Scalable Importance Tempering and Bayesian Variable Selection

Giacomo Zanella[*] and Gareth O. Roberts[†]

May 3, 2018

Abstract

We propose a Monte Carlo algorithm to sample from high-dimensional probability distributions that combines Markov chain Monte Carlo (MCMC) and importance sampling. We provide a careful theoretical analysis, including guarantees on robustness to high-dimensionality, explicit comparison with standard MCMC 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 allows to perform fast and reliable fully Bayesian inferences with tens of thousands regressors.

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).

There have been many attempts to embed importance sampling within Monte Carlo schemes for Bayesian analysis, going back to Smith and Gelfand [1992] and beyond. However, except where Sequential Monte Carlo approaches can be adopted, pure Markov chain based schemes (i.e. ones 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 [Gramacy et al., 2010]. 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. Robustness to high-dimensionality is achieved by exploiting the fact that commonly used schemes, proceed by sequential lower-dimensional updates. In particular, we propose an importance sampling generalisation of the Gibbs Sampler, which we call Tempered Gibbs Sampling (TGS).

[*] Department of Decision Sciences, BIDSA and IGIER, Bocconi University, via Roentgen 1, 20136 Milan, Italy. giacomo.zanella@unibocconi.it

[†] Department of Statistics, University of Warwick, Coventry, CV4 7AL, UK. gareth.o.roberts@warwick.ac.uk

Through an appropriately designed tempering mechanism, TGS circumvents the main limitations of standard Gibbs Sampling (GS), such as the slow mixing induced by strong posterior correlations and the inability to escape local modes. It also avoids the requirement to visit all coordinates sequentially, instead iteratively making state-informed decisions as to which coordinate 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 coordinate that is currently being updated, and compensates for the overdispersion induced by the tempered update by choosing to update components which are in the tail of their conditional distributions more frequently. The resulting dynamics often 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 standard simulated tempering) and thus avoids the tuning issues related to choosing the number of levels and collection of temperatures, as well as the heavy computational burden induced by running one chain for each temperature level.

We apply the novel sampling scheme to Bayesian Variable selection problems, observing multiple orders of magnitude improvements compared to alternative Monte Carlo schemes. For example, TGS allows to perform reliable, fully Bayesian inference for spike and slab models with over ten thousand regressors in less than two minutes using a simple R implementation and a single desktop computer.

The paper structure is as follows. The TGS scheme is introduced in Section 2. There we provide basic validity results and intuition on the potential improvement given by the the novel scheme, together with an illustrative example. In Section 3 we develop a careful analysis of the proposed scheme. First we show that, unlike common tempering schemes, TGS is robust to high-dimensionality of the target as the coordinate-wise tempering mechanism 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 ones where GS and TGS will perform similarly. In Section 4 we provide a detailed application to Bayesian Variable selection problems. We review our findings in Section 5 with the proofs to our main results being in the subsequent appendix.

2 The Tempered Gibbs Sampling scheme

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

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

Algorithm TGS. *At each iteration of the Markov chain do:*

1. (Coordinate selection) *Sample i from $\{1, \dots, d\}$ proportionally to $p_i(\boldsymbol{x})$.*

2. (Tempered update) *Sample $x_i \sim g(x_i|\boldsymbol{x}_{-i})$.*

3. (Importance weighting) *Assign to the new state $\boldsymbol{x}$ a weight $w(\boldsymbol{x}) = Z(\boldsymbol{x})^{-1}$.*

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

$$Z(\boldsymbol{x}) \text{ is bounded below, and bounded above on compact sets.} \tag{2}$$

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

$$\hat{h}_n^{TGS} = \frac{\sum_{t=1}^{n} w_t h(\boldsymbol{x}^{(t)})}{\sum_{t=1}^{n} w_t} \to \int_{\mathcal{X}} h(\boldsymbol{x}) f(\boldsymbol{x}) d\boldsymbol{x}\,, \qquad \text{as } n \to \infty\,, \tag{3}$$

almost surely for every f-integrable function $h : \mathcal{X} \to \mathbb{R}$. Here $w_t = Z(\boldsymbol{x}^{(t)})^{-1}$.

Proof. Reversibility w.r.t. $f(\boldsymbol{x})Z(\boldsymbol{x})$ can be checked as in Proposition 6 in Appendix A.3. Representing $f(\boldsymbol{x})Z(\boldsymbol{x})$ as a mixture of d probability densities on $\mathcal{X}$ we have

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

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

$$\int h(\boldsymbol{x})f(\boldsymbol{x})d\boldsymbol{x} = \int h(\boldsymbol{x})w(\boldsymbol{x})f(\boldsymbol{x})Z(\boldsymbol{x})d\boldsymbol{x}\ .$$

Therefore from Theorem 17.0.1 of Meyn and Tweedie [1993] applied to both numerator and denominator, (3) holds since by hypothesis TGS is fZ-irreducible so that $\{\boldsymbol{x}^{(t)}\}_t$ is ergodic. □

We note that fZ-irreducibility of TGS can be established in specific examples using standard techniques, see for example Roberts and Smith [1994]. Moreover under (2) conditions from that paper which imply 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 $\{g(x_i|\boldsymbol{x}_{-i})\}_{i,\boldsymbol{x}_{-i}}$. Possible choices of these include tempered conditionals of the form

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

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

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

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

The theory developed in Section 3 will provide insight into which choice for $g(x_i|\boldsymbol{x}_{-i})$ leads to effective Monte Carlo methods. Moreover, we shall see that the mixed conditionals in (5) are robust and efficient choices in general.

The modified conditionals needs to be tractable, as we need to sample from them and evaluate their density. In many cases (e.g. exponential models), if the original full conditionals $f(x_i|\boldsymbol{x}_{-i})$ are tractable, then also the densities of the form $f^{(\beta)}(x_i|\boldsymbol{x}_{-i})$ are. Thus, TGS can typically be implemented in the same contexts of GS.

TGS has various potential advantages over GS. First it makes an “informed choice” on which variable to update, choosing with higher probability coordinates 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 $g(x_i|\boldsymbol{x}_{-i})$. Finally, as we will see in the next sections, the invariant distribution $f(\boldsymbol{x})Z(\boldsymbol{x})$ has potentially much less correlation among variables compared to the original distribution $f(\boldsymbol{x})$.

2.1 Illustrative example. Consider the following illustrative example, where the target is a bivariate Gaussian with correlation $\rho = 0.999$. Posterior distributions with such strong correlations naturally arise in Bayesian modeling, e.g. in the context of hierarchical linear models with a large number of observations. The left of Figure 1 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 $(3, 3)$. The center and right of Figure 1 display the first 200 iterations of TGS with modified conditionals given by (4) and (5), respectively, and $\beta = 1 - \rho^2$. Now the tempered conditional distributions of TGS allow the chains to move freely around the state space despite correlation. However, the vanilla version of TGS, which uses tempered conditionals as in (4), spends the majority of its time outside the region of high probability under the target. This results in high variability of the importance weights $w_t = Z(\boldsymbol{x}^{(t)})^{-1}$ (represented by the size of the black dots in Figure 1) and low Effective Sample Size (ESS) of the associated importance sampling procedure. Finally TGS-mixed, which uses tempered conditionals as in (5), achieves both a fast mixing and a high importance sampling ESS. For example, for the simulations of Figure 1, the empirical ESS associated to TGS and TGS-mix, defined as $(\sum_{t=1}^{200} w_t)^2 / (\sum_{t=1}^{200} w_t^2)$, were 11.6 and 106.4, 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 coordinate rescalings or translations. More precisely, the MCMC algorithms obtained by applying TGS to the original target $f(\boldsymbol{x})$ or to the target obtained by applying any bijective transformation to a coordinate x_i are equivalent. 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 coordinates.*

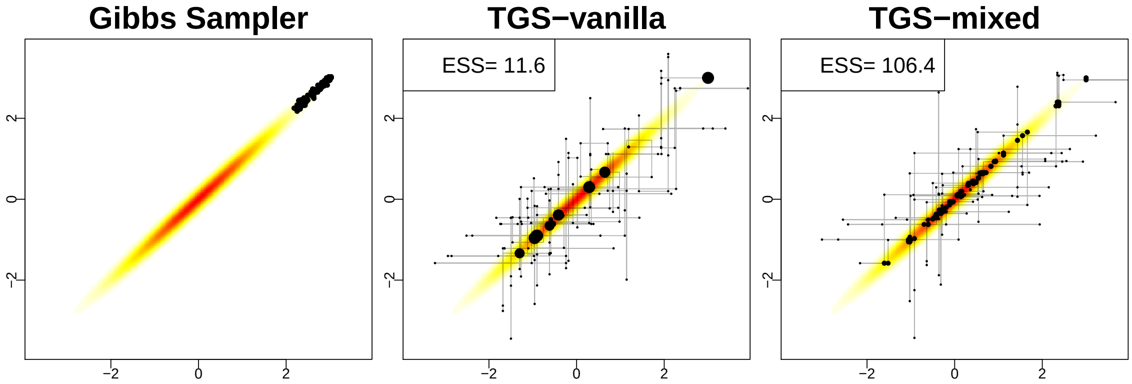

Figure 1: Comparison of GS with two versions of TGS. The sizes of the black dots are proportional to the importance weights $(w_t)_{t=1}^T$. ESS refers to the importance sampling effective sample size, defined as $(\sum_{t=1}^T w_t)^2/(\sum_{t=1}^T w_t^2)$.

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

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

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

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

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

where $Z(\boldsymbol{x}) = \frac{1}{d} \sum_{i=1}^d \frac{g(x_i|\boldsymbol{x}_{-i})}{f(x_i|\boldsymbol{x}_{-i})}$*. TGS can be seen as a scheme that targets* $\tilde{f}$ *by alternating sampling from* $\tilde{f}(i|\boldsymbol{x})$ *and* $\tilde{f}(x_i|i, \boldsymbol{x}_{-i})$*, and then corrects for the difference between* $\tilde{f}$ *and* f *with* $\frac{1}{Z(\boldsymbol{x})}$*. A direct consequence of this extended target interpretation is that the marginal distribution of* i *is uniform, meaning that each coordinate gets 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 to the GS one. The second aim is to provide understanding on which situations will be favorable to TGS and which one will not. The main message is that the performances of TGS are never significantly worse than the GS ones while, depending on the situation, can be much better.

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

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

In order to ensure robustness of TGS, we want the constant c to be finite and not too large. This can be easily achieved in practice. For example setting $g(x_i|\boldsymbol{x}_{-i})$ as in (5) we are guaranteed to have $c \leq 2$. More generally, choosing $g(x_i|\boldsymbol{x}_{-i}) = \frac{1}{1+\epsilon} f(x_i|\boldsymbol{x}_{-i}) + \frac{\epsilon}{1+\epsilon} f^{(\beta)}(x_i|\boldsymbol{x}_{-i})$ we obtain $c \leq 1+\epsilon$. The important aspect to note here is that (6) involves only ratios of one-dimensional densities rather than d-dimensional ones (more precisely densities over $\mathcal{X}_i$ rather than over $\mathcal{X}$).

3.1 Robustness to high-dimensionality. A major concern with classical importance tempering schemes is that they often collapse in high-dimensional scenarios (see e.g. Owen, 2013, Sec.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. On the contrary, the TGS algorithm is robust to high-dimensional scenarios. This can be quantified by looking at the variance of the importance weights $w_t = Z(\boldsymbol{x}^{(t)})^{-1}$ induced by TGS or, equivalently, at the Effective Sample Size defined as $ESS = n/(1 + Var(w_t))$.

Proposition 2. *Given $\boldsymbol{x}^{(t)} \sim f(\boldsymbol{x})Z(\boldsymbol{x})$ and $w_t = Z(\boldsymbol{x}^{(t)})^{-1}$, we have*

$$Var\,(w_t) \leq c - 1 \quad \textit{and} \quad ESS = \frac{n}{1 + Var\,(w_t)} \geq \frac{n}{c} \,,$$

with c defined in (6).

Proof. Equation (6) implies $p_i(\boldsymbol{x}) \geq c^{-1}$ for every $\boldsymbol{x} \in \mathcal{X}$ and thus $Z(\boldsymbol{x}^{(t)}) \geq c^{-1}$. Thus, using $w_t = Z(\boldsymbol{x}^{(t)})^{-1}$ we obtain

$$Var\,(w_t) = \mathbb{E}[w_t^2] - \mathbb{E}[w_t]^2 = \int_{\mathcal{X}} \frac{f(\boldsymbol{x})}{Z(\boldsymbol{x})} d\boldsymbol{x} - 1 \leq c - 1 \,.$$

□

Proposition 2 implies that, regardless of the dimensionality of the state space, the variance of the importance weights induced by TGS is upper bounded by $c-1$. Therefore, if $g(x_i|\boldsymbol{x}_{-i})$ are chosen to be the mixed conditionals in (5) one is guaranteed to have $Var\,(w_t) \leq 1$ and $ESS \geq n/2$. This is coherent with the empirical ESS observed in the illustrative example of Section 2.1.

An even stronger property of TGS than the bound in Proposition 2 is that, under appropriate assumptions, $Var\,(w_t)$ converges to 0 as $d \to \infty$. The underlying reason is that w_t depends on an average of d terms, namely $\frac{1}{d}\sum_{i=1}^{d} p_i(\boldsymbol{x}^{(t)})$, and the increase of dimensionality has a stabilizing effect on the latter. If, for example, the target has a product structure $f(\boldsymbol{x}) = \prod_{i=1}^{d} f(x_i)$ one can show that $Var\,(w_t)$ converges to 0 as $d \to \infty$.

Proposition 3. *Suppose $f(\boldsymbol{x}) = \prod_{i=1}^{d} f(x_i)$ and $g(x_i|\boldsymbol{x}_{-i}) = g(x_i)$ where f and g are univariate probability density functions. If $\sup_{x_i} f(x_i)/g(x_i) < \infty$, then*

$$Var\,(w_t) \to 0 \qquad \textit{and} \qquad ESS \to n \qquad \textit{as } d \to \infty \,.$$

Proof. By assumption we have $w_t = d\left(\sum_{i=1}^d \frac{g(x_i)}{f(x_i)}\right)^{-1}$ and

$$\mathbb{E}[w_t^2] = \int_{\mathcal{X}} d\left(\sum_{i=1}^d \frac{g(x_i)}{f(x_i)}\right)^{-1} f(\boldsymbol{x})d\boldsymbol{x}\,.$$

If $\boldsymbol{x} \sim f(\boldsymbol{x})$ then $g(x_i)/f(x_i)$ are independent and identically distributed random variables with mean 1 and thus by the Law of Large Numbers $d\left(\sum_{i=1}^d \frac{g(x_i)}{f(x_i)}\right)^{-1}$ converges almost surely to 1 as $d \to \infty$. Also, (6) implies $d\left(\sum_{i=1}^d \frac{g(x_i)}{f(x_i)}\right)^{-1} \leq c$. Thus by the Bounded Convergence Theorem $\mathbb{E}[w_t^2] \to 1$ as $d \to \infty$. Since $\mathbb{E}[w_t] = 1$ it follows $Var\,(w_t) \to 0$. □

Proposition 3 makes the assumption of product structure for simplicity and illustrative purposes. However we expect the same result to hold under much milder condition, such as some local Markov structure on the d components sufficient to have $\frac{1}{d}\sum_{i=1}^d p_i(\boldsymbol{x}^{(t)})$ converging to a constant as $d \to \infty$. By contrast, recall that the importance weights associated to classical tempering in the context of Proposition 3 depend on the product of d terms and have a variance that grows exponentially with d.

3.2 Explicit comparison with standard Gibbs Sampling. We now compare the efficiency of the Monte Carlo estimators produced by TGS with the ones produced by classical GS. For any function $h \in L^2(\mathcal{X}, f)$ the TGS estimator is $\hat{h}_n^{TGS}$ as defined in (3), while the GS one is $\hat{h}_n^{GS} = \frac{1}{n}\sum_{t=1}^n h(\boldsymbol{y}^{(t)})$, where $\boldsymbol{y}^{(1)}, \boldsymbol{y}^{(2)}, \dots$ is the $\mathcal{X}$-valued Markov chain generated by GS. We measure efficiency in terms of asymptotic variances, which are defined as

$$\text{var}(h, TGS) = \lim_{n\to\infty} n\,\text{var}(\hat{h}_n^{TGS}), \qquad \text{var}(h, GS) = \lim_{n\to\infty} n\,\text{var}(\hat{h}_n^{GS}).$$

The following theorem shows that the performances of TGS can never be worse than the ones of GS by a factor larger than c^2.

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

$$var(h, TGS) \leq c^2 var(h, GS) + c^2 var(h(X))\,. \tag{7}$$

where $X \sim f$.

In most non-trivial scenarios, var$(h(X))$ is negligible with respect to var(h, GS), because the the asymptotic variance obtained by GS is typically much larger than the one of an i.i.d. sampler. In such cases we can interpret (7) as saying that the asymptotic variance of TGS is at most c^2 times the ones of GS plus a smaller order term.

The proof of Theorem 1 relies on a Peskun ordering argument between continuous-time versions of the Markov chains $\boldsymbol{x}^{(t)}$ and $\boldsymbol{y}^{(t)}$ to guarantee that the mixing induced by TGS can never be significantly worse than the one of GS. The proof is in Appendix A.1.

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

Corollary 1. *Let $\epsilon, \beta > 0$. If $g(x_i|\boldsymbol{x}_{-i}) = \frac{1}{1+\epsilon}f(x_i|\boldsymbol{x}_{-i}) + \frac{\epsilon}{1+\epsilon}f^{(\beta)}(x_i|\boldsymbol{x}_{-i})$ then*

$$var(h, TGS) \leq (1+\epsilon)^2 var(h, GS) + (1+\epsilon)^2 var(h(X))\,,$$

where $X \sim f$.

By choosing ϵ not too large, we have theoretical guarantees that TGS is not doing more than $(1+\epsilon)^2$ times worse than GS. Choosing ϵ too small, however, will reduce the potential benefit obtained with TGS, with TGS collapsing to GS for $\epsilon = 0$, so that optimising 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 (5), which is a safe and robust choice both in terms of importance sampling ESS and of Markov chain mixing.

3.3 TGS and correlation structure. Theorem 1 implies that, under suitable choices of $g(x_i|\boldsymbol{x}_{-i})$, TGS never provides significantly worse (i.e. worse by more than a controllable constant factor) performances than GS. On the other hand, TGS performances can be dramatically better than standard GS. The underlying reason is that the tempering mechanism can dramatically speed up the convergence of the TGS Markov chain $\boldsymbol{x}^{(t)}$ to its stationary distribution $f(\boldsymbol{x})Z(\boldsymbol{x})$ by reducing correlations in the target. In fact, the covariance structure of $f(\boldsymbol{x})Z(\boldsymbol{x})$ is substantially different from the one of the original target $f(\boldsymbol{x})$ and this can avoid the sampler from getting stuck in situations where GS would. Figure 2 displays the original target f and the modified one fZ for a bivariate Gaussian with increasing correlation. Here the modified conditionals are defined as in (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 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 $\rho \to 1$.

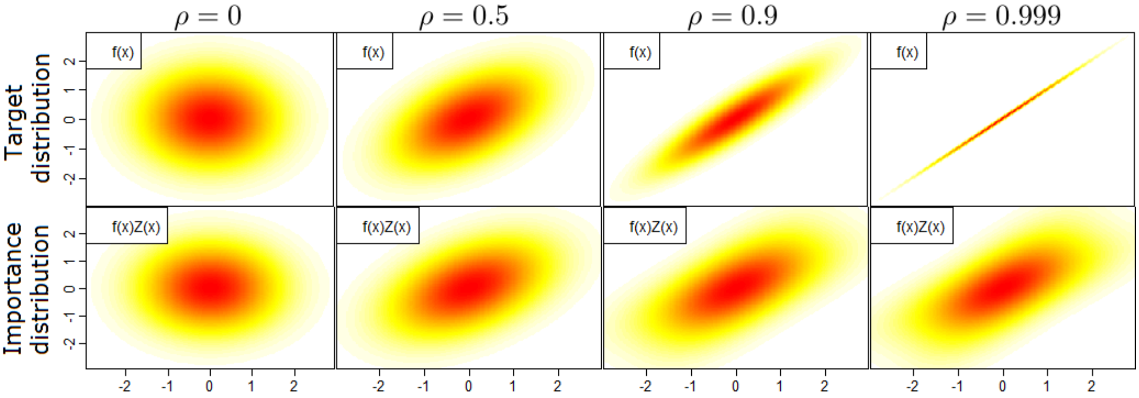

Figure 2: Comparison between $f(\boldsymbol{x})$ and $f(\boldsymbol{x})Z(\boldsymbol{x})$, first and second row respectively, for increasing correlation. Here f is a symmetric bivariate normal with correlation ρ and $g = f^{(\beta)}$ with $\beta = 1 - \rho^2$.

Note that 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 one targeting f. For example, in the Gaussian case the covariance matrix of $f^{(\beta)}$ is simply β times the one 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 one 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 to TGS.

3.4 Convergence analysis in the bivariate case. In general, $\boldsymbol{x}^{(t)}$ 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 deinitialising chains from Roberts and Rosenthal [2001] we can obtain rather explicit understanding of the convergence behaviour of $\boldsymbol{x}^{(t)}$ in the bivariate case. The results suggest that, for appropriate choices of modified conditionals, the mixing time of $\boldsymbol{x}^{(t)}$ is uniformly bounded regardless of the correlation structure of the target. This has to be contrasted with the chain induced by GS, whose mixing time diverges to infinity as the target's correlation goes to 1.

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

1. Sample $i^{(t+1)}$ from $\{1, \ldots, d\}$ proportionally to $(p_1(\boldsymbol{x}^{(t)}), \ldots, p_d(\boldsymbol{x}^{(t)}))$,
2. Sample $x_{i^{(t+1)}}^{(t+1)} \sim g(x_{i^{(t+1)}} | \boldsymbol{x}_{-i^{(t+1)}} = \boldsymbol{x}_{-i^{(t+1)}}^{(t)})$ and set $\boldsymbol{x}_{-i^{(t+1)}}^{(t+1)} = \boldsymbol{x}_{-i^{(t+1)}}^{(t)}$.

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

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

where $\boldsymbol{x}_{-i^{(t)}}^{(t)}$ denotes the vector $\boldsymbol{x}^{(t)}$ without the $i^{(t)}$-th component. The following proposition shows that the process $(\boldsymbol{z}^{(t)}, i^{(t)})_{t=0}^{\infty}$ is Markovian and contains all the information needed to characterise the convergence to stationarity of $\boldsymbol{x}^{(t)}$. The proof is in Appendix A.2.

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

$$\mathcal{L}(\boldsymbol{x}^{(t)}, i^{(t)} | \boldsymbol{x}^{(0)}, i^{(0)}, \boldsymbol{z}^{(t)}, i^{(t)}) = \mathcal{L}(\boldsymbol{x}^{(t)}, i^{(t)} | \boldsymbol{z}^{(t)}, i^{(t)}) \qquad t \geq 1 , \tag{8}$$

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

$$\|\mathcal{L}(\boldsymbol{x}^{(t)} | \boldsymbol{x}^{(0)} = \boldsymbol{x}_*) - fZ\|_{TV} = \|\mathcal{L}(\boldsymbol{z}^{(t)}, i^{(t)} | \boldsymbol{x}^{(0)} = \boldsymbol{x}_*) - \pi\|_{TV} , \tag{9}$$

where $\| \cdot \|_{TV}$ denotes total variation distance and π is the stationary distribution of $(\boldsymbol{z}^{(t)}, i^{(t)})$.

Note that the conditioning on $\boldsymbol{x}^{(0)}$ in (9) is equivalent to conditioning on $(\boldsymbol{x}^{(0)}, i^{(0)})$, because the distribution of $(\boldsymbol{x}^{(t)}, i^{(t)})$ for $t > 1$ is independent of $i^{(0)}$.

Proposition 4 implies that the convergence to stationarity of $\boldsymbol{x}^{(t)}$ is fully determined by that of $(\boldsymbol{z}^{(t)}, i^{(t)})$. In some situations, by looking at the chain $(\boldsymbol{z}^{(t)}, i^{(t)})$ rather than $\boldsymbol{x}^{(t)}$, 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 $f_1(z) = \int_{\mathbb{R}} f(z, x_2)dx_2$ and $f_2(z) = \int_{\mathbb{R}} f(x_2, z)dx_2$ are the target marginals. In order to keep notation light and have results that are easier to interpret, here we further assume exchangeability, i.e. $f(x_1, x_2) = f(x_2, x_1)$, while Lemma 3 in Appendix A.2 considers the generic case. The simplification 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 $\boldsymbol{x}_* = (z_*, z_*) \in \mathbb{R}^2$ we have*

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

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

$$P(z'|z) = r(z)\delta_{(z)}(z') + q(z'|z)\alpha_b(z'|z) , \tag{10}$$

where $r(z) = 1 - \int_{\mathbb{R}} \alpha_b(z'|z)q(z'|z)dz'$, $\alpha_b(z'|z) = \frac{f_1(z')q(z|z')}{f_1(z)q(z'|z)+f_1(z')q(z|z')}$ and $q(z'|z) = g(x_i = z'|x_{-i} = z)$.

The transition kernel in (10) coincides with the one 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 coordinate 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 a rather 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 f to have standard normal marginals, and thus be 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 (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 $(\boldsymbol{x}^{(t)})_{t=0}^{\infty}$, 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 autoregressive process $z^{(t+1)}|z^{(t)} \sim N(\rho z, 1-\rho^2)$. The latter discussion provides theoretical insight for the behaviour 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 TGS work and when does it not? The previous two sections showed that in the bivariate case TGS can induce much faster mixing compared to GS. A natural question is how much this extends to the case $d > 2$. In this section we provide insight into when TGS substantially outperform GS and when instead they are comparable (we know by Theorem 1 that TGS cannot mix substantially slower than GS). The latter depends on the correlation structure of the target with intuition being as follows. When sampling from a d-dimensional target $(x_1, \ldots, x_d)$, the tempering mechanism of TGS allows 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. On the other hand, 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 w.l.o.g. the d-dimensional target to have 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 $\Sigma_{2i-1,2i} = \rho$ for $i = 1, \ldots, \frac{d}{2}$ 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 $\Sigma_{ij} = -\frac{\rho}{n-1}$ for all $i \neq j$. In all scenarios, as $\rho \to 1$ the target distribution collapse to some singular distribution and the GS convergence properties deteriorate (see Roberts and Sahu [1997] for related results).

Figure 3 reports the (estimated) asymptotic variance of the estimators of the coordinates mean (i.e. $h(\boldsymbol{x}) = x_i$, the value of i is irrelevant) for $d = 10$. We compare GS with two versions of TGS. The first has mixed conditionals as in (5), with $\beta = 1 - \rho^2$. Note that, 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 $g(x_i|\boldsymbol{x}_{-i})$ to be a t-distribution centred at $\mathbb{E}[x_i|\boldsymbol{x}_{-i}]$, with scale $\sqrt{\text{var}(x_i|\boldsymbol{x}_{-i})}$ and shape $\nu = 0.2$. As expected, the asymptotic variance of the estimators obtained with GS deteriorate in all cases. On the contrary, 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 to GS. In all cases, the two versions of TGS perform quite similarly, with 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 variable are strongly correlated to be implemented.

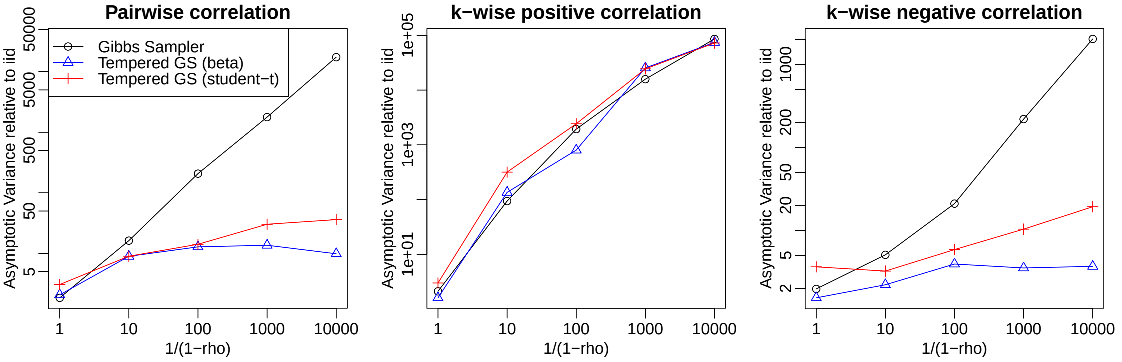

Figure 3: Log-log plots of estimated asymptotic variances for GS compared to two versions of TGS on Gaussian targets with difference covariance structures.

The reason for the presence or lack of improvements given by TGS lies in the different geometrical structure 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 good choice of β), otherwise it will not.

The intuition developed here will be useful in the Bayesian Variable Selection application of Section 4.

3.6 Controlling the frequency of coordinate updating.

In the extended target interpretation 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 absence of prior information on the structure of the problem under consideration, the latter is a desirable robustness properties as it prevents the algorithm for updating some coordinates too often and ignoring others. However, in some contexts, we may want to invest more computational effort in updating some coordinates rather than others (see for example the Bayesian Variable Selection problems discussed below). This can be done by multiplying the selection probability $p_i(\boldsymbol{x})$ for some weight function $w_i(\boldsymbol{x}_{-i})$, obtaining $p_i(\boldsymbol{x}) = w_i(\boldsymbol{x}_{-i})\frac{g(x_i|\boldsymbol{x}_{-i})}{f(x_i|\boldsymbol{x}_{-i})}$ while leaving the rest of the algorithm unchanged. We call the resulting algorithm weighted Tempered Gibbs Sampling (wTGS).

Algorithm wTGS. *At each iteration of the Markov chain*

1. *Sample i from $\{1, \ldots, d\}$ proportionally to*

$$p_i(\boldsymbol{x}) = w_i(\boldsymbol{x}_{-i})\frac{g(x_i|\boldsymbol{x}_{-i})}{f(x_i|\boldsymbol{x}_{-i})},$$

2. *Sample $x_i \sim g(x_i|\boldsymbol{x}_{-i})$,*
3. *Weight the new state $\boldsymbol{x}$ with a weight $Z(\boldsymbol{x})^{-1}$ where $Z(\boldsymbol{x}) = \frac{1}{d}\sum_{i=1}^{d} p_i(\boldsymbol{x})$.*

Clearly, TGS can be seen as a special case of wTGS with $w_i(\boldsymbol{x}_{-i}) \equiv 1$. As shown by the following proposition, such an operation does not impact the validity of the algorithm and it results in having a marginal distribution over the updated component i proportional to $\mathbb{E}[w_i(\boldsymbol{x}_{-i})]$, where $\boldsymbol{x} \sim f$. See Appendix A.3 for proof.

Proposition 6. *The Markov chain $\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \ldots$ induced by the wTGS scheme is reversible with respect to $f(\boldsymbol{x})Z(\boldsymbol{x})$. The frequency of updating of the i-th coordinate is proportional to $\mathbb{E}_{\boldsymbol{x}\sim f}[w_i(\boldsymbol{x}_{-i})]$.*

Therefore, by controlling $\mathbb{E}[w_i(\boldsymbol{x}_{-i})]$, we can modify the frequency with which we update each coordinate. In Section 4 we show an application of wTGS to Bayesian Variable Selection 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 Bayesian Variable Selection models typically possess the kind of pairwise and/or negative dependence structures conjectured to be conducive to successful application of TGS in Subsection 3.5. Therefore, in this section we provide a detailed application of TGS to sampling from the posterior distribution of Gaussian Bayesian Variable Selection models. This is a widely used class of models where posterior inferences are computationally challenging due to the presence 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 appendix B.3 for more details).

4.1 Model specification. Bayesian Variable Selection (BVS) models provide a natural and coherent framework to select a subset of explanatory variables in linear regression contexts (see e.g. Chipman et al. [2001]). In standard linear regression, an $n \times 1$ response vector Y is modeled 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 $\gamma = (\gamma_1, \dots, \gamma_p) \in \{0,1\}^p$ is introduced to indicate which regressor is included in the model and which one is not ($\gamma_i = 1$ indicates that the i-th regressor is included in the model and $\gamma_i = 0$ that it is excluded). The resulting model can be written as

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

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 the selected regressors and Σ_γ is the $|\gamma| \times |\gamma|$ prior covariance matrix for the $|\gamma|$ selected regressors. Here $|\gamma| = \sum_{i=1}^p \gamma_i$ denotes the number of "active" regressors. The covariance Σ_γ is typically chosen to be equal to a positive multiple of $(X_\gamma^T X_\gamma)^{-1}$ or the identity matrix, i.e. $\Sigma_\gamma = c(X_\gamma^T X_\gamma)^{-1}$ or $\Sigma_\gamma = c\mathbb{I}_{|\gamma|}$ for fixed $c > 0$. The binary vector γ is given a prior distribution $p(\gamma)$ on $\{0,1\}^p$, for example assuming

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

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 3. *One can also add an intercept to the linear model obtaining $Y|\beta_\gamma, \gamma, \sigma^2, \alpha \sim N(\alpha + X_\gamma\beta_\gamma, \sigma^2)$. If such intercept is given a flat prior, $p(\alpha) \propto 1$, the latter is equivalent to centering Y, X_1, ..., X_p to have zero mean.*

Under this model set-up, the continuous hyperparameters β and σ can be analytically integrated and one is left with an explicit expression for $p(\gamma|Y)$. Sampling from such $\{0,1\}^p$-valued distribution allows to perform full posterior inferences for the BVS models 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 Gibbs Sampling on the p components $(\gamma_1, \dots, \gamma_p)$, repeatedly choosing $j \in \{1, \dots, p\}$ either in a random or deterministic scan fashion and then updating $\gamma_i \sim p(\gamma_i|Y, \gamma_{-i})$.

4.2 TGS for Bayesian Variable Selection. We apply TGS to the problem of sampling from $\gamma \sim p(\gamma|Y)$. For every value of i and γ_{-i}, we set the tempered conditional distributions $g_i(\gamma_i|\gamma_{-i})$ to be the uniform distribution over $\{0,1\}$. It is easy to check that the supremum c defined in (6) is upper bounded by 2 and thus we are 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 conditional on i and γ_{-i}, with its Metropolised version (see e.g. Liu [1996]). The resulting specific instance of TGS is the following.

Algorithm TGS. *At each iteration of the Markov chain*

1. Sample i from $\{1, \dots, p\}$ proportionally to $p_i(\gamma) = \frac{1}{2p(\gamma_i|\gamma_{-i}, Y)}$.

2. Switch γ_i to $1 - \gamma_i$.

3. Weight the new state γ with a weight $Z(\gamma)^{-1}$ where $Z(\gamma) = \frac{1}{p}\sum_{i=1}^d p_i(\gamma)$.

In step 1 above, $p(\gamma_i|\gamma_{-i}, Y)$ denotes the probability that γ_i takes its current value conditional on the current value of γ_{-i} and on the observed data Y.

4.3 Weighted version. As discussed in Section 3.6, TGS updates each coordinate 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 gets 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 using non-uniform weight functions $w_i(\gamma_{-i})$. For example choosing $w_i(\gamma_{-i}) = p(\gamma_i = 1|\gamma_{-i}, Y)$ leads to a frequency of updating of the i-th component proportional to the marginal inclusion probability itself $\mathbb{E}[w_i(\gamma_{-i})] = p(\gamma_i = 1|Y)$, see e.g. Proposition 6 in the appendix. Here $p(\gamma_i = 1|Y)$ denotes the (marginal) posterior probability that γ_i equals 1, while $p(\gamma_i = 1|\gamma_{-i}, Y)$ denotes the probability of the same event conditional on both the observed data Y and the current value of γ_{-i}. Note that with wTGS we can obtain a frequency of updating proportional to $p(\gamma_i = 1|Y)$ without knowing its actual value, 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 favors exploration, as it forces the sampler to explores new regions of the space by flipping variables with low conditional inclusion probability. On the other hand, choosing a frequency of updating that focuses on variables with high conditional inclusion probability favors 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 $w_i(\gamma_{-i})$ described above and the uniform TGS, obtained by setting $w_i(\gamma_{-i}) = p(\gamma_i = 1|\gamma_{-i}, Y) + \frac{k}{p}$ with k being a fixed parameter (in our simulations we used $k = 5$). Such choice leads to frequencies of updating given by a mixture of the uniform distribution over $\{1, \dots, p\}$ and the distribution proportional to $p(\gamma_i = 1|Y)$. More precisely we have $\mathbb{E}[w_i(\gamma_{-i})] = \alpha \frac{p(\gamma_i=1|Y)}{\sum_{j=1}^{p} p(\gamma_j=1|Y)} + (1-\alpha)\frac{1}{p}$, where $\alpha = \frac{\sum_{j=1}^{p} p(\gamma_j=1|Y)}{k+\sum_{j=1}^{p} p(\gamma_j=1|Y)}$. The resulting scheme is the following (see above for the definition of $p(\gamma_i = 1|\gamma_{-i}, Y)$).

Algorithm wTGS. *At each iteration of the Markov chain*

```
1. Sample i from $\{1, \dots, p\}$ proportionally to $p_i(\gamma) = \frac{p(\gamma_i=1|\gamma_{-i},Y)+k/p}{2p(\gamma_i|\gamma_{-i},Y)}$.
2. Switch $\gamma_i$ to $1-\gamma_i$.
3. Weight the new state $\gamma$ with a weight $Z(\gamma)^{-1}$ where $Z(\gamma) = \frac{1}{p}\sum_{i=1}^{d} p_i(\gamma)$.
```

4.4 Efficient implementation and Rao-Blackwellisation. TGS provides substantially improved convergence properties with a mild computational overhead. In fact the main 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 vectorised operations as described in Appendix B. Such efficient implementation is crucial to the successful application of TGS schemes. Interestingly, $\{p(\gamma_i|Y, \gamma_{-i})\}_{i=1}^{p}$ are the same quantity needed to compute Rao-Blackwellised estimators of the marginal Posterior Inclusion Probabilities (PIPs) $\{p(\gamma_i = 1|Y)\}_{i=1}^{p}$. Therefore, using TGS allows to implement Rao-Blackwellised estimators of PIPs (for all $i \in \{1, \dots, p\}$ at each flip) without extra cost. See Appendix B.2 for more details.

4.5 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, while models including both variables or none of the two will have a low posterior probability. As a result, the Gibbs Sampler (GS) will get stuck in one of the two local modes corresponding to one variable being active and the other inactive.

Figure 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 the next section (namely Scenario 1 with SNR=3). The left and center plots in the figure show the traceplots

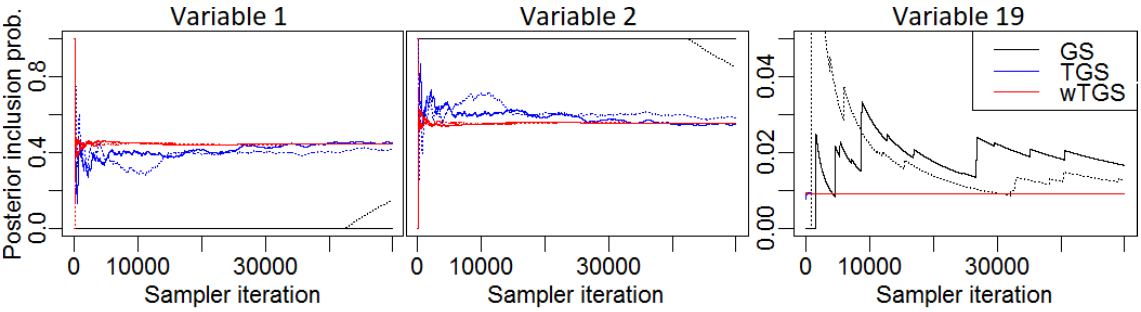

Figure 4: Running estimates of PIPs for variables 1, 2 and 243 produced by GS, TGS and wTGS over two runs. Solid lines correspond to the first run and dotted line to the second one. Thinning is used so that all schemes have the same cost per iteration.

of the estimates for the PIP of variables 1 and 2 over two runs (represented by solid and dotted lines, respectively) for GS, TGS and wTGS. 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 50000 iterations, after a burn in of 5000 iterations. GS was run for the same CPU time, performing multiple moves per iteration so that the cost per iteration matched the one of TGS and wTGS. For both runs GS got stuck in the mode corresponding to $(\gamma_1, \gamma_2) = (0, 1)$ for more than 40000 iterations, when eventually (during the second run, black dotted line) it managed to flip to $(\gamma_1, \gamma_2) = (1, 0)$. On the contrary, both TGS and wTGS manage to move frequently between the two modes and indeed the resulting estimates of PIPs for both variables seem to converge to the limiting 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 19 in this simulated data has a PIP of roughly 1%. In this case the variable is rarely included in the model and the frequency-based estimators have a high variability, while the Rao-Blackwellised ones produce nearly-instantaneous good estimates, see Figure 4 right.

Consider then an analogous simulated dataset 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 Figure 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 produce estimates converging dramatically faster to the limiting values. This is well explained by Proposition 6, which implies that TGS flips each variable every $1/p$ iterations on average, while wTGS has frequency of flipping proportional to $\mathbb{E}[w_i(\gamma_{-i})]$, 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 PIP, such as coordinate 243 which has a PIP of roughly 2% (see Figure 5 right).

To summarise, the main improvements of TGS and wTGS are due to:

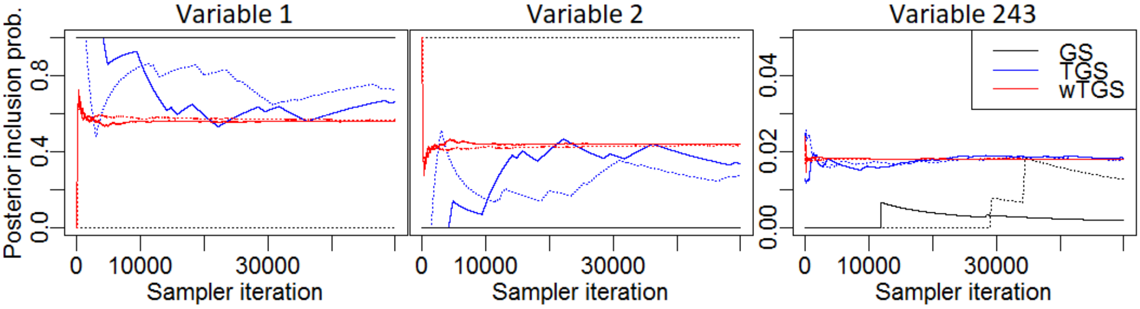

Figure 5: Analogous to Figure 4 with $p = 1000$ and $n = 500$.

(i) tempering reducing correlation and helping to move across modes (see Figure 4 left and center);
(ii) Rao-Blackwellisation producing more stable estimators (see Figures 4-5 right);
(iii) weighting mechanism of wTGS allowing to focus computation on relevant variables (see Figure 5 left and center).

The conclusions of this illustrative example would not change if one considers a scenario involving k strongly correlated variables, with $k > 2$.

4.6 Simulated data. In this section we provide a quantitative comparison between GS, TGS and wTGS under different simulated scenarios (see Section 4.7 and Appendix B.3 for real data examples and comparison with alternative samplers from the literature). Data are generated as $Y \sim N(X\beta^*, \sigma^2)$ with $\sigma^2 = 1$, $\beta^* = \text{SNR}\sqrt{\frac{\sigma^2 \log(p)}{n}}\beta_0^*$, 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 to 5. The values of β_0^* and $\Sigma^{(X)}_{ij}$ for $i \neq j$ vary depending on the considered scenario. In particular, we consider the following situations:

1. *Two strongly correlated variables:* $\beta_0^* = (1, 0, \ldots, 0)$, $\Sigma^{(X)}_{12} = \Sigma^{(X)}_{21} = 0.99$, $\Sigma^{(X)}_{ij} = 0$ otherwise.
2. *Batches of correlated variables:* $\beta_0^* = (3, 3, -2, 3, 3, -2, 0, \ldots, 0)$, $\Sigma^{(X)}_{ij} = 0.9$ if $i, j \in \{1, 2, 3\}$ or $i, j \in \{4, 5, 6\}$ and $\Sigma^{(X)}_{ij} = 0$ otherwise.
3. *Uncorrelated variables:* $\beta_0^* = (2, -3, 2, 2, -3, 3, -2, 3, -2, 3, 0, \ldots, 0)$, $\Sigma^{(X)}_{ij} = 0$ for all $i \neq j$.

Scenarios analogous to the ones above have been previously considered in the literature. For example, Titsias and Yau [2017, Sec.3.2.3] consider a scenario similar to 1, Wang et al. [2011, Ex.4] and Huang et al. [2016, Sec4.2] one similar to 2 and Yang et al. [2016] one analogous to 3. 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 Metropolised version of GS, like we did for TGS and wTGS. In order to provide a quantitative comparison we consider a standard measure of relative efficiency, being the ratio of the estimators' effective sample sizes over computational times. More precisely, we define the relative efficiency of TGS over GS as

$$\frac{\text{Eff}_{TGS}}{\text{Eff}_{GS}} = \frac{\text{ess}_{TGS}/T_{TGS}}{\text{ess}_{GS}/T_{GS}} = \frac{\sigma^2_{GS} T_{GS}}{\sigma^2_{TGS} T_{TGS}}\,, \tag{11}$$

where σ^2_{GS} and σ^2_{TGS} are the variances of the Monte Carlo estimators produced by GS and TGS, respectively, while T_{GS} and T_{TGS} are the CPU time required to produce such estimators. An analogous measure is used for the relative efficiency of wTGS over GS. For each simulated dataset, we computed the relative efficiency defined by (11) for each PIP estimator, thus getting p values, one for each variable. Table 1 reports the median of such p values for each dataset under consideration. The variances in (11), such as σ^2_{GS} and σ^2_{TGS}, were estimated with the sample variances of the PIP estimates obtained with different runs of each scheme.

		TGS-vs-GS				wTGS-vs-GS			
		SNR				SNR			
	(p,n)	0.5	1	2	3	0.5	1	2	3
scen.1	(100,50)	4.0e5	2.4e4	2.0e4	6.6e4	2.1e6	2.6e5	3.4e5	1.9e5
	(200,200)	1.0e6	4.2e6	4.9e5	2.1e6	1.6e7	5.3e7	1.0e7	2.4e7
	(1000,500)	1.3e6	1.2e6	1.1e6	2.2e6	7.8e7	9.3e7	6.5e7	1.1e8
scen.2	(100,50)	1.0e4	2.9e3	1.7e3	3.9e4	1.5e5	4.1e4	9.3e3	1.6e5
	(200,200)	1.1e5	1.0e5	8.2e3	1.4e7	1.8e6	2.8e6	1.5e5	3.2e6
	(1000,500)	4.6e5	9.2e4	6.7e5	2.1e6	3.3e7	1.1e7	1.1e7	1.5e7
scen.3	(100,50)	2.5e3	4.2e3	7.7e3	7.4e4	2.9e4	3.9e4	8.0e3	1.5e4
	(200,200)	9.1e4	4.3e4	2.8e7	3.5e6	1.0e6	3.1e5	2.9e6	8.0e5
	(1000,500)	9.8e4	5.9e5	1.1e7	2.1e7	7.0e6	4.4e6	7.6e6	1.0e7

Table 1: Median improvement over variables of TGS and wTGS relative to GS for simulated data. Scenarios 1 to 3, indicated on the leftmost column, are described in Section 4.6. Notation: 1.4e5= 1.4×10^5.

From Table 1 it can be seen that both TGS and wTGS provide orders of magnitude improvement in efficiency compared to 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 simulated datasets the fraction of variables having non-negligible PIP is small (as it 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 to Rao-Blackwelliezd versions (see right of Figures 4 and 5) 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 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 variable and thus we had no natural (finite) estimate of the variance in (11). 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 infinity, second case) and thus excluding those values from Table 2 is conservative and plays in favor of GS. The mean improvements reported in Table 2 are significantly smaller than the one in Table 1 but still potentially very large, with ranges of improvement being $(1.4, 2.5 \times 10^6)$ for TGS and$(1.8 \times 10^1, 1.9 \times 10^4)$ for wTGS. Note that values are never below 1, meaning that in these simulations TGS or wTGS are always more efficient than GS, and that wTGS

		TGS-vs-GS				wTGS-vs-GS			
		SNR				SNR			
	(p,n)	0.5	1	2	3	0.5	1	2	3
scen.1	(100,50)		7.2e1	1.8e1	2.8e2		5.8e2	4.2e2	3.1e3
	(200,200)	4.9e3		6.6e1	1.9e2	1.1e4		1.8e3	1.6e4
	(1000,500)	2.7e2	6.3e2	1.4	8.1e1	8.8e3	2.5e4	5.8e2	1.9e4
scen.2	(100,50)	4.8	1.4e1	3.3	2.0e1	1.3e2	2.4e2	1.8e1	1.4e2
	(200,200)	8.6e1	4.7e1	3.4	2.5e6	2.3e3	2.1e3	6.0e1	4.1e2
	(1000,500)	4.6e1	3.7e1	1.3e1	4.5e2	1.1e4	7.6e3	1.1e3	1.8e4
scen.3	(100,50)	2.7	5.3	9.2		2.5e1	6.7e1	2.1e1	
	(200,200)	1.1e2	6.6e1			1.3e3	4.6e2		
	(1000,500)	1.6e1	6.8e2			1.1e3	9.4e3		

Table 2: Mean improvement of TGS and wTGS relative to GS over variables with PIP>0.05. Same simulation set-ups as in Table 1. Empty values corresponds to large values with no reliable estimate available (see Section 4.6 for discussion).

is more efficient than TGS in most scenarios. Also, especially for wTGS, the improvement over GS gets larger as p increases.

The value of c in the prior covariance matrix has a large impact 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 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 considered. In Section 4.7 we will consider both $c = n$ and $c = \max\{n, p^2\}$.

4.7 Real data. In this section we consider three real datasets with increasing number of covariates. We compare wTGS to GS and the Hamming Ball (HB) sampler, a recently proposed sampling scheme designed for posterior distributions over discrete spaces, including BVS models [Titsias and Yau, 2017]. We consider three real datasets, which we refer to as DLD data, TGFB172 data and TGFB data. The DLD data comes from a genomic study by Yuan et al. [2016] based on RNA sequencing and has a moderate number of regressors, $p = 57$ and $n = 192$. The version of the dataset we used is freely available from the supplementary material of Rossell and Rubio [2017]. See Section 6.5 therein for a short description of the dataset and the inferential questions of interest. The second and third datasets are human microarray gene expression data in colon cancer patients from Calon et al. [2012]. The TGFB172 data, which has $p = 172$ and $n = 262$, is obtained as a subset of the TGFB data, for which $p = 10172$ and $n = 262$. These two datasets are are described in Section 5.3 of Rossell and Telesca [2017] and are freely available from the corresponding supplementary material.

For the BVS model of Section 4.1, the computational cost per iteration is mainly driven by p and not sensitive to n. In fact, once you precompute the $p \times p$ and $1 \times p$ matrices X^TX and y^TX, there is no operation in the sampler that has direct dependence on the value of n. Thus a dataset with a large value of p, like the TGFB data, represents a computationally challenging scenario, regardless of having a low value of n.

We performed 20 independent runs of each algorithm for each dataset 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 datasets, respectively, discarding the first 10% of samples as burnin. The number of iterations of GS and HBS were chosen to have the same runtime of wTGS. To assess the reliability of each algorithm, we compare results obtained over different runs by plotting each PIP estimate over the ones obtained with different runs of the same algorithm. The results are displayed in Figure 6. Points close to the diagonal indicate estimates in accordance with each other across runs, while point far from the diagonal indicate otherwise. It can be seen that wTGS provides substantially more reliable estimates for all combinations of dataset and value of c under consideration and that the efficiency improvement increases with the number of regressors p.

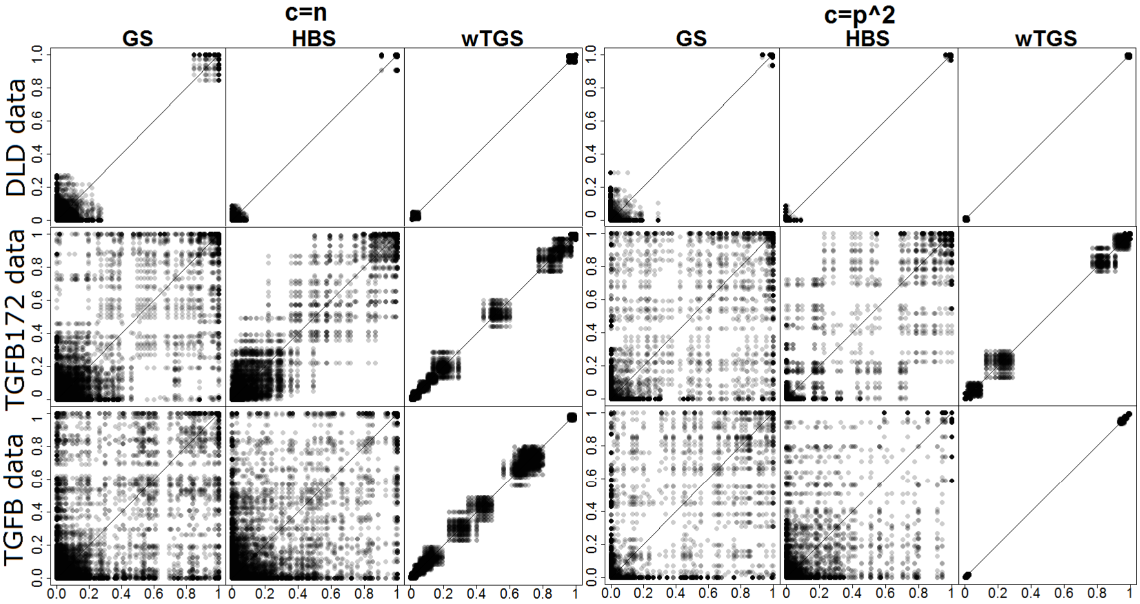

Figure 6: Comparison of GS, HBS and wTGS (columns) on three real datasets (rows) for $c = n$ and $c = p^2$. Points close to the diagonal lines indicate estimates agreeing across different runs.

All computations were performed on the same desktop computer with 16GB of RAM and an i7 Intel processor, using the R programming language [R Core Team, 2017]. The R code to implement the various samplers under consideration is freely available at `https://github.com/gzanella/TGS`. For the largest dataset under consideration (p=10172) wTGS took an average of 115 seconds for each run shown in Figure 6. We performed further experiments, in order to compare the wTGS performances with the ones of available R packages for BVS and some alternative methodology from the literature. The results, reported in Appendix B.3, suggest that wTGS provides state of the art performances for fitting spike and slab BVS models like the ones of Section 4.1.

5 Discussion

We have introduced a novel Gibbs sampler variant, demonstrating its considerable potential both in toy examples as well as more realistic Bayesian Variable Selection 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 optimised (for example 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 2 of TGS, where i is sampled proportionally to $p_i(\boldsymbol{x})$, with a Metropolised version as in [Liu, 1996], where the new value $i^{(t+1)}$ is proposed from $\{1,\dots,d\}\backslash\{i^{(t)}\}$ proportionally to $p_{i^{(t+1)}}(\boldsymbol{x})$ for $i^{(t+1)} \neq i^{(t)}$. This would effectively reduce the probability of repeatedly updating the same coordinate 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 "block-wise" 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 $\{p_i(\boldsymbol{x})\}_{i=1}^d$ 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 [2017] for a discussion of similar block-wise implementations. Also, computing $p_i(\boldsymbol{x})$ exactly may be infeasible in some contexts, and thus it would be interesting to design a version of TGS where the terms $p_i(\boldsymbol{x})$ 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 ones (see for example Roberts and Rosenthal [2015]).

Finally, one could design schemes where the conditional distributions of k coordinates $(x_{i_1},\dots,x_{i_k})$, rather than one, are tempered at the same time. A natural approach would be to start from the TGS interpretation in Remark 2 and define some extended target on $\mathcal{X} \times \{1,\dots,d\}^k$. This would allow to achieve good mixing in a larger class of target distributions (compared to the ones 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 reparametrisations [Gelfand et al., 1995, Hills and Smith, 1992]. Although these can work very well in some specific models (see e.g. Zanella and Roberts [2017], Papaspiliopoulos et al. [2018]), the generic implementations requires the ability to perform Gibbs Sampling 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 reparametrisation methods are not effective if the covariance structure of the target changes with location. Further alternative methodology to overcome strong correlations in Gibbs Sampling include the recently proposed adaptive MCMC approach of Duan et al. [2017] in the context of data augmentation models.

Given the results Section 4, it would be interesting to explore the use of the methodology 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.

Acknowledgments GZ supported by the European Research Council (ERC) through StG "N-BNP" 306406. GOR acknowledges support from EPSRC through grants EP/K014463/1 (i-Like) and EP/K034154/1 (EQUIP).

A Proofs and additional results

A.1 Proof of Theorem 1. First we provide a lemma relating the asymptotic variances of continuous-time processes and the corresponding jump chains. Let $(X_t)_{t\geq 0}$ be a continuous-time pure jump Markov chain with invariant measure $f(\boldsymbol{x})$, jump chain $(\boldsymbol{x}^{(n)})_{n=1}^{\infty}$ and holding times $(W_n)_{n=1}^{\infty}$. The continuous and discrete time chains are related as $X_t = \boldsymbol{x}^{(n(t))}$, where $n(t) = \sup\{n \in \mathbb{N} : \sum_{i=1}^{n} W_i \leq t\}$. Assume $X_0 \sim f$ and $\mathbb{E}[W_1] < \infty$. Denote by Q the jumping rate of X_t and by P the transition kernel of $\boldsymbol{x}^{(n)}$. We consider the two following asymptotic variances

$$var(h, Q) = \lim_{t\to\infty} t \operatorname{var}\left(\frac{1}{t}\int_0^t h(X_s)ds\right) \tag{12}$$

and

$$var(h, P) = \lim_{n\to\infty} n \operatorname{var}\left(\frac{\sum_{i=1}^{n} \mathbb{E}[W_i|\boldsymbol{x}^{(i)}]h(\boldsymbol{x}^{(i)})}{\sum_{i=1}^{n} \mathbb{E}[W_i|\boldsymbol{x}^{(i)}]}\right).$$

The following lemma provides an explicit connection between $var(h, Q)$ and $var(h, P)$.

Lemma 1.

$$\frac{var(h, Q)}{\mathbb{E}\left[W_1\right]} = var(h, P) + \mathbb{E}\left[h\left(\boldsymbol{x}^{(1)}\right)^2 \frac{var(W_1|\boldsymbol{x}^{(1)})}{\mathbb{E}[W_1]^2}\right].$$

Proof. By definition it holds

$$var(h, Q) = \lim_{t\to\infty} \operatorname{var}\left(\frac{1}{\sqrt{t}}\left((t - \sum_{i=1}^{n(t)-1} W_i)h(\boldsymbol{x}^{(n(t))}) + \sum_{i=1}^{n(t)-1} W_i h(\boldsymbol{x}^{(i)})\right)\right).$$

Using $n(t) \to \infty$ and $\frac{n(t)}{t} \to \frac{1}{\mathbb{E}[W_1]}$ almost surely as $t \to \infty$, it can be seen that the latter equals

$$\lim_{t\to\infty} \operatorname{var}\left(\frac{\sqrt{n(t)}}{\sqrt{t}}\frac{1}{\sqrt{n(t)}}\sum_{i=1}^{n(t)-1} W_i h(\boldsymbol{x}^{(i)})\right) = \lim_{n\to\infty} \operatorname{var}\left(\frac{1}{\sqrt{\mathbb{E}[W_1]}}\frac{1}{\sqrt{n}}\sum_{i=1}^{n} W_i h(\boldsymbol{x}^{(i)})\right),$$

which implies

$$var(h, Q) = \mathbb{E}[W_1] \lim_{n\to\infty} n \operatorname{var}\left(\frac{1}{n}\sum_{i=1}^{n} \frac{W_i}{\mathbb{E}[W_1]} h(\boldsymbol{x}^{(i)})\right). \tag{13}$$

Then, using the Law of Total Variance with conditioning on $(\boldsymbol{x}^{(t)})_{t=1}^{\infty}$, we obtain

$$\begin{aligned}
&\operatorname{var}\left(\frac{1}{n}\sum_{i=1}^{n} \frac{W_i}{\mathbb{E}[W_1]} h(\boldsymbol{x}^{(i)})\right) \\
&= \operatorname{var}\left(\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} \frac{W_i}{\mathbb{E}[W_1]} h(\boldsymbol{x}^{(i)}) \middle| (\boldsymbol{x}^{(t)})_{t=1}^{\infty}\right]\right) + \mathbb{E}\left[\operatorname{var}\left(\frac{1}{n}\sum_{i=1}^{n} \frac{W_i}{\mathbb{E}[W_1]} h(\boldsymbol{x}^{(i)}) \middle| (\boldsymbol{x}^{(t)})_{t=1}^{\infty}\right)\right] \\
&= \operatorname{var}\left(\frac{1}{n}\sum_{i=1}^{n} \frac{\mathbb{E}[W_i|\boldsymbol{x}^{(i)}]}{\mathbb{E}[W_1]} h(\boldsymbol{x}^{(i)})\right) + \frac{1}{n}\mathbb{E}\left[h(\boldsymbol{x}^{(1)})^2 \frac{\operatorname{var}\left(W_1|\boldsymbol{x}^{(1)}\right)}{\mathbb{E}[W_1]^2}\right].
\end{aligned}$$

Combining the last equation with (13) we get

$$\frac{var(h,Q)}{\mathbb{E}[W_1]} = \lim_{n\to\infty} n\,\mathrm{var}\left(\frac{1}{n}\sum_{i=1}^{n}\frac{\mathbb{E}[W_i|\boldsymbol{x}^{(i)}]}{\mathbb{E}[W_1]}h(\boldsymbol{x}^{(i)})\right) + \mathbb{E}\left[h(\boldsymbol{x}^{(1)})^2\frac{\mathrm{var}\left(W_1|\boldsymbol{x}^{(1)}\right)}{\mathbb{E}[W_1]^2}\right].$$

The thesis follows noting that

$$\lim_{n\to\infty} n\,\mathrm{var}\left(\frac{1}{n}\sum_{i=1}^{n}\frac{\mathbb{E}[W_i|\boldsymbol{x}^{(i)}]}{\mathbb{E}[W_1]}h(\boldsymbol{x}^{(i)})\right) = \lim_{n\to\infty} n\,\mathrm{var}\left(\frac{\sum_{i=1}^{n}\mathbb{E}[W_i|\boldsymbol{x}^{(i)}]h(\boldsymbol{x}^{(i)})}{\sum_{i=1}^{n}\mathbb{E}[W_i|\boldsymbol{x}^{(i)}]}\right),$$

which follows from

$$n\,\mathrm{var}\left(\frac{\sum_{i=1}^{n}\mathbb{E}[W_i|\boldsymbol{x}^{(i)}]h(\boldsymbol{x}^{(i)})}{\sum_{i=1}^{n}\mathbb{E}[W_i|\boldsymbol{x}^{(i)}]}\right) = \mathrm{var}\left(\frac{n}{\sum_{i=1}^{n}\mathbb{E}[W_i|\boldsymbol{x}^{(i)}]}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbb{E}[W_i|\boldsymbol{x}^{(i)}]h(\boldsymbol{x}^{(i)})\right)$$

and the fact that $\frac{n}{\sum_{i=1}^{n}\mathbb{E}[W_i|\boldsymbol{x}^{(i)}]} \to \frac{1}{\mathbb{E}[W_i]}$ almost surely as $n\to\infty$. □

Proof of Theorem 1. Let $(\boldsymbol{y}^{(t)})_{t=1}^{\infty}$ and $(\boldsymbol{x}^{(t)})_{t=1}^{\infty}$ be the discrete-time Markov chains induced by GS and TGS respectively. Their transition kernels P_{GS} and P_{TGS} are

$$\begin{aligned}P_{GS}(\boldsymbol{x},\boldsymbol{y}) =& \frac{1}{d}\sum_{i=1}^{d} f(y_i|\boldsymbol{x}_{-i})\mathbb{1}(\boldsymbol{x}_{-i}=\boldsymbol{y}_{-i}),\\ P_{TGS}(\boldsymbol{x},\boldsymbol{y}) =& \frac{1}{d\,Z(\boldsymbol{x})}\sum_{i=1}^{d}\frac{g(x_i|\boldsymbol{x}_{-i})}{f(x_i|\boldsymbol{x}_{-i})}g(y_i|\boldsymbol{x}_{-i})\mathbb{1}(\boldsymbol{x}_{-i}=\boldsymbol{y}_{-i}),\end{aligned}$$

where $\mathbb{1}$ denotes the indicator function. Let $(Y_t)_{t\geq 0}$ and $(X_t)_{t\geq 0}$ be continuous-time jump processes with jumping rates P_{GS} and $c^2 Z(\boldsymbol{x})P_{TGS}$ respectively. More precisely, define two kernels Q_{GS} and Q_{TGS} as follows: $Q_{GS}(\boldsymbol{x},\boldsymbol{y}) = P_{GS}(\boldsymbol{x},\boldsymbol{y})$ for $\boldsymbol{x}\neq\boldsymbol{y}$ and $Q_{GS}(x,x) = P_{GS}(x,x) - \sum_{y\neq x}P_{GS}(\boldsymbol{x},\boldsymbol{y})$; $Q_{TGS}(\boldsymbol{x},\boldsymbol{y}) = c^2 Z(\boldsymbol{x})P_{TGS}(\boldsymbol{x},\boldsymbol{y})$ for $\boldsymbol{x}\neq\boldsymbol{y}$ and $Q_{TGS}(x,x) = c^2 Z(\boldsymbol{x})(P_{TGS}(x,x) - \sum_{y\neq x}P_{TGS}(\boldsymbol{x},\boldsymbol{y}))$. Then $(Y_t)_{t\geq 0}$ and $(X_t)_{t\geq 0}$ are continuous-time Markov chains with generators of the form

$$P_{GS}^{CT}f(\boldsymbol{x}) = \int_{\mathcal{X}} f(y)Q_{GS}(\boldsymbol{x},\boldsymbol{y})d\boldsymbol{y} \qquad \text{and} \qquad P_{TGS}^{CT}f(\boldsymbol{x}) = \int_{\mathcal{X}} f(y)Q_{TGS}(\boldsymbol{x},\boldsymbol{y})d\boldsymbol{y},$$

for any $f\in L^2(\mathcal{X},f)$. By construction, for every $\boldsymbol{x}\neq\boldsymbol{y}$ we have either $Q_{GS}(\boldsymbol{x},\boldsymbol{y}) = Q_{TGS}(\boldsymbol{x},\boldsymbol{y}) = 0$ or there exist $i\in\{1,\ldots,d\}$ such that

$$\frac{Q_{TGS}(\boldsymbol{x},\boldsymbol{y})}{Q_{GS}(\boldsymbol{x},\boldsymbol{y})} = c^2\frac{g(x_i|\boldsymbol{x}_{-i})}{f(x_i|\boldsymbol{x}_{-i})}\frac{g(y_i|\boldsymbol{x}_{-i})}{f(y_i|\boldsymbol{x}_{-i})} \geq 1,$$

where the inequality holds by definition of c. If follows that $Q_{TGS}(\boldsymbol{x},A\backslash\{\boldsymbol{x}\}) \geq Q_{GS}(\boldsymbol{x},A\backslash\{\boldsymbol{x}\})$ for every measurable $A\subseteq\mathcal{X}$ and thus by Theorem 6 of [Leisen and Mira, 2008] we have $var(h,Q_{TGS}) \leq var(h,Q_{GS})$ with $var(h,Q_{TGS})$ and $var(h,Q_{GS})$ defined as in (12). Moreover, from Lemma 1 we have

$$\mathrm{var}(h,Q_{GS}) = \mathrm{var}(h,GS) + \mathbb{E}_{X\sim f}\left[h(X)^2\right]$$

and

$$\mathrm{var}(h,Q_{TGS}) = \frac{1}{c^2}\left(\mathrm{var}(h,TGS) + \mathbb{E}\left[h(\boldsymbol{x}^{(1)})^2\mathrm{var}\left(c^2 W_1^{TGS}|\boldsymbol{x}^{(1)}\right)\right]\right),$$

where W_1^{TGS} is the holding time associated to $\boldsymbol{x}^{(1)}$. Combining the last two equations with $var(h, Q_{TGS}) \leq var(h, Q_{GS})$ we get

$$\begin{aligned}\text{var}(h, TGS) &\leq c^2\text{var}(h, GS) + c^2\mathbb{E}_{X\sim f}\left[h(X)^2\right] - \mathbb{E}\left[h(\boldsymbol{x}^{(1)})^2\text{var}\left(c^2 W_1^{TGS}|\boldsymbol{x}^{(1)}\right)\right]\\ &\leq c^2\text{var}(h, GS) + c^2\mathbb{E}_{X\sim f}\left[h(X)^2\right]\,,\end{aligned}$$

as desired. □

A.2 Proofs for Section 3.4. First we state a simple fact about total variation distance that is used in the proofs below. We provide a proof for completeness.

Lemma 2. *Let μ and ν be two probability measure on a product space $\mathcal{Y}_1 \times \mathcal{Y}_2$ with marginals on $\mathcal{Y}_1$ denoted by μ_1 and ν_1, respectively. If the conditional distributions of μ and ν on $\mathcal{Y}_2$ coincide, i.e. $\mu_2(\cdot|y_1) = \nu_2(\cdot|y_1)$, for every $y_1 \in \mathcal{Y}_1$ then $\|\mu - \nu\|_{TV} = \|\mu_1 - \nu_1\|_{TV}$.*

Proof. The inequality $\|\mu_1 - \nu_1\|_{TV} \leq \|\mu - \nu\|_{TV}$ follows directly from the definition of total variation distance. For the reverse inequality, note that for every $A \subseteq \mathcal{Y}_1 \times \mathcal{Y}_2$ we have $\mu(A) - \nu(A) = \int_{\mathcal{Y}_1} h\, d\mu_1 - \int_{\mathcal{Y}_1} h\, d\nu_1$ where $h(y_1) = \mu_2(A_{y_1}|y_1) = \nu_2(A_{y_1}|y_1) \leq 1$ with $A_{y_1} = \{y_2 \in \mathcal{Y}_2 : (y_1, y_2) \in A\}$. It follows $|\mu(A) - \nu(A)| \leq \sup_{|h|\leq 1} |\int_{\mathcal{Y}_1} h\, d\mu_1 - \int_{\mathcal{Y}_1} h\, d\nu_1| = \|\mu_1 - \nu_1\|_{TV}$ and thus $\|\mu - \nu\|_{TV} \leq \|\mu_1 - \nu_1\|_{TV}$. □

Proof of Proposition 4. First we prove the Markovianity of $(\boldsymbol{x}^{(t)}_{-i^{(t)}}, i^{(t)})_{t=1}^{\infty}$. Let $t \geq 1$ and denote conditional densities by $p(\cdot|\cdot)$. By construction we have $p(x^{(t)}_{i^{(t)}}|(\boldsymbol{x}^{(s)}_{-i^{(s)}}, i^{(s)})_{s=0}^{t}) = g(x_{i^{(t)}} = x^{(t)}_{i^{(t)}}|\boldsymbol{x}_{-i^{(t)}} = \boldsymbol{x}^{(t)}_{-i^{(t)}})$ and thus $\mathcal{L}(x^{(t)}_{i^{(t)}}|(\boldsymbol{x}^{(s)}_{-i^{(s)}}, i^{(s)})_{s=0}^{t}) = \mathcal{L}(x^{(t)}_{i^{(t)}}|\boldsymbol{x}^{(t)}_{-i^{(t)}}, i^{(t)})$. Using the latter equality and the Markovianity of $(\boldsymbol{x}^{(s)}, i^{(s)})_{s=1}^{\infty}$, we have

$$\begin{aligned}&p(\boldsymbol{x}^{(t+1)}_{-i^{(t+1)}}, i^{(t+1)}|(\boldsymbol{x}^{(s)}_{-i^{(s)}}, i^{(s)})_{s=0}^{t})\\ &= \int_{\mathcal{X}_{i^{(t)}}} p(\boldsymbol{x}^{(t+1)}_{-i^{(t+1)}}, i^{(t+1)}|x^{(t)}_{-i^{(t)}} = x, (\boldsymbol{x}^{(s)}_{-i^{(s)}}, i^{(s)})_{s=0}^{t})p(x^{(t)}_{i^{(t)}} = x|(\boldsymbol{x}^{(s)}_{-i^{(s)}}, i^{(s)})_{s=0}^{t})dx\\ &= \int_{\mathcal{X}_{i^{(t)}}} p(\boldsymbol{x}^{(t+1)}_{-i^{(t+1)}}, i^{(t+1)}|x^{(t)}_{-i^{(t)}} = x, \boldsymbol{x}^{(t)}_{-i^{(t)}}, i^{(t)})p(x^{(t)}_{-i^{(t)}} = x|\boldsymbol{x}^{(t)}_{-i^{(t)}}, i^{(t)})dx\,, \qquad (14)\end{aligned}$$

which implies that the distribution of $(\boldsymbol{x}^{(t+1)}_{-i^{(t+1)}}, i^{(t+1)})$ is independent $(\boldsymbol{x}^{(s)}_{-i^{(s)}}, i^{(s)})_{s=0}^{t-1}$ given $(\boldsymbol{x}^{(t)}_{-i^{(t)}}, i^{(t)})$.

The deinitialising property in (8) follows from the fact that $(\boldsymbol{z}^{(t)}, i^{(t)}) = (\boldsymbol{x}^{(t)}_{-i^{(t)}}, i^{(t)})$ with probability one, while the distribution of $x^{(t)}_{i^{(t)}}$ given $(\boldsymbol{x}^{(t)}_{-i^{(t)}}, i^{(t)})$ equals $g(x_{i^{(t)}}|\boldsymbol{x}_{-i^{(t)}} = \boldsymbol{x}^{(t)}_{-i^{(t)}})$; meaning that both the distribution on the left and the right hand side of (8) equal

$$\delta_{(\boldsymbol{z}^{(t)}, i^{(t)})}(x^{(t)}_{i^{(t)}}, i^{(t)})g(x_{i^{(t)}}|\boldsymbol{x}_{-i^{(t)}} = \boldsymbol{x}^{(t)}_{-i^{(t)}})\,.$$

We now prove (9). Denote by $\tilde{f}$ the stationary distribution of $(\boldsymbol{x}^{(t)}, i^{(t)})_{t=0}^{\infty}$, as in Remark 2. Since $(\boldsymbol{x}^{(t)}_{-i^{(t)}}, i^{(t)})$ is deinitializing for $(\boldsymbol{x}^{(t)}, i^{(t)})$ from (8) and can be obtained as a function of $(\boldsymbol{x}^{(t)}, i^{(t)})$, it follows by Corollary 2 of Roberts and Rosenthal [2001] that

$$\|\mathcal{L}(\boldsymbol{x}^{(t)}, i^{(t)}|\boldsymbol{x}^{(0)} = \boldsymbol{x}_*) - \tilde{f}\|_{TV} = \|\mathcal{L}(\boldsymbol{x}^{(t)}_{-i^{(t)}}, i^{(t)}|\boldsymbol{x}^{(0)} = \boldsymbol{x}_*) - \pi\|_{TV}\,,$$

for every $t \geq 1$. Moreover, $\mathcal{L}(\boldsymbol{x}^{(t)}, i^{(t)}|\boldsymbol{x}^{(0)})$ and $\tilde{f}$ are probability distributions on $\mathcal{X} \times \{1, \dots, d\}$ with equal conditional distributions on $\{1, \dots, d\}$, namely given by $\mathcal{L}(i|\boldsymbol{x}) \propto (p_1(\boldsymbol{x}), \dots, p_d(\boldsymbol{x}))$ for every $\boldsymbol{x} \in \mathcal{X}$. Therefore, by Lemma 2 it follows that the total variation distance between $\mathcal{L}(\boldsymbol{x}^{(t)}, i^{(t)}|\boldsymbol{x}^{(0)})$ and $\tilde{f}$ equals the one of their marginal distributions on $\mathcal{X}$, resulting in

$$\|\mathcal{L}(\boldsymbol{x}^{(t)}, i^{(t)}|\boldsymbol{x}^{(0)} = \boldsymbol{x}_*) - \tilde{f}\|_{TV} = \|\mathcal{L}(\boldsymbol{x}^{(t)}|\boldsymbol{x}^{(0)} = \boldsymbol{x}_*) - fZ\|_{TV} .$$

Combining the two displayed equations above we obtain the desired result. □

The following lemma describes the evolution of $(z^{(t)}, i^{(t)})_{t=0}^{\infty}$ when $\mathcal{X} = \mathbb{R}^2$ without assuming exchangeability. To avoid confusion, we use the following notation for the densities of the tempered conditionals $g(x_i|x_{-i})$: for every $i \in \{1, 2\}$ and $z, z' \in \mathbb{R}$ denote the tempered conditional $g(x_i = z'|x_{-i} = z)$ by $g_i(z'|z)$. As for Proposition 5, we denote target distribution and marginals by f, f_1 and f_2.

Lemma 3. *Let $\mathcal{X} = \mathbb{R}^2$. The process $(z^{(t)}, i^{(t)})_{t=0}^{\infty}$ is a $\mathbb{R} \times \{1, 2\}$-valued Markov chain with stationary distribution $\pi(z, i) = \frac{1}{2}f_2(z)\mathbb{1}(i = 1) + \frac{1}{2}f_1(z)\mathbb{1}(i = 2)$ and transition kernel*

$$P(z', i'|z, i) = r(z, i)\delta_{(z,i)}(z', i') + g_i(z'|z)\mathbb{1}(i \neq i')\alpha_b(z', i'|z, i) , \tag{15}$$

where $r(z, i) = 1 - \sum_{i' \neq i} \int_{\mathbb{R}} \alpha_b(z', i'|z, i)q(z', i'|z, i)dz'$ and $\alpha_b(z', i'|z, i)$ is an acceptance probability equal to $\frac{\pi(z',i')q(z,i|z',i')}{\pi(z,i)q(z',i'|z,i)+\pi(z',i')q(z,i|z',i')}$.

The transition kernel in (15) coincides with the one of an accept-reject algorithm with proposal distribution $q(z', i'|z, i) = g_i(z'|z)\mathbb{1}(i \neq i')$. Lemma 3 implies that the evolution of $(z^{(t)}, i^{(t)})_{t=0}^{\infty}$ depends on f only through the marginals f_1 and f_2, rather than on the joint distribution $f(x_1, x_2)$ explicitly.

Proof of Lemma 3. From (14), the transition from $(x_{-i^{(t)}}^{(t)}, i^{(t)})$ to $(x_{-i^{(t+1)}}^{(t+1)}, i^{(t+1)})$ is obtained performing the following steps

1'. Sample $x_{i^{(t)}}^{(t)} \sim g(x_{i^{(t)}}|x_{-i^{(t)}} = x_{-i^{(t)}}^{(t)})$,

2'. Sample $i^{(t+1)}$ from $\{1, \dots, d\}$ proportionally to $(p_1(\boldsymbol{x}^{(t)}), \dots, p_d(\boldsymbol{x}^{(t)}))$,

3'. Set $x_{-i^{(t+1)}}^{(t+1)} = x_{-i^{(t+1)}}^{(t)}$.

It follows that $i^{(t+1)} = i^{(t)}$ implies $x_{-i^{(t+1)}}^{(t+1)} = x_{-i^{(t)}}^{(t)}$ and thus $(z^{(t+1)}, i^{(t+1)}) = (z^{(t)}, i^{(t)})$ if and only if $i^{(t+1)} = i^{(t)}$. Given $\boldsymbol{x}^{(t)} = \boldsymbol{x}$ and $i^{(t)} = i$, the probability of sampling $i^{(t+1)} = i^{(t)}$ is

$$\begin{aligned}\frac{p_i(\boldsymbol{x})}{p_i(\boldsymbol{x}) + p_{i'}(\boldsymbol{x})} =& \frac{g_i(x_i|x_{i'})/f(x_i|x_{i'})}{g_i(x_i|x_{i'})/f(x_i|x_{i'}) + g_{i'}(x_{i'}|x_i)/f(x_{i'}|x_i)} \\ =& \frac{f_{i'}(x_{i'})g_i(x_i|x_{i'})}{f_{i'}(x_{i'})g_i(x_i|x_{i'}) + f_{i'}(x_{i'})g_i(x_i|x_{i'})} = 1 - \alpha_b(x_{i'}, i'|x_i, i) ,\end{aligned}$$

where i' is the index in $\{1, 2\}$ different from i. The probability of $i^{(t+1)} = i^{(t)}$ given $x_{-i^{(t)}}^{(t)} = z$ and $i^{(t)} = i$ can then be obtained integrating out $x_{i^{(t)}}^{(t)}$ and equals $r(z, i)$. The term $g_i(z'|z)\mathbb{1}(i \neq i')\alpha_b(z', i'|z, i)$ in the kernel in (15) follows easily from steps 1'-3' above. □

Proof of Proposition 5. By construction (see steps 1 and 2 of TGS), $(z^{(t)}, i^{(t)})$ for $t \geq 1$ is conditionally independent of $\boldsymbol{x}^{(0)}$ given $(z^{(0)}, i^{(1)})$. Therefore $\mathcal{L}(z^{(t)}, i^{(t)}|\boldsymbol{x}^{(0)} = \boldsymbol{x}_*) = \mathcal{L}(z^{(t)}, i^{(t)}|(z^{(0)}, i^{(1)}) \sim \mathcal{L}(z^{(0)}, i^{(1)}|\boldsymbol{x}^{(0)} = \boldsymbol{x}_*))$. Since $\mathcal{L}(z^{(0)}, i^{(1)}|\boldsymbol{x}^{(0)} = (z_*, z_*)) = \delta_{z_*} \otimes \mathrm{U}(\{1,2\})$, where δ_{z_*} denotes a delta measure on z_* and $\mathrm{U}(\{1,2\})$ a uniform distribution on $\{1,2\}$, we obtain from equation (9) in Proposition 4

$$\|\mathcal{L}(\boldsymbol{x}^{(t)}|\boldsymbol{x}^{(0)} = \boldsymbol{x}_*) - fZ\|_{TV} = \|\mathcal{L}(z^{(t)}, i^{(t)}|z^{(0)} = z_*, i^{(1)} \sim \mathrm{U}(\{1,2\})) - \pi\|_{TV} \,.$$

Consider then the two distributions $\mathcal{L}(z^{(t)}, i^{(t)}|z^{(0)} = z_*, i^{(1)} \sim \mathrm{U}(\{1,2\}))$ and π on $\mathbb{R} \times \{1,2\}$. From Lemma 3 $\pi(z,i) = \frac{1}{2}f_2(z)\mathbb{1}(i=1) + \frac{1}{2}f_1(z)\mathbb{1}(i=2)$ and thus the distribution of i conditional on z under π is uniform on $\{1,2\}$ for any z because $f_1 = f_2$. Also the conditional distribution of $i^{(t)}$ conditional on $z^{(t)}$ under $\mathcal{L}(z^{(t)}, i^{(t)}|z^{(0)} = z_*, i^{(1)} \sim \mathrm{U}(\{1,2\}))$ is uniform on $\{1,2\}$ for any value of $z^{(t)}$ because

$$\begin{aligned}
&Pr(i^{(t)} = 1|z^{(t)}, z^{(0)} = z_*, i^{(1)} \sim \mathrm{U}(\{1,2\})) \\
&= \frac{1}{2}Pr(i^{(t)} = 1|z^{(t)}, z^{(0)} = z_*, i^{(1)} = 1) + \frac{1}{2}Pr(i^{(t)} = 1|z^{(t)}, z^{(0)} = z_*, i^{(1)} = 2) \\
&= \frac{1}{2}Pr(i^{(t)} = 2|z^{(t)}, z^{(0)} = z_*, i^{(1)} = 2) + \frac{1}{2}Pr(i^{(t)} = 2|z^{(t)}, z^{(0)} = z_*, i^{(1)} = 1) \\
&= Pr(i^{(t)} = 2|z^{(t)}, z^{(0)} = z_*, i^{(1)} \sim \mathrm{U}(\{1,2\}))\,,
\end{aligned}$$

where the equality between the second and third line follows from exchangeability of f. Therefore, since $\mathcal{L}(z^{(t)}, i^{(t)}|z^{(0)} = z_*, i^{(1)} \sim \mathrm{U}(\{1,2\}))$ and π have same conditionals over $\{1,2\}$, Lemma 2 implies that

$$\|\mathcal{L}(z^{(t)}, i^{(t)}|z^{(0)} = z_*, i^{(1)} \sim \mathrm{U}(\{1,2\})) - \pi\|_{TV} = \|\mathcal{L}(z^{(t)}|z^{(0)} = z_*, i^{(1)} \sim \mathrm{U}(\{1,2\})) - f_1\|_{TV} \,.$$

Finally, $\mathcal{L}(z^{(t)}|z^{(0)} = z_*, i^{(1)} \sim \mathrm{U}(\{1,2\})) = \mathcal{L}(z^{(t)}|z^{(0)} = z_*)$ implies the desired result. □

A.3 Proof of Proposition 6. Let $P(\boldsymbol{x}, \boldsymbol{y})$ be the transition probability of wTGS. Clearly $P(\boldsymbol{x}, \boldsymbol{y}) = 0$ unless $\boldsymbol{x}_{-i} = \boldsymbol{y}_{-i}$ for some $i \in \{1, \dots, p\}$. If $\boldsymbol{x}_{-i} = \boldsymbol{y}_{-i}$, by construction it holds

$$P(\boldsymbol{x}, \boldsymbol{y}) = \frac{p_i(\boldsymbol{x})}{\sum_j p_j(\boldsymbol{x})} g(y_i|\boldsymbol{x}_{-i}) = \frac{w_i(\boldsymbol{x}_{-i})g(x_i|\boldsymbol{x}_{-i})}{d\,Z(\boldsymbol{x})f(x_i|\boldsymbol{x}_{-i})} g(y_i|\boldsymbol{x}_{-i})\,.$$

From the latter equality and exploiting $\boldsymbol{x}_{-i} = \boldsymbol{y}_{-i}$ we obtain

$$\frac{P(\boldsymbol{x}, \boldsymbol{y})}{P(\boldsymbol{y}, \boldsymbol{x})} = \frac{Z(\boldsymbol{y})}{Z(\boldsymbol{x})}\frac{f(y_i|\boldsymbol{y}_{-i})}{f(x_i|\boldsymbol{x}_{-i})} = \frac{Z(\boldsymbol{y})f(\boldsymbol{y})}{Z(\boldsymbol{x})f(\boldsymbol{x})}\,,$$

which implies reversibility with respect to $Z(\boldsymbol{x})f(\boldsymbol{x})$, up to proportionality.

The frequency of updating of the i-th coordinate coincides with the marginal distribution of i in the joint extended target of $(\boldsymbol{x}, i)$, which is

$$\tilde{f}(\boldsymbol{x}, i) = \frac{w_i(\boldsymbol{x}_{-i})f(\boldsymbol{x}_{-i})g(x_i|\boldsymbol{x}_{-i})}{\sum_j \mathbb{E}_{\boldsymbol{x}\sim f(\boldsymbol{x})}[w_j(\boldsymbol{x}_{-j})]}\,. \tag{16}$$

Integrating over $\boldsymbol{x}$ we obtain the marginal distribution of i, namely

$$\begin{aligned}
\tilde{f}(i) &= \int_{\mathcal{X}} \frac{w_i(\boldsymbol{x}_{-i})f(\boldsymbol{x}_{-i})g(x_i|\boldsymbol{x}_{-i})}{\sum_j \mathbb{E}_{\boldsymbol{x}\sim f(\boldsymbol{x})}[w_j(\boldsymbol{x}_{-j})]} d\boldsymbol{x} \\
&= \frac{\int_{\mathcal{X}_{-i}} w_i(\boldsymbol{x}_{-i})f(\boldsymbol{x}_{-i})d\boldsymbol{x}_{-i}}{\sum_j \mathbb{E}_{\boldsymbol{x}\sim f(\boldsymbol{x})}[w_j(\boldsymbol{x}_{-j})]} = \frac{\mathbb{E}_{\boldsymbol{x}\sim f(\boldsymbol{x})}[w_i(\boldsymbol{x}_{-i})]}{\sum_j \mathbb{E}_{\boldsymbol{x}\sim f(\boldsymbol{x})}[w_j(\boldsymbol{x}_{-j})]}\,.
\end{aligned}$$

B Implementation of TGS for Bayesian Variable Selection

B.1 Efficient computation of conditional probabilities. Given $\gamma \in \{0,1\}^p$ we are interested in computing $\{p(\gamma_i|Y,\gamma_{-i})\}_{i=1}^p$ efficiently. This can be done by computing the ratios $\{\frac{p(\gamma_i|Y,\gamma_{-i})}{p(1-\gamma_i|Y,\gamma_{-i})}\}_{i=1}^p$ and then using $p(\gamma_i|Y,\gamma_{-i}) = \frac{p(\gamma_i|Y,\gamma_{-i})}{p(1-\gamma_i|Y,\gamma_{-i})}\left(1+\frac{p(\gamma_i|Y,\gamma_{-i})}{p(1-\gamma_i|Y,\gamma_{-i})}\right)^{-1}$. If $\Sigma_\gamma = c(X_\gamma^T X_\gamma)^{-1}$ then

$$\frac{p(\gamma_i = 1|Y,\gamma_{-i})}{p(\gamma_i = 0|Y,\gamma_{-i})} = \frac{p(\gamma_i = 1|\gamma_{-i})}{p(\gamma_i = 0|\gamma_{-i})}\frac{1}{(1+c)^{1/2}}\left(\frac{S(\gamma^0)}{S(\gamma^1)}\right), \tag{17}$$

where γ^1 is given by γ_{-i} with $\gamma_i = 1$ and γ^0 is given by γ_{-i} with $\gamma_i = 1$, and

$$S(\gamma) = y^T y - \frac{c}{1+c} y^T X_\gamma (X_\gamma^T X_\gamma)^{-1} X_\gamma^T y\,,$$

see for example Smith and Kohn [1996], Chipman et al. [2001]. We now provide simple linear algebra results to compute (17) efficiently. Define $F = (X_{\gamma^0}^T X_{\gamma^0})^{-1}$, $v = X^T y$ and $v_{\gamma^0} = (v_j)_{j\,:\,\gamma_i^0=1}$. Also define $A = X^T X$ and $a_i = (A_{ji})_{j\,:\,\gamma_i^0=1}$. Then it holds

$$S(\gamma^1) = S(\gamma^0) - \frac{c}{1+c} d_i (v_\gamma^T F a_i - v_i)^2 \tag{18}$$

where $d_i = (A_{ii} - a_i^T F a_i)^{-1}$. Equation (18) allows to compute $S(\gamma^1)$ efficiently given $S(\gamma^0)$. To further facilitate computation, define the Cholesky decomposition $L = Chol(F)$, i.e. a lower triangular matrix such that such that $F_\gamma = LL^T$, and write d_i as

$$d_i = \sum_{j\in I} (a_i L)_j^2\,. \tag{19}$$

The latter allows to compute $(d_i)_{i\,:\,\gamma_i=0}$ efficiently noting that $(d_i)_i$ are the squared ℓ^2-norms of the rows of the matrix BL, where B is the $p \times |\gamma|$ matrix made of the columns of A corresponding to variables included in γ. The expressions in (17)-(19) allows to compute the probabilities $\{p_i(\gamma)\}_{i=1}^p$ needed by TGS in a fully vectorised way, resulting in a mild computational overhead even for large values of p. An R code implementation is available at `https://github.com/gzanella/TGS`.

Proof of (18). Define $v = X^T y$ and $A = X^T X$, $F_\gamma = (X_\gamma^T X_\gamma)^{-1}$ and $L_\gamma = Chol(F_\gamma)$ a lower triangular matrix such that such that $F_\gamma = L_\gamma L_\gamma^T$. Then

$$y^T X_\gamma F_\gamma X_\gamma^T y = v_\gamma^T L_\gamma L_\gamma^T v_\gamma = \|v_\gamma^T L_\gamma\|_2^2\,.$$

Also, assuming that we add the j-th variable at the end of γ^0, we have

$$F_{\gamma^1} = \begin{pmatrix} F_{\gamma^0} + d_j F_{\gamma^0} a_j a_j^T F_{\gamma^0} & -d_j F_{\gamma^0} a_j \\ -d_j a_j^T F_{\gamma^0}^T & d_j \end{pmatrix} \tag{20}$$

where $d_j = (a_{jj} - a_j^T L_{\gamma^0} L_{\gamma^0}^T a_j)^{-1}$. The equality in (20) is easy to check noting that $F_{\gamma^1} A_{\gamma^1} = \mathbb{I}_{|\gamma^1|}$ using $F_{\gamma^0} = A_{\gamma^0}^{-1}$ and

$$A_{\gamma^1} = \begin{pmatrix} A_{\gamma^0} & a_j \\ a_j^T & a_{jj} \end{pmatrix}.$$

It follows that

$$\begin{aligned} v_{\gamma^1}^T F_{\gamma^1} v_{\gamma^1} =& \left(\begin{array}{cc} v_{\gamma^0}^T & v_j \end{array} \right) F_{\gamma^1} \left(\begin{array}{c} v_{\gamma^0} \\ v_j \end{array} \right) \\ =& v_{\gamma^0}^T F_{\gamma^1} v_{\gamma^0} + d_j (v_{\gamma^0}^T F_{\gamma^0} a_j)^2 - 2 d_j v_{\gamma^0}^T F_{\gamma^0} a_j + d_j v_j^2 \\ =& v_{\gamma^0}^T F_{\gamma^1} v_{\gamma^0} + d_j (v_{\gamma^0}^T F_{\gamma^0} a_j - v_j)^2 \, . \end{aligned}$$

Equation (18) follows. □

B.2 Rao-Blackwellised estimators. Given T samples $(\gamma^{(t)})_{t=1}^T$ from $p(\gamma|Y)$ the standard Monte Carlo estimate of $p(\gamma_i = 1|Y)$ based on frequencies is $T^{-1} \sum_{t=1}^T \mathbb{1}(\gamma_i^{(t)} = 1)$, where $\mathbb{1}$ denotes the indicator function. An improved Rao-Blackwellised estimator is given by $T^{-1} \sum_{t=1}^T p(\gamma_i = 1|Y, \gamma_{-i} = \gamma_{-i}^{(t)})$, see Robert and Casella [2004, Sec 9.3] for more details on Rao-Blackwellisation for the Gibbs Sampler. For TGS and wTGS, we need to take into account of the importance weights $w_t = Z(\gamma^{(t)})^{-1}$ and so the estimators become $\sum_{t=1}^T w_t p(\gamma_i = 1|Y, \gamma_{-i} = \gamma_{-i}^{(t)})/(\sum_{t=1}^T w_t)$. Note that, having already computed $\{p(\gamma_i = \gamma_i^{(t)}|Y, \gamma_{-i} = \gamma_{-i}^{(t)})\}_{i=1}^p$ in step 1 of TGS and wTGS, the latter estimators can be computed at no extra cost. See also Ghosh and Clyde [2011], Guan and Stephens [2011], Rossell et al. [2017] and Griffin et al. [2018] for other examples of Rao-Blackwellisation in the context of Bayesian Variable Selection.

B.3 Further experiments and references for BVS computation. We performed additional simulations and comparisons on computational methods for discrete spike and slab BVS models to assess the competitiveness of the wTGS scheme implemented in Section 4. There is a large literature on the topic and the few references reported here are far from exhaustive. We tested some available *R* packages to fit BVS models. First we tried to fit the model under consideration to the TGFB data using the *BayesVarSel* [Garcia-Donato and Forte, 2017] and *BoomSpikeSlab* [Scott, 2017] *R* packages. Both packages implement the Gibbs Sampler (as in George and McCulloch, 1997) and *BoomSpikeSlab* further implements the data augmentation scheme of Ghosh and Clyde [2011]. We run both implementations for 2 hours without being able to obtain reliable estimates for the PIPs. We then considered the *mombf* [Rossell et al., 2017] *R* package, which implements a deterministic-scan Gibbs Sampler and uses simple Rao-Blackwellisation for the PIPs estimation (i.e. Rao-Blackwellised estimation of $p(\gamma_i|Y, \gamma_{-i})$ only for the γ_i that is currently updated). We found the *mombf* implementation of Gibbs Sampling to be more scalable to large p, as we managed to obtain reliable estimates for the PIPs on the TGFB data (comparable to the one of wTGS in Figure 6) in roughly 8 minutes. We thus implemented in R the same scheme (deterministic-scan Gibbs Sampler with simple Rao-Blackwellisation) obtaining performances comparable to the ones of GS in Figure 6 and a cost per iteration of one to two orders of magnitude higher than the *mombf* implementation. This suggests that the fact of *mombf* having performances close to the one of wTGS is mainly due to the use of a lower-level programming language, in this case C++.

In terms of alternative methodology, the most scalable scheme we found in the literature is a specialised adaptive MCMC sampler recently proposed by Griffin et al. [2018]. Nonetheless, the authors report a runtime of 2.5 hours (with a MATLAB implementation and similar processor) to obtain reliable PIPs estimates for a dataset with $p = 22576$. For comparison, our R implementation of wTGS requires 5 to 10 minutes to produce reliable estimates for a simulated dataset of the same size. The larger computational burden required by their scheme could be due to the need of running multiple chains simultaneously (they suggest to use 25 chains) during the adaptation phase to learn the large number of adaptation parameters. The authors also implemented a Metropolis-Hastings Add-Delete-Swap scheme, the Hamming Ball sampler of Titsias and Yau [2017] and the adaptive sampler of Ji and Schmidler [2013] on the same large p dataset, reporting them to provide significantly worse performances. Note that, for strongly correlated variables, Metropolis-Hastings schemes that include "swap" moves as in, e.g., Brown et al. [1998] can help moving across the modes, e.g. the two modes corresponding to $(\gamma_1, \gamma_2) = (1, 0)$ and $(\gamma_1, \gamma_2) = (0, 1)$ in the illustrative example in Section 4.5. However, the probability of picking the two correlated variables in the proposal will be low, especially in a large p context, and thus the addition of swap moves would alleviate but not solve the problem (see also simulation study in Griffin et al., 2018).

References

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

A. Calon, E. Espinet, S. Palomo-Ponce, D. V. Tauriello, M. Iglesias, M. V. Céspedes, M. Sevillano, C. Nadal, P. Jung, X. H.-F. Zhang, et al. Dependency of colorectal cancer on a TGF-β-driven program in stromal cells for metastasis initiation. *Cancer cell*, 22 (5):571–584, 2012.

C. Chimisov, K. Latuszynski, and G. Roberts. Adapting the gibbs sampler. *arXiv preprint arXiv:1801.09299*, 2018.

H. A. Chipman, E. I. George, and R. E. McCulloch. The Practical Implementation of Bayesian Model Selection. *Institute of Mathematical Statistics Lecture Notes-Monograph Series*, 38:65, 2001.

L. L. Duan, J. E. Johndrow, and D. B. Dunson. Scaling up Data Augmentation MCMC via Calibration. *arXiv preprint arXiv:1703.03123*, 2017.

C. Fernandez, E. Ley, and M. F. Steel. Benchmark priors for bayesian model averaging. *Journal of Econometrics*, 100(2):381–427, 2001.

G. Garcia-Donato and A. Forte. *BayesVarSel: Bayes Factors, Model Choice and Variable Selection in Linear Models*, 2017. URL `https://CRAN.R-project.org/package=BayesVarSel`. R package version 1.7.1.

A. E. Gelfand, S. K. Sahu, and B. P. Carlin. Efficient parametrisations for normal linear mixed models. *Biometrika*, 82(3):479–488, 1995.

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

C. J. Geyer and E. A. Thompson. Annealing Markov chain Monte Carlo with applications to ancestral inference. *Journal of the American Statistical Association*, 90(431):909–920, 1995.

J. Ghosh and M. A. Clyde. Rao–blackwellization for bayesian variable selection and model averaging in linear and binary regression: A novel data augmentation approach. *Journal of the American Statistical Association*, 106(495):1041–1052, 2011.

R. Gramacy, R. Samworth, and R. King. Importance tempering. *Statistics and Computing*, 20(1):1–7, 2010.

J. Griffin, K. Latuszynski, and M. Steel. In Search of Lost (Mixing) Time: Adaptive Markov chain Monte Carlo schemes for Bayesian variable selection with very large p. *arXiv preprint arXiv:1708.05678*, 2018.

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

S. E. Hills and A. F. Smith. Parameterization issues in bayesian inference. *Bayesian statistics*, 4:227–246, 1992.

X. Huang, J. Wang, and F. Liang. A Variational Algorithm for Bayesian Variable Selection. *arXiv preprint arXiv:1602.07640*, 2016.

C. Ji and S. C. Schmidler. Adaptive markov chain Monte Carlo for Bayesian variable selection. *Journal of Computational and Graphical Statistics*, 22(3):708–728, 2013.

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

F. Leisen and A. Mira. An extension of peskun and tierney orderings to continuous time markov chains. *Statistica Sinica*, pages 1641–1651, 2008.

J. S. Liu. Peskun's theorem and a modified discrete-state gibbs sampler. *Biometrika*, 83 (3), 1996.

E. Marinari and G. Parisi. Simulated tempering: a new Monte Carlo scheme. *EPL (Europhysics Letters)*, 19(6):451, 1992.

S. P. Meyn and R. L. Tweedie. *Markov chains and stochastic stability.* Communications and Control Engineering Series. Springer-Verlag London, Ltd., London, 1993. ISBN 3-540-19832-6. doi: 10.1007/978-1-4471-3267-7. URL `http://dx.doi.org/10.1007/978-1-4471-3267-7`.

A. B. Owen. *Monte Carlo theory, methods and examples.* 2013. available at http://statweb.stanford.edu/ owen/mc/.

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

R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2017. URL `https://www.R-project.org/`.

C. Robert and G. Casella. Monte Carlo Statistical Methods. 2004.

G. O. Roberts and J. S. Rosenthal. Markov Chains and De-initializing Processes. *Scandinavian Journal of Statistics*, 28(3):489–504, 2001.

G. O. Roberts and J. S. Rosenthal. Surprising convergence properties of some simple gibbs samplers under various scans. *International Journal of Statistics and Probability*, 5(1): 51, 2015.

G. O. Roberts and S. K. Sahu. Updating schemes, correlation structure, blocking and parameterization for the Gibbs sampler. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 59(2):291–317, 1997.

G. O. Roberts and A. F. M. Smith. Simple conditions for the convergence of the Gibbs sampler and Metropolis-Hastings algorithms. *Stochastic Process. Appl.*, 49(2):207–216, 1994. ISSN 0304-4149. doi: 10.1016/0304-4149(94)90134-1. URL `http://dx.doi.org/`

10.1016/0304-4149(94)90134-1.

D. Rossell and F. J. Rubio. Tractable bayesian variable selection: beyond normality. *Journal of the American Statistical Association*, (just-accepted), 2017.

D. Rossell and D. Telesca. Nonlocal priors for high-dimensional estimation. *Journal of the American Statistical Association*, 112(517):254–265, 2017.

D. Rossell, J. D. Cook, D. Telesca, and P. Roebuck. *mombf: Moment and Inverse Moment Bayes Factors*, 2017. URL https://CRAN.R-project.org/package=mombf. R package version 1.9.5.

S. L. Scott. *BoomSpikeSlab: MCMC for Spike and Slab Regression*, 2017. URL https://CRAN.R-project.org/package=BoomSpikeSlab. R package version 0.9.0.

A. F. Smith and A. E. Gelfand. Bayesian statistics without tears: a sampling–resampling perspective. *The American Statistician*, 46(2):84–88, 1992.

M. Smith and R. Kohn. Nonparametric regression using bayesian variable selection. *Journal of Econometrics*, 75(2):317–343, 1996.

M. K. Titsias and C. Yau. The Hamming ball sampler. *Journal of the American Statistical Association*, pages 1–14, 2017.

S. Wang, B. Nan, S. Rosset, and J. Zhu. Random lasso. *The annals of applied statistics*, 5(1):468, 2011.

Y. Yang, M. J. Wainwright, and M. I. Jordan. On the computational complexity of high-dimensional bayesian variable selection. *The Annals of Statistics*, 44(6):2497–2532, 2016.

T. Yuan, X. Huang, M. Woodcock, M. Du, R. Dittmar, Y. Wang, S. Tsai, M. Kohli, L. Boardman, T. Patel, et al. Plasma extracellular rna profiles in healthy and cancer patients. *Scientific reports*, 6:19413, 2016.

G. Zanella. Informed proposals for local mcmc in discrete spaces. *arXiv preprint arXiv:1711.07424*, 2017.

G. Zanella and G. O. Roberts. Analysis of the Gibbs Sampler for Gaussian hierarchical models via multigrid decomposition. *arXiv preprint arXiv:1703.06098*, 2017.

A. Zellner. 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*, pages 233–243, 1986.