KNU BAYESIAN LAB

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

Joint work with Gyuhyeong Goh (KNU)

Gyeongmin Park

Department of Statistics, Kyungpook National University,
Daegu, Republic of Korea

BK21 Workshop, June 23, 2026

Introduction

Background: High-Dimensional Bayesian Regression

Linear Regression Model

- Consider the linear regression model:

$$\mathbf{y} = \mathbf{X}\beta + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_n)$$

  where $\mathbf{y} \in \mathbb{R}^n$, $\mathbf{X} \in \mathbb{R}^{n \times p}$, and $\beta \in \mathbb{R}^p$.
- We focus on the massive and high-dimensional setting where both n and p are large.
- In such settings, it is generally assumed that only a small fraction of the covariates actively affects the response variable.

The Gaussian Spike-and-Slab Prior (1/2)

Prior Formulation (George and McCulloch, 1993)

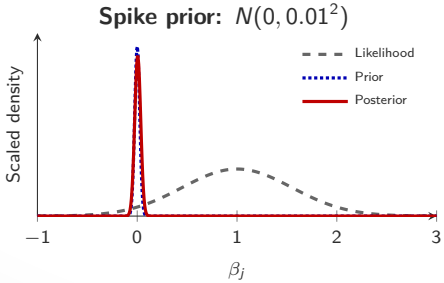
- To induce structural sparsity, we use the Gaussian spike-and-slab model with latent indicators $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p)^\top \in \{0,1\}^p$:

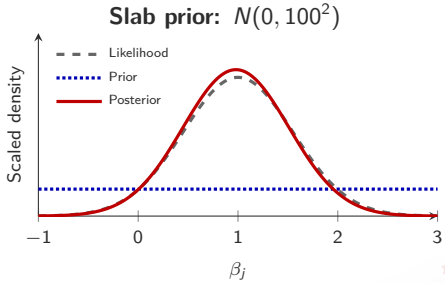
$$
\begin{aligned}
\mathbf{y} \mid \beta, \sigma^2 &\sim N_n(\mathbf{X}\beta, \sigma^2 \mathbf{I}_n) \\
\beta_j \mid \sigma^2, \gamma_j &\overset{\text{ind}}{\sim} N(0, \sigma^2 d_{\gamma_j}), \quad j = 1, \ldots, p \\
\sigma^2 &\sim \text{IG}(a_0, b_0) \\
\gamma_j &\overset{\text{iid}}{\sim} \text{Ber}(w), \quad j = 1, \ldots, p
\end{aligned}
$$

- $d_{\gamma_j} = (1-\gamma_j)\tau_0^2 + \gamma_j \tau_1^2$, where τ_0^2 and τ_1^2 represent the variance of the spike and the slab, respectively; typically τ_0^2 is very small and τ_1^2 is very large (e.g., $\tau_0^2 = 0.0001$ and $\tau_1^2 = 1000$).
- Let $\mathbf{D}_\gamma = \text{diag}(d_{\gamma_1}, \ldots, d_{\gamma_p})$, $d_0 = \tau_0^2$, and $d_1 = \tau_1^2$.

The Gaussian Spike-and-Slab Prior (2/2)

- If $\gamma_j = 0$ (spike), the concentrated prior around zero dominates the likelihood and strongly shrinks the posterior for β_j toward zero.
- If $\gamma_j = 1$ (slab), the prior is nearly flat/non-informative over plausible values, so the likelihood largely drives β_j without strong prior restriction.





- Averaging posterior draws of γ_j gives the PIP for each β_j.

Exact Gibbs Sampler

Let $\boldsymbol{\Sigma}_\gamma^{-1} = \mathbf{X}^\top \mathbf{X} + \mathbf{D}_\gamma^{-1}$ be the unscaled conditional precision matrix. The exact Gibbs sampler repeats the following three steps.

1. Update the regression coefficients:

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

2. Update the variance:

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

3. Update the latent indicators:

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

where $p_j = \frac{w\phi(\beta_j; 0, \sigma^2 d_1)}{w\phi(\beta_j; 0, \sigma^2 d_1) + (1-w)\phi(\beta_j; 0, \sigma^2 d_0)}$.

Computational Bottleneck

Let $\boldsymbol{\Sigma}_\gamma^{-1} = \mathbf{X}^\top \mathbf{X} + \mathbf{D}_\gamma^{-1}$ be the unscaled conditional precision matrix. The exact Gibbs sampler repeats the following three steps.

1. Update the regression coefficients:

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

Why this is expensive

- Computing $\boldsymbol{\Sigma}_\gamma$ requires inverting the $p \times p$ precision matrix $\boldsymbol{\Sigma}_\gamma^{-1}$.
- This repeated inversion costs $\mathcal{O}(p^3)$ or $\mathcal{O}(n^3)$ operations.
- Sampling from this multivariate normal distribution also requires a Cholesky factorization, adding another $\mathcal{O}(p^3)$ cost.
- Scalable Gibbs samplers reduce this cost using matrix identities, active-set updates, or approximate posterior targets.

Existing Method (1/2): Efficient Normal Sampling Strategy

Target:

$$\beta \mid \mathbf{y}, \sigma^2, \gamma \sim N(\boldsymbol{\Sigma}_\gamma \mathbf{X}^\top \mathbf{y}, \sigma^2 \boldsymbol{\Sigma}_\gamma), \qquad \boldsymbol{\Sigma}_\gamma^{-1} = \mathbf{X}^\top \mathbf{X} + \mathbf{D}_\gamma^{-1}.$$

- The efficient normal sampling strategy of Bhattacharya et al. (2016) uses the Woodbury identity:

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

- Sampling:

$$\mathbf{u} \sim N(\mathbf{0}, \sigma^2 \mathbf{D}_\gamma), \quad \boldsymbol{\delta} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_n), \quad (\mathbf{I}_n + \mathbf{X}\mathbf{D}_\gamma \mathbf{X}^\top)\mathbf{w} = \mathbf{y} - (\mathbf{X}\mathbf{u} + \boldsymbol{\delta}),$$

$$\beta = \mathbf{u} + \mathbf{D}_\gamma \mathbf{X}^\top \mathbf{w}.$$

- Cost: $\mathcal{O}(n^2 p)$ to form $\mathbf{X}\mathbf{D}_\gamma \mathbf{X}^\top$.

Existing Method (2/2): Scalable Spike-and-Slab

- Scalable Spike-and-Slab (S^3) (Biswas et al., 2022) updates the active model algebra incrementally as variables enter or leave the model.
- For example, when new active variables are added,

$$\mathbf{C}_{\text{new}} = \begin{bmatrix} \mathbf{C} & \mathbf{V} \\ \mathbf{V}^\top & \mathbf{D} \end{bmatrix},$$

and the inverse is updated by the Schur complement:

$$\mathbf{C}_{\text{new}}^{-1} = \begin{bmatrix} \mathbf{C}^{-1} + \mathbf{L}\mathbf{W}^{-1}\mathbf{L}^\top & -\mathbf{L}\mathbf{W}^{-1} \\ -\mathbf{W}^{-1}\mathbf{L}^\top & \mathbf{W}^{-1} \end{bmatrix},$$

where $\mathbf{L} = \mathbf{C}^{-1}\mathbf{V}$ and $\mathbf{W} = \mathbf{D} - \mathbf{V}^\top\mathbf{C}^{-1}\mathbf{V}$.

- Cost: under sparse active models, the main reported cost is $\mathcal{O}(np)$ per sweep, with additional small active-set inverse update costs.

These methods work well in small-n, large-p settings ($n \ll p$),
but may be difficult in large-n, large-p regimes ($n \gtrsim p$).

The Sketching-based Blocked Gibbs Sampler

Gaussian Sketching for Data Compression

- As discussed above, many efficient methods exploit a small-n, large-p structure ($n \ll p$).
- For massive data, we aim to recover this computational advantage by replacing the row dimension n with a much smaller sketch size k.
- A Gaussian sketching matrix $\mathbf{S} \in \mathbb{R}^{k \times n}$ compresses the original data from n rows to k rows (Mahoney, 2011; Ahfock et al., 2021):

$$\tilde{\mathbf{X}} = \mathbf{S}\mathbf{X}, \qquad \tilde{\mathbf{y}} = \mathbf{S}\mathbf{y}, \qquad k \ll n.$$

- In this work, $S_{ij} \overset{\text{iid}}{\sim} N(0, 1/k)$, so the sketch preserves the original scale in expectation:

$$\mathbb{E}_{\mathbf{S}}(\mathbf{S}^\top \mathbf{S}) = \mathbf{I}_n.$$

- Gaussian sketching has already been widely used in statistics, especially for randomized regression and least-squares approximation (Pilanci and Wainwright, 2015; Ahfock et al., 2021):

$$\hat{\beta}_{\text{sk}} = \arg\min_{\beta} \|\mathbf{S}(\mathbf{y} - \mathbf{X}\beta)\|_2^2.$$

Sketched Posterior in Bayesian Regression

Journal of Machine Learning Research 26 (2025) 1-28 Submitted 8/22; Revised 7/25; Published 8/25

Sketching in High-Dimensional Regression With Big Data Using Gaussian Scale Mixture Priors

Rajarshi Guhaniyogi RAJGUHANIYOGI@TAMU.EDU
Department of Statistics
Texas A & M University
College Station, TX 77843-3143, USA

Aaron Wolfe Scheffler AARON.SCHEFFLER@UCSF.EDU
Department of Epidemiology & Biostatistics
University of California
San Francisco, CA 94158, USA

Editor: Ji Zhu

Abstract

Bayesian computation of high-dimensional linear regression models with popular Gaussian scale mixture prior distributions using Markov Chain Monte Carlo (MCMC) or its variants can be extremely slow or completely prohibitive due to the heavy computational cost that grows in the order of p^3, with p as the number of features. Although a few recently developed algorithms allow computational efficiency in presence of a small to moderately large sample size, the computational issues are considerably less explored when sample size n is also large, except for a few recent articles. In this article we propose a sketching approach to compress the n original samples by a random linear transformation to $m << n$ samples in p dimensions, and compute Bayesian regression with Gaussian scale mixture prior distributions with the randomly compressed response vector and feature matrix. Our proposed approach yields computational complexity growing in the cubic order of m. Our detailed empirical investigation with the Horseshoe prior from the class of Gaussian scale mixture priors shows closely similar inference and a considerable reduction in per iteration computation time of the proposed approach compared to the regression with the full sample. One notable contribution of this article is to derive posterior contraction rate for high-dimensional feature coefficient with a general class of shrinkage priors on the coefficients under data compression/sketching. In particular, we characterize the dimension of the compressed response vector m as a function of the sample size, number of features and sparsity in the regression to guarantee accurate estimation of feature coefficients asymptotically, even after data sketching.

Keywords: Bayesian inference, Data sketching, Gaussian scale mixture priors, High-dimensional linear regression, Posterior convergence.

1. Introduction

Of late, due to the technological advances in a variety of disciplines, we routinely encounter data with a large number of features. In such settings, it is commonly of interest to consider the high-dimensional linear regression model

$$y = \boldsymbol{x}'\boldsymbol{\beta} + \epsilon, \quad (1)$$

- Guhaniyogi and Scheffler (2025) compress the full data,

$$(\mathbf{y}, \mathbf{X}) \to (\mathbf{Sy}, \mathbf{SX}),$$

  and then fit a Bayesian regression model on the sketched data.
- The paper studies Gaussian scale-mixture priors and gives posterior contraction results under data compression.
- With a standard small-n sampler on the sketched data, the cost is roughly $\mathcal{O}(k^2 p)$. (Using S^3 (Biswas et al., 2022), $\mathcal{O}(kp)$.)
- However, density-level agreement with the exact Gibbs posterior is not addressed in the paper; in our simulations, the sketched posterior did not closely track the exact one.

Our goal: use sketching while keeping posterior inference close to the exact Gibbs sampler.

Exact Gibbs Sampler

Let $\boldsymbol{\Sigma}_\gamma^{-1} = \mathbf{X}^\top \mathbf{X} + \mathbf{D}_\gamma^{-1}$ be the unscaled conditional precision matrix. The exact Gibbs sampler repeats the following three steps.

1. Update the regression coefficients:

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

2. Update the variance:

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

3. Update the latent indicators:

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

where $p_j = \frac{w\phi(\beta_j; 0, \sigma^2 d_1)}{w\phi(\beta_j; 0, \sigma^2 d_1) + (1-w)\phi(\beta_j; 0, \sigma^2 d_0)}$.

Blocked Gibbs Sampler

Let

$$\mathcal{A} = \{j : \gamma_j = 1\}, \qquad \mathcal{I} = \{j : \gamma_j = 0\}, \qquad \mathbf{X} = [\mathbf{X}_{\mathcal{A}}\ \mathbf{X}_{\mathcal{I}}], \quad \beta = (\beta_{\mathcal{A}}^\top, \beta_{\mathcal{I}}^\top)^\top.$$

1. Update the regression coefficients:

$$\beta_{\mathcal{I}} \mid \mathbf{y}, \sigma^2, \beta_{\mathcal{A}}, \gamma \sim N(\boldsymbol{\Sigma}_{\mathcal{I}} \mathbf{X}_{\mathcal{I}}^\top (\mathbf{y} - \mathbf{X}_{\mathcal{A}} \beta_{\mathcal{A}}), \sigma^2 \boldsymbol{\Sigma}_{\mathcal{I}}), \quad \boldsymbol{\Sigma}_{\mathcal{I}}^{-1} = \mathbf{X}_{\mathcal{I}}^\top \mathbf{X}_{\mathcal{I}} + \tau_0^{-2} \mathbf{I},$$
$$\beta_{\mathcal{A}} \mid \mathbf{y}, \sigma^2, \beta_{\mathcal{I}}, \gamma \sim N(\boldsymbol{\Sigma}_{\mathcal{A}} \mathbf{X}_{\mathcal{A}}^\top (\mathbf{y} - \mathbf{X}_{\mathcal{I}} \beta_{\mathcal{I}}), \sigma^2 \boldsymbol{\Sigma}_{\mathcal{A}}), \quad \boldsymbol{\Sigma}_{\mathcal{A}}^{-1} = \mathbf{X}_{\mathcal{A}}^\top \mathbf{X}_{\mathcal{A}} + \tau_1^{-2} \mathbf{I}.$$

2. Update the variance:

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

3. Update the latent indicators:

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

Sketching in the Blocked Sampler (1/2)

Let

$$\tilde{\mathbf{X}} = \mathbf{S}\mathbf{X}, \qquad \tilde{\mathbf{y}} = \mathbf{S}\mathbf{y}, \qquad \text{where} S_{ij} \overset{\text{iid}}{\sim} N(0, 1/k), \quad k \ll n.$$

1. Update the regression coefficients:

$$\beta_{\mathcal{I}} \mid \mathbf{y}, \sigma^2, \beta_{\mathcal{A}}, \gamma \sim N(\boldsymbol{\Sigma}_{\mathcal{I}} \mathbf{X}_{\mathcal{I}}^\top (\mathbf{y} - \mathbf{X}_{\mathcal{A}} \beta_{\mathcal{A}}), \sigma^2 \tilde{\boldsymbol{\Sigma}}_{\mathcal{I}}), \quad \tilde{\boldsymbol{\Sigma}}_{\mathcal{I}}^{-1} = \tilde{\mathbf{X}}_{\mathcal{I}}^\top \tilde{\mathbf{X}}_{\mathcal{I}} + \tau_0^{-2} \mathbf{I},$$
$$\beta_{\mathcal{A}} \mid \mathbf{y}, \sigma^2, \beta_{\mathcal{I}}, \gamma \sim N(\boldsymbol{\Sigma}_{\mathcal{A}} \mathbf{X}_{\mathcal{A}}^\top (\mathbf{y} - \mathbf{X}_{\mathcal{I}} \beta_{\mathcal{I}}), \sigma^2 \boldsymbol{\Sigma}_{\mathcal{A}}), \quad \boldsymbol{\Sigma}_{\mathcal{A}}^{-1} = \mathbf{X}_{\mathcal{A}}^\top \mathbf{X}_{\mathcal{A}} + \tau_1^{-2} \mathbf{I}.$$

2. Update the variance:

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

3. Update the latent indicators:

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

Sketching in the Blocked Sampler (2/2)

Let

$$\tilde{\mathbf{X}} = \mathbf{S}\mathbf{X}, \qquad \tilde{\mathbf{y}} = \mathbf{S}\mathbf{y}, \qquad \text{where} S_{ij} \overset{\text{iid}}{\sim} N(0, 1/k), \quad k \ll n.$$

1. Update the regression coefficients:

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

$$\beta_{\mathcal{A}} \mid \mathbf{y}, \sigma^2, \beta_{\mathcal{I}}, \gamma \sim \underbrace{N(\boldsymbol{\Sigma}_{\mathcal{A}}\mathbf{X}_{\mathcal{A}}^\top(\mathbf{y} - \mathbf{X}_{\mathcal{I}}\beta_{\mathcal{I}}), \sigma^2\boldsymbol{\Sigma}_{\mathcal{A}})}_{\text{via Algorithm 3}}, \quad \boldsymbol{\Sigma}_{\mathcal{A}}^{-1} = \mathbf{X}_{\mathcal{A}}^\top\mathbf{X}_{\mathcal{A}} + \tau_1^{-2}\mathbf{I}.$$

2. Update the variance:

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

3. Update the latent indicators:

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

Algorithms for Block-wise Updates (1/3)

Let

$$\tilde{\mathbf{X}} = \mathbf{S}\mathbf{X}, \qquad \tilde{\mathbf{y}} = \mathbf{S}\mathbf{y}, \qquad \text{where} S_{ij} \overset{\text{iid}}{\sim} N(0, 1/k), \quad k \ll n.$$

1. Update the regression coefficients:

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

$$\beta_{\mathcal{A}} \mid \mathbf{y}, \sigma^2, \beta_{\mathcal{I}}, \gamma \sim \underbrace{N(\boldsymbol{\Sigma}_{\mathcal{A}}\mathbf{X}_{\mathcal{A}}^\top(\mathbf{y} - \mathbf{X}_{\mathcal{I}}\beta_{\mathcal{I}}), \sigma^2\boldsymbol{\Sigma}_{\mathcal{A}})}_{\text{via Algorithm 3}}, \quad \boldsymbol{\Sigma}_{\mathcal{A}}^{-1} = \mathbf{X}_{\mathcal{A}}^\top\mathbf{X}_{\mathcal{A}} + \tau_1^{-2}\mathbf{I}.$$

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

Input: current active set $\mathcal{A}$ and active coefficients $\beta_{\mathcal{A}}$.
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: $\mu_{\mathcal{I}} = (\mathbf{v}_{\mathcal{I}} - \mathbf{M}_{\mathcal{I},\mathcal{A}}\beta_{\mathcal{A}}) + \mathbf{M}_{\mathcal{I},\mathcal{A}}([\mathbf{G}_1]_{\mathcal{A},\mathcal{A}})^{-1}(\mathbf{v}_{\mathcal{A}} - \mathbf{M}_{\mathcal{A},\mathcal{A}}\beta_{\mathcal{A}})$.

Algorithms for Block-wise Updates (2/3)

Let

$$\tilde{\mathbf{X}} = \mathbf{S}\mathbf{X}, \qquad \tilde{\mathbf{y}} = \mathbf{S}\mathbf{y}, \qquad \text{where} S_{ij} \overset{\text{iid}}{\sim} N(0, 1/k), \quad k \ll n.$$

1. Update the regression coefficients:

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

$$\beta_{\mathcal{A}} \mid \mathbf{y}, \sigma^2, \beta_{\mathcal{I}}, \gamma \sim \underbrace{N(\boldsymbol{\Sigma}_{\mathcal{A}}\mathbf{X}_{\mathcal{A}}^\top(\mathbf{y} - \mathbf{X}_{\mathcal{I}}\beta_{\mathcal{I}}), \sigma^2\boldsymbol{\Sigma}_{\mathcal{A}})}_{\text{via Algorithm 3}}, \quad \boldsymbol{\Sigma}_{\mathcal{A}}^{-1} = \mathbf{X}_{\mathcal{A}}^\top\mathbf{X}_{\mathcal{A}} + \tau_1^{-2}\mathbf{I}.$$

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

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

Algorithms for Block-wise Updates (3/3)

Let

$$\tilde{\mathbf{X}} = \mathbf{S}\mathbf{X}, \qquad \tilde{\mathbf{y}} = \mathbf{S}\mathbf{y}, \qquad \text{where} S_{ij} \overset{\text{iid}}{\sim} N(0, 1/k), \quad k \ll n.$$

1. Update the regression coefficients:

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

$$\beta_{\mathcal{A}} \mid \mathbf{y}, \sigma^2, \beta_{\mathcal{I}}, \gamma \sim \underbrace{N(\boldsymbol{\Sigma}_{\mathcal{A}}\mathbf{X}_{\mathcal{A}}^\top(\mathbf{y} - \mathbf{X}_{\mathcal{I}}\beta_{\mathcal{I}}), \sigma^2\boldsymbol{\Sigma}_{\mathcal{A}})}_{\text{via Algorithm 3}}, \quad \boldsymbol{\Sigma}_{\mathcal{A}}^{-1} = \mathbf{X}_{\mathcal{A}}^\top\mathbf{X}_{\mathcal{A}} + \tau_1^{-2}\mathbf{I}.$$

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

Input: newly sampled inactive coefficients $\beta_{\mathcal{I}}$ and current active set $\mathcal{A}$.
Precompute: $\mathbf{G}_3 = \mathbf{X}^\top\mathbf{X} + \tau_1^{-2}\mathbf{I}_p$, $\mathbf{X}^\top\mathbf{X} = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^\top$, $\mathbf{E} = \mathbf{V}\boldsymbol{\Lambda}^{1/2}$.
Output: for $\mathbf{z}_1 \sim N(\mathbf{0}, \mathbf{I}_p)$ and $\mathbf{z}_2 \sim N(\mathbf{0}, \mathbf{I}_{|\mathcal{A}|})$, $\beta_{\mathcal{A}} = ([\mathbf{G}_3]_{\mathcal{A},\mathcal{A}})^{-1}\{\mathbf{r}_{\mathcal{A}}^* + \sigma(\mathbf{E}_{\mathcal{A}}\mathbf{z}_1 + \tau_1^{-1}\mathbf{z}_2)\}$, where $\mathbf{r}_{\mathcal{A}}^* = (\mathbf{X}^\top\mathbf{y})_{\mathcal{A}} - (\mathbf{X}^\top\mathbf{X})_{\mathcal{A},\mathcal{I}}\beta_{\mathcal{I}}$.

The inverses $([\mathbf{G}_1]_{\mathcal{A},\mathcal{A}})^{-1}$, $([\mathbf{G}_2]_{\mathcal{A},\mathcal{A}})^{-1}$, and $([\mathbf{G}_3]_{\mathcal{A},\mathcal{A}})^{-1}$ are updated incrementally as $\mathcal{A}$ changes.

Incremental Inverse Updates

Maintains $([\mathbf{G}_1]_{\mathcal{A},\mathcal{A}})^{-1}$, $([\mathbf{G}_2]_{\mathcal{A},\mathcal{A}})^{-1}$, and $([\mathbf{G}_3]_{\mathcal{A},\mathcal{A}})^{-1}$.

These are standard Schur-complement block inverse updates.

- **Variables are added**
  Add a block to $\mathbf{C}$:

$$\mathbf{C}_+ = \begin{bmatrix} \mathbf{C} & \mathbf{V} \\ \mathbf{V}^\top & \mathbf{D} \end{bmatrix}.$$

  With $\mathbf{L} = \mathbf{C}^{-1}\mathbf{V}$ and $\mathbf{W} = \mathbf{D} - \mathbf{V}^\top\mathbf{C}^{-1}\mathbf{V}$,

$$\mathbf{C}_+^{-1} = \begin{bmatrix} \mathbf{C}^{-1} + \mathbf{L}\mathbf{W}^{-1}\mathbf{L}^\top & -\mathbf{L}\mathbf{W}^{-1} \\ -\mathbf{W}^{-1}\mathbf{L}^\top & \mathbf{W}^{-1} \end{bmatrix}.$$

Cost: $\mathcal{O}(|\mathcal{A}|^2\delta_t + \delta_t^3)$ for δ_t changed variables.

- **Variables are removed**
  After reordering, write the current inverse as

$$\mathbf{C}_+^{-1} = \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^\top & \mathbf{E} \end{bmatrix}.$$

  Then the remaining inverse is

$$\mathbf{C}^{-1} = \mathbf{A} - \mathbf{B}\mathbf{E}^{-1}\mathbf{B}^\top.$$

Variance and Indicator Updates

Updates

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

- Sketched residual quadratic form:

$$\|\tilde{\mathbf{y}} - \tilde{\mathbf{X}}\beta\|_2^2$$

- Cost: $\mathcal{O}(kp)$ instead of $\mathcal{O}(np)$ from the data or $\mathcal{O}(p^2)$ using cached Gram matrices.

Total arithmetic cost: $\mathcal{O}(kp + |\mathcal{A}|p + |\mathcal{A}|^2\delta_t + \delta_t^3)$ per iteration.

When $|\mathcal{A}| < k \ll \min(n, p)$, the leading term is $\mathcal{O}(kp)$.

Theoretical Results

Inactive Covariance Approximation

Theorem. Inactive covariance bound

Let

$$\mathbf{\Sigma}_{\mathcal{I}}^{-1} = \mathbf{X}_{\mathcal{I}}^{\top}\mathbf{X}_{\mathcal{I}} + \tau_0^{-2}\mathbf{I}, \qquad \tilde{\mathbf{\Sigma}}_{\mathcal{I}}^{-1} = \tilde{\mathbf{X}}_{\mathcal{I}}^{\top}\tilde{\mathbf{X}}_{\mathcal{I}} + \tau_0^{-2}\mathbf{I}.$$

For fixed $\mathbf{X}_{\mathcal{I}}$ and a Gaussian sketch $\mathbf{S}$, for any $\delta \in (0,1)$, with probability at least $1-\delta$,

$$\|\tilde{\mathbf{\Sigma}}_{\mathcal{I}} - \mathbf{\Sigma}_{\mathcal{I}}\|_F \le \frac{\tau_0^4}{\sqrt{k\delta}} \left[\{\mathrm{tr}(\mathbf{X}_{\mathcal{I}}^{\top}\mathbf{X}_{\mathcal{I}})\}^2 + \|\mathbf{X}_{\mathcal{I}}^{\top}\mathbf{X}_{\mathcal{I}}\|_F^2\right]^{1/2}.$$

Asymptotic implication. Suppose $\lambda_{\max}(n^{-1}\mathbf{X}_{\mathcal{I}}^{\top}\mathbf{X}_{\mathcal{I}}) = O(1)$ (the largest eigenvalue remains bounded), and let $r_{\mathcal{I}} = \mathrm{rank}(\mathbf{X}_{\mathcal{I}})$. If $n\tau_0^2 = o(1)$, as in the shrinking-spike scaling of Narisetty and He (2014), then for fixed δ,

$$\|\tilde{\mathbf{\Sigma}}_{\mathcal{I}} - \mathbf{\Sigma}_{\mathcal{I}}\|_F = O\left(\frac{\tau_0^4 n r_{\mathcal{I}}}{\sqrt{k}}\right) = O\left((n\tau_0^2)^2 \frac{r_{\mathcal{I}}}{n\sqrt{k}}\right) = o(1)$$

Sketched Residual Approximation Error Bound

Theorem. Sketched residual approximation error bound

For a fixed β independent of the sketch, let $\mathbf{r} = \mathbf{y} - \mathbf{X}\beta$. Then, for any $\epsilon \in (0, 1)$,

$$P\left(\left|\frac{\|\mathbf{Sr}\|_2^2 - \|\mathbf{r}\|_2^2}{\|\mathbf{r}\|_2^2}\right| \leq \epsilon\right) \geq 1 - 2\exp\left(-\frac{k\epsilon^2}{8}\right).$$

This follows from the Gaussian sketching property

$$\mathbf{Sr} \mid \mathbf{r} \sim N\left(\mathbf{0}, \frac{\|\mathbf{r}\|_2^2}{k}\mathbf{I}_k\right),$$

and hence

$$\frac{\|\mathbf{Sr}\|_2^2}{\|\mathbf{r}\|_2^2} \overset{d}{=} \frac{C}{k}, \qquad C \sim \chi_k^2.$$

The last step uses the Laurent–Massart chi-square inequality (Laurent and Massart, 2000).

Simulation Studies

Simulation Setup

Data Generation:

- $n = 10{,}000$, $p \in \{1000, 2000\}$, with AR(1) design $\Sigma_{ij} = 0.5^{|i-j|}$.
- $s_0 = 20$ active coefficients are generated from $U(0.1, 1)$, with SNR $= 1$.
- Each setting is repeated 10 times.

Sketching Dimensions:

- For $p = 1000$, use $k \in \{100, 150\}$; for $p = 2000$, use $k \in \{100, 200\}$.
- We use $s_0 \log p$ as a reference scale from Guhaniyogi and Scheffler (2025): $20 \log(1000) \approx 138$ and $20 \log(2000) \approx 152$.

Hyperparameters:

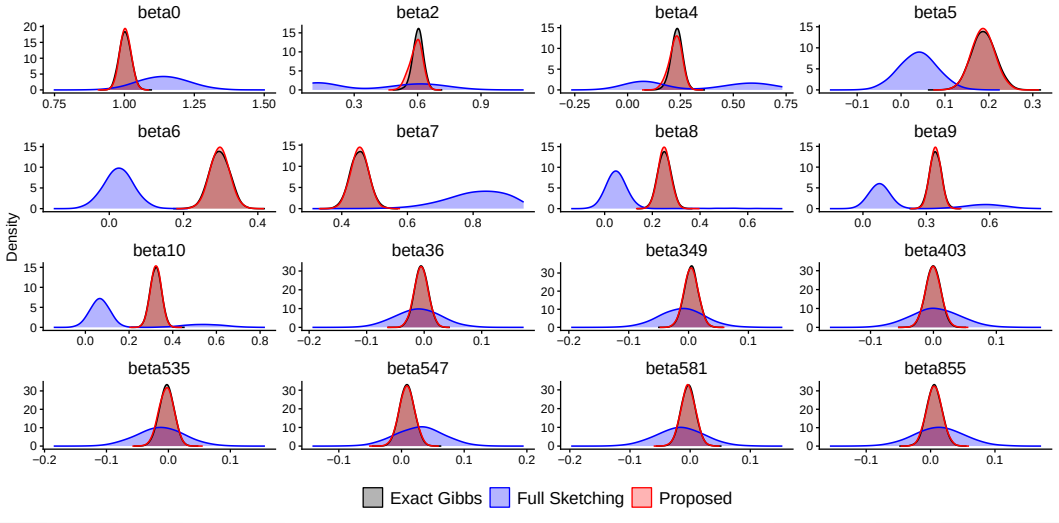
- Spike variance: $\tau_0^2 = n^{-1.1}$, satisfying $n\tau_0^2 = o(1)$ (Narisetty and He, 2014)
- Slab variance: $\tau_1^2 = p^{2.2}/n$

Variable Selection and Execution Time

- Compare exact Gibbs, full sketching (Guhaniyogi and Scheffler, 2025) with S^3 (Biswas et al., 2022), and the proposed block-wise sketched Gibbs sampler.

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)

Posterior Densities



Application to Real-World Data

MovieLens 1M Dataset Analysis

Data Setup:

- Collaborative filtering dataset with user ratings for movies
- Sampled $n = 290,580$ ratings, encoded into $p = 7,911$ features (users and movies)
- Massive and high-dimensional setting where both n and p are extremely large

Experimental Settings:

- Sketching dimension $k = 2000$ (based on $1.1 \times d_{eff}$ heuristic)
- MCMC chains run for 10,000 iterations

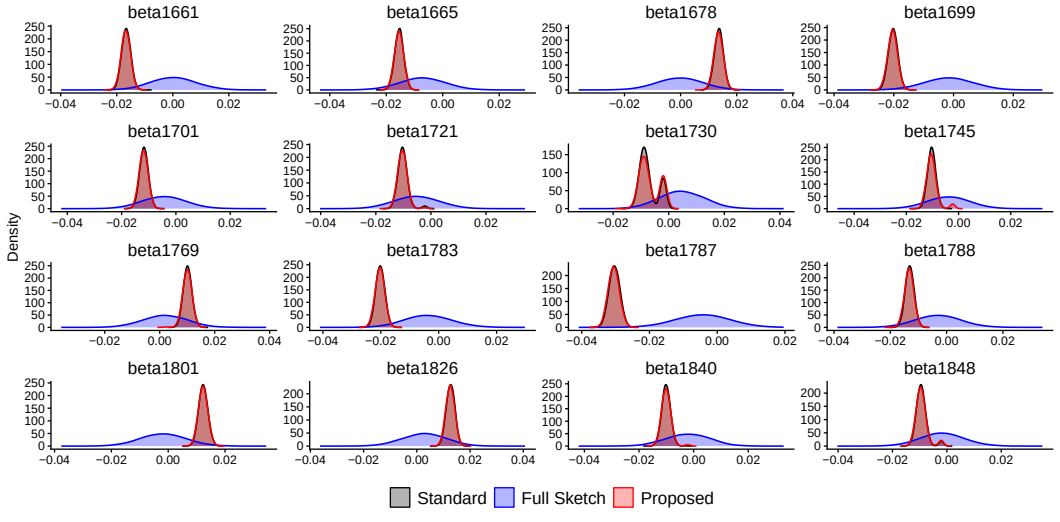
Execution Time & Scalability

Method	Time (sec)	Speedup
Exact Gibbs	99,580.4 ($\approx$ 27.7h)	1.0$\times$
Proposed	**2,142.6 ($\approx$ 35.7m)**	**46.4$\times$**

Key Findings:

- The massive $p = 7,911$ matrix inversion bottleneck is successfully bypassed
- Reduces computation time from over a day to just half an hour
- Demonstrates practical scalability for massive real-world data

Posterior Densities for MovieLens Data



Conclusion

Concluding Remarks

Summary of Contributions:

- **Algorithm:** Proposed a sketching-based blocked Gibbs sampler separating active/inactive variables
- **Scalability:** Reduced the per-iteration algebraic complexity from $O(p^3)$ to $O(kp)$ under sparse active sets
- **Presentation:** This work is scheduled for a poster presentation at the 2026 ISBA World Meeting in Nagoya, Japan

Future Directions:

- (if possible) C++ implementation (e.g., `RcppEigen`) to eliminate memory copy overhead in R

Key References I

- Ahfock, D. C., W. J. Astle, and S. Richardson (2021). "Statistical properties of sketching algorithms". In: *Biometrika* 108.2, pp. 283–297. DOI: 10.1093/biomet/asaa062.
- 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.

Key References II

- Laurent, Béatrice and Pascal Massart (2000). "Adaptive Estimation of a Quadratic Functional by Model Selection". In: *The Annals of Statistics* 28.5, pp. 1302–1338. DOI: 10.1214/aos/1015957395.
- Mahoney, Michael W. (2011). "Randomized Algorithms for Matrices and Data". In: *Foundations and Trends in Machine Learning* 3.2, pp. 123–224. DOI: 10.1561/2200000035.
- 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.
- Pilanci, Mert and Martin J. Wainwright (2015). "Randomized Sketches of Convex Programs With Sharp Guarantees". In: *IEEE Transactions on Information Theory* 61.9, pp. 5096–5115. DOI: 10.1109/TIT.2015.2450722.

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

Joint work with Gyuhyeong Goh (KNU)

Gyeongmin Park

Department of Statistics, Kyungpook National University,
Daegu, Republic of Korea

BK21 Workshop, June 23, 2026