
Bayesian Variable Selection in a Million Dimensions

Martin Jankowiak

`martin@basis.ai`

Broad Institute Basis Research Institute

Abstract

Bayesian variable selection is a powerful tool for data analysis, as it offers a principled method for variable selection that accounts for prior information and uncertainty. However, wider adoption of Bayesian variable selection has been hampered by computational challenges, especially in difficult regimes with a large number of covariates P or non-conjugate likelihoods. To scale to the large P regime we introduce an efficient Markov Chain Monte Carlo scheme whose cost per iteration is sublinear in P (though linear in the number of data points). In addition we show how this scheme can be extended to generalized linear models for count data, which are prevalent in biology, ecology, economics, and beyond. In particular we design efficient algorithms for variable selection in binomial and negative binomial regression, which includes logistic regression as a special case. In experiments we demonstrate the effectiveness of our methods, including on cancer and maize genomic data.

set of models, with distinct models including distinct sets of covariates.

The Bayesian formulation of this approach, known as Bayesian variable selection in the literature, offers a powerful set of techniques for realizing Occam’s razor in this setting (George and McCulloch, 1993, 1997; Chipman et al., 2001; Dellaportas et al., 2002; O’Hara et al., 2009). Despite the intuitive appeal of this approach, approximating the resulting posterior distribution can be computationally challenging. A principal reason for this is the astronomical size of the model space that results whenever there are more than a few dozen covariates. Indeed for P covariates the total number of distinct models, namely 2^P , exceeds the estimated number of atoms in the known universe ($\sim 10^{80}$) for $P \gtrsim 266$. Since, however, large P is common in many application areas, e.g. genetics, it is essential for methods to scale to this regime. In addition for many models of interest non-conjugate likelihoods make it infeasible to integrate out real-valued model parameters, resulting in a challenging high-dimensional inference problem defined on a transdimensional mixed discrete/continuous latent space.

In this work we develop efficient Markov Chain Monte Carlo (MCMC) methods for Bayesian variable selection. Our contributions include the following:

1 Introduction

Generalized linear models are ubiquitous throughout applied statistics and data analysis (McCullagh and Nelder, 2019). One reason for their popularity is their immediate interpretability, which results from the introduction of explicit parameters that encode how the observed response depends on each covariate. In the scientific setting this interpretability is of central importance. Indeed model fit is often a secondary concern, and the primary goal is to identify a *parsimonious* explanation of the observed data. This is naturally viewed as a model selection problem, in particular one in which the model space is defined as a nested

1. We introduce an efficient MCMC sampler for large P whose cost per iteration is sublinear in P .
2. We develop efficient MCMC samplers for two generalized linear models for count data: i) binomial regression and the special case of logistic regression; and ii) negative binomial regression.
3. We show how the algorithmic strategies in 1 and 2 can be combined and how they can accommodate inference over the prior inclusion probability.
4. We provide a unit-tested, easy-to-use, GPU-enabled, open source implementation of our methods at the following URL:

<https://github.com/BasisResearch/millipede>

2 Background

2.1 Problem Setup

Consider linear regression with $X \in \mathbb{R}^{N \times P}$ and $Y \in \mathbb{R}^N$ and define the following space of models:

$$\begin{aligned} \text{[inclusion variables]} \quad & \gamma_i \sim \text{Bernoulli}(h) & (1) \\ \text{[noise variance]} \quad & \sigma_\gamma^2 \sim \text{InvGamma}(\tfrac{1}{2}\nu_0, \tfrac{1}{2}\nu_0\lambda_0) \\ \text{[coefficients]} \quad & \beta_\gamma \sim \mathcal{N}(0, \sigma_\gamma^2 \tau^{-1} \mathbb{1}_{|\gamma|}) \\ \text{[responses]} \quad & Y_n \sim \mathcal{N}(\beta_\gamma \cdot X_{n\gamma}, \sigma_\gamma^2) \end{aligned}$$

where $i = 1, \dots, P$ and $n = 1, \dots, N$. Here each $\gamma_i \in \{0, 1\}$ controls whether the coefficient β_i and the i^{th} covariate are included ($\gamma_i = 1$) or excluded ($\gamma_i = 0$) from the model. In the following we use γ to refer to the vector $(\gamma_1, \dots, \gamma_P)$. The hyperparameter $h \in (0, 1)$ controls the overall level of sparsity; in particular hP is the expected number of covariates included a priori. The $|\gamma|$ coefficients $\beta_\gamma \in \mathbb{R}^{|\gamma|}$ are governed by a Normal prior with precision proportional to $\tau > 0$.¹ Here $|\gamma| \in \{0, 1, \dots, P\}$ denotes the total number of included covariates. The response Y_n is generated from a Normal distribution with variance governed by an Inverse Gamma prior.² Note that we do not include a bias term in Eqn. 1, but doing so may be desirable in practice. An attractive feature of Eqn. 1 is that it explicitly reasons about variable inclusion and allows us to define *posterior inclusion probabilities* or PIPs:

$$\text{PIP}(i) \equiv \text{p}(\gamma_i = 1 | \mathcal{D}) \in [0, 1] \quad (2)$$

where $\mathcal{D} = \{X, Y\}$ is the observed dataset.

2.2 Inference

Conjugacy in Eqn. 1 implies that the coefficients β and the variance σ^2 can be integrated out, resulting in a discrete inference problem over $\{0, 1\}^P$ (Chipman et al., 2001). Inference over $\{0, 1\}^P$ readily admits a Gibbs sampling scheme; however, the resulting sampler is notoriously slow in high dimensions. For example, consider the scenario in which the two covariates corresponding to $i = 1$ and $i = 2$ are highly correlated and each on its own is sufficient for explaining the responses Y . In this scenario the posterior concentrates on models with $\gamma = (1, 0, 0, \dots)$ and $\gamma = (0, 1, 0, \dots)$. Single-site Gibbs updates w.r.t. γ_i will move between the two modes very infrequently, since they are separated by low probability models like $\gamma = (0, 0, 0, \dots)$.

A recently developed inference algorithm—Tempered Gibbs Sampling (TGS) (Zanella and Roberts, 2019)—utilizes *coordinatewise tempering* to cope with this kind

¹We usually drop the γ subscript on β_γ and σ_γ^2 to simplify the notation.

²Throughout we take the limit $\nu_0 \rightarrow 0$ and $\lambda_0 \rightarrow 0$, which corresponds to an improper prior $p(\sigma^2) \propto \sigma^{-2}$.

of problematic stickiness. In the following we describe a variant of TGS called wTGS that is particularly well-suited to Bayesian variable selection (Zanella and Roberts, 2019). As we will see, this algorithm will serve as a powerful substrate for building MCMC samplers for Bayesian variable selection that can accommodate large P as well as count-based likelihoods.

2.3 wTGS

Consider the (unnormalized) target distribution

$$\begin{aligned} f(\gamma, i) &\equiv p(\gamma | \mathcal{D}) \frac{\eta(\gamma_{-i}) U(\gamma_i)}{p(\gamma_i | \gamma_{-i}, \mathcal{D})} & (3) \\ &= U(\gamma_i) \eta(\gamma_{-i}) p(\gamma_{-i} | \mathcal{D}) & (4) \end{aligned}$$

where we have introduced an auxiliary variable $i \in \{1, \dots, P\}$. Here $U(\cdot)$ is the uniform distribution on $\{0, 1\}$ and γ_{-i} denotes all components of γ apart from γ_i . Finally $\eta(\gamma_{-i})$ is an additional weighting factor to be defined below. The key property of Eqn. 3 is that for any i the distribution over γ_i is uniform and factorizes across $\{\gamma_i, \gamma_{-i}\}$. wTGS proceeds by defining a sampling scheme for the target Eqn. 3 that utilizes Gibbs updates w.r.t. i and Metropolized-Gibbs updates w.r.t. γ_i .

i -updates If we marginalize i from Eqn. 3 we obtain

$$f(\gamma) = p(\gamma | \mathcal{D}) \phi(\gamma) \quad (5)$$

where we define

$$\phi(\gamma) \equiv \sum_{i=1}^P \frac{\frac{1}{2} \eta(\gamma_{-i})}{p(\gamma_i | \gamma_{-i}, \mathcal{D})} \quad (6)$$

As is clear from Eqn. 5, $\phi(\gamma)^{-1}$ is an *importance weight* that can be used to obtain samples from the non-tempered target of interest, i.e. $p(\gamma | \mathcal{D})$. Additionally Eqn. 3 implies that we can do Gibbs updates w.r.t. i using the distribution³

$$f(i | \gamma) = \frac{1}{\phi(\gamma)} \frac{\frac{1}{2} \eta(\gamma_{-i})}{p(\gamma_i | \gamma_{-i}, \mathcal{D})} \quad (7)$$

γ -updates The auxiliary variable i is used to control which component of γ we update in each step. Since the posterior conditional w.r.t. γ_i is the uniform distribution $U(\gamma_i)$, Metropolized-Gibbs (Liu, 1996) updates w.r.t. γ_i result in deterministic flips that are accepted with probability one: $\gamma_i \rightarrow 1 - \gamma_i$.

Weighting factor η To finish specifying wTGS we need to define the weighting factor $\eta(\gamma_{-i})$ in Eqn. 3:

$$\eta(\gamma_{-i}) = p(\gamma_i = 1 | \gamma_{-i}, \mathcal{D}) + \frac{\epsilon}{P} \quad (8)$$

³Note that Eqn. 6-7-8 depends on conditional PIPs $p(\gamma_i = 1 | \gamma_{-i}, \mathcal{D})$; as discussed in Sec. A.4 these can be computed efficiently with careful linear algebra.

Here $p(\gamma_i = 1|\gamma_{-i}, \mathcal{D})$ is a conditional PIP, and ϵ trades off between exploitation ($\epsilon \rightarrow 0$) and exploration ($\epsilon \rightarrow \infty$). Indeed since the marginal $f(i)$ is given by

$$f(i) \propto \mathbb{E}_{p(\gamma_{-i}|\mathcal{D})} [\eta(\gamma_{-i})] = \text{PIP}(i) + \frac{\epsilon}{P} \quad (9)$$

this choice of η ensures that the sampler focuses its computational effort on large PIP covariates.⁴ Following Zanella and Roberts (2019) we typically choose $\epsilon = 5$ and note that we observe little sensitivity to the precise value of ϵ . For the full algorithm see Algorithm 3 in the supplement.

Rao-Blackwellization A side benefit of computing conditional PIPs in Eqn. 7 is that they can be repurposed to compute lower variance Rao-Blackwellized PIP estimates. See Sec. A.10 for details.

3 The Large P Regime: Subset wTGS

Running wTGS in the large P regime can be prohibitively expensive, since it involves computing P conditional PIPs per MCMC iteration. We would like to devise an algorithm that, like wTGS, utilizes conditional PIPs to make informed moves in γ space while avoiding this prohibitive $\mathcal{O}(P)$ computational cost.

Subset wTGS To do so we leverage a simple augmentation strategy. In effect, we introduce an auxiliary variable $S \subset \{1, \dots, P\}$ that controls which conditional PIPs are computed in a given MCMC iteration. Since we can choose the (fixed) size S of \mathcal{S} to be much less than P , this can result in significant speed-ups.

In more detail, consider the following (unnormalized) target distribution:

$$f(\gamma, i, \mathcal{S}) \equiv p(\gamma|\mathcal{D}) \frac{\eta(\gamma_{-i})U(\gamma_i)}{p(\gamma_i|\gamma_{-i}, \mathcal{D})} U(\mathcal{S}|i, \mathcal{A}) \quad (10)$$

Here \mathcal{S} ranges over all the subsets of $\{1, \dots, P\}$ of size S that also contain a fixed ‘anchor’ set $\mathcal{A} \subset \{1, \dots, P\}$ of size $A < S$. Moreover $U(\mathcal{S}|i, \mathcal{A})$ is the uniform distribution over all size S subsets of $\{1, \dots, P\}$ that contain both i and \mathcal{A} .⁵ We choose the same weighting function η as in wTGS (see Eqn. 8). In practice we adapt \mathcal{A} during burn-in, but for now the reader can suppose that \mathcal{A} is chosen at random. Subset wTGS proceeds by defining a sampling scheme for the target distribution Eqn. 10 that utilizes Gibbs updates w.r.t. i and \mathcal{S} and Metropolized-Gibbs updates w.r.t. γ_i .

⁴See Sec. A.6 for intuition and motivation behind tempering and Sec. A.7 for further discussion of the mix of exploration and exploitation that it enables.

⁵This is reminiscent of the Hamming ball construction in Tittias and Yau (2017).

i -updates Marginalizing i from Eqn. 10 yields

$$f(\gamma, \mathcal{S}) = p(\gamma|\mathcal{D})\phi(\gamma, \mathcal{S}) \quad (11)$$

where we define

$$\phi(\gamma, \mathcal{S}) \equiv \sum_{i \in \mathcal{S}} \frac{\frac{1}{2}\eta(\gamma_{-i})}{p(\gamma_i|\gamma_{-i}, \mathcal{D})} U(\mathcal{S}|i, \mathcal{A}) \quad (12)$$

and have leveraged that $U(\mathcal{S}|i, \mathcal{A}) = 0$ if $i \notin \mathcal{S}$. Crucially, computing $\phi(\gamma, \mathcal{S})$ is $\mathcal{O}(S)$ instead of $\mathcal{O}(P)$. We can do Gibbs updates w.r.t. i using the distribution

$$f(i|\gamma, \mathcal{S}) \propto \frac{\eta(\gamma_{-i})}{p(\gamma_i|\gamma_{-i}, \mathcal{D})} U(\mathcal{S}|i, \mathcal{A}) \quad (13)$$

γ -updates Just as for wTGS we utilize Metropolized-Gibbs updates w.r.t. γ_i that result in deterministic flips $\gamma_i \rightarrow 1 - \gamma_i$. Likewise the marginal $f(i)$ is proportional to $\text{PIP}(i) + \frac{\epsilon}{P}$ so that the sampler focuses computational effort on large PIP covariates.

\mathcal{S} -updates \mathcal{S} is updated with Gibbs moves, $\mathcal{S} \sim U(\cdot|i, \mathcal{A})$. For the full algorithm see Algorithm 1.

Importance weights The Markov chain in Algorithm 1 targets the auxiliary distribution Eqn. 10. To obtain samples from the desired posterior $p(\gamma|\mathcal{D})$ we reweight each sample (γ, \mathcal{S}) with an importance weight $\tilde{\rho} = \phi(\gamma, \mathcal{S})^{-1}$ and discard \mathcal{S} ; see Eqn. 11. Crucially, the importance weights are upper bounded and exhibit only moderate variance. Ultimately this moderate variance can be traced to the coordinatewise tempering, which keeps the tempering to a modest level; see Sec. A.8 for additional discussion. Moreover, we can show that samples obtained with Algorithm 1 can be used to estimate posterior quantities of interest:

Proposition 1 *The Subset wTGS estimator satisfies*

$$\sum_{t=1}^T \rho^{(t)} h(\gamma^{(t)}) \rightarrow \mathbb{E}_{p(\gamma|\mathcal{D})} [h(\gamma)] \quad \text{as } T \rightarrow \infty \quad (14)$$

almost surely for every test function $h(\gamma) : \{0, 1\}^P \rightarrow \mathbb{R}$, where $\rho^{(t)} \propto \phi^{-1}(\gamma^{(t)}, \mathcal{S}^{(t)})$ are normalized weights. Moreover, we can use a (partially) Rao-Blackwellized PIP estimator in Eqn. 14. See Sec. A.9 in the supplement for the proof and additional details.

4 Binomial Regression: PG-wTGS

For simplicity we focus on the binomial regression case, leaving a discussion of the negative binomial case to Sec. A.14 in the supplement. Let $X \in \mathbb{R}^{N \times P}$, $C \in \mathbb{Z}_{>0}^N$, and $Y \in \mathbb{Z}_{\geq 0}^N$ with $Y \leq C$ and consider the following

Algorithm 1: We outline the main steps in Subset wTGS. See Sec. 3 for details. Subset wTGS reduces to wTGS in the limit $S \rightarrow P$, in which case \mathcal{S} becomes redundant. Superscripts indicate MCMC iterations.

Input: Dataset $\mathcal{D} = \{X, Y\}$ with P covariates; prior inclusion probability h ; prior precision τ ; subset size S ; anchor set size A ; total number of MCMC iterations T ; number of burn-in iterations T_{burn} .

Output: Approximate weighted posterior samples $\{\rho^{(t)}, \gamma^{(t)}\}_{t=T_{\text{burn}}+1}^T$

- 1 Let $\gamma^{(0)} = (0, \dots, 0)$ and choose \mathcal{A} to be the A covariate indices exhibiting the largest correlations with Y .
- 2 Choose $i^{(0)}$ randomly from $\{1, \dots, P\}$ and $\mathcal{S}^{(0)} \sim U(\cdot | i^{(0)}, \mathcal{A})$.
- 3 **for** $t = 1, \dots, T$ **do**
- 4 Sample $i^{(t)} \sim f(\cdot | \gamma^{(t-1)}, \mathcal{S}^{(t-1)})$ using Eqn. 13
- 5 Let $\gamma^{(t)} = \text{flip}(\gamma^{(t-1)} | i^{(t)})$ where $\text{flip}(\gamma | i)$ flips the i^{th} coordinate of γ : $\gamma_i \rightarrow 1 - \gamma_i$.
- 6 Sample $\mathcal{S}^{(t)} \sim U(\cdot | i^{(t)}, \mathcal{A})$ and compute the unnormalized weight $\tilde{\rho}^{(t)} = \phi(\gamma^{(t)}, \mathcal{S}^{(t)})^{-1}$ using Eqn. 12.
- 7 If $t \leq T_{\text{burn}}$ adapt \mathcal{A} using the scheme described in Sec. A.13.
- 8 Compute the normalized weights $\rho^{(t)} = \frac{\tilde{\rho}^{(t)}}{\sum_{s>T_{\text{burn}}} \tilde{\rho}^{(s)}}$ for $t = T_{\text{burn}} + 1, \dots, T$.
- 9 **return** $\{\rho^{(t)}, \gamma^{(t)}\}_{t=T_{\text{burn}}+1}^T$

space of generalized linear models:

[inclusion variables]	$\gamma_i \sim \text{Bernoulli}(h)$	(15)
[bias term]	$\beta_0 \sim \mathcal{N}(0, \tau_{\text{bias}}^{-1})$	
[coefficients]	$\beta_\gamma \sim \mathcal{N}(0, \tau^{-1} \mathbb{1}_{ \gamma })$	
[success logits]	$\psi_n \equiv \beta_0 + \beta_\gamma \cdot X_{n\gamma}$	
[responses]	$Y_n \sim \text{Binomial}(C_n, \sigma(\psi_n))$	

where $i = 1, \dots, P$ and $n = 1, \dots, N$. Note that we introduce a bias term β_0 governed by a Normal prior with precision $\tau_{\text{bias}} > 0$; we assume that β_0 is always included in the model.⁶ The response Y_n is generated from a Binomial distribution with total count C_n and success probability $\sigma(\psi_n)$, where $\sigma(\cdot)$ denotes the logistic function $\sigma(x) \equiv \{1 + \exp(-x)\}^{-1}$. This reduces to logistic regression with binary responses if $C_n = 1$ for all n .

4.1 Pòlya-Gamma augmentation

wTGS relies on conditional PIPs to construct informed moves; unfortunately these cannot be computed in closed form for non-conjugate likelihoods like that in Eqn. 15. To get around this we introduce Pòlya-Gamma auxiliary variables, which rely on the identity

$$\frac{(e^\psi)^a}{(1 + e^\psi)^b} = \frac{1}{2^b} e^{(a - \frac{1}{2}b)\psi} \mathbb{E}_{\text{PG}(\omega|b,0)} [\exp(-\frac{1}{2}\omega\psi^2)]$$

noted by Polson et al. (2013). Here $a, \psi \in \mathbb{R}$, $b > 0$, and $\text{PG}(\omega|b, 0)$ is the Pòlya-Gamma distribution, which has support on the positive real axis. Using this identity we can introduce a N -dimensional vector of Pòlya-Gamma (PG) variates ω governed by the prior $\omega_n \sim \text{PG}(C_n, 0)$ and

⁶Note that β_0 has an implicit γ subscript that we elide (different models have different biases). For simplicity we take $\tau = \tau_{\text{bias}}$ throughout.

rewrite the Binomial likelihood in Eqn. 15 as follows

$$\begin{aligned} p(Y_n | C_n, \sigma(\psi_n)) &\propto \sigma(\psi_n)^{Y_n} (1 - \sigma(\psi_n))^{C_n - Y_n} & (16) \\ &= \frac{(\exp(-\psi_n))^{C_n - Y_n}}{(1 + \exp(-\psi_n))^{C_n}} = \frac{(\exp(\psi_n))^{Y_n}}{(1 + \exp(\psi_n))^{C_n}} \end{aligned}$$

so that each likelihood term in Eqn. 15 is replaced with

$$\exp(\kappa_n \psi_n - \frac{1}{2} \omega_n \psi_n^2) \quad \text{with} \quad \kappa_n \equiv Y_n - \frac{1}{2} C_n \quad (17)$$

This augmentation leaves the marginal distribution w.r.t. (γ, β) unchanged. Crucially each factor in Eqn. 17 is Gaussian w.r.t. β , with the consequence that Pòlya-Gamma augmentation establishes conjugacy.

4.2 PG-wTGS

We can now adapt wTGS to our setting. The augmented target distribution in Sec. 4.1 is given by

$$p(Y | \beta, \gamma, \omega, X, C) p(\beta) p(\gamma) p(\omega | C) \propto p(\beta, \gamma, \omega | \mathcal{D}) \quad (18)$$

where we define $\mathcal{D} \equiv \{X, Y, C\}$. We marginalize out β to obtain

$$p(Y | \gamma, \omega, X, C) p(\gamma) p(\omega | C) \propto p(\gamma, \omega | \mathcal{D}) \quad (19)$$

Thanks to PG augmentation we can compute $p(Y | \gamma, \omega, X, C)$ in closed form. Next we introduce an auxiliary variable $i \in \{0, 1, 2, \dots, P\}$ that controls which variables, if any, are tempered (note the additional state $i = 0$). We define the (unnormalized) target distribution $f(\gamma, \omega, i)$ as follows:

$$p(\gamma, \omega | \mathcal{D}) \left\{ \delta_{i0} \xi + \frac{1}{P} \sum_{j=1}^P \frac{\delta_{ij} \eta(\gamma_{-j}, \omega) U(\gamma_j)}{p(\gamma_j | \gamma_{-j}, \omega, \mathcal{D})} \right\} \quad (20)$$

Here $\xi > 0$ is a hyperparameter whose choice we discuss below. We note two important features of Eqn. 20.

Algorithm 2: We outline the main steps in PG-wTGS. See Sec. 4 for details.

Input: Dataset $\mathcal{D} = \{X, Y, C\}$ with P covariates; prior inclusion probability h ; prior precision τ ; total number of MCMC iterations T ; number of burn-in iterations T_{burn} ; hyperparameter $\xi > 0$ (optional)

Output: Approximate weighted posterior samples $\{\rho^{(t)}, \gamma^{(t)}, \omega^{(t)}\}_{t=T_{\text{burn}}+1}^T$

- 1 Let $\gamma^{(0)} = (0, \dots, 0)$ and $\omega^{(0)} \sim \text{PG}(C, 0)$.
- 2 **for** $t = 1, \dots, T$ **do**
- 3 Sample $i^{(t)} \sim f(\cdot | \gamma^{(t-1)}, \omega^{(t-1)})$ using Eqn. 22.
- 4 If $i^{(t)} > 0$ let $\omega^{(t)} = \omega^{(t-1)}$ and $\gamma^{(t)} = \text{flip}(\gamma^{(t-1)} | i^{(t)})$.
- 5 Otherwise if $i^{(t)} = 0$ let $\gamma^{(t)} = \gamma^{(t-1)}$ and sample $\omega'^{(t)} \sim p(\cdot | \gamma^{(t-1)}, \hat{\beta}(\gamma^{(t-1)}, \omega^{(t-1)}), \mathcal{D})$. Set $\omega^{(t)} = \omega'^{(t)}$ with probability $\alpha(\omega^{(t)} \rightarrow \omega'^{(t)} | \gamma^{(t)})$ given in Eqn. 25 (accept); otherwise set $\omega^{(t)} = \omega^{(t-1)}$ (reject).
- 6 Compute the unnormalized weight $\tilde{\rho}^{(t)} = \phi(\gamma^{(t)}, \omega^{(t)})^{-1}$ using Eqn. 21.
- 7 If ξ is not provided and $t \leq T_{\text{burn}}$ adapt ξ using the scheme described in Sec. A.12.
- 8 Compute the normalized weights $\rho^{(t)} = \frac{\tilde{\rho}^{(t)}}{\sum_{s>T_{\text{burn}}} \tilde{\rho}^{(s)}}$ for $t = T_{\text{burn}} + 1, \dots, T$.
- 9 **return** $\{\rho^{(t)}, \gamma^{(t)}, \omega^{(t)}\}_{t=T_{\text{burn}}+1}^T$

First, by construction when $i > 0$ the posterior conditional w.r.t. γ_i is the uniform distribution $U(\gamma_i)$. Second, as we discuss in more detail in Sec. A.4, the posterior conditional $p(\gamma_i | \gamma_{-i}, \omega, \mathcal{D})$ in Eqn. 20 can be computed in closed form thanks to PG augmentation. This is important because computing $p(\gamma_i | \gamma_{-i}, \omega, \mathcal{D})$ is necessary for importance weighting and Rao-Blackwellization. We proceed to construct a sampler for the target distribution Eqn. 20 that utilizes Gibbs updates w.r.t. i , Metropolized-Gibbs updates w.r.t. γ_i , and Metropolis-Hastings updates w.r.t. ω .

i -updates If we marginalize i from Eqn. 20 we obtain $f(\gamma, \omega) = p(\gamma, \omega | \mathcal{D}) \phi(\gamma, \omega)$ where we define

$$\phi(\gamma, \omega) \equiv \xi + \frac{1}{P} \sum_{i=1}^P \frac{\frac{1}{2} \eta(\gamma_{-i}, \omega)}{p(\gamma_i | \gamma_{-i}, \omega, \mathcal{D})} \quad (21)$$

Evidently $\phi(\gamma, \omega)^{-1}$ is the importance weight that is used to obtain samples from the non-tempered target Eqn. 19. Moreover we can do Gibbs updates w.r.t. i using the distribution

$$f(i | \gamma, \omega) \propto \delta_{i0} \xi + \frac{1}{P} \sum_{j=1}^P \delta_{ij} \frac{\frac{1}{2} \eta(\gamma_{-j}, \omega)}{p(\gamma_j | \gamma_{-j}, \omega, \mathcal{D})} \quad (22)$$

To better understand the behavior of the auxiliary variable i , we compute the marginal distribution w.r.t. i for the special case $\eta(\cdot) = 1$,

$$f(i) \propto \delta_{i0} \xi + \frac{1}{P} \sum_{j=1}^P \delta_{ij} \quad (23)$$

which clarifies that ξ controls how often we visit $i = 0$.

γ -updates Whenever $i > 0$ we do a Metropolized-Gibbs update of γ_i , resulting in a flip $\gamma_i \rightarrow 1 - \gamma_i$.

ω -updates Whenever $i = 0$ we update ω . To do so we use a simple proposal that can be computed in closed form. Importantly $f(\gamma, \omega, i = 0)$ is not tempered by construction so we can rely on the conjugate structure that is made manifest when we condition on a value of β . In more detail, we first compute the *mean* of the conditional posterior $p(\beta | \gamma, \omega, \mathcal{D})$ of Eqn. 18:

$$\hat{\beta}(\gamma, \omega) \equiv \mathbb{E}_{p(\beta | \gamma, \omega, \mathcal{D})} [\beta] \quad (24)$$

Using this (deterministic) value we then form the conditional posterior distribution $p(\omega' | \gamma, \hat{\beta}, \mathcal{D})$, which is a Pölya-Gamma distribution whose parameters are readily computed. We then sample a proposal $\omega' \sim p(\cdot | \gamma, \hat{\beta}, \mathcal{D})$ and compute the corresponding MH acceptance probability $\alpha(\omega \rightarrow \omega' | \gamma)$. The proposal is then accepted with probability $\alpha(\omega \rightarrow \omega' | \gamma)$; otherwise it is rejected. The acceptance probability can be computed in closed form and is given by

$$\alpha(\omega \rightarrow \omega' | \gamma) = \min \left(1, \frac{p(Y | \gamma, \omega', X, C)}{p(Y | \gamma, \omega, X, C)} \times \frac{p(Y | \gamma, \omega, \hat{\beta}(\gamma, \omega'), X, C) p(Y | \gamma, \hat{\beta}(\gamma, \omega), X, C)}{p(Y | \gamma, \omega', \hat{\beta}(\gamma, \omega), X, C) p(Y | \gamma, \hat{\beta}(\gamma, \omega'), X, C)} \right) \quad (25)$$

Eqn. 25 is readily computed; conveniently there is no need to compute the PG density, which can be challenging in some regimes. See Sec. A.11 for details.

We note that the proposal distribution $p(\omega' | \gamma, \hat{\beta}(\gamma, \omega), \mathcal{D})$ can be thought of as an approximation to the posterior conditional $p(\omega' | \gamma, \mathcal{D}) = \int d\beta p(\omega' | \gamma, \beta, \mathcal{D}) p(\beta | \gamma, \mathcal{D})$ that would be used in a Gibbs update. Since this latter density is intractable, we instead opt for this tractable option. One might worry that the resulting acceptance probability could be low, since ω is N -dimensional and N can be large. However, $p(\omega' | \gamma, \beta, \mathcal{D})$ only depends on β through $\psi_n = \beta_\gamma \cdot X_{n\gamma}$; the induced posterior over ψ_n is typically somewhat narrow, since the ψ_n are pinned down by the

observed data, and consequently $p(\omega'|\gamma, \hat{\beta}, \mathcal{D})$ is a reasonably good approximation to the exact posterior conditional. In practice we observe high mean acceptance probabilities $\sim 50\% - 95\%$ for all the experiments in this work,⁷ even for $N \gg 10^2$.

Weighting factor η We choose $\eta(\gamma_{-j}, \omega) = p(\gamma_j = 1|\gamma_{-j}, \omega) + \frac{\epsilon}{P}$. For the full algorithm see Algorithm 2.

5 Further Extensions

We briefly discuss how to accommodate inference over the prior inclusion probability h in Eqn. 1 and Eqn. 15. It is natural to place a Beta prior on h (Steel and Ley, 2007), since in the non-tempered setting this allows for conjugate Gibbs updates of h . Unfortunately this is spoiled by the tempering in wTGS, in particular the denominator in Eqn. 3, which exhibits complex dependence on h . We adopt a simple workaround. For example in the case of linear regression, Eqn. 1, we add an additional state $i = 0$ to wTGS analogous to the $i = 0$ state in PG-wTGS in Sec. 4. By construction when $i = 0$ the target distribution is not tempered, thus allowing for conjugate updates of h . See Algorithm 4 in the supplement for details. Subset wTGS and PG-wTGS are also readily combined; see Algorithm 5 in the supplement for details.

6 Related Work

Some of the earliest approaches to Bayesian variable selection (BVS) were introduced by George and McCulloch (1993, 1997). Chipman et al. (2001) provide an in-depth discussion of BVS for linear regression and CART models. Nikooienejad et al. (2016) and Shin et al. (2018) advocate the use of non-local priors in the BVS setting. (Garcia-Donato and Martinez-Beneito, 2013) compare non-MCMC search methods for BVS to MCMC-based methods. Zanella and Roberts (2019) introduce Tempered Gibbs Sampling (TGS) and apply it to BVS for linear regression. Griffin et al. (2021) introduce an efficient adaptive MCMC method for BVS in linear regression. Wan and Griffin (2021) extend this approach to logistic regression and accelerated failure time models. We include this approach (ASI) in our empirical validation in Sec. 7. Tian et al. (2019) introduce an approach to BVS for logistic regression that relies on joint credible regions. Lamnisis et al. (2009) develop transdimensional MCMC chains for BVS in probit regression. Dellaportas et al. (2002) and O'Hara et al. (2009) review various methods for BVS. Polson et al. (2013) introduce Pölya-Gamma augmentation and use it to develop efficient Gibbs samplers.

⁷For the experiment in Sec. A.16.3 N varies between 100 and 4000 and P varies between 134 and 69092 and the average acceptance prob. ranges between 49% and 89%.

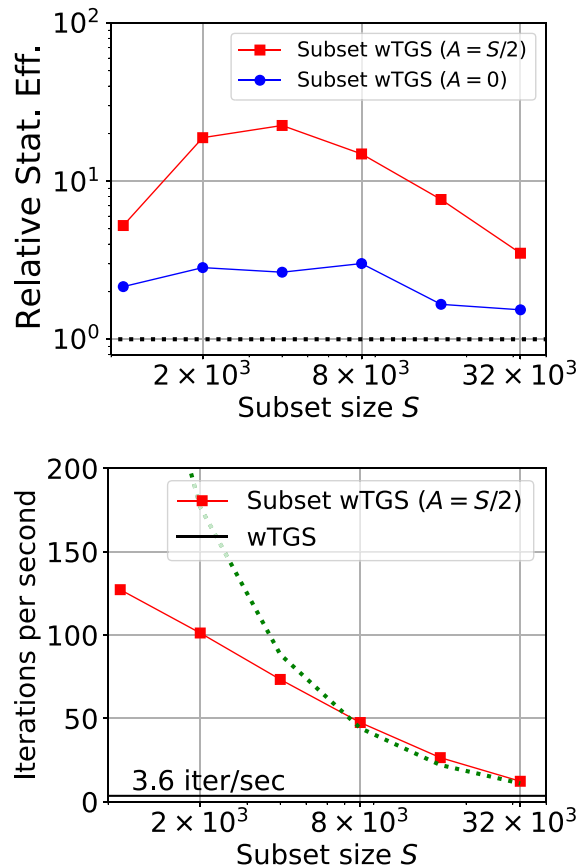


Figure 1: We report results for the experiment in Sec. 7.1 with $P = 98385$. **(Top)** We depict the relative statistical efficiency of Subset wTGS with subset size S compared to wTGS. Here statistical efficiency is defined such that it is proportional to both estimator variance and to runtime; see Sec. A.17 for details. **(Bottom)** We depict the number of iterations per second (IPS) for Subset wTGS as a function of S . The green dotted line depicts the IPS that would be expected if the latter scaled like S^{-1} .

7 Experiments

We validate the performance of Algorithms 1 & 2 on synthetic and real world data. We implement all algorithms using PyTorch (Paszke et al., 2017) and the `polygamma` package for sampling from Pölya-Gamma distributions (Bleki, 2021). See Sec. A.16-A.17 for additional experimental details and experiments (e.g. negative binomial results in Sec. A.16.4-A.16.5) and Sec. A.5 for discussion of computational complexity.

7.1 Subset wTGS performance for large P

We conduct two semi-synthetic experiments using maize genomic data from Romay et al. (2013) that have also been analyzed by Zeng and Zhou (2017); Biswas et al. (2022).

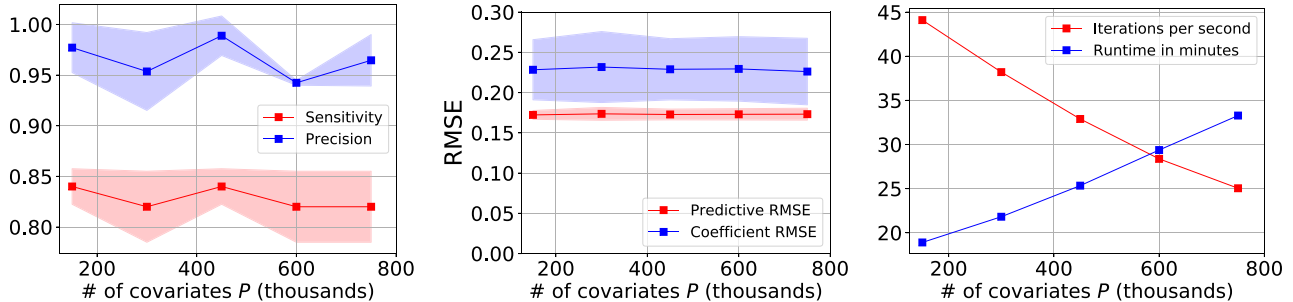


Figure 2: We report results for the Subset wTGS experiment in Sec. 7.1 with P ranging from 1.5×10^5 to 7.5×10^5 and $S = 16384$. We report 95% confidence intervals across 5 train/test splits. **(Left)** We depict the sensitivity and precision obtained if covariates with PIPs above a threshold of 0.5 are declared hits. **(Middle)** We depict the root mean squared error (RMSE) for held out responses Y^* and the inferred posterior mean over coefficients β . **(Right)** We depict runtimes obtained on a NVIDIA Tesla T4 GPU for collecting 50k MCMC samples.

This dataset serves as a good benchmark for our method, since it has large P ($P = 98385$), moderately large N ($N = 2267$), and complex correlation structure in the covariates X . We use synthetic responses Y so we have access to ground truth. Note that we do not include comparisons to ASI (Griffin et al., 2021), since we were unable to obtain results using ASI that were remotely competitive with wTGS.

In the first experiment we examine statistical efficiency and runtime, see Fig. 1. We find that Subset wTGS exhibits large speed-ups over wTGS and that these speed-ups translate to improved statistical efficiency. Indeed Subset wTGS with anchor set size $A = \frac{1}{2}S$ —which is our default choice—exhibits a relative statistical efficiency $\sim 20x$ larger than wTGS. Subset wTGS with no anchor set ($A = 0$) exhibits somewhat marginal efficiency improvements above wTGS, highlighting the importance of the anchor set \mathcal{A} in Algorithm 1. For an additional experiment that varies the anchor set size A see Fig. 11 in the supplement.

We note that the lower panel in Fig. 1 indicates that GPU resources are approximately saturated for $S \sim 4000$; increasing S beyond this point results in linear increases in runtime per MCMC iteration. Not surprisingly, the largest relative gain in statistical efficiency is observed precisely at this saturation point. Consequently we advise users to choose S to saturate available computational resources. Indeed we expect Subset wTGS to deliver significant performance gains precisely in the regime where wTGS saturates computational resources and choosing $S \ll P$ results in many more MCMC iterations for a fixed runtime budget. Thus when P is moderate, performance gains may be limited (see Fig. 12 in the supplement).

Next we demonstrate the feasibility of scaling Subset wTGS to $P \sim 10^6$, see Fig. 2 for results. To extend the maize data to $P > 98385$ we append random covari-

ates drawn from a unit Normal distribution. Thanks to the $O(S)$ iteration cost of Subset wTGS, GPU memory is the main bottleneck to accommodating large(r) datasets. Indeed the time needed to obtain 50k MCMC samples for $P = 7.5 \times 10^5$ is ~ 35 minutes on a Tesla T4 GPU. By contrast wTGS does not scale to this regime unless conditional PIPs are computed sequentially in batches (we estimate a ~ 24 hour runtime). We find high precision and sensitivity in identifying causal covariates across the entire range of P considered, highlighting the value of scalable Bayesian variable selection algorithms.

7.2 PG-wTGS and correlated covariates

We consider simulated Binomial regression datasets in which two covariates ($i \in \{1, 2\}$) are highly correlated and each alone can explain the response. This can be a challenging regime, since it is easy to get stuck in one mode and fail to explore the other mode. We consider four datasets with $32 \leq N \leq 512$, $32 \leq P \leq 4096$, and $C_n = 10$ for all data points. See Fig. 3 for the results.

To better understand the performance of PG-wTGS, we consider two variants, PG-TGS and PG-wGS, which do without weighting by $\eta(\gamma_{-j}, \omega)$ and tempering, respectively. In addition we compare to ASI (Wan and Griffin, 2021), an adaptive MCMC scheme that also uses Pölya-Gamma augmentation.

We see that PG-wGS does poorly on all datasets, including the smallest one with $P = 32$ covariates. PG-TGS does well for $P = 32$ and $P = 128$ but exhibits large variance for $P \geq 1024$. By contrast PG-wTGS yields low-variance PIP estimates in all cases, demonstrating the benefits of η -weighting and tempering. ASI estimates exhibit low variance for $P = 32$ (apart from a single outlier) but are high variance for larger P . This outcome is easy to understand. Since ASI adapts its proposal distribution during warmup using a running estimate of each PIP, it is

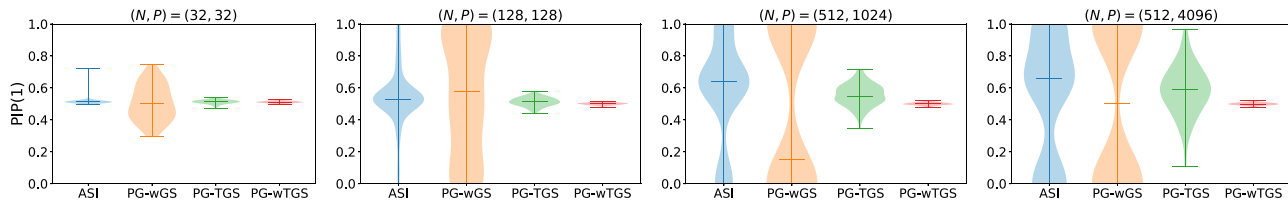


Figure 3: We depict violin plots for PIP (posterior inclusion probability) estimates for the first covariate obtained with 10^5 MCMC samples for four different methods on four datasets with varying numbers of data points N and covariates P . Horizontal bars denote the minimum, median, and maximum PIP estimates obtained from 100 independent MCMC runs. See Sec. 7.2 for details and Fig. 8 in the supplement for corresponding trace plots.

vulnerable to a rich-get-richer phenomenon in which covariates with large initial PIP estimates tend to crowd out covariates with which they are highly correlated. In the present case the result is that the ASI PIP estimates for the first two covariates are strongly anti-correlated. That this anti-correlation is ultimately due to suboptimal adaptation is easily verified. For example for $P = 1024$ ($P = 4096$) the Pearson correlation coefficient between the difference of final PIP estimates, i.e. $\text{PIP}(1) - \text{PIP}(2)$, and the difference of the corresponding initial PIP estimates that define the proposal distribution is 0.904 (0.998), respectively.

7.3 PG-wTGS and cancer data

We consider data collected from 900+ cancer cell lines in the Cancer Dependency Map project (Meyers et al., 2017; Behan et al., 2019; Pacini et al., 2021). Each cell line has been subjected to a loss-of-function genetic screen that uses CRISPR-Cas9 genetic editing to identify genes essential for cancer proliferation and survival. Genes identified by such screens are thought to be promising candidates for genetic vulnerabilities that can be used to guide the development of treatment strategies and novel therapeutics.

In more detail, we consider a subset of the data that includes $N = 907$ cell lines and $P = 17273$ covariates, with each covariate encoding the RNA expression level for a given gene. We consider two gene knockouts: DUSP4 and HNF1B.⁸ For each knockout the dataset contains a real-valued response that encodes the effect of knocking out that particular gene. We binarize this response variable by using the 20% quantile as a cutoff.

What makes this dataset particularly challenging is that the covariates are strongly correlated. For example, DUSP4 RNA expression exhibits a correlation greater than 0.40 (0.30) with 19 (203) other covariates, respectively. Similarly the HNF1B covariate has a correlation greater than 0.70 (0.50) with 2 (33) other covariates, respectively. In Fig. 4 we compare PIP estimates obtained with PG-wTGS

⁸This choice serves as a sanity check, since for both knockouts the RNA expression level of the corresponding gene is known to be highly predictive of cell viability.

and ASI, in both cases comparing to estimates obtained with long PG-wTGS runs. The much lower variance of PG-wTGS estimates as compared to ASI estimates is apparent. Indeed the mean absolute PIP error in the top hits is about $\sim 5x$ larger for ASI (see Table 2 in the supplemental materials for details and a list of all the top hits).

7.4 Inferring the inclusion probability h

We consider an application of Bayesian variable selection to viral transmission (Jankowiak et al., 2022). Each covariate encodes the presence or absence of a particular mutation in a virus like SARS-CoV-2, only a small number of which are assumed to affect viral fitness. Modeling viral transmission as a diffusion process results in a tractable Gaussian likelihood. We consider a virus with $P = 3000$ mutations and simulate a pandemic occurring in 30 geographic regions. We place a Beta prior on h and vary the number of causal mutations (i.e. those with non-zero effects) and investigate whether the inferred posterior inclusion probability h reflects the true number of causal mutations. As we would expect, see Fig. 5, this is indeed the case. See Sec. A.17 for additional details.

8 Discussion

We have shown that Bayesian variable selection can be efficiently scaled to $P \sim 10^6$ and can accommodate count-based likelihoods. Given the extremely large N and P that can be found in some genomics datasets, an interesting direction for future work would be to devise algorithms that can support N and P in the tens of millions. Doing so would likely require new algorithmic ideas (e.g. deterministic screening of covariates) as well as linear algebra speed-ups (e.g. incrementally caching computations as γ space is explored). Furthermore, while the augmentation strategy used to accommodate non-conjugate likelihoods in PG-wTGS can be generalized to other likelihoods that admit similar auxiliary variable constructions, other inference strategies are likely required to accommodate a broader set of non-conjugate likelihoods.

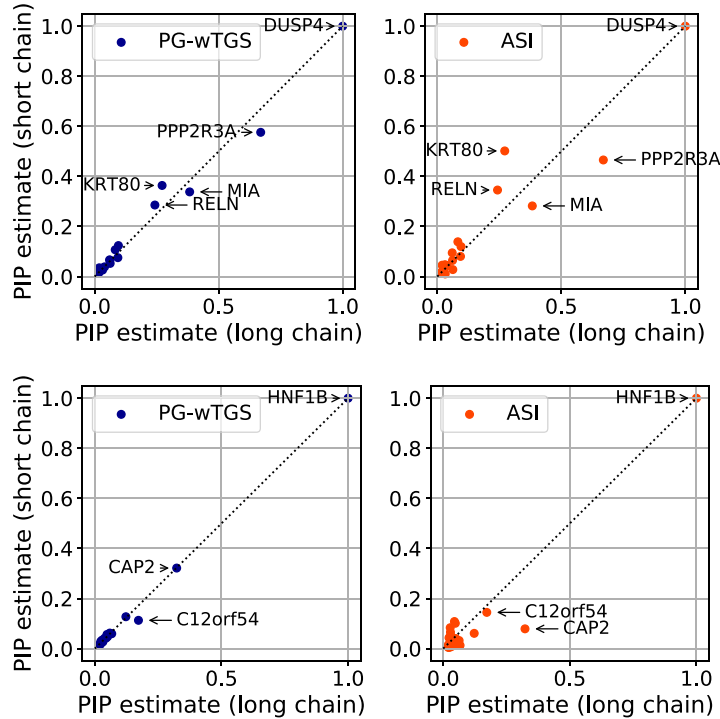


Figure 4: We compare PIP estimates obtained from PG-wTGS and ASI chains with 2.5×10^5 samples to a long PG-wTGS chain with 5×10^6 samples. For each method we depict the top 20 PIPs from the long chain paired with estimates from the short chain. The PG-wTGS estimates are significantly more accurate than is the case for ASI. See Sec. 7.3 for details and Fig. 9-10 in the supplement for additional comparisons.

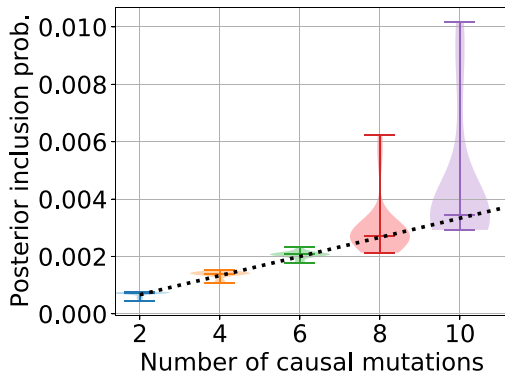


Figure 5: We depict posterior estimates of the inclusion probability h for the experiment in Sec. 7.4. Middle horizontal bars depict median posterior estimates across 20 simulations. The black dotted line indicates the proportion of mutations that are causal.

9 Acknowledgments

We thank Jim Griffin for clarifying some of the details of the methodology described in Wan and Griffin (2021) and Zolisa Bleki for help with the `polygamma` package. We also thank James McFarland, Joshua Dempster, and Ashir

Borah for help with DepMap data. We thank Niloy Biswas for sharing the genomic data we used in our experiments in Sec. 7.1. Finally we kindly thank Giacomo Zanella for interesting discussion about Bayesian variable selection.

References

- Fiona M Behan, Francesco Iorio, Gabriele Picco, Emanuel Gonçalves, Charlotte M Beaver, Giorgia Migliardi, Rita Santos, Yanhua Rao, Francesco Sassi, Marika Pinnelli, et al. Prioritization of cancer therapeutic targets using CRISPR–Cas9 screens. *Nature*, 568(7753):511–516, 2019.
- Niloy Biswas, Lester Mackey, and Xiao-Li Meng. Scalable Spike-and-Slab. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 2021–2040. PMLR, 17–23 Jul 2022.
- Zolisa Bleki. `polygamma`: An efficient and flexible sampler of the Polya-Gamma distribution with a NumPy/SciPy compatible interface, May 2021. URL <https://pypi.org/project/polygamma/>.
- Hugh Chipman, Edward I George, Robert E McCulloch, Merlise Clyde, Dean P Foster, and Robert A Stine. The

- practical implementation of Bayesian model selection. *Lecture Notes-Monograph Series*, pages 65–134, 2001.
- A Philip Dawid. Present position and potential developments: Some personal views statistical theory the prequential approach. *Journal of the Royal Statistical Society: Series A (General)*, 147(2):278–290, 1984.
- Petros Dellaportas, Jonathan J Forster, and Ioannis Ntzoufras. On Bayesian model and variable selection using MCMC. *Statistics and Computing*, 12(1):27–36, 2002.
- Gonzalo Garcia-Donato and Miguel A Martinez-Beneito. On sampling strategies in Bayesian variable selection problems with large model spaces. *Journal of the American Statistical Association*, 108(501):340–352, 2013.
- Edward I George and Robert E McCulloch. Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, 88(423):881–889, 1993.
- Edward I George and Robert E McCulloch. Approaches for Bayesian variable selection. *Statistica sinica*, pages 339–373, 1997.
- JE Griffin, KG Łatuszyński, and MFJ Steel. In search of lost mixing time: adaptive Markov chain Monte Carlo schemes for Bayesian variable selection with very large p. *Biometrika*, 108(1):53–69, 2021.
- JM Hilbe and WH Greene. Count response regression models, in (eds) cr rao, jp miller, and dc rao, epidemiology and medical statistics, 2007.
- Joseph M Hilbe. *Negative binomial regression*. Cambridge University Press, 2011.
- Martin Jankowiak, Fritz H Obermeyer, and Jacob E Lemieux. Inferring selection effects in SARS-CoV-2 with Bayesian viral allele selection. *PLoS genetics*, 18(12):e1010540, 2022.
- Demetris Lamnisos, Jim E Griffin, and Mark FJ Steel. Transdimensional sampling algorithms for Bayesian variable selection in classification problems with many more variables than observations. *Journal of Computational and Graphical Statistics*, 18(3):592–612, 2009.
- Jun S Liu. Peskun’s theorem and a modified discrete-state Gibbs sampler. *Biometrika*, 83(3), 1996.
- Peter McCullagh and John A Nelder. *Generalized linear models*. Routledge, 2019.
- Robin M Meyers, Jordan G Bryan, James M McFarland, Barbara A Weir, Ann E Sizemore, Han Xu, Neekesh V Dharia, Phillip G Montgomery, Glenn S Cowley, Sasha Pantel, et al. Computational correction of copy number effect improves specificity of CRISPR–Cas9 essentiality screens in cancer cells. *Nature genetics*, 49(12):1779–1784, 2017.
- Sean P Meyn and Richard L Tweedie. *Markov chains and stochastic stability*. Springer Science & Business Media, 2012.
- Amir Nikooienejad, Wenyi Wang, and Valen E Johnson. Bayesian variable selection for binary outcomes in high-dimensional genomic studies using non-local priors. *Bioinformatics*, 32(9):1338–1345, 2016.
- Robert B O’Hara, Mikko J Sillanpää, et al. A review of Bayesian variable selection methods: what, how and which. *Bayesian analysis*, 4(1):85–117, 2009.
- Clare Pacini, Joshua M Dempster, Isabella Boyle, Emanuel Gonçalves, Hanna Najgebauer, Emre Karakoc, Dieudonne van der Meer, Andrew Barthorpe, Howard Lightfoot, Patricia Jaaks, et al. Integrated cross-study datasets of genetic dependencies in cancer. *Nature communications*, 12(1):1–14, 2021.
- Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An Imperative Style, High-Performance Deep Learning Library. 2017.
- Nicholas G Polson, James G Scott, and Jesse Windle. Bayesian inference for logistic models using Pólya–Gamma latent variables. *Journal of the American Statistical Association*, 108(504):1339–1349, 2013.
- Maria C Romay, Mark J Millard, Jeffrey C Glaubitz, Jason A Peiffer, Kelly L Swarts, Terry M Casstevens, Robert J Elshire, Charlotte B Acharya, Sharon E Mitchell, Sherry A Flint-Garcia, et al. Comprehensive genotyping of the USA national maize inbred seed bank. *Genome biology*, 14(6):1–18, 2013.
- Minsuk Shin, Anirban Bhattacharya, and Valen E Johnson. Scalable Bayesian variable selection using nonlocal prior densities in ultrahigh-dimensional settings. *Statistica Sinica*, 28(2):1053, 2018.
- Mark FJ Steel and Eduardo Ley. *On the effect of prior assumptions in Bayesian model averaging with applications to growth regression*. The World Bank, 2007.
- Yiqing Tian, Howard D Bondell, and Alyson Wilson. Bayesian variable selection for logistic regression. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 12(5):378–393, 2019.
- Michalis K Titsias and Christopher Yau. The Hamming ball sampler. *Journal of the American Statistical Association*, 112(520):1598–1611, 2017.
- Kitty Yuen Yi Wan and Jim E Griffin. An adaptive MCMC method for Bayesian variable selection in logistic and accelerated failure time regression models. *Statistics and Computing*, 31(1):1–11, 2021.
- Giacomo Zanella and Gareth Roberts. Scalable importance tempering and Bayesian variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 81(3):489–517, 2019.
- Ping Zeng and Xiang Zhou. Non-parametric genetic prediction of complex traits with latent Dirichlet process

regression models. *Nature communications*, 8(1):1–11,
2017.

Algorithm 3: We outline the main steps in wTGS (Zanella and Roberts, 2019). See Sec. 2.3 for discussion. Note that we use superscripts to indicate MCMC iterations.

Input: Dataset $\mathcal{D} = \{X, Y\}$ with P covariates; prior inclusion probability h ; prior precision τ ; total number of MCMC iterations T ; number of burn-in iterations T_{burn}

Output: Approximate weighted posterior samples $\{\rho^{(t)}, \gamma^{(t)}\}_{t=T_{\text{burn}}+1}^T$

- 1 Let $\gamma^{(0)} = (0, \dots, 0)$.
- 2 **for** $t = 1, \dots, T$ **do**
- 3 Sample $i^{(t)} \sim f(\cdot | \gamma^{(t-1)})$ using Eqn. 7
- 4 Let $\gamma^{(t)} = \text{flip}(\gamma^{(t-1)} | i^{(t)})$ where $\text{flip}(\gamma | i)$ flips the i^{th} coordinate of γ : $\gamma_i \rightarrow 1 - \gamma_i$.
- 5 Compute the unnormalized weight $\tilde{\rho}^{(t)} = \phi(\gamma^{(t)})^{-1}$ using Eqn. 6.
- 6 Compute the normalized weights $\rho^{(t)} = \frac{\tilde{\rho}^{(t)}}{\sum_{s>T_{\text{burn}}} \tilde{\rho}^{(s)}}$ for $t = T_{\text{burn}} + 1, \dots, T$.
- 7 **return** $\{\rho^{(t)}, \gamma^{(t)}\}_{t=T_{\text{burn}}+1}^T$

A Appendix

This appendix is organized as follows. In Sec. A.1 we discuss societal impact. In Sec. A.2 we discuss how we infer the inclusion probability h . In Sec. A.3 we discuss how we combine Subset wTGS and PG-wTGS. In Sec. A.4 we discuss conditional marginal log likelihood computations. In Sec. A.5 we discuss computational complexity. In Sec. A.6 we motivate the tempering scheme that underlies wTGS. In Sec. A.7 we discuss the nature of the local moves made by wTGS. In Sec. A.8 we briefly discuss the role played by importance weighting in our MCMC methods. In Sec. A.9 we provide a proof of Proposition 1. In Sec. A.10 we discuss Rao-Blackwellized PIP estimators. In Sec. A.11 we discuss ω -updates. In Sec. A.12 we discuss ξ -adapation. In Sec. A.13 we discuss how we adapt the anchor set \mathcal{A} . In Sec. A.14 we discuss the modifications of PG-wTGS that are needed to accommodate negative binomial likelihoods. In Sec. A.15 we include additional figures and tables accompanying the experimental results in Sec. 7. In Sec. A.16 we report additional experimental results. In Sec. A.17 we discuss experimental details.

A.1 Societal impact

We do not anticipate any negative societal impact from the methods described in this work, although we note that they inherit the risks that are inherent to any algorithm that can be used for hypothesis testing and/or prediction. In more detail there is the possibility of the following risks. First, predictive algorithms can be deployed in ways that disadvantage vulnerable groups in a population. Even if these effects are unintended, they can still arise if deployed algorithms are poorly vetted with respect to their fairness implications. The same applies to any hypotheses investigated with a variable selection algorithm, especially if variables are correlated with indicators that encode the identity of vulnerable groups. Second, algorithms that offer uncertainty quantification may be misused by users who place unwarranted confidence in the uncertainties produced by the algorithm. This can arise, for example, in the presence of undetected covariate shift.

A.2 Inferring the inclusion probability h

Consider the following (unnormalized) target distribution

$$f(\gamma, i, h) \equiv p(\gamma | h, \mathcal{D}) p(h | \alpha_h, \beta_h) \left\{ \delta_{i0} \xi + \frac{1}{P} \sum_{j=1}^P \frac{\delta_{ij} \eta(\gamma_{-j}, h) U(\gamma_j)}{p(\gamma_j | \gamma_{-j}, h, \mathcal{D})} \right\} \quad (26)$$

where we have introduced a hyperparameter $\xi > 0$ and $\alpha_h > 0$ and $\beta_h > 0$ parameterize the prior over h . We define the inverse importance weight

$$\phi(\gamma, h) \equiv \xi + \frac{1}{P} \sum_{i=1}^P \frac{\frac{1}{2} \eta(\gamma_{-i}, h)}{p(\gamma_i | \gamma_{-i}, h, \mathcal{D})} \quad (27)$$

We can do i updates using the Gibbs distribution

$$f(i | \gamma, h) \propto \delta_{i0} \xi + \frac{1}{P} \sum_{j=1}^P \delta_{ij} \frac{\frac{1}{2} \eta(\gamma_{-j}, h)}{p(\gamma_j | \gamma_{-j}, h, \mathcal{D})} \quad (28)$$

Algorithm 4: We extend wTGS (see Algorithm 3) to allow inference over the inclusion probability h , with h governed by a $\text{Beta}(\alpha_h, \beta_h)$ prior. See Sec. A.2 for additional discussion. Note that we use superscripts to indicate MCMC iterations.

Input: Dataset $\mathcal{D} = \{X, Y\}$ with P covariates; prior precision τ ; hyperparameters (α_h, β_h) ; total number of MCMC iterations T ; number of burn-in iterations T_{burn} ; hyperparameter $\xi > 0$ (optional)

Output: Approximate weighted posterior samples $\{\rho^{(t)}, \gamma^{(t)}, h^{(t)}\}_{t=T_{\text{burn}}+1}^T$

- 1 Let $\gamma^{(0)} = (0, \dots, 0)$ and $h^{(0)} \sim \text{Beta}(\alpha_h, \beta_h)$.
 - 2 **for** $t = 1, \dots, T$ **do**
 - 3 Sample $i^{(t)} \sim f(\cdot | \gamma^{(t-1)}, h^{(t-1)})$ using Eqn. 28.
 - 4 If $i^{(t)} > 0$ let $\gamma^{(t)} = \text{flip}(\gamma^{(t-1)} | i^{(t)})$ and $h^{(t)} = h^{(t-1)}$.
 - 5 Otherwise if $i^{(t)} = 0$ let $\gamma^{(t)} = \gamma^{(t-1)}$ and $h^{(t)} \sim f(\cdot | \gamma^{(t)}, i = 0)$ using Eqn. 29.
 - 6 Compute the unnormalized weight $\tilde{\rho}^{(t)} = \phi(\gamma^{(t)}, h^{(t)})^{-1}$ using Eqn. 27.
 - 7 If ξ is not provided and $t \leq T_{\text{burn}}$ adapt ξ using the scheme described in Sec. A.12.
 - 8 Compute the normalized weights $\rho^{(t)} = \frac{\tilde{\rho}^{(t)}}{\sum_{s>T_{\text{burn}}} \tilde{\rho}^{(s)}}$ for $t = T_{\text{burn}} + 1, \dots, T$.
 - 9 **return** $\{\rho^{(t)}, \gamma^{(t)}, h^{(t)}\}_{t=T_{\text{burn}}+1}^T$
-

When $i = 0$ we can do h updates using the Gibbs distribution

$$f(h | \gamma, i = 0) = \text{Beta}(\alpha = \alpha_h + |\gamma|, \beta = \beta_h + P - |\gamma|) \quad (29)$$

where $|\gamma|$ is the number of covariates included in the model in the current iteration. See Algorithm 4 for a complete description of wTGS with inference over h .

The above discussion assumes the linear regression case, Eqn. 1. To accommodate count-based likelihoods we simply use the untempered $i = 0$ state to make h updates *and* ω updates in succession (and in random order).

A.3 Subset PG-wTGS

We show how to combine the algorithmic ideas from Sec. 3 and Sec. 4, i.e. how to scale Bayesian variable selection with a count-based likelihood to large P . The target distribution is

$$f(\gamma, i, \omega, \mathcal{S}) = p(\gamma, \omega | \mathcal{D}) \left\{ \delta_{i0} \xi + \frac{1}{P} \sum_{j=1}^P \frac{\delta_{ij} \eta(\gamma_{-j}, \omega) U(\gamma_j)}{p(\gamma_j | \gamma_{-j}, \omega, \mathcal{D})} \right\} U(\mathcal{S} | i, \mathcal{A}) \quad (30)$$

where we assume \mathcal{S} ranges over size S subsets of $\{0, \dots, P\}$ and that $0 \in \mathcal{A}$. We define

$$f(i | \gamma, \omega, \mathcal{S}) \propto \left(\delta_{i0} \xi + \frac{1}{P} \sum_{j \in \mathcal{S}, j > 0} \delta_{ij} \frac{\frac{1}{2} \eta(\gamma_{-j}, \omega)}{p(\gamma_j | \gamma_{-j}, \omega, \mathcal{D})} \right) U(\mathcal{S} | i, \mathcal{A}) \quad (31)$$

and

$$\phi(\gamma, \omega, \mathcal{S}) \equiv \left(\xi \times U(\mathcal{S} | 0, \mathcal{A}) + \frac{1}{P} \sum_{i \in \mathcal{S}, i > 0} \frac{\frac{1}{2} \eta(\gamma_{-i}, \omega)}{p(\gamma_i | \gamma_{-i}, \omega, \mathcal{D})} U(\mathcal{S} | i, \mathcal{A}) \right) \quad (32)$$

The algorithm then follows the same logic as in Subset wTGS and PG-wTGS; see Algorithm 5 for a complete description.

We note an important implementation detail that is common to Algorithm 1 and Algorithm 5. Here we deal with the case of Algorithm 1 for concreteness. Besides the value of zero, the probability $U(\mathcal{S} | i, \mathcal{A})$ takes on two possible values:

$$\begin{aligned} U(\mathcal{S} | i, \mathcal{A}) &= \frac{(S - A)!(P - S)!}{(P - A)!} && \text{if } i \in \mathcal{A} \\ U(\mathcal{S} | i, \mathcal{A}) &= \frac{(S - A - 1)!(P - S)!}{(P - A - 1)!} && \text{if } i \notin \mathcal{A} \end{aligned} \quad (33)$$

Since, however, we always use normalized weights $\{\rho^{(t)}\}$ when computing approximate posterior expectations, any overall constant factor in $U(\mathcal{S} | i, \mathcal{A})$ is irrelevant. Consequently we only need to keep track of the ratio of the two values in Eqn. 33, namely $\frac{S-A}{P-A}$. In particular there is no need to compute factorials.

Algorithm 5: We outline the main steps in Subset PG-wTGS, which combines Algorithm 1 & 2. See Sec A.3 for additional discussion. Note that we use superscripts to indicate MCMC iterations.

Input: Dataset $\mathcal{D} = \{X, Y, C\}$ with P covariates; prior inclusion probability h ; prior precision τ ; subset size S ; anchor set size A ; total number of MCMC iterations T ; number of burn-in iterations T_{burn} ; hyperparameter $\xi > 0$ (optional)

Output: Approximate weighted posterior samples $\{\rho^{(t)}, \gamma^{(t)}, \omega^{(t)}\}_{t=T_{\text{burn}}+1}^T$

- 1 Let $\gamma^{(0)} = (0, \dots, 0)$ and $\omega^{(0)} \sim \text{PG}(C, 0)$.
- 2 Choose \mathcal{A} to include $\{0\}$ as well as the $A - 1$ covariate indices exhibiting the largest correlations with the response Y .
- 3 Choose $i^{(0)}$ randomly from $\{1, \dots, P\}$ and $S^{(0)} \sim U(\cdot | i^{(0)}, \mathcal{A})$.
- 4 **for** $t = 1, \dots, T$ **do**
- 5 Sample $i^{(t)} \sim f(\cdot | \gamma^{(t-1)}, \omega^{(t-1)}, S^{(t-1)})$ using Eqn. 31.
- 6 If $i^{(t)} > 0$ let $\omega^{(t)} = \omega^{(t-1)}$ and $\gamma^{(t)} = \text{flip}(\gamma^{(t-1)} | i^{(t)})$.
- 7 Otherwise if $i^{(t)} = 0$ let $\gamma^{(t)} = \gamma^{(t-1)}$ and sample $\omega^{(t)} \sim p(\cdot | \gamma^{(t-1)}, \hat{\beta}(\gamma^{(t-1)}, \omega^{(t-1)}), \mathcal{D})$. Set $\omega^{(t)} = \omega^{(t)}$ with probability $\alpha(\omega^{(t)} \rightarrow \omega'^{(t)} | \gamma^{(t)})$ given in Eqn. 25. Otherwise set $\omega^{(t)} = \omega^{(t-1)}$.
- 8 Compute the unnormalized weight $\tilde{\rho}^{(t)} = \phi(\gamma^{(t)}, \omega^{(t)}, S^{(t)})^{-1}$ using Eqn. 32.
- 9 If ξ is not provided and $t \leq T_{\text{burn}}$ adapt ξ using the scheme described in Sec. A.12.
- 10 Compute the normalized weights $\rho^{(t)} = \frac{\tilde{\rho}^{(t)}}{\sum_{s>T_{\text{burn}}} \tilde{\rho}^{(s)}}$ for $t = T_{\text{burn}} + 1, \dots, T$.
- 11 **return** $\{\rho^{(t)}, \gamma^{(t)}, \omega^{(t)}\}_{t=T_{\text{burn}}+1}^T$

A.4 Efficient linear algebra for the (conditional) marginal log likelihood

Here we focus on computing the marginal log likelihood in the case of count-based likelihoods as required for Algorithm 2. The linear algebra required for the linear regression case is essentially identical. See Chipman et al. (2001); Zanella and Roberts (2019) for discussion of the linear case.

The conditional marginal log likelihood $\log p(Y|X, C, \gamma, \omega)$ can be computed in closed form where, up to irrelevant constants, we have

$$\begin{aligned} \log p(Y|X, C, \gamma, \omega) &= \frac{1}{2} \mathcal{Z}_{\gamma+1}^T (X_{\gamma+1}^T \Omega X_{\gamma+1} + \tau \mathbb{1}_{\gamma+1})^{-1} \mathcal{Z}_{\gamma+1} \\ &\quad - \frac{1}{2} \log \det(X_{\gamma+1}^T \Omega X_{\gamma+1} + \tau \mathbb{1}_{\gamma+1}) - \frac{1}{2} \log \det(\tau^{-1} \mathbb{1}_{\gamma+1}) \end{aligned} \quad (34)$$

where $\mathcal{Z} \in \mathbb{R}^{P+1}$ with $\mathcal{Z}_j = \sum_{n=1}^N \kappa_n X_{n,j}$ for $j = 1, \dots, P$ and the final component $\mathcal{Z}_{P+1} = \sum_{n=1}^N \kappa_n$ corresponds to the bias. Here and elsewhere X is augmented with a column of all ones where necessary and $\kappa_n \equiv Y_n - \frac{1}{2} C_n$, Ω is the $N \times N$ diagonal matrix formed from ω , and $\gamma + 1$ is used to refer to the active indices in γ as well as the bias, which is always included in the model by assumption. Using a Cholesky decomposition the quantity in Eqn. 34 can be computed in $\mathcal{O}(|\gamma|^3 + |\gamma|^2 N)$ time. If done naively this becomes expensive in cases where Eqn. 34 needs to be computed for many values of γ (as is needed e.g. to compute Rao-Blackwellized PIPs). Luckily, and as is done by (Zanella and Roberts, 2019) and others in the literature, the computational cost can be reduced significantly since we can exploit the fact that in practice we always consider ‘neighboring’ values of γ and so we can leverage rank-1 update structure where appropriate. In the following we provide the formulae necessary for doing so. We keep the derivation generic and consider the case of adding arbitrarily many variables to γ even though in practice we only make use of the rank-1 formulae.

In more detail we proceed as follows. Let \mathcal{I} be the active indices in γ together with the bias index $P + 1$ (i.e. we conveniently augment X by an all-ones feature column in the following). Let \mathcal{K} be a non-empty set of indices not in \mathcal{I} and let $\mathcal{I}_{\mathcal{K}} = \mathcal{I} \cup \mathcal{K}$. We let $\mathcal{X} = \Omega^{\frac{1}{2}} X$ and rewrite $F_{\mathcal{I}_{\mathcal{K}}} \equiv (\mathcal{X}_{\mathcal{I}_{\mathcal{K}}}^T \mathcal{X}_{\mathcal{I}_{\mathcal{K}}} + \tau \mathbb{1}_{\mathcal{I}_{\mathcal{K}}})^{-1}$ in terms of $F_{\mathcal{I}} \equiv (\mathcal{X}_{\mathcal{I}}^T \mathcal{X}_{\mathcal{I}} + \tau \mathbb{1}_{\mathcal{I}})^{-1}$ as follows:

$$F_{\mathcal{I}_{\mathcal{K}}} = \begin{pmatrix} \mathcal{X}_{\mathcal{I}}^T \mathcal{X}_{\mathcal{I}} + \tau \mathbb{1}_{\mathcal{I}} & \mathcal{X}_{\mathcal{I}}^T \mathcal{X}_{\mathcal{K}} \\ \mathcal{X}_{\mathcal{K}}^T \mathcal{X}_{\mathcal{I}} & \mathcal{X}_{\mathcal{K}}^T \mathcal{X}_{\mathcal{K}} + \tau \mathbb{1}_{\mathcal{K}} \end{pmatrix}^{-1} = \begin{pmatrix} F_{\mathcal{I}} + F_{\mathcal{I}} \mathcal{X}_{\mathcal{I}}^T \mathcal{X}_{\mathcal{K}} G_{\mathcal{K}} \mathcal{X}_{\mathcal{K}}^T \mathcal{X}_{\mathcal{I}} F_{\mathcal{I}} & -F_{\mathcal{I}} \mathcal{X}_{\mathcal{I}}^T \mathcal{X}_{\mathcal{K}} G_{\mathcal{K}} \\ -G_{\mathcal{K}} \mathcal{X}_{\mathcal{K}}^T \mathcal{X}_{\mathcal{I}} F_{\mathcal{I}} & G_{\mathcal{K}} \end{pmatrix} \quad (35)$$

where $G_{\mathcal{K}}^{-1} \equiv \mathcal{X}_{\mathcal{K}}^T \mathcal{X}_{\mathcal{K}} + \tau \mathbb{1}_{\mathcal{K}} - \mathcal{X}_{\mathcal{K}}^T \mathcal{X}_{\mathcal{I}} F_{\mathcal{I}} \mathcal{X}_{\mathcal{I}}^T \mathcal{X}_{\mathcal{K}}$.

To efficiently compute the quadratic term in Eqn. 34 we need to compute $\mathcal{Z}_{\mathcal{I}_{\mathcal{K}}}^T F_{\mathcal{I}_{\mathcal{K}}} \mathcal{Z}_{\mathcal{I}_{\mathcal{K}}}$ in terms of $\mathcal{Z}_{\mathcal{I}}^T F_{\mathcal{I}} \mathcal{Z}_{\mathcal{I}}$. Write

$\mathcal{Z}_{\mathcal{I}\mathcal{K}} = (\mathcal{Z}_{\mathcal{I}}, \mathcal{Z}_{\mathcal{K}})$ so we have

$$\mathcal{Z}_{\mathcal{I}\mathcal{K}}^{\top} F_{\mathcal{I}\mathcal{K}} \mathcal{Z}_{\mathcal{I}\mathcal{K}} = \begin{pmatrix} \mathcal{Z}_{\mathcal{I}} \\ \mathcal{Z}_{\mathcal{K}} \end{pmatrix}^{\top} \begin{pmatrix} F_{\mathcal{I}} + F_{\mathcal{I}} \mathcal{X}_{\mathcal{I}}^{\top} \mathcal{X}_{\mathcal{K}} G_{\mathcal{K}} \mathcal{X}_{\mathcal{K}}^{\top} \mathcal{X}_{\mathcal{I}} F_{\mathcal{I}} & -F_{\mathcal{I}} \mathcal{X}_{\mathcal{I}}^{\top} \mathcal{X}_{\mathcal{K}} G_{\mathcal{K}} \\ -G_{\mathcal{K}} \mathcal{X}_{\mathcal{K}}^{\top} \mathcal{X}_{\mathcal{I}} F_{\mathcal{I}} & G_{\mathcal{K}} \end{pmatrix} \begin{pmatrix} \mathcal{Z}_{\mathcal{I}} \\ \mathcal{Z}_{\mathcal{K}} \end{pmatrix} \quad (36)$$

$$= \mathcal{Z}_{\mathcal{I}}^{\top} F_{\mathcal{I}} \mathcal{Z}_{\mathcal{I}} + \mathcal{Z}_{\mathcal{I}}^{\top} F_{\mathcal{I}} \mathcal{X}_{\mathcal{I}}^{\top} \mathcal{X}_{\mathcal{K}} G_{\mathcal{K}} \mathcal{X}_{\mathcal{K}}^{\top} \mathcal{X}_{\mathcal{I}} F_{\mathcal{I}} \mathcal{Z}_{\mathcal{I}} \quad (37)$$

$$- \mathcal{Z}_{\mathcal{I}}^{\top} F_{\mathcal{I}} \mathcal{X}_{\mathcal{I}}^{\top} \mathcal{X}_{\mathcal{K}} G_{\mathcal{K}} \mathcal{Z}_{\mathcal{K}} - \mathcal{Z}_{\mathcal{K}}^{\top} G_{\mathcal{K}} \mathcal{X}_{\mathcal{K}}^{\top} \mathcal{X}_{\mathcal{I}} F_{\mathcal{I}} \mathcal{Z}_{\mathcal{I}} + \mathcal{Z}_{\mathcal{K}}^{\top} G_{\mathcal{K}} \mathcal{Z}_{\mathcal{K}} \quad (38)$$

$$= \tilde{\mathcal{Z}}_{\mathcal{I}}^{\top} \tilde{\mathcal{Z}}_{\mathcal{I}} + (\tilde{\mathcal{X}}_{\mathcal{K}}^{\top} \tilde{\mathcal{X}}_{\mathcal{I}} \tilde{\mathcal{Z}}_{\mathcal{I}})^{\top} (\tilde{\mathcal{X}}_{\mathcal{K}}^{\top} \tilde{\mathcal{X}}_{\mathcal{I}} \tilde{\mathcal{Z}}_{\mathcal{I}}) - 2(\tilde{\mathcal{X}}_{\mathcal{K}}^{\top} \tilde{\mathcal{X}}_{\mathcal{I}} \tilde{\mathcal{Z}}_{\mathcal{I}})^{\top} \tilde{\mathcal{Z}}_{\mathcal{K}} + \tilde{\mathcal{Z}}_{\mathcal{K}}^{\top} \tilde{\mathcal{Z}}_{\mathcal{K}} \quad (39)$$

$$= \tilde{\mathcal{Z}}_{\mathcal{I}}^{\top} \tilde{\mathcal{Z}}_{\mathcal{I}} + \left\| \tilde{\mathcal{X}}_{\mathcal{K}}^{\top} \tilde{\mathcal{X}}_{\mathcal{I}} \tilde{\mathcal{Z}}_{\mathcal{I}} - \tilde{\mathcal{Z}}_{\mathcal{K}} \right\|^2. \quad (40)$$

where $\|\cdot\|$ is the 2-norm in $\mathbb{R}^{|\mathcal{K}|}$ and we define

$$L_{\mathcal{I}} L_{\mathcal{I}}^{\top} = \mathcal{X}_{\mathcal{I}}^{\top} \mathcal{X}_{\mathcal{I}} + \tau \mathbb{1}_{\mathcal{I}} = F_{\mathcal{I}}^{-1} \quad (41)$$

$$L_{\mathcal{K}} L_{\mathcal{K}}^{\top} = \mathcal{X}_{\mathcal{K}}^{\top} \mathcal{X}_{\mathcal{K}} + \tau \mathbb{1}_{\mathcal{K}} - \mathcal{X}_{\mathcal{K}}^{\top} \mathcal{X}_{\mathcal{I}} F_{\mathcal{I}} \mathcal{X}_{\mathcal{I}}^{\top} \mathcal{X}_{\mathcal{K}} = \mathcal{X}_{\mathcal{K}}^{\top} \mathcal{X}_{\mathcal{K}} + \tau \mathbb{1}_{\mathcal{K}} - \mathcal{X}_{\mathcal{K}}^{\top} \tilde{\mathcal{X}}_{\mathcal{I}} \tilde{\mathcal{X}}_{\mathcal{I}}^{\top} \mathcal{X}_{\mathcal{K}} = G_{\mathcal{K}}^{-1}$$

$$\tilde{\mathcal{Z}}_{\mathcal{I}} \equiv L_{\mathcal{I}}^{-1} \mathcal{Z}_{\mathcal{I}} \quad \tilde{\mathcal{Z}}_{\mathcal{K}} \equiv L_{\mathcal{K}}^{-1} \mathcal{Z}_{\mathcal{K}} \quad \tilde{\mathcal{X}}_{\mathcal{I}} \equiv \mathcal{X}_{\mathcal{I}} L_{\mathcal{I}}^{-\top} \quad \tilde{\mathcal{X}}_{\mathcal{K}} \equiv \mathcal{X}_{\mathcal{K}} L_{\mathcal{K}}^{-\top}$$

Here $L_{\mathcal{I}}$ and $L_{\mathcal{K}}$ are Cholesky factors. This can be rewritten as

$$\mathcal{Z}_{\mathcal{I}\mathcal{K}}^{\top} F_{\mathcal{I}\mathcal{K}} \mathcal{Z}_{\mathcal{I}\mathcal{K}} = \tilde{\mathcal{Z}}_{\mathcal{I}}^{\top} \tilde{\mathcal{Z}}_{\mathcal{I}} + \|W_{\mathcal{K}}\|^2 \quad \text{with} \quad W_{\mathcal{K}} \equiv L_{\mathcal{K}}^{-1} \left(\mathcal{X}_{\mathcal{K}}^{\top} \tilde{\mathcal{X}}_{\mathcal{I}} \tilde{\mathcal{Z}}_{\mathcal{I}} - \mathcal{Z}_{\mathcal{K}} \right) \quad (42)$$

Together these formulae can be used to compute the quadratic term efficiently.

Next we turn to the log determinant in Eqn. 34. We begin by noting that

$$\log \det \left(\mathcal{X}_{\mathcal{I}\mathcal{K}}^{\top} \mathcal{X}_{\mathcal{I}\mathcal{K}} + \tau \mathbb{1}_{\mathcal{I}\mathcal{K}} \right) + \log \det \left(\tau^{-1} \mathbb{1}_{\mathcal{I}\mathcal{K}} \right) = \log \det(\Omega) + \log \det \left(X_{\mathcal{I}\mathcal{K}} X_{\mathcal{I}\mathcal{K}}^{\top} / \tau + \Omega^{-1} \right) \quad (43)$$

and

$$\log \det \left(X_{\mathcal{I}\mathcal{K}} X_{\mathcal{I}\mathcal{K}}^{\top} / \tau + \Omega^{-1} \right) = \log \det \left(X_{\mathcal{I}} X_{\mathcal{I}}^{\top} / \tau + \Omega^{-1} \right) + \log \det \left(\mathbb{1}_{\mathcal{K}} / \tau \right) \quad (44)$$

$$+ \log \det \left(\tau \mathbb{1}_{\mathcal{K}} + X_{\mathcal{K}}^{\top} \left(X_{\mathcal{I}} X_{\mathcal{I}}^{\top} / \tau + \Omega^{-1} \right)^{-1} X_{\mathcal{K}} \right) \quad (45)$$

which together imply

$$\begin{aligned} & \{ \log \det \left(\mathcal{X}_{\mathcal{I}\mathcal{K}}^{\top} \mathcal{X}_{\mathcal{I}\mathcal{K}} + \tau \mathbb{1}_{\mathcal{I}\mathcal{K}} \right) + \log \det \left(\tau^{-1} \mathbb{1}_{\mathcal{I}\mathcal{K}} \right) \} - \\ & \{ \log \det \left(\mathcal{X}_{\mathcal{I}}^{\top} \mathcal{X}_{\mathcal{I}} + \tau \mathbb{1}_{\mathcal{I}} \right) + \log \det \left(\tau^{-1} \mathbb{1}_{\mathcal{I}} \right) \} = \log \det \left(X_{\mathcal{I}\mathcal{K}} X_{\mathcal{I}\mathcal{K}}^{\top} / \tau + \Omega^{-1} \right) - \log \det \left(X_{\mathcal{I}} X_{\mathcal{I}}^{\top} / \tau + \Omega^{-1} \right) \\ & = \log \det \left(\mathbb{1}_{\mathcal{K}} + \tau^{-1} X_{\mathcal{K}}^{\top} \left(X_{\mathcal{I}} X_{\mathcal{I}}^{\top} / \tau + \Omega^{-1} \right)^{-1} X_{\mathcal{K}} \right) \end{aligned}$$

While these equations can be used to compute the log determinant reasonably efficiently, they exhibit cubic computational complexity w.r.t. N . So instead we write

$$\begin{aligned} \det \left(\mathcal{X}_{\mathcal{I}\mathcal{K}}^{\top} \mathcal{X}_{\mathcal{I}\mathcal{K}} + \tau \mathbb{1}_{\mathcal{I}\mathcal{K}} \right) &= \det \begin{pmatrix} \mathcal{X}_{\mathcal{I}}^{\top} \mathcal{X}_{\mathcal{I}} + \tau \mathbb{1}_{\mathcal{I}} & \mathcal{X}_{\mathcal{I}}^{\top} \mathcal{X}_{\mathcal{K}} \\ \mathcal{X}_{\mathcal{K}}^{\top} \mathcal{X}_{\mathcal{I}} & \mathcal{X}_{\mathcal{K}}^{\top} \mathcal{X}_{\mathcal{K}} + \tau \mathbb{1}_{\mathcal{K}} \end{pmatrix} \quad (46) \\ &= \det \left(\mathcal{X}_{\mathcal{K}}^{\top} \mathcal{X}_{\mathcal{K}} + \tau \mathbb{1}_{\mathcal{K}} - \mathcal{X}_{\mathcal{K}}^{\top} \mathcal{X}_{\mathcal{I}} \left(\mathcal{X}_{\mathcal{I}}^{\top} \mathcal{X}_{\mathcal{I}} + \tau \mathbb{1}_{\mathcal{I}} \right)^{-1} \mathcal{X}_{\mathcal{I}}^{\top} \mathcal{X}_{\mathcal{K}} \right) \times \det \left(\mathcal{X}_{\mathcal{I}}^{\top} \mathcal{X}_{\mathcal{I}} + \tau \mathbb{1}_{\mathcal{I}} \right) \\ &= \det \left(G_{\mathcal{K}}^{-1} \right) \times \det \left(\mathcal{X}_{\mathcal{I}}^{\top} \mathcal{X}_{\mathcal{I}} + \tau \mathbb{1}_{\mathcal{I}} \right) \end{aligned}$$

This form is convenient because it relies on the term $G_{\mathcal{K}}$ that we in any case need to compute the quadratic form. Similarly $\det \left(\mathcal{X}_{\mathcal{I}}^{\top} \mathcal{X}_{\mathcal{I}} + \tau \mathbb{1}_{\mathcal{I}} \right)$ is easily computed from the Cholesky factor $L_{\mathcal{I}}$.

Above we considered the case of turning on covariates, i.e. $\gamma_i = 0 \rightarrow \gamma_i = 1$. Since we assume that $|\gamma| \ll P$ these computations tend to dominate the computational cost. However, we must also consider the case of turning off covariates, i.e. $\gamma_i = 1 \rightarrow \gamma_i = 0$. To efficiently compute the required terms we make extensive use of the following identity. Let A , B , C , and D be appropriate $(M-1) \times (M-1)$, $(M-1) \times 1$, $1 \times (M-1)$, and 1×1 matrices, respectively. Then the identity

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix}^{-1} = \begin{pmatrix} \tilde{A} & \tilde{B} \\ \tilde{C} & \tilde{D} \end{pmatrix} \quad \implies \quad A^{-1} = \tilde{A} - \tilde{B} \tilde{D}^{-1} \tilde{C} \quad (47)$$

can be used to cheaply compute A^{-1} if the inverse of the block matrix $((A, B), (C, D))$ is available. In other words once we have computed $F_{\mathcal{I}} \equiv (\mathcal{X}_{\mathcal{I}}^T \mathcal{X}_{\mathcal{I}} + \tau \mathbb{1}_{\mathcal{I}})^{-1}$ using the Cholesky factor $L_{\mathcal{I}}$ we can use submatrices of $F_{\mathcal{I}}$ to cheaply compute the inverse of submatrices of $\mathcal{X}_{\mathcal{I}}^T \mathcal{X}_{\mathcal{I}} + \tau \mathbb{1}_{\mathcal{I}}$, which are precisely the quantities we need to compute Eqn. 34 for downdates of γ . In particular once the quadratic term has been computed, we can compute the log determinant by again appealing to Eqn. 46, using Eqn. 47 to compute $F_{\mathcal{I}}$ in $G_{\mathcal{K}}^{-1} = \mathcal{X}_{\mathcal{K}}^T \mathcal{X}_{\mathcal{K}} + \tau \mathbb{1}_{\mathcal{K}} - \mathcal{X}_{\mathcal{K}}^T \mathcal{X}_{\mathcal{I}} F_{\mathcal{I}} \mathcal{X}_{\mathcal{I}}^T \mathcal{X}_{\mathcal{K}}$ for redefinitions of \mathcal{I} and \mathcal{K} appropriate to a downdate.

A.5 Computational complexity

The primary computational cost in Subset wTGS, PG-wTGS, ASI, and the other MCMC algorithms considered in the main text arises in computing conditional PIPs of the form $p(\gamma_j = 1 | \gamma_{-j}, \mathcal{D})$ (linear regression case) or $p(\gamma_j = 1 | \gamma_{-j}, \omega, \mathcal{D})$ (count-based likelihood case) for $j = 1, \dots, P$, the principal ingredient for which are conditional marginal log likelihoods as in Eqn. 34. In the case of PG-wTGS, PG-TGS, PG-wGS, and ASI the next largest computational cost is usually sampling Pölya-Gamma variables, although this is $\mathcal{O}(N)$ and so the cost is moderate in most cases. For PG-wTGS, PG-TGS, PG-wGS, and ASI computing the MH acceptance probability (e.g. Eqn. 25 for the case of PG-wTGS) is another subdominant but non-negligible cost.

The precise computational cost of computing $p(\gamma_j = 1 | \gamma_{-j}, \mathcal{D})$ and $p(\gamma_j = 1 | \gamma_{-j}, \omega, \mathcal{D})$ depends on the details of how the formulae in Sec. A.4 are implemented. For example in the linear regression setting it can be advantageous to pre-compute $X^T X$ if the result can be stored in memory. In our experiments we do so whenever this is feasible (for a mid-grade GPU this is typically possible for $P \lesssim 4 \times 10^4$). In the case of PG-wTGS where ω changes every few MCMC iterations, pre-computing $\mathcal{X}^T \mathcal{X}$ is not advantageous. Note that to avoid possible accumulation of numerical errors we do not compute $F_{\mathcal{I}}$ or other quantities using computations from the previous MCMC iteration, although doing so is possible in principle for the linear regression case (see e.g. Zanella and Roberts (2019)).

We emphasize that in all cases the computational cost of a MCMC iteration is linear in the number of data points N . Thus in practice on commonly available hardware the $P \gtrsim 10^5$ regime is only viable for moderate $N \lesssim 10^4$. This is both because of memory constraints of storing X but also because $|\gamma|$ tends to increase (slowly) as P increases.⁹ Nevertheless running Subset wTGS in the regime with $P \gtrsim 10^5$ and $N \gtrsim 10^4$ should become increasingly viable in the near future, especially on a multi-GPU setup that can fit X in memory.

Linear regression case (wTGS) Using the various rank-1 update/downdate formulae from Sec. A.4 the result is $\mathcal{O}(|\gamma|NP + N|\gamma|^2 + |\gamma|^3)$ computational complexity per MCMC iteration if pre-computing $X^T X$ is not possible. If pre-computing $X^T X$ is possible the computational complexity per MCMC iteration is instead $\mathcal{O}(P|\gamma|^2 + |\gamma|^3)$ along with a one-time $\mathcal{O}(NP^2)$ cost to compute $X^T X$.

Linear regression case (Subset wTGS) Using the various rank-1 update/downdate formulae from Sec. A.4 the result is $\mathcal{O}(|\gamma|NS + N|\gamma|^2 + |\gamma|^3)$ computational complexity per MCMC iteration if pre-computing $X^T X$ is not possible. If pre-computing $X^T X$ is possible the computational complexity per MCMC iteration is instead $\mathcal{O}(S|\gamma|^2 + |\gamma|^3)$ along with a one-time $\mathcal{O}(NP^2)$ cost to compute $X^T X$.

PG-wTGS for Binomial and Negative Binomial regression Using the various rank-1 update/downdate formulae from Sec. A.4 the result is $\mathcal{O}(|\gamma|NP + N|\gamma|^2 + |\gamma|^3)$ computational complexity per MCMC iteration with $i > 0$ and $\mathcal{O}(N + N|\gamma|^2 + |\gamma|^3)$ per MCMC iteration with $i = 0$. For Subset PG-wTGS the computational complexity per MCMC iteration with $i > 0$ is instead $\mathcal{O}(|\gamma|NS + N|\gamma|^2 + |\gamma|^3)$.

We note that the asymptotic formulae reported above are somewhat misleading in practice, since most of the necessary tensor ops are highly-parallelizable and very efficiently implemented on modern hardware. For this reason Fig. 12 and Fig. 13 are particularly useful for understanding the runtime in practice, since the various parts of the computation will be more or less expensive depending on the precise regime and the underlying low-level implementation and hardware.

⁹Note this expectation may be violated if the response Y cannot be adequately modeled with a sparse set of covariates. In this regime all wTGS-like algorithms are computationally expensive and Bayesian variable selection is likely poorly suited to analyzing the relationship between X and Y . This regime is probably better served by other regularization approaches, e.g. ridge regression.

	$\gamma = 0$	$\gamma = 1$
$p(\gamma_i = 1 \mathcal{D}) \approx 0$	$\epsilon/P \ll 1$	$\frac{\epsilon/P}{p(\gamma_i=1 \mathcal{D})} \gg 1$
$p(\gamma_i = 1 \mathcal{D}) \approx 1$	$\frac{1}{1-p(\gamma_i=1 \mathcal{D})} \gg 1$	≈ 1

Table 1: We explore how the quantity $\frac{\eta(\gamma_{-i})}{p(\gamma_i|\gamma_{-i},\mathcal{D})} = \frac{p(\gamma_i=1|\mathcal{D})+\frac{\epsilon}{P}}{p(\gamma_i|\gamma_{-i},\mathcal{D})}$ varies as a function of γ and $\text{PIP}(i) = p(\gamma_i = 1|\mathcal{D})$ under the approximation $p(\gamma_i|\gamma_{-i},\mathcal{D}) \approx \gamma_i p(\gamma_i = 1|\mathcal{D}) + (1 - \gamma_i)(1 - p(\gamma_i = 1|\mathcal{D}))$. We further assume that either $p(\gamma_i = 1|\mathcal{D}) \ll \epsilon/P$ or $1 - p(\gamma_i = 1|\mathcal{D}) \ll \epsilon/P$. Off-diagonal entries in the table correspond to ‘greedy’ moves that are given large weight by wTGS. See Sec. A.7 for discussion.

A.6 TGS motivation: binary variables and Metropolized-Gibbs

To provide intuition for the Tempered Gibbs Sampling (TGS) strategy that underlies wTGS, we consider a single latent binary variable x governed by the probability distribution $p(x) = \text{Bernoulli}(q)$. A Gibbs sampler for this distribution simply samples $x \sim p$ in each iteration of the Markov chain. An alternative strategy is to employ a so-called Metropolized-Gibbs move w.r.t. x (Liu, 1996). For binary x this results in a proposal distribution that is deterministic in the sense that it always proposes a flip: $0 \rightarrow 1$ or $1 \rightarrow 0$. The corresponding Metropolis-Hastings (MH) acceptance probability for a move $x \rightarrow x'$ is given by

$$\alpha(x \rightarrow x') = \begin{cases} \min(1, \frac{q}{1-q}) & \text{if } x = 0 \\ \min(1, \frac{1-q}{q}) & \text{if } x = 1 \end{cases} \quad (48)$$

As is well known, this update rule is more statistically efficient than the corresponding Gibbs move (Liu, 1996). For our purposes, however, what is particularly interesting is the special case where $q = \frac{1}{2}$. In this case the acceptance probability in Eqn. 48 is identically equal to one. Consequently the Metropolized-Gibbs chain is deterministic:

$$\dots \rightarrow 0 \rightarrow 1 \rightarrow 0 \rightarrow 1 \rightarrow 0 \rightarrow 1 \rightarrow \dots \quad (49)$$

Indeed this Markov chain can be described as *maximally non-sticky*. This shows why building tempering into inference algorithms for binary latent variable models like that in Bayesian variable selection might be an attractive strategy for avoiding the stickiness of a vanilla Gibbs sampler.

A.7 The nature of local moves in wTGS

wTGS samples an auxiliary variable i controlled by the Gibbs update in Eqn. 7. To better understand how wTGS and its variants Subset wTGS and PG-wTGS are designed to efficiently explore regions of high posterior mass it is important to take a closer look at the form of these i updates. To do so we compute $\frac{\eta(\gamma_{-i})}{p(\gamma_i|\gamma_{-i},\mathcal{D})} \propto f(i|\gamma)$ in four regimes, see Table 1. We see that if covariate i is not included in the model ($\gamma = 0$) and has a small PIP covariate i will be chosen to be updated only infrequently and, furthermore, that the probability of i being chosen depends on ϵ ; thus ϵ controls the amount of exploration. By contrast if i has a large PIP and is currently excluded from the model ($\gamma = 0$) or if i has a small PIP and is currently included in the model ($\gamma = 1$), then $f(i|\gamma) \sim \mathcal{O}(1)$, with the consequence that i is likely to be flipped in the next move. This reflects the greedy nature of wTGS, which focuses much of its computational budget on turning on likely covariates and/or turning off unlikely covariates (i.e. un/likely under the posterior). Finally, if i has a large PIP and is currently on ($\gamma = 1$) it will occasionally be turned off (especially if no other covariates satisfy the ‘greedy’ condition described in the previous two sentences), which promotes exploration in and around posterior modes. In particular if covariate i is turned off and covariate i is highly correlated with j then turning off i allows for the possibility that j is turned on instead in the next MCMC iteration; indeed there will be a $\sim 50\%$ chance of doing so if i and j are the only covariates that satisfy the greedy condition. Taken together the behavior of $f(i|\gamma)$ reflected in Table 1 results in a satisfying balance between exploration and exploitation.

A.8 Importance weights

Importance weights $\rho \sim \phi^{-1}$ in wTGS and its variants (see e.g. Eqn. 6 and Algorithm 1) are bounded from above. For example for wTGS in the linear regression case we have

$$\phi(\gamma) = \frac{1}{2} \sum_{i=1}^P \frac{\text{PIP}(i) + \epsilon/P}{p(\gamma_i|\gamma_{-i}, \mathcal{D})} \geq \frac{\epsilon}{2} \quad (50)$$

with the consequence that (unnormalized) importance weights are bounded from above by $\frac{2}{\epsilon}$; note that in experiments we typically use $\epsilon = 5$. We also note that the bound in Eqn. 50 is somewhat loose. In practice the variance of importance weights normalized so that $\sum_{t=1}^T \rho^{(t)} = T$ is $\mathcal{O}(1)$; see the rightmost panels in Fig. 6-7 for variances observed in practice.

A.9 Proof of Proposition 1

In the main text we made use of an auxiliary variable representation in which the state i is explicitly included in the state space. For the present purpose it is more convenient to think of Subset wTGS, Algorithm 1, as acting on the space $\{0, 1\}^P \times \mathcal{P}$, where \mathcal{P} is the set of all subsets of $\{1, \dots, P\}$ of size S that contain the anchor set \mathcal{A} . The transition kernel can be written as

$$K((\gamma, \mathcal{S}) \rightarrow (\gamma', \mathcal{S}')) = \sum_{i \in \mathcal{S}} f(i|\gamma, \mathcal{S}) \delta(\gamma' - \text{flip}(\gamma|i)) U(\mathcal{S}'|i, \mathcal{A}) \quad (51)$$

where $f(i|\gamma, \mathcal{S})$ is the posterior conditional probability in Eqn. 13 and $\delta(\cdot)$ is the Dirac delta function. We first show that K is reversible w.r.t. the auxiliary target $f(\gamma, \mathcal{S}) = p(\gamma|\mathcal{D})\phi(\gamma, \mathcal{S})$, see Eqn. 11. As is evident from Eqn. 51, K is zero unless γ and γ' differ in exactly one coordinate—call it i —so that we have $\gamma_{-i} = \gamma'_{-i}$. Thus for non-zero K we have

$$K((\gamma, \mathcal{S}) \rightarrow (\gamma', \mathcal{S}')) = f(i|\gamma, \mathcal{S}) U(\mathcal{S}'|i, \mathcal{A}) \quad (52)$$

$$= \phi(\gamma, \mathcal{S})^{-1} \frac{\frac{1}{2}\eta(\gamma_{-i})}{p(\gamma_i|\gamma_{-i}, \mathcal{D})} U(\mathcal{S}|i, \mathcal{A}) U(\mathcal{S}'|i, \mathcal{A}) \quad (53)$$

which implies that

$$\frac{K((\gamma, \mathcal{S}) \rightarrow (\gamma', \mathcal{S}'))}{K((\gamma', \mathcal{S}') \rightarrow (\gamma, \mathcal{S}))} = \frac{\phi(\gamma', \mathcal{S}') p(\gamma'_i|\gamma'_{-i}, \mathcal{D})}{\phi(\gamma, \mathcal{S}) p(\gamma_i|\gamma_{-i}, \mathcal{D})} \quad (54)$$

$$= \frac{\phi(\gamma', \mathcal{S}') p(\gamma'|\mathcal{D})}{\phi(\gamma, \mathcal{S}) p(\gamma|\mathcal{D})} \quad (55)$$

$$= \frac{f(\gamma', \mathcal{S}')}{f(\gamma, \mathcal{S})} \quad (56)$$

where we used that $p(\gamma'_{-i}|\mathcal{D}) = p(\gamma_{-i}|\mathcal{D})$. Since reversibility is trivially satisfied if $K((\gamma, \mathcal{S}) \rightarrow (\gamma', \mathcal{S}'))$ is zero, we have thus shown that K is reversible w.r.t. $f(\gamma, \mathcal{S})$ and therefore f -invariant. Since our state space is finite and $f(i|\gamma, \mathcal{S}) > 0$ if $i \in \mathcal{S}$ it is also clear that our Markov chain is both irreducible and Harris recurrent. Thus our Markov chain satisfies the conditions of Theorem 17.0.1 in Meyn and Tweedie (2012) so that the Law of Large Numbers holds for any test function $h(\gamma, \mathcal{S}) : \{0, 1\}^P \times \mathcal{P} \rightarrow \mathbb{R}$. In particular for any test function $h(\gamma) : \{0, 1\}^P \rightarrow \mathbb{R}$ we can apply the Law of Large Numbers twice, once to $h\phi^{-1}$ and once to ϕ^{-1} (note that ϕ is bounded away from zero and bounded from above). If we let Z_f be the partition function of $f(\gamma, \mathcal{S})$, i.e. $Z_f \equiv \sum_{\gamma, \mathcal{S}} f(\gamma, \mathcal{S})$, then

$$\frac{1}{T} \sum_{t=1}^T h(\gamma^{(t)}) \phi^{-1}(\gamma^{(t)}, \mathcal{S}^{(t)}) \rightarrow \mathbb{E}_{f(\gamma, \mathcal{S})/Z_f} [h(\gamma) \phi^{-1}(\gamma, \mathcal{S})] = \mathbb{E}_{p(\gamma|\mathcal{D})} [h(\gamma)] / Z_f \quad (57)$$

and

$$\frac{1}{T} \sum_{t=1}^T \phi^{-1}(\gamma^{(t)}, \mathcal{S}^{(t)}) \rightarrow \mathbb{E}_{f(\gamma, \mathcal{S})/Z_f} [\phi^{-1}(\gamma, \mathcal{S})] = Z_f^{-1} \quad (58)$$

It follows that

$$\frac{\frac{1}{T} \sum_{t=1}^T h(\gamma^{(t)}) \phi^{-1}(\gamma^{(t)}, \mathcal{S}^{(t)})}{\frac{1}{T} \sum_{t=1}^T \phi^{-1}(\gamma^{(t)}, \mathcal{S}^{(t)})} \rightarrow \mathbb{E}_{p(\gamma|\mathcal{D})} [h(\gamma)] \quad (59)$$

or equivalently utilizing normalized weights $\{\rho^{(t)}\}$

$$\sum_{t=1}^T \rho^{(t)} h(\gamma^{(t)}) \rightarrow \mathbb{E}_{p(\gamma|\mathcal{D})} [h(\gamma)] \text{ as } T \rightarrow \infty \quad (60)$$

This finishes the proof of the central claim of Proposition 1. For the specific claim about Rao-Blackwellized PIP estimators see the next section.

A.10 Rao-Blackwellized PIP estimators

A naive estimator for $\text{PIP}(i) = p(\gamma_i = 1|\mathcal{D})$ directly uses weighted samples $\{(\rho^{(t)}, \gamma^{(t)})\}$ provided by Algorithm 3:

$$\text{PIP}(i) \approx \sum_t \rho^{(t)} \gamma_i^{(t)} \quad (61)$$

However, since wTGS and its variants compute conditional PIPs as part of inference, it is preferable to use a lower variance Rao-Blackwellized estimator instead:

$$\text{PIP}(i) \approx \sum_t \rho^{(t)} p(\gamma_i = 1|\gamma_{-i}^{(t)}, \mathcal{D}) \quad (62)$$

We use the appropriate version of Eqn. 62 in all experiments. In the case of Subset wTGS, Algorithm 1, only S conditional PIPs are computed in each MCMC iteration. Using the analog of Eqn. 62 would inflate the computational cost from $\mathcal{O}(S)$ to $\mathcal{O}(P)$, entirely defeating the purpose of Subset wTGS. Thus for Subset wTGS we use a partially Rao-Blackwellized estimator instead:

$$\text{PIP}(i) \approx \sum_t \rho^{(t)} \left\{ \mathcal{I}(i \in \mathcal{S}^{(t)}) p(\gamma_i = 1|\gamma_{-i}^{(t)}, \mathcal{D}) + \mathcal{I}(i \notin \mathcal{S}^{(t)}) \gamma_i^{(t)} \right\} \quad (63)$$

where $\mathcal{I}(\cdot)$ is an indicator function. In other words we use conditional PIPs if they are computed as part of inference (because $i \in \mathcal{S}$) and otherwise use raw γ samples. It is easy to see that the estimator in Eqn. 63 is unbiased, since the test statistic under consideration factorizes between γ and \mathcal{S} . Indeed if we let $q(\mathcal{S})$ denote the uniform distribution on \mathcal{P} and $\zeta = \mathbb{E}_{q(\mathcal{S})} [\mathcal{I}(i \in \mathcal{S})]$ then the proof in Sec. A.9 makes it clear that the partially Rao-Blackwellized estimator in Eqn. 63 converges to

$$\mathbb{E}_{p(\gamma|\mathcal{D})} \mathbb{E}_{q(\mathcal{S})} [\mathcal{I}(i \in \mathcal{S}) p(\gamma_i = 1|\gamma_{-i}, \mathcal{D}) + (1 - \mathcal{I}(i \in \mathcal{S})) \gamma_i] \quad (64)$$

$$= \zeta \mathbb{E}_{p(\gamma|\mathcal{D})} [p(\gamma_i = 1|\gamma_{-i}, \mathcal{D})] + (1 - \zeta) \mathbb{E}_{p(\gamma|\mathcal{D})} [\gamma_i] \quad (65)$$

$$= \zeta \text{PIP}(i) + (1 - \zeta) \text{PIP}(i) = \text{PIP}(i) \quad (66)$$

It is also evident that Eqn. 63 is lower variance than the raw estimator Eqn. 61.

A.11 ω -update in PG-wTGS

The acceptance probability for the ω -update in Sec. 4.2 is given by

$$\alpha(\omega \rightarrow \omega'|\gamma) = \min \left(1, \frac{p(Y|\gamma, \omega', X, C) p(\gamma) p(\omega'|C) p(\omega|\gamma, \hat{\beta}(\gamma, \omega'), \mathcal{D})}{p(Y|\gamma, \omega, X, C) p(\gamma) p(\omega|C) p(\omega'|\gamma, \hat{\beta}(\gamma, \omega), \mathcal{D})} \right) \quad (67)$$

where the ratio of proposal densities is given by

$$\frac{p(\omega|\gamma, \hat{\beta}(\gamma, \omega'), \mathcal{D})}{p(\omega'|\gamma, \hat{\beta}(\gamma, \omega), \mathcal{D})} = \frac{p(Y|\gamma, \omega, \hat{\beta}(\gamma, \omega'), X, C) p(\gamma) p(\omega|C) p(\hat{\beta}(\gamma, \omega'))}{\int d\hat{\omega} p(Y|\gamma, \hat{\omega}, \hat{\beta}(\gamma, \omega'), X, C) p(\gamma) p(\hat{\omega}|C) p(\hat{\beta}(\gamma, \omega'))} \times \left\{ \frac{p(Y|\gamma, \omega', \hat{\beta}(\gamma, \omega), X, C) p(\gamma) p(\omega'|C) p(\hat{\beta}(\gamma, \omega))}{\int d\hat{\omega} p(Y|\gamma, \hat{\omega}, \hat{\beta}(\gamma, \omega), X, C) p(\gamma) p(\hat{\omega}|C) p(\hat{\beta}(\gamma, \omega))} \right\}^{-1} \quad (68)$$

Simplifying we have that the ratio in $\alpha(\omega \rightarrow \omega' | \gamma)$ is given by

$$\begin{aligned} & \frac{p(Y|\gamma, \omega', X, C)}{p(Y|\gamma, \omega, X, C)} \frac{p(Y|\gamma, \omega, \hat{\beta}(\gamma, \omega'), X, C)}{\int d\hat{\omega} p(Y|\gamma, \hat{\omega}, \hat{\beta}(\gamma, \omega'), X, C)p(\hat{\omega}|C)} \frac{\int d\hat{\omega} p(Y|\gamma, \hat{\omega}, \hat{\beta}(\gamma, \omega), X, C)p(\hat{\omega}|C)}{p(Y|\gamma, \omega', \hat{\beta}(\gamma, \omega), X, C)} \\ &= \frac{p(Y|\gamma, \omega', X, C)}{p(Y|\gamma, \omega, X, C)} \frac{p(Y|\gamma, \omega, \hat{\beta}(\gamma, \omega'), X, C)}{p(Y|\gamma, \hat{\beta}(\gamma, \omega'), X, C)} \frac{p(Y|\gamma, \hat{\beta}(\gamma, \omega), X, C)}{p(Y|\gamma, \omega', \hat{\beta}(\gamma, \omega), X, C)} \\ &= \frac{p(Y|\gamma, \omega', X, C)}{p(Y|\gamma, \omega, X, C)} \frac{p(Y|\gamma, \omega, \hat{\beta}(\gamma, \omega'), X, C)}{p(Y|\gamma, \omega', \hat{\beta}(\gamma, \omega), X, C)} \frac{p(Y|\gamma, \hat{\beta}(\gamma, \omega), X, C)}{p(Y|\gamma, \hat{\beta}(\gamma, \omega'), X, C)} \end{aligned}$$

which is Eqn. 25 in the main text. Here

$$\frac{p(Y|\gamma, \omega, \hat{\beta}(\gamma, \omega'), X, C)}{p(Y|\gamma, \omega', \hat{\beta}(\gamma, \omega), X, C)} = \frac{\exp(\kappa \cdot \hat{\psi}(\gamma, \omega') - \frac{1}{2}\omega \cdot \hat{\psi}(\gamma, \omega')^2)}{\exp(\kappa \cdot \hat{\psi}(\gamma, \omega) - \frac{1}{2}\omega' \cdot \hat{\psi}(\gamma, \omega)^2)} \quad (69)$$

and

$$\frac{p(Y|\gamma, \hat{\beta}(\gamma, \omega), X, C)}{p(Y|\gamma, \hat{\beta}(\gamma, \omega'), X, C)} = \frac{\prod_n \exp(\hat{\psi}(\gamma, \omega)_n)^{Y_n}}{\prod_n (1 + \exp(\hat{\psi}(\gamma, \omega)_n))^{C_n}} \frac{\prod_n (1 + \exp(\hat{\psi}(\gamma, \omega')_n))^{C_n}}{\prod_n \exp(\hat{\psi}(\gamma, \omega')_n)^{Y_n}} \quad (70)$$

where

$$(\hat{\psi}(\gamma, \omega))_n \equiv \hat{\beta}(\gamma, \omega)_0 + \hat{\beta}(\gamma, \omega)_\gamma \cdot X_{n\gamma} \quad (71)$$

and

$$\hat{\beta}(\gamma, \omega) = (X_{\gamma+1}^T \Omega X_{\gamma+1} + \tau \mathbb{1}_{|\gamma|+1})^{-1} X_{\gamma+1}^T \kappa \in \mathbb{R}^{|\gamma|+1} \quad (72)$$

where as in Sec. A.4 X is here augmented with a column of all ones. As detailed in (Polson et al., 2013) the (approximate) Gibbs proposal distribution that results from conditioning on $\hat{\beta}$ is given by a Pòlya-Gamma distribution determined by C and $\hat{\psi}$:

$$p(\omega' | \gamma, \hat{\beta}(\gamma, \omega), \mathcal{D}) = \text{PG}(\omega' | C, \hat{\psi}(\gamma, \omega)) \quad (73)$$

In practice we do without the MH rejection step for ω in the early stages of burn-in to allow the MCMC chain to more quickly reach probable states.

A.12 ξ -adaptation in PG-wTGS and other wTGS variants

Here we discuss how $\xi > 0$ in Eqn. 20 can be adapted during burn-in. The same adaptation scheme (*mutatis mutandis*) can also be used for Algorithm 4, where the $i = 0$ state is introduced to allow for h -updates.

The magnitude of ξ controls the frequency of ω updates. Ideally ξ is such that an $\mathcal{O}(1)$ fraction of MCMC iterations result in a ω update, with the remainder of the computational budget being spent on γ updates. Typically this can be achieved by choosing ξ in the range $\xi \sim 1 - 5$. Here we describe a simple scheme for choosing ξ adaptively during burn-in to achieve the desired behavior.

We introduce a hyperparameter $f_\omega \in (0, 1)$ that controls the desired ω update frequency. Here f_ω is normalized such that $f_\omega = 1$ corresponds to a situation in which all updates are ω updates, i.e. all states in the MCMC chain are in the $i = 0$ state (something that would be achieved by taking $\xi \rightarrow \infty$). Since ω updates are of somewhat less importance for obtaining accurate PIP estimates than γ updates, we recommend a somewhat moderate value of f_ω , e.g. $f_\omega \sim 0.1 - 0.4$. For all experiments in this paper we use $f_\omega = 0.25$.

Our adaptation scheme proceeds as follows. We initialize $\xi^{(0)} = 5$. At iteration t during the burn-in a.k.a. warm-up phase we update $\xi^{(t)}$ as follows:

$$\xi^{(t+1)} = \xi^{(t)} + \frac{f_\omega - \frac{\xi^{(t)}}{\phi(\gamma^{(t)}, \omega^{(t)})}}{\sqrt{t+1}} \quad (74)$$

By construction this update aims to achieve that a fraction f_ω of MCMC states satisfy $i = 0$, since the quantity

$$\phi(\gamma, \omega) = \xi + \frac{1}{P} \sum_{i=1}^P \frac{\frac{1}{2}\eta(\gamma_{-i}, \omega)}{p(\gamma_i|\gamma_{-i}, \omega, \mathcal{D})} \quad (75)$$

encodes the total probability mass assigned to states $i = 0$ and $i > 0$.

A.13 Anchor set \mathcal{A} adaptation in Subset wTGS

We adopt a simple adaptation scheme for the anchor set \mathcal{A} . During burn-in we keep a running PIP estimate for each covariate using the partially Rao-Blackwellized estimator described in Sec. A.10. Periodically—in our experiments every 100 iterations—we update \mathcal{A} to be the A covariates exhibiting the largest PIPs according to the current running PIP estimate. At the end of the burn-in period the anchor set is updated one last time and remains fixed thereafter.

A.14 PG-wTGS for Negative Binomial regression

We specify in more detail how we can accommodate the negative binomial likelihood using Pölya-Gamma augmentation. Using the identity

$$\frac{(e^\psi)^a}{(1 + e^\psi)^b} = \frac{1}{2^b} e^{(a - \frac{1}{2}b)\psi} \mathbb{E}_{\text{PG}(\omega|b,0)} [\exp(-\frac{1}{2}\omega\psi^2)] \quad (76)$$

we write

$$\begin{aligned} \text{NegBin}(Y_n|\psi_n, \nu) &= \frac{\Gamma(Y_n + \nu)}{\Gamma(Y_n + 1)\Gamma(\nu)} \left(\frac{\exp(\psi_n + \psi_0 - \log \nu)}{1 + \exp(\psi_n + \psi_0 - \log \nu)} \right)^{Y_n} \left(\frac{1}{1 + \exp(\psi_n + \psi_0 - \log \nu)} \right)^\nu \\ &\propto \frac{1}{2^\nu} e^{\frac{1}{2}(Y_n - \nu)(\psi_n + \psi_0 - \log \nu)} \mathbb{E}_{p(\omega_n|Y_n + \nu, 0)} [\exp(-\frac{1}{2}\omega_n(\psi_n + \psi_0 - \log \nu)^2)] \end{aligned} \quad (77)$$

where as before $\psi_n = \beta_0 + \beta_\gamma \cdot X_{n\gamma}$ and ψ_0 is a user-specified offset. Here $\nu > 0$ controls the overdispersion of the negative binomial likelihood. We note that by construction the mean of $\text{NegBin}(Y_n|\psi_n, \nu)$ is given by $\exp(\psi_n + \psi_0)$. Thus ψ_0 (which can potentially depend on n) can be used to specify a prior mean for Y . This is equivalent to adjusting the prior mean of the bias β_0 in the case of constant ψ_0 .

Comparing to Sec. 4.1 we see that κ_n is now given by $\kappa_n = \frac{1}{2}(Y_n - \nu)$. When computing $\log p(Y|X, \gamma, \omega, \nu, \psi_0)$ the quantity \mathcal{Z} now becomes $\mathcal{Z}_j = \sum_n X_{n,j} (\kappa_n - \omega_n(\psi_0 - \log \nu))$, see Sec. A.4. One also picks up an additional factor of

$$\exp(\kappa \cdot (\psi_0 - \log \nu) - \frac{1}{2}\omega \cdot (\psi_0 - \log \nu)^2) \quad (78)$$

In particular we have the formula

$$\begin{aligned} \log p(Y|X, \gamma, \omega, \nu, \psi_0) &= \frac{1}{2} \mathcal{Z}_{\gamma+1}(\omega, \nu)^\top (X_{\gamma+1}^\top \Omega X_{\gamma+1} + \tau \mathbb{1}_{\gamma+1})^{-1} \mathcal{Z}_{\gamma+1}(\omega, \nu) \\ &\quad - \frac{1}{2} \log \det(X_{\gamma+1}^\top \Omega X_{\gamma+1} + \tau \mathbb{1}_{\gamma+1}) - \frac{1}{2} \log \det(\tau^{-1} \mathbb{1}_{\gamma+1}) \\ &\quad + \sum_n (\log \Gamma(Y_n + \nu) - \log \Gamma(\nu) - \log 2^\nu) \\ &\quad + \sum_n \kappa_n(\nu) (\psi_0 - \log \nu) - \frac{1}{2} \sum_n \omega_n (\psi_0 - \log \nu)^2 \end{aligned} \quad (79)$$

In our experiments we infer ν , which we assume to be unknown. For simplicity we put a flat (i.e. improper) prior on $\log \nu$, although other choices are easily accommodated. To do so we modify the ω update described in Sec. A.11 to a joint $(\omega, \log \nu)$ update. In more detail we use a simple gaussian random walk proposal for $\log \nu$ with a user-specified scale (we use 0.03 in our experiments). Conditioned on a proposal $\log \nu'$ we then sample a proposal ω' . Similar to the binomial likelihood case, we do this by computing

$$\hat{\beta}(\gamma, \omega, \nu) \equiv \mathbb{E}_{p(\beta|\gamma, \omega, \nu, \mathcal{D})} [\beta] \quad (80)$$

and use a proposal distribution $\omega' \sim p(\cdot|\gamma, \hat{\beta}, \nu', \mathcal{D})$. In the negative binomial case the formula for $\hat{\beta}$ in Eqn. 72 becomes

$$\hat{\beta}(\gamma, \omega, \nu) = (X_{\gamma+1}^\top \Omega X_{\gamma+1} + \tau \mathbb{1}_{|\gamma|+1})^{-1} X_{\gamma+1}^\top (\kappa - \omega(\psi_0 - \log \nu)) \in \mathbb{R}^{|\gamma|+1} \quad (81)$$

Additionally the proposal distribution is given by

$$p(\omega' | \gamma, \hat{\beta}(\gamma, \omega, \nu), \nu', \mathcal{D}) = \text{PG}(\omega' | Y + \nu', \hat{\psi}(\gamma, \omega, \nu)) \quad (82)$$

The acceptance probability can then be computed as in Sec. A.11, although in this case the resulting formulae are somewhat more complicated because of the need to keep track of ν and ν' as well as the fact that there is less scope for cancellations so that we need to compute quantities like $\Gamma(\nu)$. Happily, just like in the binomial regression case, the acceptance probability can be computed without recourse to the Pölya-Gamma density. In more detail the acceptance probability can be computed with help of the following expressions.

$$\alpha(\omega, \nu \rightarrow \omega', \nu' | \gamma) = \min(1, \tilde{\alpha}(\omega, \nu \rightarrow \omega', \nu' | \gamma)) \quad (83)$$

where $\tilde{\alpha}(\omega, \nu \rightarrow \omega', \nu' | \gamma)$ is given by

$$\tilde{\alpha}(\omega, \nu \rightarrow \omega', \nu' | \gamma) = \tilde{\alpha}_1 \times \tilde{\alpha}_2 \times \tilde{\alpha}_3 \quad (84)$$

with

$$\tilde{\alpha}_1 = 2^{N(\nu' - \nu)} \frac{p(Y|X, \gamma, \omega', \nu', \psi_0)}{p(Y|X, \gamma, \omega, \nu, \psi_0)} \quad (85)$$

$$\tilde{\alpha}_2 = \frac{\prod_n \left(e^{\hat{\psi}_n(\gamma, \omega, \nu) + \psi_0 - \log \nu'} \right)^{Y_n} \prod_n \left(1 + e^{\hat{\psi}_n(\gamma, \omega', \nu') + \psi_0 - \log \nu} \right)^{Y_n + \nu}}{\prod_n \left(e^{\hat{\psi}_n(\gamma, \omega', \nu') + \psi_0 - \log \nu} \right)^{Y_n} \prod_n \left(1 + e^{\hat{\psi}_n(\gamma, \omega, \nu) + \psi_0 - \log \nu'} \right)^{Y_n + \nu'}} \quad (86)$$

$$\tilde{\alpha}_3 = \frac{\prod_n e^{\kappa_n(\nu)(\hat{\psi}_n(\gamma, \omega', \nu') + \psi_0 - \log \nu)}}{\prod_n e^{\kappa_n(\nu')(\hat{\psi}_n(\gamma, \omega, \nu) + \psi_0 - \log \nu')}} \frac{\prod_n e^{-\frac{1}{2}\omega_n(\hat{\psi}_n(\gamma, \omega', \nu') + \psi_0 - \log \nu)^2}}{\prod_n e^{-\frac{1}{2}\omega'_n(\hat{\psi}_n(\gamma, \omega, \nu) + \psi_0 - \log \nu')^2}} \quad (87)$$

The correctness of these formulae can be checked numerically by comparing to the Pölya-Gamma density in regimes where the density can be easily and reliably computed. This is equally true for the binomial likelihood case.

A.15 Additional figures and tables

Additional figures for the first experiment in Sec. 7.1 are depicted in Fig. 6-7. Additional trace plots for the experiment in Sec. 7.2 are depicted in Fig. 8. In Table 2 we report PIP estimates for top hits in the cancer experiment in Sec. 7.3; we also include Fig. 9 and Fig. 10, where the latter is a companion of Fig. 4.

A.16 Additional experiments

A.16.1 Subset wTGS and anchor set size A

We run an additional experiment in which we vary the anchor set size A to complement the experimental results presented in Sec. 7.1. We use the same maize data with $P = 98385$ but generate an additional (statistically independent) semi-synthetic dataset from the same generative process. We report the results in Fig. 11, which complements the top panel in Fig. 1.

A.16.2 Subset wTGS and cancer data

We run an additional experiment to complement the experimental results presented in Sec. 7.1. We use the same cancer dataset as in Sec. 7.3 (so $N = 907$ and $P = 17273$) except we look at the gene ZEB2. We also use the continuous response as provided in the dataset (i.e. without quantization). See Fig. 12 for results.

A.16.3 PG-wTGS runtime

In Fig. 13 we depict MCMC iteration times for PG-wTGS for various values of N and P . To make the benchmark realistic we use semi-synthetic data derived from the DUSP4 cancer dataset ($N = 907$, $P = 17273$) used in Sec. 7.3. In particular for $N \neq 907$ and $P \neq 17273$ we subsample and/or add noisy data point replicates and/or add random covariates as needed. As discussed in more detail in Sec. A.5, PG-wTGS, PG-wGS, PG-TGS, and ASI all have similar runtimes, since each is

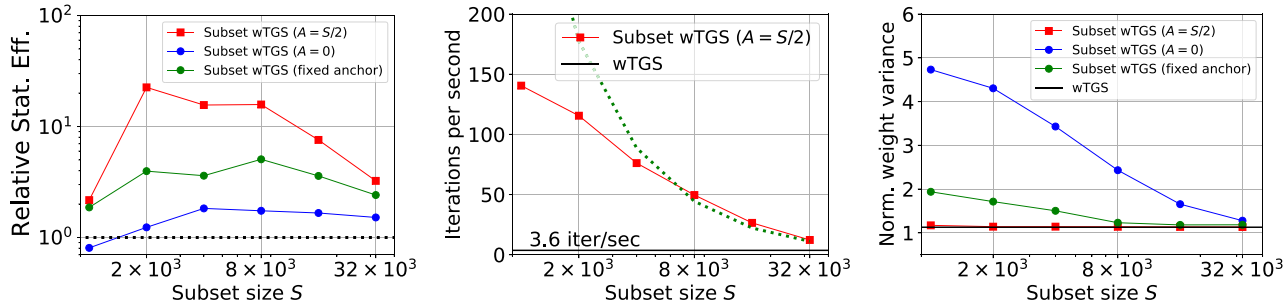


Figure 6: We report additional results for the experiment in Sec. 7.1 with $P = 98385$. These results are directly analogous to the results reported in the main text except we use synthetic data with a lower signal to noise ratio. In addition to Subset wTGS with our default setting of $A = S/2$, we also report results with $A = 0$ (i.e. no anchor set) and a fixed anchor set with $A = S/2$ (i.e. the anchor set is not updated during burn-in; instead \mathcal{A} is chosen with the initialization heuristic in Algorithm 1). See Sec. A.17 for additional details. **(Left)** We depict the relative statistical efficiency of Subset wTGS with subset size S compared to wTGS. As in the main text, we restrict our focus to covariates with PIPs above a threshold of 0.001. Subset wTGS with $A = S/2$ exhibits large gains in statistical efficiency relative to wTGS. The benefits of adapting \mathcal{A} are also apparent: we expect \mathcal{A} adaptation to be particularly important for noisy datasets. **(Middle)** We depict the number of iterations per second (IPS) for Subset wTGS as a function of S . The green dotted line depicts the IPS that would be expected if the latter scaled like S^{-1} . **(Right)** We depict the variance of weights $\{\rho^{(t)}\}$ obtained by running Subset wTGS. For the purposes of this figure, the weights are normalized so that the mean weight is equal to unity. We see that Subset wTGS with $A = 0$ has substantially higher variance. Moreover, in this regime (which is characterized by larger observation noise) the variance of Subset wTGS weights for $A = S/2$ is nearly identical to the variance of wTGS weights; importantly, both these variances are moderate.

dominated by the $\mathcal{O}(P)$ cost of computing $p(\gamma_j = 1 | \gamma_{-j}, \omega, \mathcal{D})$ for $j = 1, \dots, P$. As can be seen in Fig. 13 for any given N the iteration time is lower on CPU for small P , but GPU parallelization is advantageous for sufficiently large P .¹⁰ We also note that the computational complexity of PG-wTGS is no worse than linear in N , with the consequence that PG-wTGS can be applied to datasets with large N and large P in practice, at least if the sparsity assumption holds (i.e. most variables are excluded in the posterior: $|\gamma| \ll P$).

A.16.4 Hospital visit data and negative binomial regression

We consider a hospital visit dataset with $N = 1798$ considered in (Hilbe, 2011) and gathered from Arizona Medicare data. The response variable is length of hospital stay for patients undergoing a particular class of heart procedure and ranges between 1 and 53 days. We expect the hospital stay to exhibit significant dispersion and so we use a negative binomial likelihood. There are three binary covariates: sex (female/male), admission type (elective/urgent), and age (over/under 75). To make the analysis more challenging we add 97 superfluous covariates drawn i.i.d. from a unit Normal distribution so that $P = 100$.

Running PG-wTGS on the full dataset we find strong evidence for inclusion of two of the covariates: sex (PIP ≈ 0.95) and admission type (PIP ≈ 1.0). The corresponding coefficients are negative (-0.15 ± 0.02) and positive (0.63 ± 0.03), respectively.¹¹ This corresponds to shorter hospital stays for males and longer hospital stays for urgent admissions. In Fig. 14 (left) we depict trace plots for a few latent variables, each of which is consistent with good mixing; see also Fig. 15 for a zoomed-in view.

Next we hold-out half of the dataset in order to assess the quality of the model-averaged predictive distribution. We use the mean predicted hospital stay to rank the held-out patients and then partition them into two groups of equal size. Comparing this predicted partition to the observed partition of patients into short- and long-stay patients, we find a classification accuracy of 66.6%. In Fig. 14 (right) we depict a more fine-grained predictive diagnostic, namely Dawid’s Probability Integral Transform (PIT) (Dawid, 1984). Since the PIT values are approximately uniformly distributed, we conclude that the predictive distribution is reasonably well-calibrated, although probably somewhat overdispersed.

¹⁰All Pölya-Gamma sampling is done on CPU.

¹¹Each estimate is conditioned upon inclusion of the corresponding covariate in the model.

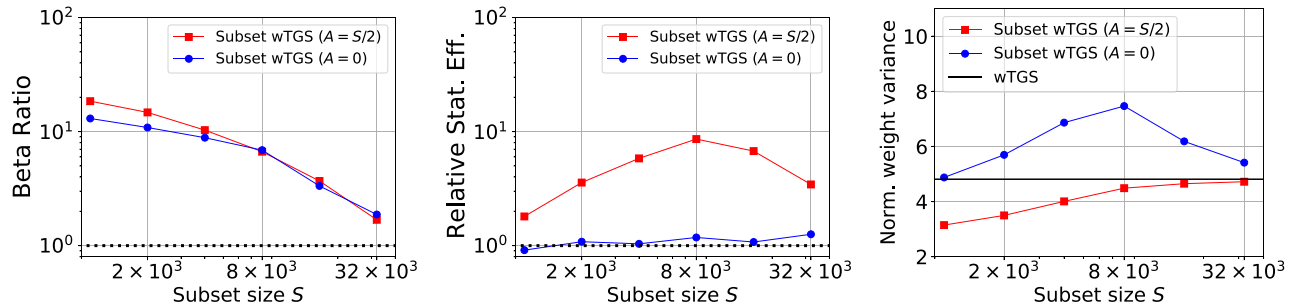


Figure 7: We report additional results for the experiment in Sec. 7.1 with $P = 98385$. See Sec. 7 and Sec. A.17 for additional details. **(Left)** We report a statistical efficiency ratio (relative to wTGS) for the inferred coefficients β . **(Middle)** This figure is identical to the top panel in Fig. 1, except the reported statistical efficiency is w.r.t. *all* covariates, instead of those with PIP above 0.001. Since the resulting (relative) statistical efficiency is somewhat less than that in Fig. 1, we see that some of the improved statistical efficiency of Subset wTGS is due to a variance trade-off between large PIP and small PIP covariates. We hypothesize that this is largely driven by the partial Rao-Blackwellization of the PIP estimator used in Subset wTGS: since low PIP covariates i are in \mathcal{S} only infrequently, the corresponding PIP estimate does not benefit much from Rao-Blackwellization and is consequently higher variance. Since there is generally no compelling need to obtain precise PIP estimates of low PIP covariates (e.g. distinguishing a PIP of 1.1×10^{-5} from 1.2×10^{-5}), this trade-off is well worth the resulting speed-ups. **(Right)** We depict the variance of weights $\{\rho^{(t)}\}$ obtained by running Subset wTGS. For the purposes of this figure, the weights are normalized so that the mean weight is equal to unity. We see that Subset wTGS with $A = 0$ has substantially higher variance, which explains its suboptimal performance. Importantly the variance of Subset wTGS with $A = \frac{1}{2}S$ is moderate and is indeed less than that of wTGS. We also note that a variance of 4 corresponds to a situation in which roughly half the weights are equal to 2 and the remainder are very small; this is indeed approximately what we observe in practice. A non-negligible number of low weight samples is the price we pay for the exploration enabled by tempering.

A.16.5 Health survey data and negative binomial regression

We consider the German health survey with $N = 1127$ considered in (Hilbe and Greene, 2007). The response variable is the annual number of visits to the doctor and ranges from 0 to 40 with a mean of 2.35. As in Sec. A.16.4, we expect significant dispersion and thus use a negative binomial likelihood. There are two covariates: i) a binary covariate for self-reported health status (not bad/bad); and ii) an age covariate, which ranges from 20 to 60. We normalize the age covariate so that it has mean zero and standard deviation one. To make the analysis more challenging we add 198 superfluous covariates drawn i.i.d. from a unit Normal distribution so that $P = 200$.

Running PG-wTGS on the full dataset we find strong evidence for inclusion of the health status covariate (PIP ≈ 1.0). The health status coefficient is positive (1.15 ± 0.10), suggesting that patients whose health is self-reported as bad have $e^{1.15} \approx 3.17$ times as many visits to the doctor as compared to those who report otherwise. This is consistent with the raw empirical ratio, which is about 3.16. We find that the data are very overdispersed and infer the dispersion parameter to be $\nu = 0.99 \pm 0.07$. See Sec. A.14 for additional details on PG-wTGS for negative binomial regression.

A.17 Experimental details

Large P experiments For both experiments we create semi-synthetic datasets as follows. We first shuffle the covariate indices. Next we divide the covariates into 20 approximately equally sized blocks. Within each block we compute the correlation between each pair of covariates and randomly select a pair with absolute correlation between 0.5 and 0.9; we then randomly choose one of the two indices. In this way we select 20 covariates, each of which exhibits non-trivial correlations with at least one other covariate. We then draw 20 coefficients from the uniform distribution on $[-1.0, -0.1] \cup [0.1, 1.0]$. We then use our synthetic coefficient vector β^* with 20 non-zero coefficients to generate a response Y_n as $Y_n = \beta^* \cdot X_n + \epsilon_n$ for $n = 1, \dots, N$ and where $\epsilon_n \sim \mathcal{N}(0, \sigma_0^2)$ is i.i.d. gaussian noise. We generate two datasets: one with $\sigma_0 = 0.5$ (these results are reported in Fig. 1 in the main text) and one with $\sigma_0 = 2.5$ (these results are reported in Fig. 6).

For the first experiment with $P = 98385$ we use all $N = 2267$ datapoints and a single fixed dataset. For the second experiment with $P > 98385$ we also use a single fixed dataset, but run experiments for 5 train/test splits, where half the

Gene	PG-wTGS-5M	PG-wTGS-250k	ASI-250k	Gene	PG-wTGS-5M	PG-wTGS-250k	ASI-250k
DUSP4	1.000	1.000 / 1.000	1.000 / 1.000	HNF1B	1.000	1.000 / 1.000	1.000 / 1.000
PPP2R3A	0.669	0.576 / 0.647	0.466 / 0.318	CAP2	0.323	0.322 / 0.354	0.079 / 0.068
MIA	0.383	0.338 / 0.365	0.282 / 0.138	C12orf54	0.172	0.113 / 0.152	0.145 / 0.005
KRT80	0.272	0.364 / 0.296	0.502 / 0.482	AQP1	0.122	0.127 / 0.128	0.061 / 0.096
RELN	0.243	0.286 / 0.219	0.346 / 0.502	FAM43B	0.067	0.059 / 0.050	0.013 / 0.077
ZNF132	0.096	0.124 / 0.093	0.120 / 0.142	KLRF1	0.063	0.059 / 0.065	0.032 / 0.034
TRIM51	0.094	0.075 / 0.099	0.080 / 0.053	ARMC4	0.059	0.062 / 0.035	0.027 / 0.120
ZNF471	0.083	0.107 / 0.081	0.139 / 0.181	SERPINE1	0.050	0.047 / 0.052	0.040 / 0.003
S100B	0.063	0.053 / 0.068	0.028 / 0.022	CLIC6	0.049	0.050 / 0.057	0.013 / 0.094
ZNF571	0.062	0.065 / 0.058	0.064 / 0.065	GSDME	0.048	0.056 / 0.050	0.101 / 0.066
ZNF304	0.060	0.067 / 0.069	0.095 / 0.068	UGCG	0.044	0.042 / 0.034	0.108 / 0.093
ZNF772	0.040	0.039 / 0.034	0.044 / 0.028	NEK6	0.039	0.041 / 0.041	0.019 / 0.005
RXRG	0.032	0.026 / 0.025	0.010 / 0.028	SERPINA10	0.032	0.027 / 0.026	0.007 / 0.017
ZNF17	0.031	0.033 / 0.037	0.047 / 0.040	ECH1	0.029	0.028 / 0.022	0.054 / 0.029
ZNF134	0.026	0.026 / 0.029	0.026 / 0.033	KIF1C	0.029	0.034 / 0.024	0.066 / 0.081
KRT7	0.025	0.025 / 0.021	0.016 / 0.206	S100A4	0.028	0.029 / 0.025	0.083 / 0.009
ZNF71	0.020	0.022 / 0.014	0.045 / 0.023	MSANTD3	0.023	0.029 / 0.023	0.005 / 0.006
CCIN	0.019	0.035 / 0.037	0.019 / 0.026	PLIN3	0.023	0.019 / 0.020	0.043 / 0.007
ZNF419	0.018	0.019 / 0.017	0.018 / 0.006	IL4R	0.021	0.018 / 0.021	0.016 / 0.003
ZMYM3	0.017	0.016 / 0.027	0.014 / 0.036	SHBG	0.020	0.016 / 0.022	0.005 / 0.027

Table 2: These tables are companions to Fig. 4, Fig. 9, and Fig. 10. We depict PIP estimates for DUSP4 (left) and HNF1B (right). In each case we include the result from a PG-wTGS run with five million samples as well as two shorter runs from PG-wTGS and ASI (with the two results separated by a slash). We depict the top 20 genes as determined by the long PG-wTGS run. The much lower variance and higher accuracy of PG-wTGS are apparent. Indeed if we take the top 20 PIP estimates from the long run as truth then we can compute the mean absolute error (MAE) of the short run estimates. The resulting MAEs are 0.007 (0.014) for PG-wTGS and 0.043 (0.061) for ASI for HNF1B (DUSP4), respectively. In other words the ASI MAE is about five times larger than the PG-wTGS MAE.

data is held-out for testing. As described in the main text, to obtain a dataset with $P > 98385$ covariates we augment the maize data with covariates drawn i.i.d. from a unit Normal distribution. We set the prior inclusion probability $h = 10/P$, the prior precision to $\tau = 10^{-4}$, and $\epsilon = 5$.

The relative statistical efficiency reported in Fig. 2 is defined as a ratio of effective samples sizes per unit time, which is equivalent to a ratio of time-normalized variances. It is computed as follows:

$$\frac{\text{StatEff}(\text{Subset wTGS})}{\text{StatEff}(\text{wTGS})} = \frac{\sigma_{\text{wTGS}}^2 T_{\text{wTGS}}}{\sigma_{\text{Subset wTGS}}^2 T_{\text{Subset wTGS}}} \quad (88)$$

where e.g. T_{wTGS} is the runtime of wTGS and σ_{wTGS}^2 is the corresponding variance for the estimator of interest. In Fig. 2 the estimator of interest is the sum of PIPs over all covariates with a PIP that exceeds a threshold of 0.001, of which there are 53. To determine these “relevant” covariates and compute reference PIPs to compute the required variance in Eqn. 88, we run 10 independent wTGS chains with 50k samples each and compute a mean PIP across the 10 chains (this requires about 40 hours of GPU compute). Note that these long chains are independent of the shorter chains used to assess the statistical efficiency of each method. For each short chain we collect 20k post-adaptation samples, except for wTGS where we collect 10k. In all cases there are 5k burn-in iterations. For each method we run 10 independent chains; the resulting variability determines the variance in Eqn. 88. Together with the runtime, this allows us to compute the (relative) statistical efficiency.

Runtime results are obtained with a NVIDIA Tesla T4 GPU. The predictive and coefficient RMSEs reported in Fig. 2 are normalized by the standard deviation of Y and the euclidean norm of β^* , respectively, for interpretability: with this normalization a RMSE less than unity is a strict improvement over guessing zero.

PG-wGS/PG-TGS/PG-wTGS/ASI For experiments with count-based likelihoods (unless specified otherwise) we set the prior precision $\tau = 0.01$ and $h = 5/P$. We choose the exploration parameter ϵ that enters $\eta(\cdot)$ to be $\epsilon = 5$. We use the ξ -adaptation scheme described in Sec. A.12.

We note that PG-TGS uses $\eta(\cdot) = 1$ but still utilizes Metropolized-Gibbs moves to update γ_i ; these moves result in

deterministic flips because of tempering. By contrast PG-wGS uses the same weighting function $\eta(\cdot)$ as in PG-wTGS but there is no tempering, with the consequence that γ_i still undergoes Metropolized-Gibbs moves but the acceptance probability is no longer identically equal to one. See Eqn. 48 for the resulting acceptance probability.

ASI has several hyperparameters which we set as follows. We set the exponent λ_{ASI} that controls adaptation to $\lambda_{\text{ASI}} = 0.75$. We set $\epsilon_{\text{ASI}} = 0.1/P$ as suggested by the authors. We target an acceptance probability of $\tau_{\text{ASI}} = 0.25$.

Correlated covariates scenario The covariates for $p = 3, 4, \dots, P$ are generated independently from a standard Normal distribution: $X_{n,p} \sim \mathcal{N}(0, 1)$ for all n . We then generate $z \in \mathbb{R}^N$ with $z_n \sim \mathcal{N}(0, 1)$ and set $X_{n,p=1} \sim \mathcal{N}(z_n, 10^{-4})$ and $X_{n,p=2} \sim \mathcal{N}(z_n, 10^{-4})$. That is the first two covariates are almost identical apart from a small amount of noise. We then generate the responses Y_n using success logits given by $\psi_n = z_n$. The total count C_n for each data point is set to 10. Consequently the true posterior concentrates on two modes with $\gamma = (1, 0, 0, \dots)$ and $\gamma = (0, 1, 0, \dots)$. We set $h = 1/P$ and run each algorithm for 10 thousand burn-in/warmup iterations and use the subsequent 100 thousand samples for analysis.

Cancer data All chains are run for 25 thousand burn-in/warmup iterations.

Inferring h We follow the discrete time branching process simulator setup described in the supplement of Jankowiak et al. (2022). We use identical hyperparameters to those used in the reference except we vary the number of causal effects in each simulation. In addition for each simulation we choose effect sizes from the uniform distribution on $[-0.10, -0.02] \cup [0.02, 0.10]$. We choose $\alpha_h = 0.25$ and $\beta_h = 250$ to define the Beta prior over h ; this choice corresponds to a relatively broad prior with prior mean $h = 0.001$ (which corresponds to 3 causal mutations expected a priori).

We provide some intuition for the behavior observed in Fig. 5. Note that the diffusion-based likelihood that underlies Jankowiak et al. (2022) is an approximation of the underlying discrete time branching process dynamics. Consequently the model is not perfectly well specified. For this reason—and because of the inherent noisiness of the data—as h increases, there may be a tendency to push h up further, since doing so allows the model to achieve a better fit of the observed pandemic, even if some of the identified mutations may be spurious. This explains the larger tails observed for simulations with 10 causal mutations. This is a general reminder that one needs to proceed with caution when placing a prior on h ; in some cases it may be more sensible to assume fixed values of h and do a sensitivity analysis to assess sensitivity to prior assumptions.

Subset wTGS and cancer experiment The experimental details closely follow the experiment in Sec. 7.1, except in contrast the data we use here is not semi-synthetic. We run 10 independent chains with wTGS for 500k iterations each to compute reference PIPs. We then run 20 additional independent chains for each method (i.e. vanilla wTGS and Subset wTGS for various values of S) for 50k iterations; the results of these chains are then used to compute the relative statistical efficiency. To do so we use the PIP over all covariates as the estimator of interest. In all cases we allow for 10k burn-in iterations.

Runtime experiment For each value of N and P we run each MCMC chain for 2000 burn-in iterations and report iteration times averaged over a subsequent 10^4 iterations; we report results in Fig. 13.

Hospital data We run PG-wTGS for 10 thousand burn-in iterations and use the subsequent 100 thousand samples for analysis. The 899 held-out patients are chosen at random. We use a random walk proposal scale for $\log \nu$ of 0.03. We set ψ_0 to be the logarithm of the mean of the observed Y (this is equivalent to shifting the prior mean of the bias β_0 ; see Sec A.14).

Health survey data We run PG-wTGS for 10 thousand burn-in iterations and use the subsequent 100 thousand samples for analysis. We use a random walk proposal scale for $\log \nu$ of 0.03. We set ψ_0 to be the logarithm of the mean of the observed Y (this is equivalent to shifting the prior mean of the bias β_0 ; see Sec A.14).

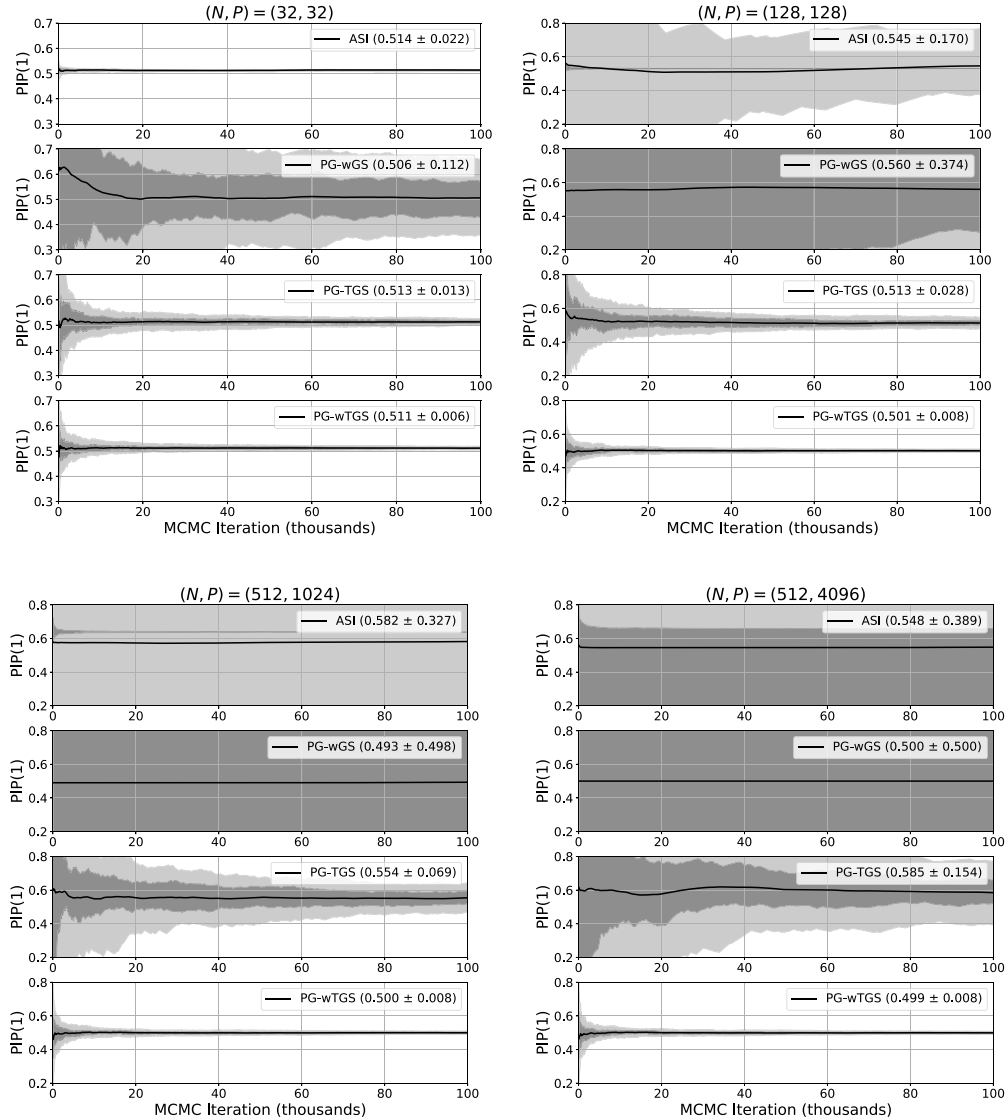


Figure 8: This is a companion figure to Fig. 3 in the main text. We depict posterior inclusion probability (PIP) estimates for the first covariate in the scenario described in Sec. 7.2 for four different MCMC methods and four different values of (N, P) . At each iteration t the PIP is computed using all samples obtained through iteration t . The mean PIP is depicted with a solid black line and light and dark grey confidence intervals denote 10%–90% and 30%–70% quantiles, respectively. The true PIP is almost exactly $\frac{1}{2}$. In each case we run 100 independent chains. For each method we also report the final PIP estimate (mean and standard deviation) in parentheses.

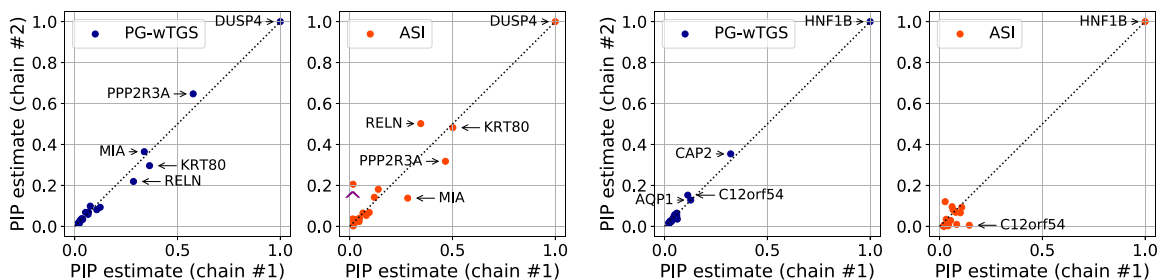


Figure 9: We depict PIP estimates for two independent MCMC chains for two cancer datasets (left: DUSP4; right: HNF1B) using two MCMC methods. For each method we depict the top 20 PIPs from chain #1 paired with the estimate from chain #2. The PG-wTGS estimates show much better inter-chain concordance. For example, the PIPs obtained with ASI for KRT7 on the DUSP4 dataset (marked with a purple caret \wedge) differ by a factor of 12.8 between the two chains, while the two PG-wTGS estimates are 0.025 and 0.021. Similarly the PIPs obtained with ASI for C12orf54 on the HNF1B dataset differ by a factor of 26.9, while the two PG-wTGS estimates are 0.113 and 0.152. See Sec. 7.3 for details.

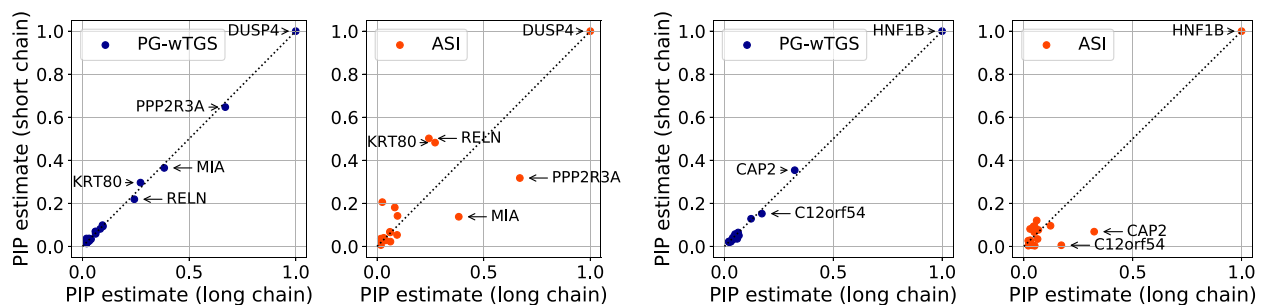


Figure 10: In this companion figure to Fig. 4 we compare PIP estimates obtained from short PG-wTGS and ASI chains with 2.5×10^5 samples to a long PG-wTGS chain with 5×10^6 samples. For each method we depict the top 20 PIPs from the long chain paired with estimates from the short chains. Note that this figure is identical to Fig. 4 except the two short chains are independent of the two short chains in Fig. 4. The PG-wTGS estimates obtained with the short chains are significantly more accurate than is the case for ASI. See Sec. 7.3 for details.

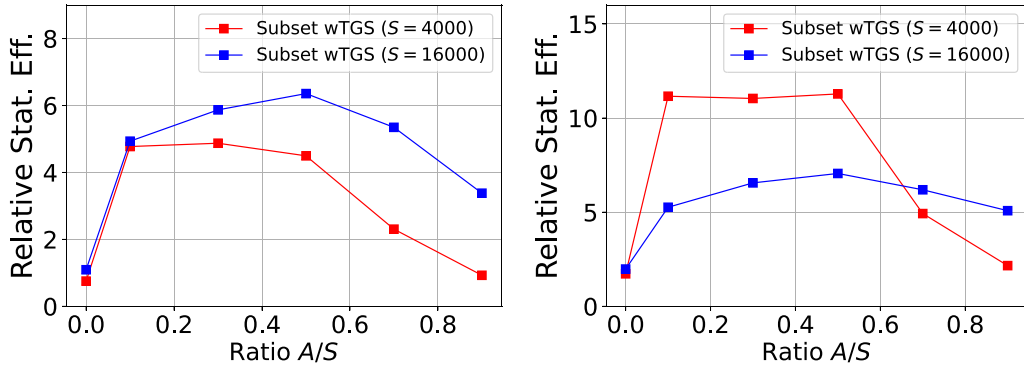


Figure 11: We explore the impact of varying the anchor set size A for two subset sizes $S \in [4000, 16000]$. The ratio A/S , which is depicted along the horizontal axis, ranges from 0 to $\frac{9}{10}$. On the vertical axis we report the relative statistical efficiency compared to wTGS. On the left the relative statistical efficiency for estimating PIPs for *all* covariates is computed using Eqn. 88, while on the right the relative statistical efficiency metric is computed for covariates with a PIP above 0.001. We see that Subset wTGS exhibits sizable efficiency gains above wTGS provided that A is not too close to 0 or S . This observation motivates our default choice of $A = \frac{S}{2}$. With this choice, provided S is sufficiently large, the anchor set \tilde{A} is expected to contain all (or at least most) covariates with non-trivial PIPs after adaptation. Moreover, with $A = \frac{S}{2}$ half the ‘PIP budget’ is spent on exploration, which is important for marginalizing out small PIP ‘background’ covariates and obtaining accurate PIP estimates for important covariates. Results are obtained with a NVIDIA Tesla K80. See Sec. A.16.1 for additional details.

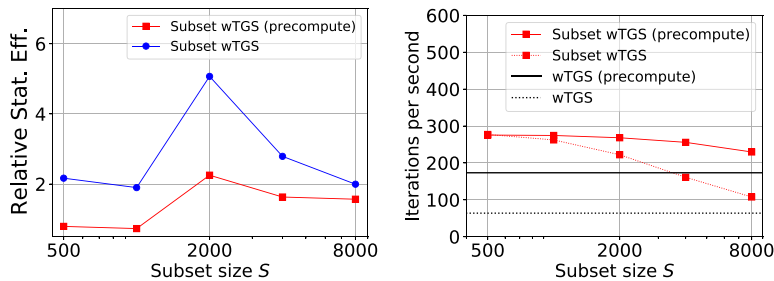


Figure 12: We explore the moderately large P regime using cancer data from a ZEB2 gene knockout experiment. Since $P = 17273$ is somewhat moderate, and since we are doing linear regression and not generalized linear regression with a non-linear link function, the covariate covariance matrix $X^T X$ can be pre-computed and stored in memory on a mid-grade GPU; this leads to iteration speed-ups of a factor of $\sim 2-3$. See Sec. A.16.2 for additional details on the experiment. **(Left)** We depict the relative statistical efficiency of Subset wTGS with subset size S compared to wTGS. The gain in statistical efficiency is largest when $X^T X$ cannot be pre-computed (blue), which would be the case for Negative Binomial and Binomial Regression. Note that the moderate gain in statistical efficiency in this regime is not surprising; the runtime results in the rightmost panel make it clear that GPU utilization is only moderate and so the speed-ups that result in switching from wTGS to Subset wTGS are limited. **(Right)** We depict the number of iterations per second (IPS) for Subset wTGS as a function of S as well as wTGS. Results are obtained with a NVIDIA Tesla T4 GPU.

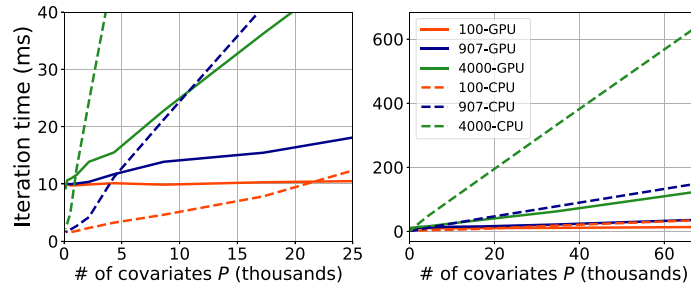


Figure 13: We depict MCMC iteration times in milliseconds for PG-wTGS on CPU and GPU as the number of covariates P is varied. We also vary the number of data points $N \in \{100, 907, 4000\}$. See Sec. A.16.3 for details. Note that the figure on the left is a magnified version of the figure on the right. The CPU has 24 cores (Intel Xeon Gold 5220R 2.2GHz) and the GPU is a NVIDIA Tesla K80 GPU.

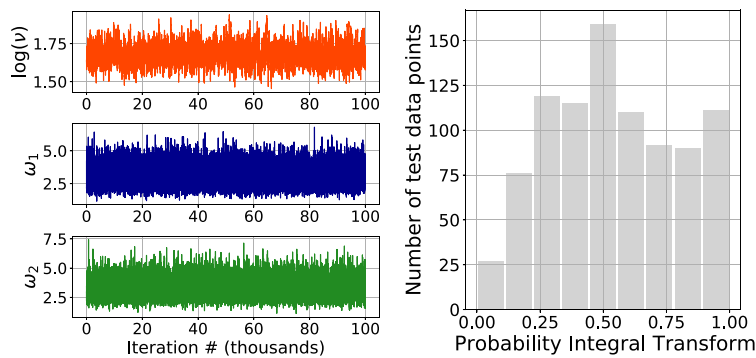


Figure 14: **Left:** We depict trace plots for $\log(\nu)$ and two randomly chosen Pòlya-Gamma variates ω_1 and ω_2 for a PG-wTGS run on the data in Sec. A.16.4. The data is quite dispersed, with the posterior mean of the dispersion parameter ν being about 5.4. **Right:** We depict the Probability Integral Transform histogram for 899 held-out test points for the hospital data in Sec. A.16.4.

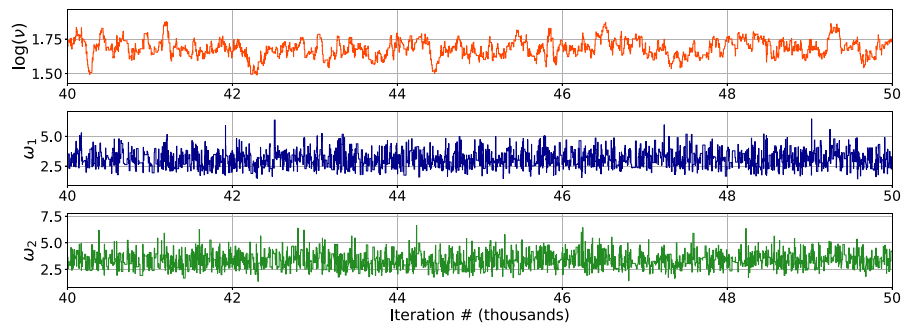


Figure 15: In this zoomed-in companion figure to Fig. 14 we depict trace plots for $\log(\nu)$ and two randomly chosen Pòlya-Gamma variates ω_1 and ω_2 for a PG-wTGS run on the hospital data in Sec. A.16.4.