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

Rajarshi Guhaniyogi

Associate Professor, Department of Statistics,

Texas A & M University, College Station, TX 77843-3143, E-mail: rajguhaniyogi@tamu.edu

Aaron Wolfe Scheffler

Assistant Professor, Department of Epidemiology & Biostatistics,

UC San Francisco, 550 16th. Street San Francisco CA 94158, E-mail: Aaron.Scheffler@ucsf.edu

April 20, 2022

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 predictors. Although a few recently developed algorithms make the computation efficient in presence of a small to moderately large sample size (with the complexity growing in the order of n^3), the computation becomes intractable when sample size n is also large. In this article we adopt the data 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 predictor matrix. Our proposed approach yields computational complexity growing in the cubic order of m. Another important motivation for this compression procedure is that it anonymizes the data by revealing little information about the original data in the course of analysis. Our detailed empirical investigation with the Horseshoe prior from the class of Gaussian scale mixture priors shows closely similar inference and a massive 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 predictor coefficient with a general class of shrinkage priors on them 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 predictors and sparsity in the regression to guarantee accurate estimation of predictor coefficients asymptotically, even after data compression. Supplementary material contains proofs of the theoretical results.

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

1 Introduction

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

$$y = \boldsymbol{x}'\boldsymbol{\beta} + \epsilon, \tag{1}$$

where $\boldsymbol{x}$ is a $p \times 1$ predictor vector, $\boldsymbol{\beta}$ is the corresponding $p \times 1$ coefficient, y is the continuous response and ϵ is the idiosyncratic error following i.i.d. $\mathrm{N}(0, \sigma^2)$. Bayesian methods for estimating $\boldsymbol{\beta}$ broadly employ two classes of prior distributions. The traditional approach is to develop a discrete mixture of prior distributions (George and McCulloch, 1997; Scott and Berger, 2010). These methods enjoy the advantage of inducing exact sparsity for a subset of parameters (allowing some components of $\boldsymbol{\beta}$ to be exactly zero a posteriori) and minimax rate of posterior contraction (Castillo *et al.*, 2015) in high dimensional regression, but face computational challenges when the number of predictors is even moderately large. As an alternative to this approach, continuous shrinkage priors (Bhadra *et al.*, 2019; Armagan *et al.*, 2013; Carvalho *et al.*, 2010; Caron and Doucet, 2008) have emerged, which can mostly be expressed as global-local scale mixtures of Gaussians (Polson and Scott, 2010) given by,

$$\begin{gathered}\beta_j|\lambda_j,\tau,\sigma \sim N(0,\sigma^2\tau^2\lambda_j^2),\ \lambda_j \sim g_1,\ \text{for } j=1,...,p\\ \tau \sim g_2,\ \sigma \sim f, \end{gathered} \tag{2}$$

where τ is known as the global shrinkage parameter and λ_j's are known as the local shrinkage parameters, g_1, g_2 and f are densities supported on $\mathbb{R}^+$. The prior structure (2) induces approximate sparsity in β_j by shrinking the null components toward zero while retaining the true signals (Polson and Scott, 2010). The global parameter τ controls the number of signals, while the local parameters λ_j dictate whether they are nulls. In this sense, the prior (2) approximates the properties of point-mass mixture priors (George and McCulloch, 1997; Scott and Berger, 2010).

Global-local priors allow parameters to be updated in blocks via a fairly automatic Gibbs sampler that leads to rapid mixing and convergence of the resulting Markov chain. In particular, letting $\boldsymbol{X}$ be the $n \times p$ predictor matrix, $\boldsymbol{y}$ be the $n \times 1$ response vector and $\boldsymbol{\Delta} = \tau^2 diag(\lambda_1, ..., \lambda_p)$, the distribution of $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)'$ conditional on $\boldsymbol{\lambda} = (\lambda_1, ..., \lambda_p)', \tau$, σ, $\boldsymbol{y}$ and $\boldsymbol{X}$ follows $N((\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{\Delta}^{-1})^{-1}\boldsymbol{X}'\boldsymbol{y}, \sigma^2(\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{\Delta}^{-1})^{-1})$, and can be updated in a block. On the other hand, λ_j's are conditionally independent and allow fairly straightforward updating using either Gibbs sampling or slice sampling. The posterior draws from $\boldsymbol{\beta}, \boldsymbol{\lambda}, \tau, \sigma$ are found to offer an accurate approximation to the operating characteristics of discrete mixture priors. However, sampling from the full conditional posterior of $\boldsymbol{\beta}$ require storing and computing the Cholesky decomposition of the $p \times p$ matrix $(\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{\Delta}^{-1})$, that necessitates

p^3 floating point operations (flops) and p^2 storage units, which can be severely prohibitive for large p. Recent work in high dimensional regressions involving small n and large p (Bhattacharya *et al.*, 2016) exploits the Woodbury matrix identity to draw from the full conditional posterior distribution of $\boldsymbol{\beta}$ by inverting only an $n \times n$ matrix. When n is large, this algorithm is embedded within an approximate MCMC sampling framework of Johndrow *et al.* (2020) to facilitate fast computation.

Following the literature on *data sketching*, we propose to compress the response vector and predictor matrix by a random linear transformation, reducing the number of records from n to m, while preserving the number and interpretation of original predictors. The compressed version of the original dataset, referred to as a *sketch*, then serves as a surrogate for a high dimensional regression analysis with a suitable Gaussian scale mixture prior on the predictor coefficients. Since the number of compressed records m is much smaller than the sample size n, it is possible to adapt existing algorithms on the compressed data for efficient estimation of posterior distribution for predictor coefficients with large number of predictors and large sample. On the theoretical front, we assume that the shrinkage priors of our interest have densities with a dominating peak around 0 and flat, heavy tails, and have sufficient mass around the true regression coefficients. We then identify conditions on the predictor matrix, the interlink between the dimensions of the random compression matrix, sample size, sparsity of the true regression coefficient vector and the number of predictors to prove optimal convergence rate of estimating the predictor coefficients asymptotically under data compression. Our empirical investigation ensures that the relevant predictors can be accurately learnt from the compressed data. Moreover, in presence of a higher degree of sparsity in the true regression model, the actual estimates of regression coefficients and predictions turn out to be very close to the corresponding quantities, when the uncompressed data are used. Another attractive predictor of this approach is that the original data are not recoverable from the compressed data, and the compressed data effectively reveal no more information than would be revealed by a completely new sample. In fact, the original uncompressed data does not need to be stored in the course of the analysis. While the core idea behind the development apply broadly to the class of global-local priors (2), for sake of concreteness our detailed empirical investigation focuses on the popular horseshoe prior (Carvalho *et al.*, 2010) which corresponds to both g_1 and g_2 in (2) being the half-Cauchy distribution. The horseshoe achieves the minimax adaptive rate of contraction when the true $\boldsymbol{\beta}$ is sparse (Van Der Pas *et al.*, 2014; Van der Pas *et al.*, 2017) and is considered to be among the state-of-the-art shrinkage priors.

In this context, it is worth mentioning the contribution of this article in light of the relevant literature of *data sketching*, where the computational task is relaxed by generating a compressed version of the original dataset which then serves as a surrogate for calculations. Sketching has become an increasingly popular research topic in the machine learning literature in the last decade or so, see Vempala (2005); Sarlos (2006); Halko *et al.* (2011); Mahoney (2011); Woodruff (2014) and references therein. Sketching has been extensively studied in the context of ridge regression, referred to as the sketched ridge regression, to identify the theoretical conditions to ensure accurate estimation of ridge regression coefficients from sketched data (Zhang *et al.*, 2013; Chen *et al.*, 2015; Wang *et al.*, 2017). Similarly, Zhou *et al.* (2008) showed that identifying the correct sparse set of relevant variables by the lasso are as effective under data sketching. Dobriban and Liu (2018) studied sketching using asymptotic random matrix theory, but only for un-regularized linear regression. Chowdhury *et al.* (2018) proposed a data-dependent algorithm in the context of estimating ridge leverage scores. Other related works include Ailon and Chazelle (2006); Drineas *et al.* (2011); Raskutti and Mahoney (2016); Ahfock *et al.* (2017); Huang (2018). To the best of our knowledge, we are the first to offer efficient and principled Bayesian computation algorithm in high dimensional linear regressions involving large n and p using data sketching. Moreover, to the best of our knowledge, the theoretical result on the posterior convergence rate of regression parameters under data sketching has not been established before.

While bearing some similarities, our current contribution differs from compressed sensing (Donoho, 2006; Candes and Tao, 2006; Eldar and Kutyniok, 2012; Yuan, 2016) in the inferential objectives. Specifically, compressed sensing solves an inverse problem by "nearly" recovering a sparse vector of responses from a smaller set of random linear transformations. In contrast, our response vector $\boldsymbol{y}$ and predictor matrix $\boldsymbol{X}$ are not necessarily sparse. Also, we do not seek to (approximately) recover $\boldsymbol{y}$ and $\boldsymbol{X}$ from their compressed counterparts, so our method is applicable to situations where preserving confidentiality of the response (and predictors) is important. Our approach is fundamentally different from Maillard and Munos (2009); Guhaniyogi and Dunson (2015, 2016) in that they compress each predictor vector, leading to an m-dimensional compressed predictors from p-dimensional predictors for each sample. In contrast, our compression framework does not alter the number of predictors in the analysis before and after compression.

The rest of the article proceeds as follows. Section 2 details out the proposed model and algorithm for efficient estimation of predictor coefficients in presence of large n and p. Section 3 offers theoretical insights into the choice of m as a function of the true sparsity, number of predictors and sample size n to obtain accurate estimation of predictor coefficients asymptotically. Section 4 empirically investigates parametric and predictive inferences from the proposed approach with the horseshoe shrinkage prior under various simulation cases. The proposed method is illustrated with the orthopedic fractures data in Section 5, followed by the concluding remarks in Section 6. The supplementary material contains proofs of the theoretical results.

2 Sketching Response Vector and Predictor Matrix for Large n

For subjects $i = 1, ..., n$, let $y_i \in \mathcal{Y}$ denote the response for subject i corresponding to the predictor $\boldsymbol{x}_i \in \mathbb{R}^p$. This article focuses on the scenario where n and p both large. Let $\boldsymbol{y} = (y_1, ..., y_n)'$ be the $n \times 1$ vector of responses and $\boldsymbol{X} = [\boldsymbol{x}_1 : \cdots : \boldsymbol{x}_p]'$ be the $n \times p$ matrix of predictors. As a first step to our proposal, we consider a sketching or data compression approach by pre-multiplying $\boldsymbol{y}$ and $\boldsymbol{X}$ with a sketching matrix $\boldsymbol{\Phi}$ of dimensions $m \times n$ with $m << n$ to construct data sketches $\tilde{\boldsymbol{y}} = \boldsymbol{\Phi y}$ and $\tilde{\boldsymbol{X}} = \boldsymbol{\Phi X}$ of dimensions $m \times 1$ and $m \times p$, respectively. The sketched/compressed response vector is related to the compressed predictor matrix according to the high dimensional linear regression model

$$\tilde{\boldsymbol{y}} = \tilde{\boldsymbol{X}}\boldsymbol{\beta} + \boldsymbol{\epsilon},\ \boldsymbol{\epsilon} \sim N(0, \sigma^2 \boldsymbol{I}), \tag{3}$$

where σ^2 is the idiosyncratic error variance. We do not estimate $\boldsymbol{\Phi}$ as a variable in the regression, rather follow the idea of data oblivious sketches to construct $\boldsymbol{\Phi}$ prior to fitting the model (3). More specifically, following the idea of Gaussian sketching (Sarlos, 2006), the elements Φ_{ij} of the $\boldsymbol{\Phi}$ matrix are drawn independently from $\mathrm{N}(0, 1/n)$. The computational complexity of obtaining the sketched data using Gaussian sketches is given by $O(mnp)$. While there are more computationally efficient data oblivious options for random projection/sketching matrix $\boldsymbol{\Phi}$, such as the Hadamard sketch (Ailon and Chazelle, 2009) and the Clarkson-Woodruff sketch (Clarkson and Woodruff, 2017), we find it to be less concerning in our framework since the computation time for fitting (3) using a Bayesian architecture far exceeds the difference in time for computing sketched data with different options of sketching matrices.

The data compression approach implemented here is a special case of the *matrix masking* technique proposed in the earlier literature on data privacy (Ting *et al.*, 2008; Zhou *et al.*, 2008; Zhao and Chen, 2019), which, although popular in machine learning, has not been given any attention in high dimensional regression, especially from a Bayesian perspective. A typical matrix masking procedure pre- and post-multiplies the data matrix $\boldsymbol{X}$ by matrices $\boldsymbol{C}$ and $\boldsymbol{D}$, respectively, and releases $\boldsymbol{CXD}$ for the ensuing analysis. The transformation is quite general, and allows the possibility of deleting records, suppressing subsets of variables and data swapping. Our proposal in this article corresponds to $\boldsymbol{C} = \boldsymbol{\Phi}$ and $\boldsymbol{D}$ as the identity matrix so as to keep the original interpretation of the predictors. Notably, even in the case of $\boldsymbol{\Phi}$ being known, the linear system $\boldsymbol{\Phi X}$ is grossly under-determined to recover $\boldsymbol{X}$ due to $m << min(n, p)$. Moreover, an upper bound of the average mutual information $\mathcal{I}(\tilde{\boldsymbol{X}}, \boldsymbol{X})/np$ per unit in the original data matrix $\boldsymbol{X}$ satisfies $Sup\, \mathcal{I}(\tilde{\boldsymbol{X}}, \boldsymbol{X})/np = O(m/n)$ (Zhou *et al.*, 2008), where supremum is taken over all possible distributions of $\boldsymbol{X}$. With m growing at a much slower rate than n, asymptotically as $n \to \infty$, the supremum over average mutual information converges to 0, intuitively meaning that the compressed data reveal no more information about the original data than could be obtained from an independent sample. It is be noted that such a bound is obtained assuming that $\boldsymbol{\Phi}$ is known. In practice, only $\tilde{\boldsymbol{X}} = \boldsymbol{\Phi X}$ (and not even $\boldsymbol{\Phi}$) is revealed to the analyst. Hence, the imposed masking of data through compression is more strict than what is revealed by this result.

Although not apparent, the ordinary high dimensional regression model in (1) bears a close connection with its computationally convenient alternative (3), especially for large n. To elaborate on it, note that pre-multiplying the high dimensional linear regression equation $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$ by $\boldsymbol{\Phi}$ results in

$$\boldsymbol{\Phi y} = \boldsymbol{\Phi X \beta} + \tilde{\boldsymbol{\epsilon}}, \;\; \tilde{\boldsymbol{\epsilon}} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{\Phi\Phi}'). \tag{4}$$

Equations (4) and (3) are similar in the mean function but differ in the error distribution. More specifically, our approach assumes components of the error vector $\boldsymbol{\epsilon}$ are i.i.d., whereas the error vector from (4) follows a $N(\boldsymbol{0}, \sigma^2\boldsymbol{\Phi\Phi}')$ distribution. Invoking Lemma 5.36 and Remark 5.40 of Vershynin (2010) guarantee that (A) $||\boldsymbol{\Phi\Phi}' - \boldsymbol{I}_m||_2 \to 0$, with probability converging to 1; and (B) all eigenvalues of $\boldsymbol{\Phi\Phi}'$ converges to 1, as $m \to \infty$, $m/n \to 0$, when $\boldsymbol{\Phi}$ is constructed using the Gaussian sketching strategy. Hence, with large n, the error distributions of (3) and (4) behave similarly with a probability close to 1. It is important to note that the $\boldsymbol{\Phi}$-transformed model (4) does not provide the computational advantage we offer using our framework, as discussed later. Besides, computing the $\boldsymbol{\Phi}$-transformed model requires the supplying the analyst with $\boldsymbol{\Phi}$ which hurts the purpose of data masking.

With prior distribution on $\boldsymbol{\beta}$ set as a Gaussian scale-mixture distribution from the class of distributions given by (2), posterior computation using a blocked Metropolis-within-Gibbs algorithm cycles through updating the full conditional distributions: (a) $\boldsymbol{\beta}|\boldsymbol{\lambda}, \sigma, \tau$, (b) $\boldsymbol{\lambda}|\boldsymbol{\beta}, \sigma, \tau$, (c) $\sigma|\boldsymbol{\lambda}, \boldsymbol{\beta}, \tau$ and (d) $\tau|\boldsymbol{\lambda}, \boldsymbol{\beta}, \sigma$. Explicit expressions for (a), (b), (c) and (d) for the horseshoe shrinkage priors can be found in (Carvalho *et al.*, 2010). While updating (b), (c) and (d) do not face any computational challenge due to big n or p, full conditional posterior updating of $\boldsymbol{\beta}|\boldsymbol{\lambda}, \sigma, \tau$ has the form given by

$$N\left(\left(\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{X}} + \boldsymbol{\Delta}^{-1}\right)^{-1}\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{y}}, \sigma^2(\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{X}} + \boldsymbol{\Delta}^{-1})^{-1}\right), \;\; \boldsymbol{\Delta} = \tau^2\text{diag}(\lambda_1, ..., \lambda_p). \tag{5}$$

The most efficient algorithm to sample from $\boldsymbol{\beta}$ (Rue, 2001) computes Cholesky decomposition of $\left(\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{X}} + \boldsymbol{\Delta}^{-1}\right)$ and employs the Cholesky factor to solve a series of linear systems to draw a sample from (5). In absence of any easily exploitable structure, computing and storing the Cholesky factor of this matrix involves $O(p^3)$ and $O(p^2)$ floating point operations, respectively (Golub and Van Loan, 2012), which leads to computational and storage bottlenecks with a large p. To overcome the computational and storage burden, we adapt the recent algorithm proposed in the context of uncompressed data with small sample size (Bhattacharya *et al.*, 2016) to our setting. The detailed steps are given as follows:

Step 1: Draw $\boldsymbol{v}_1 \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{\Delta})$ and $\boldsymbol{v}_2 \sim N(\boldsymbol{0}, \boldsymbol{I}_m)$.

Step 2: Set $\boldsymbol{v}_3 = \tilde{\boldsymbol{X}}\boldsymbol{v}_1/\sigma + \boldsymbol{v}_2$.

Step 3: Solve $(\tilde{\boldsymbol{X}}\boldsymbol{\Delta}\tilde{\boldsymbol{X}}' + \boldsymbol{I}_m)\boldsymbol{v}_4 = (\tilde{\boldsymbol{y}}/\sigma - \boldsymbol{v}_3)$.

Step 4: Set $\boldsymbol{v}_5 = \boldsymbol{v}_1 + \sigma\boldsymbol{\Delta}\tilde{\boldsymbol{X}}'\boldsymbol{v}_4$.

$\boldsymbol{v}_5$ is a draw from the full conditional posterior distribution of $\boldsymbol{\beta}$. Notably, the computational complexity of Steps 1-4 is dominated by two operations: (Operation A) computing the inverse of $(\boldsymbol{\Phi}\boldsymbol{X}\boldsymbol{\Delta}\boldsymbol{X}'\boldsymbol{\Phi}' + \boldsymbol{I}_m)$, and (Operation B) calculating $\boldsymbol{\Phi}\boldsymbol{X}\boldsymbol{\Delta}\boldsymbol{X}'\boldsymbol{\Phi}'$. (Operation A) leads to a complexity of $O(m^3)$, whereas (Operation B) incurs complexity of $O(m^2p)$. As we demonstrate in Section 4, the algorithm offers massive speed-up in computation with big p and n, since $m << min(n, p)$. Notably, an application of Bhattacharya *et al.* (2016) on the uncompressed data would have incurred computational complexity dominated by $O(n^3)$ and $O(n^2p)$. Thus, our compression approach helps speeding up computation in our empirical investigations with big n and p.

One important question arises as to how much inference is lost in lieu of the computational speed-up achieved by the data compression approach. In the sequel, we address this question both theoretically and empirically. Section 3 derives theoretical conditions on m, n, p and the sparsity of the true data generating model to show asymptotically desirable estimation of predictor coefficients. Thereafter, finite sample performance of the proposed approach is presented both in the simulation study and in the real data section.

3 Posterior Concentration Properties of the Sketching Approach

This section studies convergence properties of the data sketching approach with high dimensional shrinkage prior on predictor coefficients. In particular, we will establish the posterior contraction rate of estimating the predictor coefficient vector for the proposed model (3) under mild regularity conditions. To begin with, we define a few notations.

3.1 Notations

In what follows, we add a subscript n to the dimension of the number of predictors p_n and the dimension of the compression matrix m_n to indicate that both of them increase with the sample size n. This asymptotic paradigm is also meant to capture the fact that the number of rows of the sketching matrix m_n is smaller than the sample size n. Naturally, the response vector $\boldsymbol{y}$, predictor matrix $\boldsymbol{X}$, predictor coefficient vector $\boldsymbol{\beta}$ and the sketching matrix $\boldsymbol{\Phi}$ are also functions of n. We denote them by $\boldsymbol{y}_n$, $\boldsymbol{X}_n$, $\boldsymbol{\beta}_n$ and $\boldsymbol{\Phi}_n$, respectively. Note that the true data generating model under data sketching is given by (4). We use superscript $*$ to indicate the true parameters $\boldsymbol{\beta}_n^*$ and σ^{*2}. For simplicity in the algebraic manipulation, we assume that $\sigma^2 = \sigma^{*2}$ are both known and fixed at 1. This is a common assumption in asymptotic studies (Vaart and Zanten, 2011). Furthermore, it is known that the theoretical results obtained by assuming σ^2 as a fixed value is equivalent to those obtained by assigning a prior with a bounded support on σ^2 (Van der Vaart and van Zanten, 2009). $P_{\boldsymbol{\beta}_n^*}$ denotes probability distribution under the true data generating model (4). For vectors, we let $||\cdot||_1, ||\cdot||_2$ and $||\cdot||_\infty$ denote the L_1, L_2 and L_∞ norms, respectively. The number of nonzero elements in a vector is given by $||\cdot||_0$. The quantities $e_{min}(\boldsymbol{A})$ and $e_{max}(\boldsymbol{A})$ respectively represent the minimum and maximum eigenvalues of a square matrix $\boldsymbol{A}$. We use $\{\theta_n\}$ to denote the Bayesian posterior contraction rate which satisfies $\theta_n \to 0$. Finally, for two sequences $\{a_n\}_{n\geq 1}$ and $\{b_n\}_{n\geq 1}$, $a_n = o(b_n)$ and $a_n = O(b_n)$ imply $a_n/b_n \to 0$ and $a_n/b_n \to C$ (C is a constant), respectively, as $n \to \infty$.

3.2 Assumptions, Framework and The Main Result

For any subset of indices $\boldsymbol{\xi} \subset \{1, ..., p_n\}$, $|\boldsymbol{\xi}|$ denotes the number of elements in the index set $\boldsymbol{\xi}$. Depending on whether $\boldsymbol{A}$ is a vector or a matrix, $\boldsymbol{A}_{\boldsymbol{\xi}}$ denotes the sub-vector or the sub-matrix corresponding to the indices $\boldsymbol{\xi}$. We let $\boldsymbol{\xi}^* = \{j : \beta_{j,n}^* \neq 0\}$, i.e., $\boldsymbol{\xi}^*$ are the indices of the nonzero entries for the true predictor coefficient $\boldsymbol{\beta}_n^*$, and s_n (dependent on n) designates the number of nonzero entries in $\boldsymbol{\beta}_n^*$, i.e., $s_n = ||\boldsymbol{\beta}_n^*||_0 = |\boldsymbol{\xi}^*|$. Since the shrinkage prior on $\boldsymbol{\beta}_n$ assigns zero probability at the point zero, the exact number of nonzero elements of $\boldsymbol{\beta}_n$ is always p_n. Before rigorously studying properties of the posterior distribution, we state some regularity conditions on the design matrix $\boldsymbol{X}_n$, the compression matrix $\boldsymbol{\Phi}_n$ and the true sparsity s_n.

(A) All covariates are uniformly bounded, let $|x_{i,j}| \leq 1$, for all $i = 1, ..., n$ and $j = 1, .., p_n$.

(B) $||\boldsymbol{\Phi}_n\boldsymbol{\Phi}_n' - \boldsymbol{I}_{m_n}||_2 \leq C'\sqrt{m_n/n}$, for some constant $C' > 0$, for all large n.

(C) $s_n \log(p_n) = o(m_n)$, $m_n = o(n)$.

(D) There exists $\tilde{s}_n > 0$ such that $e_{min}(\boldsymbol{X}'_{n,\boldsymbol{\xi}}\boldsymbol{X}_{n,\boldsymbol{\xi}}/n) \geq \eta$, for some $\eta > 0$ and for all $\boldsymbol{\xi} \supset \boldsymbol{\xi}^*$ and $|\boldsymbol{\xi}| \leq \tilde{s}_n + s_n$, where $\tilde{s}_n$ satisfies $\tilde{s}_n = O(s_n)$. Here $\boldsymbol{X}_{n,\boldsymbol{\xi}}$ is the sub-matrix of $\boldsymbol{X}_n$ with column indices $\boldsymbol{\xi}$.

(A) is a common assumption in the context of compressed sensing, see Zhou *et al.* (2008). From the theory of random matrices, (B) occurs with probability at least $1 - e^{-C''m_n}$ (see Lemma 5.36 and Remark 5.40 of Vershynin (2010)). Hence (B) is a mild assumption for large n. (C) restricts the growth of the true sparsity and presents an interlink between the true sparsity, the rank of the random matrix, number of predictor coefficients and the sample size. (D) puts restriction on the smallest eigenvalue of the matrix $\tilde{\boldsymbol{X}}'_{n,\boldsymbol{\xi}}\tilde{\boldsymbol{X}}_{n,\boldsymbol{\xi}}/m_n$. Notably, Gaussian sketching approximately preserves the isometry condition (Ahfock *et al.*, 2017), so that $\exists\ \eta_0 > 0$ with the property that $e_{min}(\tilde{\boldsymbol{X}}'_{n,\boldsymbol{\xi}}\tilde{\boldsymbol{X}}_{n,\boldsymbol{\xi}}/m_n) \geq \eta_0 e_{min}(\boldsymbol{X}'_{n,\boldsymbol{\xi}}\boldsymbol{X}_{n,\boldsymbol{\xi}}/n)$ with probability f_n depending on m_n and p_n . This fact, together with assumption $\text{A}_1(3)$ in Song and Liang (2017) ensure assumption (D) to hold with a positive probability. Our next set of assumptions concern the tail behavior of the shrinkage priors of interest and the magnitude of the nonzero entries of the true coefficient $\boldsymbol{\beta}_n^*$. Let $h_{\mu_n}(x)$ denote the prior density of $\beta_{j,n}$ for all j with the set of hyper-parameters μ_n. For $a_n = \sqrt{s_n \log(p_n)/m_n}/p_n$ and for a sequence M_n nondecreasing as a function of n, we assume

(E) $\max\limits_{j \in \boldsymbol{\xi}^*} |\beta^*_{j,n}| < M_n/2$.

(F) $1 - \int_{-a_n}^{a_n} h_{\mu_n}(x)dx \leq p_n^{-(1+u)}$, for some positive constant u.

(G) $-\log(\inf\limits_{x \in [-M_n, M_n]} h_{\mu_n}(x)) = O(\log(p_n))$.

Assumption (E) restricts the growth of the nonzero entries in the true regression parameter asymptotically. Assumption (F) concerns the prior concentration, requiring that the prior density of $\beta_{j,n}$ for all j has sufficient mass within the interval $[-a_n, a_n]$. Finally, Assumption (G) essentially controls the prior density around the true predictor coefficient. Notably, Assumptions (E)-(G) are frequently used in the high dimensional Bayesian regression literature, including in Jiang (2007) and Song and Liang (2017).

Define $\mathcal{A}_n = \{\boldsymbol{\beta}_n : ||\boldsymbol{\beta}_n - \boldsymbol{\beta}_n^*||_2 > 3\theta_n\}$, $\mathcal{B}_n = \{\text{At least } \tilde{s}_n \text{ number of } |\beta_{k,n}| \geq a_n\}$, with $\tilde{s}_n = O(s_n)$, $\mathcal{C}_n = \mathcal{A}_n \cup \mathcal{B}_n$. Since the shrinkage prior assigns zero probability at point zero, the number of nonzero elements of $\boldsymbol{\beta}_n$ is p_n. Thus, the number of nonzero components of $\boldsymbol{\beta}_n$ is assessed by considering the number of $\beta_{k,n}$'s which exceeds a certain threshold a_n. Therefore, $\mathcal{B}_n$ can be viewed as a set that indicates the number of nonzero predictor coefficients. Further suppose $\pi_n(\cdot)$ and $\Pi_n(\cdot)$ are the prior and posterior densities of $\boldsymbol{\beta}_n$ with n observations respectively, so that

$$\pi_n(\boldsymbol{\beta}_n) = \prod_{j=1}^{p_n} h_{\mu_n}(\beta_{j,n}), \;\; \Pi_n(\mathcal{C}_n) = \frac{\int_{\mathcal{C}_n} f(\tilde{\boldsymbol{y}}_n|\boldsymbol{\beta}_n)\pi_n(\boldsymbol{\beta}_n)}{\int f(\tilde{\boldsymbol{y}}_n|\boldsymbol{\beta}_n)\pi_n(\boldsymbol{\beta}_n)},$$

where $f(\tilde{\boldsymbol{y}}_n|\boldsymbol{\beta}_n)$ is the joint density of $\tilde{\boldsymbol{y}}_n = \boldsymbol{\Phi}_n\boldsymbol{y}_n$ under model (3). The following theorem shows posterior contraction for the proposed model, with the proof of the theorem given in the supplementary material.

Theorem 3.1 *Under Assumptions (A)-(G), our proposed model satisfies* $E_{\boldsymbol{\beta}_n^*}(\Pi_n(\mathcal{C}_n)) \to 0$, *as* $n, m_n \to \infty$ *and* $m_n/n \to 0$ *with the posterior contraction rate* $\theta_n = E\sqrt{s_n \log(p_n)/m_n}$, *for some constant* $E > 0$.

The general result on posterior contraction in Theorem 3.1 is applied to provide posterior contraction result for the proposed data sketching approach with a class of Gaussian scale mixture prior distributions on $\beta_{j,n}$. Indeed we assume that the prior density h_{μ_n} with hyper-parameter μ_n of each $\beta_{j,n}$ is symmetric around 0 and has a polynomial tail, i.e., $h_{\mu_n}(x) \sim x^{-r}$ when $|x|$ is large, for some $r > 1$. Notably prior densities for both the horseshoe shrinkage prior (Carvalho *et al.*, 2010) and the generalized double pareto shrinkage prior (Armagan *et al.*, 2013) have polynomial tails. Theorem 3.1 can be adapted in such a setting to arrive at the following result. The proof of the result can be found in the supplementary material.

Theorem 3.2 *Let the predictor matrix* $\boldsymbol{X}_n$, *random compression matrix* $\boldsymbol{\Phi}_n$ *and the true predictor coefficients* $\boldsymbol{\beta}_n^*$ *satisfy Assumptions (A)-(G). Let the prior density with hyper-parameter* μ_n, *given by* $h_{\mu_n}(x) = h(x/\mu_n)$, *has a polynomial tail, i.e.,* $h_{\mu_n}(x) \sim x^{-r}$ *when* $|x|$

is large, for some $r > 1$*. Further assume that* $\log(M_n) = O(\log p_n)$*,* $a_n = \sqrt{s_n \log(p_n)/m_n}/p_n$*,* $\mu_n \leq a_n p_n^{-(u'+1)/(r-1)}$ *and* $\log(\mu_n) = O(\log(p_n))$*, for some* $u' > 0$*. Then the posterior contraction rate* θ_n *can be taken as* $E\sqrt{s_n \log(p_n)/m_n}$*,for some constant* $E > 0$*.*

Note that the minimax optimal posterior contraction rate without data sketching is given by $\sqrt{s_n \log(p_n)/n}$ which is $\rho_n = \sqrt{n/m_n}$ times faster that the posterior contraction rate with data sketching. In fact, ρ_n throws light on the connection between the theoretical performance of (3) with the choice of m_n. In particular, choice of $m_n = O(n/\log(n))$ maintains minimax optimal posterior contraction rate upto a $\log(n)$ factor even with data sketching, when the true number of nonzero coefficients s_n is small not to violate Assumption (C). The next section empirically studies the performance of data sketching in high dimensional regressions with various other competitors. Special emphasis is given to investigate the discrepancy in the inference on $\boldsymbol{\beta}_n$ from the full data and the sketched data to carefully assess the impact of sketching.

4 Simulation Studies

This section investigates performance of the data sketching approach (3) with the horseshoe shrinkage prior (Carvalho *et al.*, 2010) on each of the predictor coefficients β_j, referred to as the Compressed Horseshoe (CHS). While the idea of data sketching sufficiently general which allows application to any shrinkage prior, we choose Horseshoe as a state-of-the-art representative shrinkage prior to illustrate our approach. Broadly, we implement and present two different sets of simulations. In **Simulation 1**, we focus on data simulated from (1) with $n = 1000$ and $p = 10000$, where both models (1) with uncompressed data and (3) with sketched data can be fitted to analyze the difference in their posterior distributions of $\boldsymbol{\beta}$ for different choices of m and different degrees of sparsity. These simulation examples also highlight the relative computational efficiency of (3) with respect to (1). **Simulation 2** is then designed with a larger sample size $n = 5000$ and $p = 10000$ which render infeasibility in fitting the model (1) with the uncompressed data based on our available computational resources. Thus the purpose for **Simulation 2** is to assess the frequentist operating characteristics of CHS along with relevant frequentist competitors.

4.1 Simulation 1: Comparison between the performances of CHS and HS for moderate n and large p

In **Simulation 1**, we draw $n = 1000$ samples from the high dimensional linear regression model (1) with the number of predictors $p = 10000$ and the error variance $\sigma^2 = 1.5$. The p-dimensional predictor vectors $\boldsymbol{x}_i$ for each $i = 1, ..., n$ are simulated from $N(\mathbf{0}, \mathbf{\Sigma})$, with two different constructions of $\mathbf{\Sigma}$ undertaken in simulation studies.

Scenario 1: $\mathbf{\Sigma} = \boldsymbol{I}_p$, i.e., all predictors are simulated i.i.d. We refer to this as the independent correlation structure for the predictors.

Scenario 2: $\mathbf{\Sigma} = 0.5\boldsymbol{I}_p + 0.5\boldsymbol{J}_p$, where $\boldsymbol{J}_p$ is a matrix with 1 at each entry. This structure ensures that any pair of predictors have the same correlation of 0.5. We refer to this as the compound correlation structure for the predictors.

Under Scenarios 1 and 2, the p-dimensional true predictor coefficient vector is simulated with the number of nonzero entries: (a) $s = 10$; (b) $s = 30$ and (c) $s = 50$. The quantity $(1 - s/p)$ is referred to as the true sparsity of the model. The magnitude of s nonzero entries are simulated randomly from a $U(1.5, 3)$ distribution with the sign of each entry randomly assigned to be positive or negative.

According to Theorem 3.2 and the discussion following it, the choice of $m = n/\log(n) \approx 150$ should be sufficient to offer satisfactory inference when true sparsity is high. To compare the effect of data sketching on the estimation of posterior distribution of $\boldsymbol{\beta}$, we implement (1) (with the uncompressed data) and (3) with different choices of $m = 100, 200, 300, 400, 500$. The full/uncompressed data posterior distribution obtained using MCMC serves as the benchmark in our assessment of the performance of (3). Let $\pi(\beta_j|\boldsymbol{y}, \boldsymbol{X})$ be the density of the full data posterior distribution for β_j estimated using sampling and $\pi_m(\beta_j|\tilde{\boldsymbol{y}}, \tilde{\boldsymbol{X}})$ be the density of posterior distribution for β_j estimated using (3) with the compressed data, where the subscript m denotes the dimension of the sketching matrix $\boldsymbol{\Phi}$ to compute $\tilde{\boldsymbol{y}}$ and $\tilde{\boldsymbol{X}}$. We used the following metric based on the Hellinger distance to compare the accuracy of $\pi_m(\beta_j|\tilde{\boldsymbol{y}}, \tilde{\boldsymbol{X}})$ in approximating $\pi(\beta_j|\boldsymbol{y}, \boldsymbol{X})$

$$Accuracy_{j,m} = 1 - \frac{1}{2}\int_{\boldsymbol{\beta}} \left(\sqrt{\pi_m(\beta_j|\tilde{\boldsymbol{y}}, \tilde{\boldsymbol{X}})} - \sqrt{\pi(\beta_j|\boldsymbol{y}, \boldsymbol{X})}\right)^2 d\beta_j. \tag{6}$$

The metric $\text{Accuracy}_{j,m}$ satisfies $0 \leq \text{Accuracy}_{j,m} \leq 1$. The approximation of full data posterior density $\pi(\beta_j|\boldsymbol{y}, \boldsymbol{X})$ by $\pi_m(\beta_j|\tilde{\boldsymbol{y}}, \tilde{\boldsymbol{X}})$ is poor or excellent if the accuracy metric is close to 0 or 1, respectively. We present $\text{Accuracy}_{j,m}$ averaged over all predictors, given by $\text{Accuracy}_m = \frac{1}{p}\sum_{j=1}^{p} \text{Accuracy}_{j,m}$.

Simulation 1 also highlights the computational efficiency offered by the data sketching approach. Let ESS_m be the average effective sample size of $\boldsymbol{\beta}$ (out of 5000 post burn-in iterates) from (3) with $\text{rank}(\boldsymbol{\Phi}) = m$, that runs for T_m hours. We will measure the computational efficiency of our proposed approach for a specific choice of m as

$$\text{Computational Efficiency}_m = log_2\text{ESS}_m/\text{T}_m, \tag{7}$$

where ESS_m over p predictor coefficients are computed using the `coda` package in `R`. Computational efficiency of the full posterior will also be reported to provide a relative assessment. All simulations are replicated 50 times.

4.1.1 Results

Table 1 presents the Accuracy metric averaged over all predictors and all replications. The results show excellent performance of $\pi_m(\boldsymbol{\beta}|\tilde{\boldsymbol{y}}, \tilde{\boldsymbol{X}})$ in approximating $\pi(\boldsymbol{\beta}|\boldsymbol{y}, \boldsymbol{X})$ for all cases except when both the sparsity and rank of the random compression matrix are both low. This empirical observation is also supported by Theorem 3.1 which requires the degree of sparsity to grow at a much slower rate than the rank of the random compression matrix. Understandably, as m increases the accuracy becomes close to 1, with the accuracy being little impacted when the sparsity is very low. No notable difference is observed in the performance when predictors are correlated vis-a-vis when predictors are simulated independently. Since the sample size is moderate, we do not expect to see a lots of gain in terms of computational efficiency of CHS over HS. CHS with $m = 100$ appears to be around ~ 10 times computationally more efficient than HS. The computational efficiency decreases as we increase the rank of the compression matrix. The computational efficiency seems to be not severely affected by the degree of sparsity or the correlation in the predictors.

		Scenario 1			Scenario 2		
		$s = 10$	$s = 30$	$s = 50$	$s = 10$	$s = 30$	$s = 50$
Avg. Accuracy	$m = 100$	0.88	0.79	0.64	0.88	0.81	0.66
	$m = 200$	0.94	0.87	0.76	0.92	0.84	0.73
	$m = 300$	0.98	0.96	0.93	0.99	0.96	0.94
	$m = 400$	0.98	0.98	0.94	0.98	0.98	0.94
	$m = 500$	0.98	0.98	0.95	0.99	0.98	0.95
Comp. Efficiency	$m = 100$	2.83	2.81	2.81	2.86	2.81	2.83
	$m = 200$	2.01	2.03	2.03	2.02	2.04	2.03
	$m = 300$	1.28	1.30	1.30	1.32	1.31	1.32
	$m = 400$	1.03	1.02	1.06	1.03	1.02	1.05
	$m = 500$	0.85	0.86	0.86	0.86	0.87	0.84
	HS	0.32	0.31	0.31	0.31	0.32	0.32

Table 1: The first five rows present metric to estimate accuracy of estimating full posterior of $\boldsymbol{\beta}$ by the posterior of $\boldsymbol{\beta}$ with compressed data, as described in (6). We present the metric averaged over all predictors and all replications. The metric is presented for different choices of $m = 100, 200, 300, 400, 500$ and different degrees of sparsity for the true coefficient $\boldsymbol{\beta}^*$. The upper bound of the accuracy measure is 1 and a higher value represents more accuracy. We also present computational efficiency of CHS, as described in (7), for different choices of m and for different degrees of sparsity under the two different simulation scenarios. Computational efficiency of the posterior distribution of $\boldsymbol{\beta}$ with the uncompressed data (referred to as the HS) has also been presented.

4.2 Simulation 2: Comparison between CHS and its frequentist competitors with larger sample size

Simulation 2 is designed to assess performance of the proposed framework for a large p, large n setting. We follow the identical data generation scheme as **Simulation 1** with a larger sample size $n = 5000$ to construct simulated data. The large values of p and n prohibit

Bayesian model fitting of (1) using the horseshoe prior using our available computational resources. Hence, we focus on investigating frequentist operating characteristics of CHS along with its frequentist competitors in high dimensional regression. As a frequentist competitor to CHS, we implement the minimax concave penalty (MCP) method (Zhang, 2010) on the full data. Additionally, we fit MCP on randomly chosen m data points from the sample of size n, and refer to this competitor as Partial MCP (PMCP). The MCP on full data provides a benchmark for comparison with a frequentist penalized optimizer in high dimensional regression with big n and p. While MCP with the full data is likely to perform better than CHS with the compressed data, the discrepancy in performance of CHS and MCP can be seen as an indicator of loss of inference due to data sketching. On the other hand, comparison of CHS with PMCP demonstrates the inferential advantage of fitting a principled Bayesian approach with sketching that uses information from the entire sample over fitting of a frequentist penalization scheme with naive sub-sampling of m out of n data points. Although the remaining section presents excellent performance of the sketching approach with the horseshoe prior on β_j's, we expect similar performance from other Gaussian scale mixture prior distributions, such as the Generalized Double Pareto (Armagan *et al.*, 2013) prior or the normal gamma prior (Griffin *et al.*, 2010).

According to Theorem 3.2, choice of $m = n/\log(n) \approx 500$ should lead to satisfactory estimation of the regression coefficients. However, to assess how the true sparsity $(1 - s/p)$ and the rank m of the random compression matrix interplay, we fit CHS with $m = 200$ and $m = 400$ in both simulation scenarios under the three different sparsity levels corresponding to (a), (b) and (c). For MCMC-based model implementation of CHS, we discard the first 5000 samples as burn-in and draw inference based on the 5000 post burn-in samples. Both MCP and PMCP are fitted with the `R` package `ncvreg` with tuning parameters chosen using a 10-fold cross validation.

The inferential performances of the competitors are compared based on the overall mean squared error (MSE) of estimating the true predictor coefficient vector $\boldsymbol{\beta}^*$ and the mean squared error of estimating the truly nonzero predictor coefficient vector $\boldsymbol{\beta}^*_{nz}$ (referred to as the MSE$_{nz}$). These metrics are given by

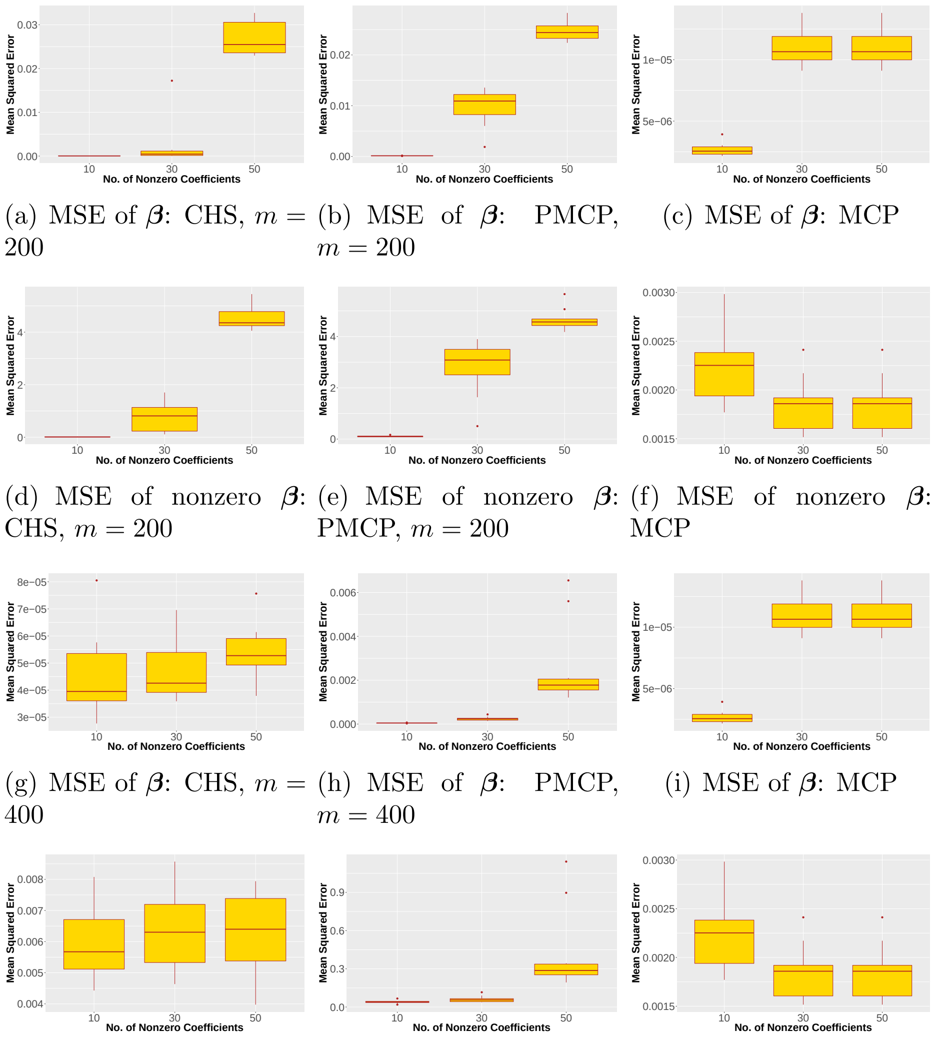
$$\text{MSE} = ||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*||_2^2/p, \quad \text{MSE}_{nz} = ||\hat{\boldsymbol{\beta}}_{nz} - \boldsymbol{\beta}^*_{nz}||_2^2/s, \tag{8}$$

where $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}_{nz}$ is a point estimate for $\boldsymbol{\beta}$ and $\boldsymbol{\beta}_{nz}$, respectively. For CHS, the point estimate is taken to be the posterior mean. Uncertainty of estimating $\boldsymbol{\beta}$ from CHS is characterized through coverage and length of 95% credible intervals averaged over all β_j's, $j = 1, ..., p$. Additionally, we report the coverage and length of 95% credible intervals averaged over truly nonzero β_j's. Since model fitting in (3) is performed with data sketches, it is not possible to draw predictive inference directly. Hence, the quantity $||\boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{X}\boldsymbol{\beta}^*||_2^2/n$ is reported to provide a rough assessment of the predictive inference from CHS. This quantity is also computed and presented for other competitors. All results presented are averaged over 50 replications.

4.2.1 Results

Figures 1 and 2 present the boxplots for MSE and MSE$_{nz}$ for all competitors under the three different sparsity levels in Scenarios 1 and 2, respectively. Understandably, MCP applied on the full data is the best performer in all simulation cases. With small to moderate value of the ratio s/m, CHS significantly outperforms PMCP, both in terms of MSE and MSE$_{nz}$. This becomes evident by comparing the performances of CHS and PMCP for $m = 400$ under all three cases (a)-(c) and for the case $m = 200, s = 10$. In fact when s/m is small, CHS is also found to offer competitive performance with MCP (refer to the results under $m = 400$). This observation is consistent with our findings in Section 4.1.1, where small values of s/m shows little discrepancy between the full posterior of $\boldsymbol{\beta}$ and posterior of $\boldsymbol{\beta}$ under data compression. As sparsity decreases and s/m becomes higher, the performance gap between CHS and PMCP narrows. This is evident from both Figures 1 and 2, corresponding to the case with $s = 30, 50$ and $m = 200$. Consistent with the point estimation of $\boldsymbol{\beta}$, Table 2 shows notable advantage of CHS over PMCP in terms of predictive inference, especially with smaller s/m. MCP on the full data is naturally found to be the superior performer among the three. We observe a similar trend in the performance, both under Scenario 1 and 2.

While accurate point estimation of $\boldsymbol{\beta}^*$ is one of our primary objectives, characterizing uncertainty is of paramount importance given the recent developments in the frequentist literature on characterizing uncertainty in high dimensional regression (Javanmard and Montanari, 2014; Van de Geer *et al.*, 2014; Zhang and Zhang, 2014). Although Bayesian procedures provide an automatic characterization of uncertainty, the resulting credible intervals may not possess the correct frequentist coverage in nonparametric/high-dimensional problems (Szabó *et al.*, 2015). To this end, an attractive adaptive property of the shrinkage priors, including horseshoe, is that the length of the intervals automatically adapt between the signal and noise variables, maintaining close to nominal coverage. It is important to see if this property is preserved under data sketching when the horseshoe prior is set on each component of $\boldsymbol{\beta}$. Table 3 shows that under $m = 400$, 95% credible intervals (CI) of all nonzero coefficients offer closely nominal coverage. While it is also true for $m = 200$ and $s = 10$, the coverage for nonzero coefficients tend to deteriorate as s/m increases. Comparing the average length of 95% CIs for all coefficients with the average length of 95% CIs of nonzero coefficients, we observe that the posterior yields much narrower CIs for coefficients corresponding to the noise predictors. As demonstrated in some of the recent literature (Bhattacharya *et al.*, 2016), the frequentist procedures of constructing confidence intervals for high dimensional parameters (Javanmard and Montanari, 2014; Van de Geer *et al.*, 2014; Zhang and Zhang, 2014) in MCP yield approximately equal sized intervals for the signals and noise variables. Additionally, the tuning parameters in the frequentist procedure require substantial tuning to arrive at satisfactory coverage for the noise (though at the cost of under-covering the signals), while our Bayesian approach is naturally auto-tuned.



(a) MSE of $\boldsymbol{\beta}$: CHS, $m = 200$

(b) MSE of $\boldsymbol{\beta}$: PMCP, $m = 200$

(c) MSE of $\boldsymbol{\beta}$: MCP

(d) MSE of nonzero $\boldsymbol{\beta}$: CHS, $m = 200$

(e) MSE of nonzero $\boldsymbol{\beta}$: PMCP, $m = 200$

(f) MSE of nonzero $\boldsymbol{\beta}$: MCP

(g) MSE of $\boldsymbol{\beta}$: CHS, $m = 400$

(h) MSE of $\boldsymbol{\beta}$: PMCP, $m = 400$

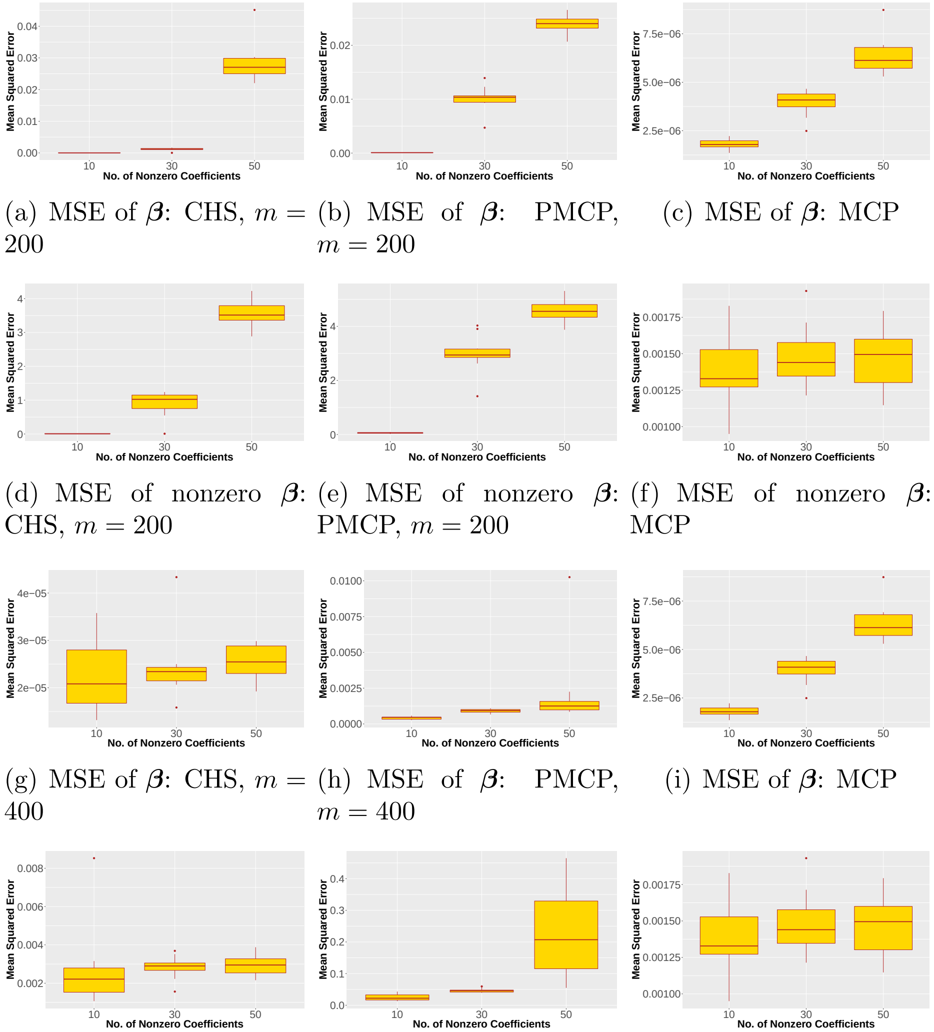
(i) MSE of $\boldsymbol{\beta}$: MCP

(j) MSE of nonzero $\boldsymbol{\beta}$: CHS, $m = 400$

(k) MSE of nonzero $\boldsymbol{\beta}$: PMCP, $m = 400$

(l) MSE of nonzero $\boldsymbol{\beta}$: MCP

Figure 1: First and third row present mean squared error (MSE) of estimating the true predictor coefficient vector $\boldsymbol{\beta}^*$ by a point estimate of $\boldsymbol{\beta}$ from CHS, PMCP and MCP for $m = 200$ and $m = 400$, respectively. Second and fourth row present mean squared error (MSE) of estimating the true nonzero coefficients in $\boldsymbol{\beta}^*$ by a point estimate of the corresponding coefficients in $\boldsymbol{\beta}$ from CHS, PMCP and MCP for $m = 200$ and $m = 400$, respectively. All figures correspond to the scenarios where the predictors are generated under the independent correlation structure (Scenario 1). Each figure shows performance of a competitor under the data generated with 10, 30 and 50 nonzero coefficients in $\boldsymbol{\beta}^*$.



(a) MSE of $\boldsymbol{\beta}$: CHS, $m = 200$

(b) MSE of $\boldsymbol{\beta}$: PMCP, $m = 200$

(c) MSE of $\boldsymbol{\beta}$: MCP

(d) MSE of nonzero $\boldsymbol{\beta}$: CHS, $m = 200$

(e) MSE of nonzero $\boldsymbol{\beta}$: PMCP, $m = 200$

(f) MSE of nonzero $\boldsymbol{\beta}$: MCP

(g) MSE of $\boldsymbol{\beta}$: CHS, $m = 400$

(h) MSE of $\boldsymbol{\beta}$: PMCP, $m = 400$

(i) MSE of $\boldsymbol{\beta}$: MCP

(j) MSE of nonzero $\boldsymbol{\beta}$: CHS, $m = 400$

(k) MSE of nonzero $\boldsymbol{\beta}$: PMCP, $m = 400$

(l) MSE of nonzero $\boldsymbol{\beta}$: MCP

Figure 2: First and third row presenting mean squared error (MSE) of estimating the true predictor coefficient vector $\boldsymbol{\beta}^*$ by a point estimate of β from CHS, PMCP and MCP for $m = 200$ and $m = 400$ respectively. Second and fourth row presenting mean squared error (MSE) of estimating the true nonzero coefficients in $\boldsymbol{\beta}^*$ by a point estimate of the corresponding coefficients in $\boldsymbol{\beta}$ from CHS, PMCP and MCP for $m = 200$ and $m = 400$ respectively. All figures correspond to the scenarios where the predictors are generated under the compound correlation structure (Scenario 2). Each figure shows performance of a competitor under the data generated with 10, 30 and 50 nonzero coefficients in $\boldsymbol{\beta}^*$.

	Scenario 1, $m = 200$			Scenario 1, $m = 400$			Scenario 2, $m = 200$			Scenario 2, $m = 400$		
Sparsity	10	30	50	10	30	50	10	30	50	10	30	50
CHS	0.62	46.56	205.67	0.51	0.57	0.61	0.53	39.22	196.78	0.47	0.59	0.64
PMCP	1.95	71.97	249.70	0.62	2.28	33.19	1.36	62.89	234.63	0.58	1.75	50.49
MCP	0.02	0.07	0.10	0.02	0.07	0.10	0.03	0.07	0.12	0.03	0.07	0.12

Table 2: Mean squared prediction error×10^3 for all the competing models under different simulation scenarios. MSPE is computed as $||\boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{X}\boldsymbol{\beta}^*||^2/n$ for all the competitors.

	Scenario 1, $m = 200$			Scenario 1, $m = 400$			Scenario 2, $m = 200$			Scenario 2, $m = 400$		
Sparsity	10	30	50	10	30	50	10	30	50	10	30	50
Coverage	0.99	0.99	0.98	0.99	0.99	0.98	0.99	0.99	0.98	0.99	0.99	0.98
Length	0.02	0.08	0.17	0.02	0.03	0.03	0.01	0.11	0.17	0.01	0.02	0.03
Coverage$_{nz}$	0.97	0.86	0.68	0.95	0.97	0.97	0.98	0.89	0.63	0.95	0.96	0.95
Length$_{nz}$	5.72	5.93	5.19	5.53	5.90	5.83	5.49	6.59	4.42	5.51	5.79	5.77

Table 3: Average coverage and average length of 95% credible intervals of β_j for CHS under different simulation cases. Here subscript nz is added when the average coverage and average lengths are calculated for truly nonzero coefficients.

5 Study of Orthopedic Fractures Data

We apply our Bayesian sketching approach to a dataset from a study of orthopedic fractures (SOF)(Cummings *et al.*, 1995) (`https://sofonline.ucsf.edu/Home/About`), a multi-center prospective cohort study funded by the National Institutes of Health to identify factors associated with fracture risk in women over the age of 65. Beginning in 1986, the SOF has accumulated over 20 years of data including repeated measures of bone mineral density, hormone levels, functional assessments, and other biometric factors related to fracture risk, osteoporosis, and aging. A primary risk factor for fracture occurrence is bone mineral density (BMD), a numerical summary of a bone's calcium and mineral content which can be obtained via dual x-ray absorptiometry (Black *et al.*, 2020). In fact, the association between BMD, a continuous measure, and fracture occurrence, a binary event, is so strong that BMD has been proposed as a surrogate endpoint for fracture occurrence in clinical trials to simultaneously increase power and decrease trial enrollment. Thus, rather than model fracture occurrence directly, our inferential goal is to identify predictors of total hip BMD (g/cm^2) at Year 10 of the SOF using baseline data (including baseline total hip BMD). The dataset records variables for $n = 4314$ observations with 63 main effects and 1953 pairwise interactions between main effects included as predictors, leading to a total of $p = 2016$ predictors in the study.

Preliminary descriptive analysis of the data shows that the response is influenced only by a few predictors, indicating a sparse regression scenario conducive to the application of the data sketching approach. Additionally, the presence of large number of predictors and large sample size in the data become suitable for the application of data sketching for efficient computation. Given that the descriptive analysis shows high sparsity in the data, Theorem 3.2 indicates a choice of $m = n/\log(n) \approx 500$ should lead to satisfactory performance of our approach. However, as in the simulation studies, we moderately perturb values of m around 500 to assess the change in the inference of $\boldsymbol{\beta}$ to the choice of m. More specifically, analysis is conducted as in the simulation studies, with CHS fitted for $m = 300, 400, 500, 600, 700, 800$. We implemented MCP with the uncompressed data as the benchmark to assess the loss in inference due to data sketching. Similar to simulation studies, the tuning parameters in MCP are chosen based on ten-fold cross validation.

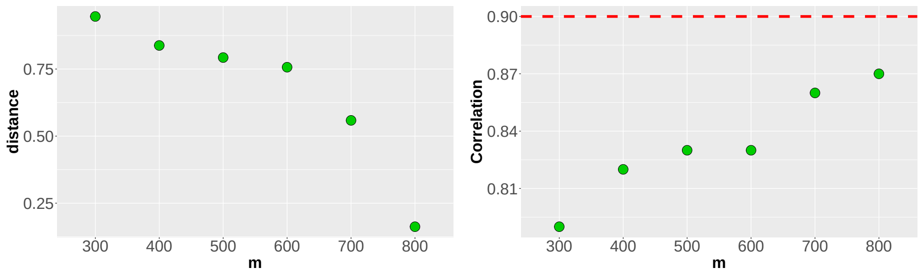
In absence of any ground truth on predictor coefficients, we evaluate point estimate of $\boldsymbol{\beta}$ from CHS by comparing it with the point estimate of $\boldsymbol{\beta}$ obtained from MCP on the uncompressed data. More specifically, standardized sum of squared distance between the two point estimates, given by $||\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{MCP}||_2^2/||\hat{\boldsymbol{\beta}}_{MCP}||_2^2$, is computed and presented for different values of m in Figure 3(a), where the point estimate from CHS is taken to be the posterior mean. As per Figure 3(a), the standardized sum of squared error distance between the two point estimates is small for $m = 500$ and it is close to zero for $m = 800$, suggesting practically equivalent performance of Bayesian sketching with uncompressed MCP in terms of estimating regression parameters. This is presumably due to high sparsity in the regression model which leads to almost identical performance of Bayesian sketching approach with uncompressed regression methods. The performance seems to improve marginally as m increases.

While the above analysis demonstrates excellent performance for estimating regression

Table 4: Average coverage and average length of 95% predictive intervals over all hold-out samples.

m	300	400	500	600	700	800
Coverage	0.97	0.97	0.95	0.95	0.96	0.95
Length	0.48	0.39	0.23	0.21	0.17	0.17

coefficients by the data sketching approach, it is instructive to compare this approach with the uncompressed MCP based on out-of-sample predictive performance. To this end, we emphasize that (3) with sketched data does not allow straightforward model based prediction at hold-out samples. However, viewing this model as a computationally efficient approximation to (1), we develop a strategy wherein the parameter estimates from (3) is employed for predicting response from (1). To be more precise, we first divide the data randomly into 4000 training and 314 hold-out samples and fit (3) on the training data to collect post burn-in MCMC samples of $\boldsymbol{\beta}$ and σ^2. Let the L post burn-in iterates of $\boldsymbol{\beta}$ are given by $\boldsymbol{\beta}^{(1)}, ..., \boldsymbol{\beta}^{(L)}$ and the corresponding iterates for σ^2 are $\sigma^{(1)2}, ...\sigma^{(L)2}$. For the predictor matrix $\boldsymbol{X}_{test}$ corresponding to the hold-out sample, we draw $\boldsymbol{y}_{test}^{(l)} \sim N(\boldsymbol{X}_{test}\boldsymbol{\beta}^{(l)}, \sigma^{(l)2}\boldsymbol{I})$, for $l = 1, ..., L$, from the uncompressed high dimensional linear regression model (1). The correlation between these predicted response and the observed response is computed and the mean of this quantity over post burn-in MCMC samples is presented in Figure 3(b). The results shows highly positive correlation close to 1 for all values of m. As expected, there is an upward trend in the correlation values, with $m = 800$ showing a correlation of 0.87 between the predicted and the observed response. For comparison, the corresponding quantity for MCP with the uncompressed data is also estimated and it turns out to be 0.91. Additionally, to assess how well calibrated the predictive estimates are, we compute coverage and length of 95% predictive interval averaged over all the hold-out samples. The results in Table 4 confirm nominal or close to nominal predictive coverage for all values of m and narrower predictive intervals with decreasing m. Overall, the CHS appears to be competitive even with smaller values of m that leads to substantially efficient computation over its uncompressed analogue.



(a) Std. Sum of Squared Error: $(||\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{MCP}||^2/||\hat{\boldsymbol{\beta}}_{MCP}||^2)$

(b) Correlation between observed and predicted response

Figure 3: Figure on the left shows the ratio of sum of squared distance between point estimate from CHS and from MCP (with uncompressed data) to the sum of squared magnitude of the point estimate from MCP. Figure on the right shows correlation between observed and predicted response corresponding to CHS for the hold out observations. The predicted response is computed using the strategy described in the text. The dotted line shows the correlation for MCP on the uncompressed data.

The horseshoe prior on predictor coefficients does not allow direct identification of influential predictors. Thus, we adopt a two-stage strategy proposed in Guha and Rodriguez (2020) where predictors with highest absolute values of estimated coefficients are identified as influential after accounting for false discovery rate (FDR) at 0.01. The analysis is conducted for $m = 500$ and it finds 135 influential predictors which include 8 main effects and 127 pairwise interaction terms, a sparse number of influential predictors considering the total number of predictors $p = 2016$. The influential main effects include weight-related factors (waist girth, BMI using knee height), age, any difficulty engaging in physical tasks (walking, heavy housework, and preparing meals) and baseline total hip BMD which are consistent with the established literature that identifies weight, limited activity, and age as predictors of future BMD (Dargent-Molina *et al.*, 2000). Of the 127 influential pairwise-interaction terms, the most common factors included in the interaction terms are degree of difficulty in walking, standing, or heavy housework as well as any fracture by age 50 which have previously been highlighted as individual predictors of total hip BMD but should be investigated as potential effect modifiers in future analyses. Further, the interaction between age and baseline total hip BMD is influential suggesting that the relationship between baseline and Year 10 total hip BMD is attenuated by patient age. Together, these results demonstrate the proposed CHS method identifies influential main effects and interaction terms that affirm clinically established associations and identify potential effect modifiers, respectively.

6 Conclusion

This article presents a data sketching/compression approach in high dimensional linear regression with Gaussian scale mixture priors. The approach is arguably the first article on the usage of data sketching to solve computational issues in Bayesian high-dimensional regression with a large sample. The proposed approach does not require storing or manipulating original data in the course of the analysis, rather the analyst can be supplied with the compressed data, which reveal little information about the original data. Simulation studies show advantage of data compression over naive sub-sampling of data, as well as competitive performance of the approach with uncompressed data, especially in presence of a high degree of sparsity. Asymptotic results throw light on the interplay of sparsity, dimension of the compression matrix, sample size and the number of predictors.

Although our approach is demonstrated with the Horseshoe shrinkage prior, it lends easy usage to any other Gaussian scale mixture prior, such as the Generalized Double Pareto (Armagan *et al.*, 2013) or the normal gamma prior (Griffin *et al.*, 2010). The data sketching approach also finds natural extension to high dimensional binary or categorical regression using the data augmentation approach. While simulation studies show promising empirical performance of such an approach, we plan to put forth effort to develop theoretical results in a similar spirit as Section 3. We also plan to extend the data sketching approach to high dimensional nonparametric regression models with big n and p.

7 Acknowledgement

The research of Rajarshi Guhaniyogi is partially supported by grants from the Office of Naval Research (ONR-BAA N000141812741) and the National Science Foundation (DMS-

1854662).

Supplementary Material

Supplementary material contains proofs of the theoretical results.

References

Ahfock, D., Astle, W. J., and Richardson, S. (2017). Statistical properties of sketching algorithms. *arXiv preprint arXiv:1706.03665*.

Ailon, N. and Chazelle, B. (2006). Approximate nearest neighbors and the fast johnson-lindenstrauss transform. In *Proceedings of the thirty-eighth annual ACM symposium on Theory of computing*, pages 557–563.

Ailon, N. and Chazelle, B. (2009). The fast johnson–lindenstrauss transform and approximate nearest neighbors. *SIAM Journal on computing*, **39**(1), 302–322.

Armagan, A., Dunson, D. B., and Lee, J. (2013). Generalized double Pareto shrinkage. *Statistica Sinica*, **23**(1), 119–143.

Bhadra, A., Datta, J., Polson, N. G., and Willard, B. (2019). Lasso meets horseshoe: A survey. *Statistical Science*, **34**(3), 405–427.

Bhattacharya, A., Chakraborty, A., and Mallick, B. K. (2016). Fast sampling with Gaussian scale mixture priors in high-dimensional regression. *Biometrika*, **103**(4), 985–991.

Black, D. M., Bauer, D. C., Vittinghoff, E., Lui, L.-Y., Grauer, A., Marin, F., Khosla, S., de Papp, A., Mitlak, B., Cauley, J. A., McCulloch, C. E., Eastell, R., Bouxsein, M. L., and Foundation for the National Institutes of Health Bone Quality Project (2020). Treatment-related changes in bone mineral density as a surrogate biomarker for fracture risk reduction: meta-regression analyses of individual patient data from multiple randomised controlled trials. *Lancet Diabetes Endocrinol*, **8**(8), 672–682.

Candes, E. J. and Tao, T. (2006). Near-optimal signal recovery from random projections: Universal encoding strategies? *IEEE transactions on information theory*, **52**(12), 5406–5425.

Caron, F. and Doucet, A. (2008). Sparse Bayesian nonparametric regression. In *Proceedings of the 25th international conference on Machine learning*, pages 88–95.

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika*, **97**(2), 465–480.

Castillo, I., Schmidt-Hieber, J., and Van der Vaart, A. (2015). Bayesian linear regression with sparse priors. *The Annals of Statistics*, **43**(5), 1986–2018.

Chen, S., Liu, Y., Lyu, M. R., King, I., and Zhang, S. (2015). Fast relative-error approximation algorithm for ridge regression. In *UAI*, pages 201–210.

Chowdhury, A., Yang, J., and Drineas, P. (2018). An iterative, sketching-based framework for ridge regression. In *International Conference on Machine Learning*, pages 989–998.

Clarkson, K. L. and Woodruff, D. P. (2017). Low-rank approximation and regression in input sparsity time. *Journal of the ACM (JACM)*, **63**(6), 1–45.

Cummings, S. R., Nevitt, M. C., Browner, W. S., Stone, K., Fox, K. M., Ensrud, K. E., Cauley, J., Black, D., and Vogt, T. M. (1995). Risk factors for hip fracture in white women. study of osteoporotic fractures research group. *N. Engl. J. Med.*, **332**(12), 767–773.

Dargent-Molina, P., Poitiers, F., Bréart, G., and EPIDOS Group (2000). In elderly women weight is the best predictor of a very low bone mineral density: evidence from the EPIDOS study. *Osteoporos. Int.*, **11**(10), 881–888.

Dobriban, E. and Liu, S. (2018). A new theory for sketching in linear regression. *arXiv preprint arXiv:1810.06089*.

Donoho, D. L. (2006). Compressed sensing. *IEEE Transactions on information theory*, **52**(4), 1289–1306.

Drineas, P., Mahoney, M. W., Muthukrishnan, S., and Sarlós, T. (2011). Faster least squares approximation. *Numerische mathematik*, **117**(2), 219–249.

Eldar, Y. C. and Kutyniok, G. (2012). *Compressed sensing: theory and applications*. Cambridge university press.

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

Golub, G. H. and Van Loan, C. F. (2012). *Matrix computations*, volume 3. JHU press.

Griffin, J. E., Brown, P. J., *et al.* (2010). Inference with normal-gamma prior distributions in regression problems. *Bayesian Analysis*, **5**(1), 171–188.

Guha, S. and Rodriguez, A. (2020). Bayesian regression with undirected network predictors with an application to brain connectome data. *Journal of the American Statistical Association*, pages 1–13.

Guhaniyogi, R. and Dunson, D. B. (2015). Bayesian compressed regression. *Journal of the American Statistical Association*, **110**(512), 1500–1514.

Guhaniyogi, R. and Dunson, D. B. (2016). Compressed Gaussian process for manifold regression. *The Journal of Machine Learning Research*, **17**(1), 2472–2497.

Halko, N., Martinsson, P.-G., and Tropp, J. A. (2011). Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM review*, **53**(2), 217–288.

Huang, Z. (2018). Near optimal frequent directions for sketching dense and sparse matrices. In *International Conference on Machine Learning*, pages 2048–2057. PMLR.

Javanmard, A. and Montanari, A. (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *The Journal of Machine Learning Research*, **15**(1), 2869–2909.

Jiang, W. (2007). Bayesian variable selection for high dimensional generalized linear models: convergence rates of the fitted densities. *The Annals of Statistics*, **35**(4), 1487–1511.

Johndrow, J. E., Orenstein, P., and Bhattacharya, A. (2020). Scalable approximate MCMC algorithms for the horseshoe prior. *Journal of Machine Learning Research*, **21**(73), 1–61.

Mahoney, M. W. (2011). Randomized algorithms for matrices and data. *arXiv preprint arXiv:1104.5557*.

Maillard, O. and Munos, R. (2009). Compressed least-squares regression. *Advances in neural information processing systems*, **22**, 1213–1221.

Polson, N. G. and Scott, J. G. (2010). Shrink globally, act locally: Sparse Bayesian regularization and prediction. *Bayesian Statistics*, **9**, 501–538.

Raskutti, G. and Mahoney, M. W. (2016). A statistical perspective on randomized sketching for ordinary least-squares. *The Journal of Machine Learning Research*, **17**(1), 7508–7538.

Rue, H. (2001). Fast sampling of Gaussian Markov random fields. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **63**(2), 325–338.

Sarlos, T. (2006). Improved approximation algorithms for large matrices via random projections. In *2006 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS'06)*, pages 143–152. IEEE.

Scott, J. G. and Berger, J. O. (2010). Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. *The Annals of Statistics*, pages 2587–2619.

Song, Q. and Liang, F. (2017). Nearly optimal Bayesian shrinkage for high dimensional regression. *arXiv preprint arXiv:1712.08964*.

Szabó, B., Van Der Vaart, A. W., and van Zanten, J. (2015). Frequentist coverage of adaptive nonparametric Bayesian credible sets. *The Annals of Statistics*, **43**(4), 1391–1428.

Ting, D., Fienberg, S. E., and Trottini, M. (2008). Random orthogonal matrix masking methodology for microdata release. *International Journal of Information and Computer Security*, **2**(1), 86–105.

Vaart, A. V. D. and Zanten, H. V. (2011). Information rates of nonparametric Gaussian process methods. *Journal of Machine Learning Research*, **12**(Jun), 2095–2119.

Van de Geer, S., Bühlmann, P., Ritov, Y., and Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, **42**(3), 1166–1202.

Van Der Pas, S., Kleijn, B., and Van Der Vaart, A. (2014). The horseshoe estimator: Posterior concentration around nearly black vectors. *Electronic Journal of Statistics*, **8**(2), 2585–2618.

Van der Pas, S., Szabó, B., and Van der Vaart, A. (2017). Adaptive posterior contraction rates for the horseshoe. *Electronic Journal of Statistics*, **11**(2), 3196–3225.

Van der Vaart, A. W. and van Zanten, J. H. (2009). Adaptive Bayesian estimation using a gaussian random field with inverse gamma bandwidth. *The Annals of Statistics*, **37**(5B), 2655–2675.

Vempala, S. S. (2005). *The random projection method*, volume 65. American Mathematical Soc.

Vershynin, R. (2010). Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*.

Wang, S., Gittens, A., and Mahoney, M. W. (2017). Sketched ridge regression: Optimization perspective, statistical perspective, and model averaging. *The Journal of Machine Learning Research*, **18**(1), 8039–8088.

Woodruff, D. P. (2014). Sketching as a tool for numerical linear algebra. *arXiv preprint arXiv:1411.4357*.

Yuan, X. (2016). Generalized alternating projection based total variation minimization for compressive sensing. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 2539–2543. IEEE.

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of statistics*, **38**(2), 894–942.

Zhang, C.-H. and Zhang, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **76**(1), 217–242.

Zhang, L., Mahdavi, M., Jin, R., Yang, T., and Zhu, S. (2013). Recovering the optimal solution by dual random projection. In *Conference on Learning Theory*, pages 135–157.

Zhao, L. and Chen, L. (2019). On the privacy of matrix masking-based verifiable (outsourced) computation. *IEEE Transactions on Cloud Computing*.

Zhou, S., Wasserman, L., and Lafferty, J. D. (2008). Compressed regression. In *Advances in Neural Information Processing Systems*, pages 1713–1720.