×10^2
(1000,500)	4.6×10	3.7×10	1.3×10	4.5×10^2	1.1×10^4	7.6×10^3	1.1×10^3	1.8×10^4
Scenario (c)								
(100,50)	2.7	5.3	9.2		2.5×10	6.7×10	2.1×10	
(200,200)	1.1×10^2	6.6×10			1.3×10^3	4.6×10^2		
(1000,500)	1.6×10	6.8×10^2			1.1×10^3	9.4×10^3		

†Same simulation set-ups as in Table 1. Empty values correspond to large values with no reliable estimate available (see Section 5.2 for discussion).

sets are described in section 5.3 of Rossell and Telesca (2017) and are freely available from the corresponding supplementary material.

If $X^{\mathrm{T}}X$ and $Y^{\mathrm{T}}X$ are precomputed, the cost per iteration of the algorithms under consideration is not sensitive to n (see Section 4.4 and section B.2 of the on-line supplement). Thus a data set with a large value of p, like the TGFB data, represents a computationally challenging scenario, regardless of having a low value of n. Moreover, low values of n have been reported to induce posterior distributions $p(\gamma|Y)$ that are less concentrated and more difficult to explore (Johnson (2013), sections 3–4). In this sense, small n–large p scenarios are among the most computationally challenging in the BVS scenario.

We performed 20 independent runs of each algorithm for each data set with both $c = n$ and $c = p^2$, recording the resulting estimates of PIPs. We ran WTGS for 500, 1000 and 30000 iterations for the DLD, TGFB172 and TGFB data sets respectively, discarding the first 10% of samples as burn-in. The number of iterations of GS and the Hamming ball sampler were chosen to have the same run time as WTGS. To assess the reliability of each algorithm, we compare results that were obtained over different runs by plotting each PIP estimate over the estimates that were obtained with different runs of the same algorithm. The results are displayed in Fig. 6. Points close to the diagonal indicate estimates in accordance with each other across runs, whereas points that are far from the diagonal indicate otherwise. It can be seen that WTGS provides substantially more reliable estimates for all combinations of data set and value of c under consideration and that the improvement in efficiency increases with the number of regressors p. Since each box in Fig. 6 contains a large number of PIP estimates (namely $p \times 20 \times 19$ points), we also provide the analogous figure that was obtained by running only two runs of each algorithm in section B.7 of the on-line supplement. The latter representation may be more familiar to the reader.

All computations reported in this section were performed on the same desktop computer with 16 Gbytes of random-access memory and an i7 Intel processor, using the R programming

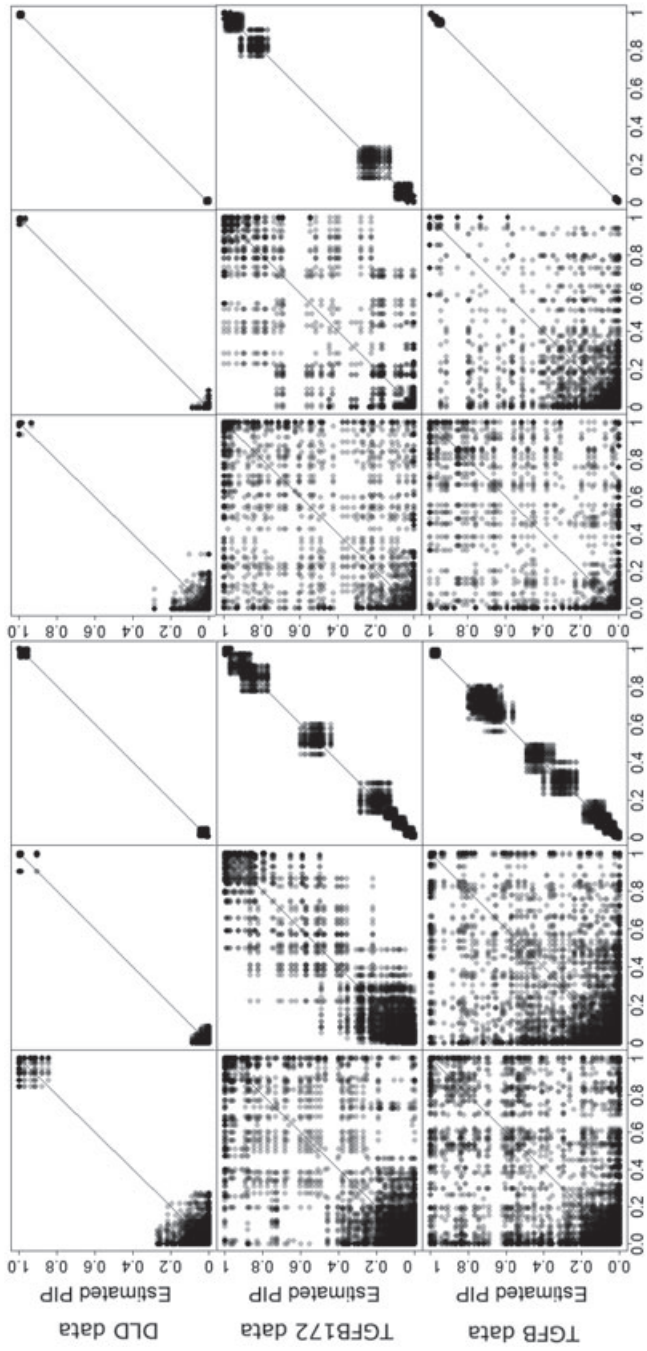

Fig. 6. Comparison of GS, the Hamming ball sampler and WTGS on three real data sets for $c = n$ and $c = p^2$ (points close to the diagonal lines indicate estimates agreeing across different runs): (a) GS, $c = n$; (b) HBS, $c = n$; (c) WTGS, $c = n$; (d) GS, p^2; (e) HBS, p^2; (f) WTGS, p^2

language (R Core Team, 2017). The R code to implement the various samplers under consideration is freely available from `https://github.com/gzanella/TGS`. For the largest data set under consideration ($p = 10172$) WTGS took an average of 115 s for each run shown in Fig. 6. We performed further experiments, to compare the WTGS performances with those of available R packages for BVS and some alternative methodology from the literature. The results, which are reported in section B.6 of the on-line supplement, suggest that WTGS provides state of the art performances for fitting spike-and-slab BVS models like those of Section 4.1.

6. Discussion

We have introduced a novel Gibbs sampler variant, demonstrating its considerable potential both in toy examples as well as more realistic BVS models, and giving underpinning theory to support the use of the method and to explain its impressive convergence properties.

TGS can be thought of as an intelligent random-scan Gibbs sampler, using current state information to inform the choice of component to be updated. In this way, the method is different from the usual random-scan method which can also have heterogeneous component updating probabilities which can be optimized (e.g. by adaptive MCMC methodology; see for example Chimisov *et al*. (2018)).

There are many potential extensions of TGS that we have not considered in this paper. For example, we could replace step (b) of TGS, where i is sampled proportionally to $p_i(\mathbf{x})$, with a Metropolized version as in Liu (1996), where the new value $i^{(t+1)}$ is proposed from $\{1, \ldots, d\} \backslash \{i^{(t)}\}$ proportionally to $p_{i^{(t+1)}}(\mathbf{x})$ for $i^{(t+1)} \neq i^{(t)}$. This would effectively reduce the probability of repeatedly updating the same co-ordinate in consecutive iterations, which, as shown in proposition 5, can be interpreted as a rejected move.

Another direction for further research might aim to reduce the cost per iteration of TGS when d is very large. For example, we could consider a 'blockwise' version of TGS, where first a subset of variables is selected at random and then TGS is applied only to such variables conditionally on the others, to avoid computing all the values of $\{p_i(\mathbf{x})\}_{i=1}^{d}$ at each iteration. The choice of the number of variables to select would then be related to a cost per iteration *versus* mixing trade-off. See section 6.4 of Zanella (2019) for a discussion of similar blockwise implementations. Also, computing $p_i(\mathbf{x})$ exactly may be infeasible in some contexts, and thus it would be interesting to design a version of TGS where the terms $p_i(\mathbf{x})$ are replaced by unbiased estimators while preserving the correct invariant distribution.

A further possibility for future research is to construct deterministic scan versions of TGS which may be of value for contexts where deterministic scan Gibbs samplers are known to outperform random-scan samplers (see for example Roberts and Rosenthal (2015)). Also, it would be useful to provide detailed methodological guidance regarding the choice of good modified conditionals $g(x_i|\mathbf{x}_{-i})$, e.g. good choices of the tempering level β, extending the preliminary results of Section 3.5.

One could design schemes where the conditional distributions of k co-ordinates are tempered at the same time, rather than a single co-ordinate. A natural approach would be to use the TGS interpretation of remark 2 and to define some extended target on $\mathcal{X} \times \{1, \ldots, d\}^k$. This would enable good mixing to be achieved in a larger class of target distributions (compared with those of Section 3.5) at the price of a larger cost per iteration.

TGS provides a generic way of mitigating the worst effects of dependence on Gibbs sampler convergence. Classical ways of reducing posterior correlations involve reparameterizations (Gelfand *et al*., 1995; Hills and Smith, 1992). Although these can work very well in some specific models (see for example Zanella and Roberts (2017) and Papaspiliopoulos *et al*. (2018)), the

generic implementations requires the ability to perform GS on generic linear transformations of the target, which is often not practical beyond the Gaussian case. For example it is not clear how to apply such methods to the BVS models of Section 4. Moreover reparameterization methods are not effective if the covariance structure of the target changes with location. Further alternative methodology to overcome strong correlations in GS include the recently proposed adaptive MCMC approach of Duan *et al*. (2018) in the context of data augmentation models.

Given the results of Sections 4 and 5, it would be interesting to explore the use of the methodology that is proposed in this paper for other BVS models, such as models with more elaborate priors (e.g. Johnson and Rossell (2012)) or binary response variables.

Acknowledgements

GZ was supported by the European Research Council through StG 'New directions in Bayesian nonparametrics' 306406. GOR acknowledges support from the Engineering and Physical Sciences Research Council through grants EP/K014463/1 ('Intractable likelihood') and EP/K034154/1 ('Enabling quantification of uncertainty for large-scale inverse problems'). GZ is affiliated to the Innocenzo Gasparini Institute for Economic Research and the Bocconi Institute for Data Science and Analytics at the Bocconi University, Milan.

References

Belloni, A. and Chernozhukov, V. (2009) On the computational complexity of MCMC-based estimators in large samples. *Ann. Statist.*, **37**, 2011–2055.

Calon, A., Espinet, E., Palomo-Ponce, S., Tauriello, D. V., Iglesias, M., Céspedes, M. V., Sevillano, M., Nadal, C., Jung, P., Zhang, X. H.-F., Byrom, D., Riera, A., Rossell, D., Mangues, R., Massagué, J., Sancho, E. and Batlle, E. (2012) Dependency of colorectal cancer on a TGF-β-driven program in stromal cells for metastasis initiation. *Cancer Cell*, **22**, 571–584.

Chimisov, C., Latuszynski, K. and Roberts, G. (2018) Adapting the Gibbs sampler. *Preprint arXiv:1801.09299*.

Chipman, H. A., George, E. I. and McCulloch, R. E. (2001) *The Practical Implementation of Bayesian Model Selection*, pp. 65–116. Beachwood: Institute of Mathematical Statistics.

Deligiannidis, G. and Lee, A. (2018) Which ergodic averages have finite asymptotic variance? *Ann. Appl. Probab.*, **28**, 2309–2334.

Duan, L. L., Johndrow, J. E. and Dunson, D. B. (2018) Scaling up data augmentation MCMC via calibration. *J. Mach. Learn. Res.*, **19**, 2575–2608.

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

Frieze, A., Kannan, R. and Polson, N. (1994) Sampling from log-concave distributions. *Ann. Appl. Probab.*, **4**, 812–837.

Gelfand, A. E., Sahu, S. K. and Carlin, B. P. (1995) Efficient parametrisations for normal linear mixed models. *Biometrika*, **82**, 479–488.

Geyer, C. J. and Thompson, E. A. (1995) Annealing Markov chain Monte Carlo with applications to ancestral inference. *J. Am. Statist. Ass.*, **90**, 909–920.

Gramacy, R., Samworth, R. and King, R. (2010) Importance tempering. *Statist. Comput.*, **20**, 1–7.

Hills, S. E. and Smith, A. F. (1992) Parameterization issues in Bayesian inference. *Baysn Statist.*, **4**, 227–246.

Huang, X., Wang, J. and Liang, F. (2016) A variational algorithm for Bayesian variable selection. *Preprint arXiv:1602.07640*.

Johnson, V. E. (2013) On numerical aspects of Bayesian model selection in high and ultrahigh-dimensional settings. *Baysn Anal.*, **8**, 741–758.

Johnson, V. E. and Rossell, D. (2012) Bayesian model selection in high-dimensional settings. *J. Am. Statist. Ass.*, **107**, 649–660.

Liu, J. S. (1996) Peskun's theorem and a modified discrete-state Gibbs sampler. *Biometrika*, **83**, 681–682.

Marinari, E. and Parisi, G. (1992) Simulated tempering: a new Monte Carlo scheme. *Eurphys. Lett.*, **19**, 451–458.

Meyn, S. P. and Tweedie, R. L. (1993) *Markov Chains and Stochastic Stability*. London: Springer.

Owen, A. B. (2013) Monte Carlo theory, methods and examples. Stanford University, Stanford. (Available from `http://statweb.stanford.edu/owen/mc/`.)

Papaspiliopoulos, O., Roberts, G. O. and Zanella, G. (2018) Scalable inference for crossed random effects models. *Preprint arXiv:1803.09460*.

R Core Team (2017) *R: a Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.
Roberts, G. O. and Rosenthal, J. S. (2001) Markov chains and de-initializing processes. *Scand. J. Statist.*, **28**, 489–504.
Roberts, G. O. and Rosenthal, J. S. (2015) Surprising convergence properties of some simple Gibbs samplers under various scans. *Int. J. Statist. Probab.*, **5**, 51–60.
Roberts, G. O. and Rosenthal, J. S. (2016) Complexity bounds for Markov chain Monte Carlo algorithms via diffusion limits. *J. Appl. Probab.*, **53**, 410–420.
Roberts, G. O. and Sahu, S. K. (1997) Updating schemes, correlation structure, blocking and parameterization for the Gibbs sampler. *J. R. Statist. Soc.* B, **59**, 291–317.
Roberts, G. O. and Smith, A. F. M. (1994) Simple conditions for the convergence of the Gibbs sampler and Metropolis-Hastings algorithms. *Stoch. Processes Appl.*, **49**, 207–216.
Rosenthal, J. S. (2003) Asymptotic variance and convergence rates of nearly-periodic Markov chain Monte Carlo algorithms. *J. Am. Statist. Ass.*, **98**, 169–177.
Rossell, D. and Rubio, F. J. (2018) Tractable Bayesian variable selection: beyond normality. *J. Am. Statist. Ass.*, **113**, 1742–1758.
Rossell, D. and Telesca, D. (2017) Nonlocal priors for high-dimensional estimation. *J. Am. Statist. Ass.*, **112**, 254–265.
Smith, A. F. and Gelfand, A. E. (1992) Bayesian statistics without tears: a sampling–resampling perspective. *Am. Statistn*, **46**, 84–88.
Smith, M. and Kohn, R. (1996) Nonparametric regression using Bayesian variable selection. *J. Econmetr.*, **75**, 317–343.
Titsias, M. K. and Yau, C. (2017) The Hamming ball sampler. *J. Am. Statist. Ass.*, **112**, 1598–1611.
Wang, S., Nan, B., Rosset, S. and Zhu, J. (2011) Random lasso. *Ann. Appl. Statist.*, **5**, 468–485.
Yang, Y., Wainwright, M. J. and Jordan, M. I. (2016) On the computational complexity of high-dimensional Bayesian variable selection. *Ann. Statist.*, **44**, 2497–2532.
Yuan, T., Huang, X., Woodcock, M., Du, M., Dittmar, R., Wang, Y., Tsai, S., Kohli, M., Boardman, L., Patel, T. and Wang, L. (2016) Plasma extracellular RNA profiles in healthy and cancer patients. *Scient. Rep.*, **6**, article 19413.
Zanella, G. (2019) Informed proposals for local MCMC in discrete spaces. *J. Am. Statist. Ass.*, to be published, doi 10.1080/01621459.2019.1585255.
Zanella, G. and Roberts, G. O. (2017) Analysis of the Gibbs Sampler for Gaussian hierarchical models via multigrid decomposition. *Preprint arXiv:1703.06098*.
Zellner, A. (1986) On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti* (eds P. Goel and A. Zellner), pp. 233–243. New York: Elsevier.

Supporting information

Additional 'supporting information' may be found in the on-line version of this article:

'Supplementary material for "Scalable importance tempering and Bayesian variable selection"'.