# Scalable Gibbs sampler using randomized sketching for massive and high-dimensional Bayesian regression

Gyeongmin Park [1] Gyuhyeong Goh [1] Dipak K. Dey [2]

[1]Department of Statistics, Kyungpook National University, Daegu, South Korea [2]Department of Statistics, University of Connecticut, Storrs, CT, USA

## Introduction

- High-dimensional Bayesian regression with spike-and-slab priors offers a unified framework for variable selection, estimation, and model uncertainty quantification via Gibbs sampling.
- However, standard Gibbs sampling becomes computationally prohibitive when the number of predictors $p$ is large since its computational complexity is $O(p^3)$ per iteration.
- Many existing scalable Bayesian methods for high-dimensional regression are tailored to the $n \ll p$ setting and may still face computational bottlenecks when $n \gg p$.
- To address these computational challenges, we propose a novel sketching-based approach for scalable Bayesian inference in massive, high-dimensional settings where $n \gg p$ but $p$ remains large.

## Bayesian regression with spike-and-slab priors

- The Bayesian regression model with spike-and-slab priors can be expressed as the following hierarchical representation (George and McCulloch 1993):

$$\begin{aligned} \text{[Likelihood]} \quad & \mathbf{y} \mid \boldsymbol{\beta}, \sigma^2 \sim N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_n), \\ \text{[Prior]} \quad & \beta_j \mid \sigma^2, \gamma_j \overset{\text{ind}}{\sim} N(0,\, \sigma^2 d_{\gamma_j}), \\ & \sigma^2 \sim \mathsf{IG}(a_0, b_0), \\ & \gamma_j \overset{\text{iid}}{\sim} \mathsf{Ber}(w), \end{aligned}$$

  where $w$ is the prior inclusion probability, $d_{\gamma_j} = (1-\gamma_j)\tau_0^2 + \gamma_j \tau_1^2$ with $d_0 = \tau_0^2$ and $d_1 = \tau_1^2$, and $\mathbf{D}_{\boldsymbol{\gamma}} = \mathrm{diag}(d_{\gamma_1}, \ldots, d_{\gamma_p})$. Let $\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}^{-1} = \mathbf{X}^\top \mathbf{X} + \mathbf{D}_{\boldsymbol{\gamma}}^{-1}$ denote the unscaled conditional precision matrix; then $\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}$ is the unscaled conditional covariance matrix.

- Posterior sampling can be performed by iterating the following Gibbs sampling steps:

  **Step 1:** Generate $\boldsymbol{\beta}$ from

$$\boldsymbol{\beta} \mid \mathbf{y},\, \sigma^2, \boldsymbol{\gamma} \sim N\left(\boldsymbol{\Sigma}_{\boldsymbol{\gamma}} \mathbf{X}^\top \mathbf{y},\, \sigma^2 \boldsymbol{\Sigma}_{\boldsymbol{\gamma}}\right).$$

  **Step 2:** Generate $\sigma^2$ from

$$\sigma^2 \mid \mathbf{y}, \boldsymbol{\beta}, \boldsymbol{\gamma} \sim \mathsf{IG}\left(a_0 + \frac{n+p}{2},\, b_0 + \frac{\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \boldsymbol{\beta}^\top \mathbf{D}_{\boldsymbol{\gamma}}^{-1} \boldsymbol{\beta}}{2}\right).$$

  **Step 3:** Generate $\gamma_j$ from

$$\gamma_j \mid \mathbf{y}, \sigma^2, \boldsymbol{\beta} \overset{\text{ind}}{\sim} \mathsf{Ber}(p_j), \quad p_j = \frac{w \phi(\beta_j; 0, \sigma^2 d_1)}{w \phi(\beta_j; 0, \sigma^2 d_1) + (1-w)\phi(\beta_j; 0, \sigma^2 d_0)},$$

  where $\phi(\cdot; 0, v)$ denotes the $N(0, v)$ density.

∗ **Remark**: Computing $\boldsymbol{\Sigma}_{\boldsymbol{\gamma}} = (\mathbf{X}^\top \mathbf{X} + \mathbf{D}_{\boldsymbol{\gamma}}^{-1})^{-1}$ incurs a computational cost of $O(p^3)$ per iteration, which becomes prohibitively expensive as $p$ increases.

## Existing scalable Gibbs samplers

- **Efficient Gaussian sampler.**(Bhattacharya, Chakraborty, and Mallick 2016) The Gaussian update can avoid direct $p \times p$ Cholesky decomposition by using the Woodbury identity:

$$(\mathbf{X}^\top \mathbf{X} + \mathbf{D}_{\boldsymbol{\gamma}}^{-1})^{-1} = \mathbf{D}_{\boldsymbol{\gamma}} - \mathbf{D}_{\boldsymbol{\gamma}} \mathbf{X}^\top (\mathbf{I}_n + \mathbf{X}\mathbf{D}_{\boldsymbol{\gamma}} \mathbf{X}^\top)^{-1} \mathbf{X}\mathbf{D}_{\boldsymbol{\gamma}}.$$

  This reduces the main Gaussian update to an $n \times n$ linear system, with cost about $\mathcal{O}(n^2 p)$.
- **Scalable Spike-and-Slab, $S^3$.**(Biswas, Mackey, and Meng 2022) The $S^3$ sampler updates the active-set covariance incrementally as predictors enter or leave. Schur-complement add/delete updates avoid repeated full matrix construction, giving about $\mathcal{O}(np)$ cost per iteration for linear regression.
- **Remaining bottleneck.** These methods are mainly useful in the $n \ll p$ setting. When $n \approx p$ or $n > p$ with large $p$, they can still be inefficient because each iteration uses full-data terms.
- **Sketched data approach.**(Guhaniyogi and Scheffler 2025) These samplers can be applied by replacing $(\mathbf{X}, \mathbf{y})$ with sketched data $(\tilde{\mathbf{X}}, \tilde{\mathbf{y}}) = (\mathbf{S}\mathbf{X}, \mathbf{S}\mathbf{y})$, where $\mathbf{S} \in \mathbb{R}^{k \times n}$ with $S_{ij} \sim N(0, 1/k)$. This reduces $n$ to $k$, but approximating the exact Gibbs sampler at the density level can be difficult.

## Proposed method: block-wise sketched Gibbs sampler

- Given the current indicator vector $\boldsymbol{\gamma}$, define the active and inactive sets

$$\mathcal{A} = \{j : \gamma_j = 1\}, \qquad \mathcal{I} = \{j : \gamma_j = 0\}.$$

  Accordingly, partition the design matrix and regression coefficients as

$$\mathbf{X} = [\mathbf{X}_{\mathcal{A}}\ \mathbf{X}_{\mathcal{I}}], \qquad \boldsymbol{\beta} = (\boldsymbol{\beta}_{\mathcal{A}}^\top, \boldsymbol{\beta}_{\mathcal{I}}^\top)^\top.$$

- To reduce the size of the data, we use a Gaussian sketching matrix $\mathbf{S} \in \mathbb{R}^{k \times n}$ with $S_{ij} \sim N(0, 1/k)$ and form the sketched data $\tilde{\mathbf{X}} = \mathbf{S}\mathbf{X}$ and $\tilde{\mathbf{y}} = \mathbf{S}\mathbf{y}$, where $k \ll \min(n, p)$.
- Instead of sketching all conditional updates, the proposed sampler applies sketching selectively to the computationally expensive inactive covariance update and the global variance update.

**Algorithm: Overall block-wise sketched Gibbs sampler**

1. **Update the regression coefficients:**

$$\boldsymbol{\beta}_{\mathcal{I}} \mid \mathbf{y}, \sigma^2, \boldsymbol{\beta}_{\mathcal{A}}, \boldsymbol{\gamma} \sim N(\underbrace{\boldsymbol{\Sigma}_{\mathcal{I}} \mathbf{X}_{\mathcal{I}}^\top (\mathbf{y} - \mathbf{X}_{\mathcal{A}} \boldsymbol{\beta}_{\mathcal{A}})}_{\text{via Algorithm 1}}, \sigma^2 \tilde{\boldsymbol{\Sigma}}_{\mathcal{I}}), \quad \underbrace{\tilde{\boldsymbol{\Sigma}}_{\mathcal{I}}^{-1} = \tilde{\mathbf{X}}_{\mathcal{I}}^\top \tilde{\mathbf{X}}_{\mathcal{I}} + \tau_0^{-2} \mathbf{I}}_{\text{via Algorithm 2}},$$

$$\boldsymbol{\beta}_{\mathcal{A}} \mid \mathbf{y}, \sigma^2, \boldsymbol{\beta}_{\mathcal{I}}, \boldsymbol{\gamma} \sim \underbrace{N(\boldsymbol{\Sigma}_{\mathcal{A}} \mathbf{X}_{\mathcal{A}}^\top (\mathbf{y} - \mathbf{X}_{\mathcal{I}} \boldsymbol{\beta}_{\mathcal{I}}), \sigma^2 \boldsymbol{\Sigma}_{\mathcal{A}})}_{\text{via Algorithm 3}}.$$

2. **Update the variance:**

$$\sigma^2 \mid \mathbf{y}, \boldsymbol{\beta}, \boldsymbol{\gamma} \sim \mathrm{IG}\left(a_0 + \frac{n+p}{2},\, b_0 + \frac{\|\tilde{\mathbf{y}} - \tilde{\mathbf{X}}\boldsymbol{\beta}\|_2^2 + \boldsymbol{\beta}^\top \mathbf{D}_{\boldsymbol{\gamma}}^{-1} \boldsymbol{\beta}}{2}\right).$$

3. **Update the latent indicators:**

$$\gamma_j \mid \mathbf{y}, \sigma^2, \boldsymbol{\beta} \overset{\text{ind}}{\sim} \mathrm{Ber}(p_j).$$

## Algorithms for block-wise updates

$\mathbf{G}_1, \mathbf{G}_2, \mathbf{G}_3$: update $([\mathbf{G}_\ell]_{\mathcal{A},\mathcal{A}})^{-1}$ via Schur complement, where $\delta_t$ is the number of changed indicators.

**Algorithm 1: Exact inactive mean update** $\mathcal{O}(p|\mathcal{A}|)$

**Input:** current active set $\mathcal{A}$ and active coefficients $\boldsymbol{\beta}_{\mathcal{A}}$.
**Precompute:** $\mathbf{H} = \mathbf{X}\mathbf{X}^\top + \tau_0^{-2}\mathbf{I}_n, \quad \mathbf{M} = \mathbf{X}^\top \mathbf{H}^{-1} \mathbf{X}, \quad \mathbf{v} = \mathbf{X}^\top \mathbf{H}^{-1} \mathbf{y}, \quad \mathbf{G}_1 = \mathbf{I}_p - \mathbf{M}.$
**Output:** $\boldsymbol{\mu}_{\mathcal{I}} = (\mathbf{v}_{\mathcal{I}} - \mathbf{M}_{\mathcal{I},\mathcal{A}} \boldsymbol{\beta}_{\mathcal{A}}) + \mathbf{M}_{\mathcal{I},\mathcal{A}} ([\mathbf{G}_1]_{\mathcal{A},\mathcal{A}})^{-1} (\mathbf{v}_{\mathcal{A}} - \mathbf{M}_{\mathcal{A},\mathcal{A}} \boldsymbol{\beta}_{\mathcal{A}})$.

**Algorithm 2: Sketched inactive covariance sampling** $\mathcal{O}(kp)$

**Input:** exact inactive mean $\boldsymbol{\mu}_{\mathcal{I}}$ and current active set $\mathcal{A}$.
**Precompute:** $\tilde{\mathbf{X}} = \mathbf{S}\mathbf{X}, \quad \mathbf{M}_0 = \tau_0^{-2}\mathbf{I}_k + \tilde{\mathbf{X}}\tilde{\mathbf{X}}^\top, \quad \mathbf{G}_2 = \mathbf{I}_p - \tilde{\mathbf{X}}^\top \mathbf{M}_0^{-1} \tilde{\mathbf{X}}.$
**Noise:** sample $\mathbf{u} \sim N(\mathbf{0}, \mathbf{I}_k)$, $\boldsymbol{\xi} \sim N(\mathbf{0}, \mathbf{I}_{|\mathcal{I}|})$, and set $\mathbf{z}_{\mathcal{I}} = \tilde{\mathbf{X}}_{\mathcal{I}}^\top \mathbf{u} + \tau_0^{-1} \boldsymbol{\xi}$.
**Compute:** $\tilde{\boldsymbol{\eta}}_{\mathcal{I}} = \tau_0^2 \mathbf{z}_{\mathcal{I}} - \tau_0^2 \tilde{\mathbf{X}}_{\mathcal{I}}^\top \mathbf{M}_0^{-1} (\tilde{\mathbf{X}}_{\mathcal{I}} \mathbf{z}_{\mathcal{I}}) - \tau_0^2 \tilde{\mathbf{X}}_{\mathcal{I}}^\top \mathbf{M}_0^{-1} \tilde{\mathbf{X}}_{\mathcal{A}} ([\mathbf{G}_2]_{\mathcal{A},\mathcal{A}})^{-1} \tilde{\mathbf{X}}_{\mathcal{A}}^\top \mathbf{M}_0^{-1} (\tilde{\mathbf{X}}_{\mathcal{I}} \mathbf{z}_{\mathcal{I}})$.
**Output:** $\boldsymbol{\beta}_{\mathcal{I}} = \boldsymbol{\mu}_{\mathcal{I}} + \sigma \tilde{\boldsymbol{\eta}}_{\mathcal{I}}, \quad \mathrm{Var}(\tilde{\boldsymbol{\eta}}_{\mathcal{I}}) = \boldsymbol{\Sigma}_{\mathcal{I}}$.

**Algorithm 3: Exact active block update** $\mathcal{O}(p|\mathcal{A}|)$

**Input:** newly sampled inactive coefficients $\boldsymbol{\beta}_{\mathcal{I}}$ and current active set $\mathcal{A}$.
**Precompute:** $\mathbf{G}_3 = \mathbf{X}^\top \mathbf{X} + \tau_1^{-2} \mathbf{I}_p$.
**Noise:** sample $\boldsymbol{\zeta}_{\mathcal{A}} \sim N(\mathbf{0}, [\mathbf{G}_3]_{\mathcal{A},\mathcal{A}})$.
**Output:** $\boldsymbol{\beta}_{\mathcal{A}} = ([\mathbf{G}_3]_{\mathcal{A},\mathcal{A}})^{-1} (\mathbf{r}_{\mathcal{A}}^* + \sigma \boldsymbol{\zeta}_{\mathcal{A}}), \quad \mathbf{r}_{\mathcal{A}}^* = (\mathbf{X}^\top \mathbf{y})_{\mathcal{A}} - (\mathbf{X}^\top \mathbf{X})_{\mathcal{A},\mathcal{I}} \boldsymbol{\beta}_{\mathcal{I}}$.

- **Overall per-iteration cost:** $\mathcal{O}(kp + p|\mathcal{A}| + |\mathcal{A}|^2 \delta_t + \delta_t^3)$ using Schur-complement updates for $\delta_t$ changed variables. Typically dominated by $\mathcal{O}(kp)$ when $|\mathcal{A}| \ll k \ll p$.

## Theorem 1. Inactive covariance approximation error bound

Let $\boldsymbol{\Sigma}_{\mathcal{I}}^{-1} = \mathbf{X}_{\mathcal{I}}^\top \mathbf{X}_{\mathcal{I}} + \tau_0^{-2} \mathbf{I}_{|\mathcal{I}|}$ and $\tilde{\boldsymbol{\Sigma}}_{\mathcal{I}}^{-1} = \tilde{\mathbf{X}}_{\mathcal{I}}^\top \tilde{\mathbf{X}}_{\mathcal{I}} + \tau_0^{-2} \mathbf{I}_{|\mathcal{I}|}$. Treating $\mathcal{I}$ as fixed independently of $\mathbf{S}$,

$$\mathbb{E}_{\mathbf{S}} \|\tilde{\boldsymbol{\Sigma}}_{\mathcal{I}} - \boldsymbol{\Sigma}_{\mathcal{I}}\|_F^2 \le \frac{\tau_0^8}{k} \left[\{\mathrm{tr}(\mathbf{X}_{\mathcal{I}}^\top \mathbf{X}_{\mathcal{I}})\}^2 + \|\mathbf{X}_{\mathcal{I}}^\top \mathbf{X}_{\mathcal{I}}\|_F^2\right].$$

If $\lambda_{\max}(n^{-1} \mathbf{X}_{\mathcal{I}}^\top \mathbf{X}_{\mathcal{I}}) = O(1)$ and the shrinking-spike condition of Narisetty and He (2014), $n\tau_0^2 = o(1)$, holds, then

$$\mathbb{E}_{\mathbf{S}} \|\tilde{\boldsymbol{\Sigma}}_{\mathcal{I}} - \boldsymbol{\Sigma}_{\mathcal{I}}\|_F^2 = O\left(\frac{\tau_0^8 n^2 r_{\mathcal{I}}^2}{k}\right) = o(1), \qquad r_{\mathcal{I}} = \mathrm{rank}(\mathbf{X}_{\mathcal{I}}).$$

## Theorem 2. Sketched residual quadratic form bound

For a fixed $\boldsymbol{\beta}$ and residual $\mathbf{r} = \mathbf{y} - \mathbf{X}\boldsymbol{\beta}$ independent of $\mathbf{S}$, for any $0 < \epsilon < 1$,

$$\mathbb{P}\left(\left|\frac{\|\mathbf{S}\mathbf{r}\|_2^2}{\|\mathbf{r}\|_2^2} - 1\right| \le \epsilon\right) \ge 1 - 2\exp\left(-\frac{k\epsilon^2}{8}\right).$$

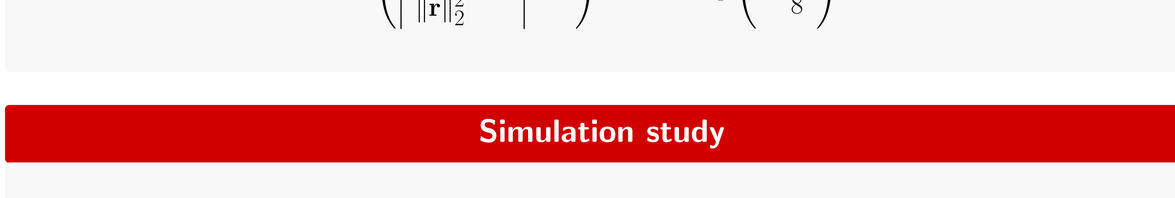
## Simulation study

- Data: $n = 10,000$, $p \in \{1000, 2000\}$, AR(1) covariance $\Sigma_{ij} = 0.5^{|i-j|}$, and SNR $= 1$. True signals have $s_0 = 20$ nonzero coefficients with $\beta_j \sim \mathrm{U}(0.1, 1)$; $\tau_0^2 = n^{-1.1}$, $\tau_1^2 = p^{2.2}/n$; 10 reps.
- Methods: exact Gibbs, full sketching (Guhaniyogi and Scheffler 2025) with $S^3$ (Biswas, Mackey, and Meng 2022), and proposed sampler. PIP denotes posterior inclusion probability.

| $p$ | $k$ | Method | Time (s) | Speedup | PIP RMSE | Jaccard |
|---|---|---|---|---|---|---|
| 1000 | – | Exact | 7678.9 (10.3) (2.1 h) | – | – | – |
| | 100 | Full sketch | 2.3 (0.0) | 3339× | 0.127 (0.001) | 0.06 (0.01) |
| | 100 | Proposed | 4.9 (0.1) | 1567× | 0.009 (0.002) | 0.96 (0.01) |
| | 150 | Full sketch | 3.3 (0.2) | 2327× | 0.126 (0.001) | 0.07 (0.01) |
| | 150 | Proposed | 5.8 (0.2) | 1324× | 0.012 (0.003) | 0.97 (0.01) |
| 2000 | – | Exact | 60978.8 (284.4) (16.9 h) | – | – | – |
| | 100 | Full sketch | 4.2 (0.2) | 14519× | 0.087 (0.001) | 0.06 (0.00) |
| | 100 | Proposed | 20.9 (0.9) | 2918× | 0.011 (0.004) | 0.95 (0.02) |
| | 200 | Full sketch | 8.4 (0.5) | 7259× | 0.086 (0.001) | 0.09 (0.01) |
| | 200 | Proposed | 23.8 (1.2) | 2562× | 0.010 (0.003) | 0.95 (0.02) |

## Real data application: MovieLens 1M

- Movie ratings data: sampled $n = 290{,}580$ ratings and encoded $p = 7{,}911$ user/movie features.
- Used sketch size $k = 2{,}000$ with $\tau_0^2 = n^{-1.1}$ and $\tau_1^2 = p^{2.2}/n$.




**Runtime.** Exact Gibbs: 99,580.4 sec (≈27.7 h); proposed: 2,142.6 sec (≈35.7 min), 46.4× faster.

## References


- Bhattacharya, Anirban, Antik Chakraborty, and Bani K. Mallick (2016). "Fast sampling with Gaussian scale mixture priors in high-dimensional regression". In: *Biometrika* 103.4, pp. 985–991. DOI: 10.1093/biomet/asw042.
- Biswas, Niloy, Lester Mackey, and Xiao-Li Meng (2022). "Scalable Spike-and-Slab". In: *Proceedings of the 39th International Conference on Machine Learning*. Vol. 162. Proceedings of Machine Learning Research. Baltimore, MD, USA: PMLR, pp. 2021–2040.
- George, Edward I. and Robert E. McCulloch (1993). "Variable Selection Via Gibbs Sampling". In: *Journal of the American Statistical Association* 88.423, pp. 881–889.
- Guhaniyogi, Rajarshi and Aaron Wolfe Scheffler (2025). "Sketching in High-Dimensional Regression With Big Data Using Gaussian Scale Mixture Priors". In: *Journal of Machine Learning Research* 26.271, pp. 1–28.
- Narisetty, Naveen Naidu and Xuming He (2014). "Bayesian variable selection with shrinking and diffusing priors". In: *The Annals of Statistics* 42.2, pp. 789–817. DOI: 10.1214/14-AOS1207.