Biometrika (2018), **103**, 1, *pp*. 1–15
Printed in Great Britain Advance Access publication on 31 July 2018

Privacy Preserving Efficient Computation in Bayesian High Dimensional Regression With Big Data Using Gaussian Scale Mixture Priors

BY RAJARSHI GUHANIYOGI

Department of Statistics, University of California Santa Cruz,
1156 High Street, Santa Cruz, CA 95064, U.S.A
rguhaniy@ucsc.edu

SUMMARY

Bayesian computation of high dimensional linear regression models with a popular Gaussian scale mixture prior distribution 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 computational efficiency greatly diminishes when sample size n is also large. In this article we propose 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 and preserves privacy 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. 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.

Some key words: Bayesian inference, Data privacy, 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

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

where x is a $p \times 1$ predictor vector, β is the corresponding $p \times 1$ coefficient, y is the continuous response and ϵ is the idiosyncratic error. Bayesian methods for estimating β broadly employ two classes of prior distributions. The traditional approach is to develop a discrete mixture of prior distributions (George & McCulloch, 1997; Scott & Berger, 2010). These methods enjoy the advantage of inducing exact sparsity for a subset of parameters and minimax rate of pos-

terior 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 (Armagan et al., 2013; Carvalho et al., 2010; Caron & Doucet, 2008) have emerged, which induce approximate sparsity in high-dimensional parameters. Such prior distributions can mostly be expressed as global-local scale mixtures of Gaussians (Polson & Scott, 2010) given by,

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

where g_1, g_2 and f are densities supported on the real line. The parameters λ_j's are referred to as the local-scale parameters specific to the predictors and τ is known as the global scale parameter that controls overall shrinkage induced by these priors. Different choices of g_1 and g_2 lead to different classes of Bayesian shrinkage priors. For example, the state-of-the-art Horseshoe shrinkage prior (Carvalho et al., 2010) is obtained by choosing g_1 and g_2 both as half Cauchy distributions.

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 X be the $n \times p$ predictor matrix, y be the $n \times 1$ response vector and $\Delta = \tau^2 diag(\lambda_1,...,\lambda_p)$, the distribution of $\beta = (\beta_1,...,\beta_p)'$ conditional on $\lambda = (\lambda_1,...,\lambda_p)', \tau,\ \sigma$, y and X follows $N((X'X+\Delta^{-1})^{-1}X'y, \sigma^2(X'X+\Delta^{-1})^{-1})$, 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 $\beta,\lambda,\tau,\sigma$ are found to offer an accurate approximation to the operating characteristics of discrete mixture priors. However, the existing algorithms (Rue, 2001) to sample from the full conditional posterior of β require storing and computing the Cholesky decomposition of the $p \times p$ matrix $(X'X+\Delta^{-1})$, that necessitates p^3 floating point operations (flops) and p^2 storage units, which can be severely prohibitive for large p. There is a recently proposed algorithm for efficient computations in high dimensional regressions involving small n and large p (Bhattacharya et al., 2016), though it is less straightforward to adapt this approach when n is also large.

The approach we develop here compresses the response vector and predictor matrix by a random linear transformation, reducing the number of records from n to m, while preserving the number of original predictors. The compressed response and predictors are then made available to 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, one can adapt existing algorithms on the compressed data for efficient estimation of posterior distribution for predictor coefficients. Theoretically, 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 coefficient. We then identify conditions on the predictor matrix, the interlink between the dimension of the random compression matrix, sample size, sparsity of the true regression coefficient vector and the number of predictors to prove consistent estimation of the predictor coefficients asymptotically. Our detailed empirical investigation with the Horseshoe shrinkage prior (Carvalho et al., 2010) ensures that the relevant predictors can be learnt from the compressed data as well as from the original uncompressed data. Moreover, in presence of a higher degree of sparsity in the true regression model, the actual estimates of parameters and predictions are as accurate as they would have been, had the uncompressed data been used. Another attractive feature 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.

Our proposal is related to compressed sensing approaches (Donoho, 2006; Candes & Tao, 2006; Eldar & Kutyniok, 2012), with an important difference. While compressed sensing approaches broadly aim at reconstructing a sparse X from a small number of its random linear combinations, we intend to reconstruct a sparse function of X only, and not the X and y themselves. In fact, from our point of view of privacy of the response vector and predictor matrices, approximately reconstructing them should be viewed as undesirable. Our approach is fundamentally different from Guhaniyogi & Dunson (2015) in that they compress each predictor vector, leading to an m-dimensional compressed predictor from a p-dimensional predictor for each sample. In contrast, our compression framework does not alter the number of predictor variables in the analysis before and after compression. A few notable articles in the machine learning literature show that major statistical procedures, such as the principal component analysis, clustering, and even identifying the correct sparse set of relevant variables by the lasso are as effective under compression (Liu et al., 2005; Zhou et al., 2008). However, we are not aware of any earlier work where the full potential of the data compression approach to enable efficient Bayesian computation with big n and p has been carefully studied both theoretically and empirically.

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 on a real data with big p and n in Section 5, followed by the concluding remarks in Section 6.

2. Compressing Response Vector and Predictor Matrices for Large n

For subjects $i = 1, ..., n$, let $y_i \in \mathcal{Y}$ denote the response for subject i corresponding to the predictor $x_i \in \mathcal{R}^p$. We focus on the scenario where $n < p$, with n and p both large. Let $y = (y_1, ..., y_n)'$ be the $n \times 1$ vector of responses and $X = [x_1 : \cdots : x_p]'$ be the $n \times p$ matrix of predictors. We consider a data compression approach in the high dimensional linear regression setting having the form

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

where σ^2 is a idiosyncratic error variance and Φ is an $m \times n$ dimensional compression matrix with $m << min(n, p)$. We do not estimate Φ as a variable in the regression, rather draw the elements Φ_{ij} of the Φ matrix independently from N$(0, 1/n)$. This is a well known method of constructing compression matrices in the literature of compressed sensing (see e.g., Eldar & Kutyniok (2012)).

The data compression approach implemented here appears to be a special case of the *matrix masking* technique proposed in the earlier privacy literature (Ting et al., 2008; Zhou et al., 2008; Zhao & Chen, 2019), which, although popular in the privacy literature, has not been given due attention theoretically, especially from a Bayesian perspective. A typical matrix masking procedure pre- and post-multiplies the data matrix X by matrices C and D, respectively, and releases CXD for the ensuing analysis. The transformation is quite general, and allows the possibility of deleting records, suppressing subsets of variables and data swapping. This article chooses $C = \Phi$ and D as the identity matrix so as to keep the original interpretation of the predictors.

However, even in the case of Φ being known, the linear system ΦX is grossly under-determined due to $m << min(n, p)$. The privacy in information theoretic terms of this compression procedure could be evaluated using an upper bound of the average mutual information $I(\Phi X, X)/np$ per unit in the original data matrix X, and showing that $Sup\, I(\Phi X, X)/np = O(m/n)$ (Zhou et al., 2008), where supremum is taken over all possible distributions of 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 Φ is known. In practice, only ΦX (and not even Φ) will be revealed to the analyst. Hence, the imposed privacy 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 see this, note that pre-multiplying the high dimensional linear regression equation $y = X\beta + \epsilon$ by Φ results in

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

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 ϵ are i.i.d., whereas the error vector from (4) follows a $N(0, \sigma^2\Phi\Phi')$ distribution. Lemma 5.36 and Remark 5.40 of Vershynin (2010) show that $||\Phi\Phi' - I_m||_2 \leq C'\sqrt{m/n}$, with probability at least $1 - e^{-C''m}$, for some constants $C', C'' > 0$. As m grows at a slower rate than n, $m/n \to 0$ asymptotically. Hence, with large n, the error distributions of (3) and (4) behave similarly with a probability close to 1.

With prior distribution on β set as a Gaussian scale-mixture distribution from the class of distributions given by (2), posterior computation cycles through updating the full conditional distributions: (a) $\beta|\lambda, \sigma, \tau$, (b) $\lambda|\beta, \sigma, \tau$, (c) $\sigma|\lambda, \beta, \tau$ and (d) $\tau|\lambda, \beta, \sigma$. While updating (b), (c) and (d) do not face any computational challenge due to big n or p, full conditional posterior updating of $\beta|\lambda, \sigma, \tau$ has the form given by

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

The most efficient algorithm to sample from β (Rue, 2001) computes Cholesky decomposition of $\left(X'\Phi'\Phi X + \Delta^{-1}\right)$ 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 & 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 Bhattacharya et al. (2016) (in the context of uncompressed data with small sample size) to our setting. The detailed steps are given as following:

Step 1: Draw $v_1 \sim N(0, \sigma^2\Delta)$ and $v_2 \sim N(0, I_m)$
Step 2: Set $v_3 = \Phi X v_1/\sigma + v_2$.
Step 3: Solve $(\Phi X\Delta X'\Phi' + I_m)v_4 = (y/\sigma - v_3)$.
Step 4: Set $v_5 = v_1 + \sigma\Delta X'\Phi' v_4$.

v_5 is a draw from the full conditional posterior distribution of β. Notably, the computational complexity of Steps 1-4 is dominated by two operations: (Operation A) computing the inverse of $(\Phi X\Delta X'\Phi' + I_m)$, and (Operation B) calculating $\Phi X\Delta X'\Phi'$. (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 even by 400 times 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 Compression Approach

This section studies convergence properties of the data compression approach with high dimensional shrinkage prior on predictor coefficients. 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 predictors p_n and the number of rows of the compression matrix m_n are respectively larger and smaller than the sample size n. Naturally, the predictor coefficient β and the compression matrix Φ are also functions of n. We denote them by β_n and Φ_n, respectively. Note that the true data generating model under data compression is given by (4). We use superscript $*$ to indicate the true parameters β_n^* and σ^{*2} in (4). For simplicity, we assume that $\sigma^2 = \sigma^{*2}$ is known and fixed at 1. This is a common assumption in asymptotic studies (Vaart & 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 et al., 2009). 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 $||\cdot||_0$. Finally, $e_{min}(A)$ and $e_{max}(A)$ respectively represent the minimum and maximum eigenvalues of the square matrix A.

3.2. *Assumptions, Framework and The Main Result*

For any subset of indices $\xi \in \{1, ..., p_n\}$, $|\xi|$ denotes the number of elements in the index set ξ. Depending on whether A is a vector or a matrix, A_ξ denotes the sub-vector or the sub-matrix corresponding to the indices ξ. We let $\xi^* = \{j : \beta_{j,n}^* \neq 0\}$, i.e., ξ^* are the indices of the nonzero entries for the true predictor coefficient β_n^*, and s_n (dependent on n) designates the number of nonzero entries in β_n^*, i.e., $s_n = ||\beta_n^*||_0 = |\xi^*|$. Since the shrinkage prior on β_n assigns zero probability at the point zero, the exact number of nonzero elements of β_n is always p_n. A meaningful comparison with the value s_n is made by considering $\tilde{s}_n$, the number of elements of β_n exceeding in absolute value a threshold a_n, which will be specified later. In other words, only elements with absolute value larger than a_n will be treated as significant and counted towards non-zero entries. Before rigorously studying properties of the posterior distribution, we state some regularity conditions on the design matrix X, the compression matrix Φ_n, the true sparsity s_n and the model determined approximate sparsity $\tilde{s}_n$.

(A) The columns X_j of the design matrix X satisfy $||X_j||_2^2 = O(n), \forall\, j = 1, ..., p_n$.

(B) $||\Phi_n\Phi_n' - I_m||_2 \leq C'\sqrt{m_n/n}$, for some constant $C' > 0$, for all large n.
(C) $s_n \log(p_n) = o(m_n)$, $s_n \log(n) = o(m_n)$.
(D) $m_n = o(n)$ and $n^{1/2+\tilde{\delta}}/m_n \to 0$ for some $\tilde{\delta} > 0$.
(E) $\tilde{s}_n = O(s_n)$.

(A) is a common assumption in the context of compressed sensing, see Zhou et al. (2008). From the theory of random compression 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 dimension of the random matrix, number of predictor coefficients and the sample size. (D) allows m_n to grow at a slower rate than n, while at the same time ensuring a faster growth than $\sqrt{n}$ for m_n. Assumptions (A) and (C) jointly impose restrictions on the compressed predictor matrix $\tilde{X} = \Phi_n X$. Following the proof of Proposition 3.6 in Zhou et al. (2008), we observe that assumptions (A) and (C) imply $e_{min}(\tilde{X}_\xi'\tilde{X}_\xi/m_n) \geq \eta$, for some $\eta > 0$ and for all $\xi \supset \xi^*$ such that $|\xi| \leq s_n + \tilde{s}_n$. We will make use of this fact in the proof of our posterior consistency result. Our next set of assumptions concern the tail behavior of the shrinkage priors of interest and the magnitude of the nonzero entries of β_n^*. Let $h_\mu(x)$ denote the prior density of $\beta_{j,n}$ for all j with the set of hyper-parameters μ. For $a_n = \sqrt{s_n \log(p_n)/m_n}/p_n$ and for a sequence M_n, we assume

(F) $\max_{j \in \xi^*} |\beta_{j,n}^*| < M_n/2$.
(G) $1 - \int_{-a_n}^{a_n} h_\mu(x)dx \leq p_n^{-(1+u)}$, for some positive constant u.
(H) $-\log(\inf_{x \in [-M_n, M_n]} h_\mu(x)) = O(\log(p_n))$.

Assumption (F) restricts the growth of the nonzero entries in the true regression parameter asymptotically. Assumption (G) 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 (H) essentially controls the prior density around the true predictor coefficient. Notably, Assumptions (F)-(H) are frequently used in the high dimensional Bayesian regression literature, including in Jiang (2007) and Song & Liang (2017).

Define $\mathcal{B}_n = \Big\{\text{At least } \tilde{s}_n \text{ absolute values of } \beta_n \text{ are greater than } a_n = \sqrt{s_n \log(p_n)/m_n}/p_n\Big\}$, $\mathcal{C}_n = \{\beta_n : ||\beta_n - \beta_n^*||_2 > \epsilon\}$ and $\mathcal{A}_n = \mathcal{B}_n \cup \mathcal{C}_n$. Further suppose $\pi_n(\cdot)$ and $\Pi_n(\cdot)$ are the prior and posterior densities of β_n with n observations respectively, so that

$$\Pi_n(\mathcal{A}_n) = \frac{\int_{\mathcal{A}_n} f(\tilde{y}|\beta_n)\pi_n(\beta_n)}{\int f(\tilde{y}|\beta_n)\pi_n(\beta_n)},$$

where $f(\tilde{y}|\beta_n)$ is the joint density of $\tilde{y} = \Phi_n y$ under model (3). This article intends to show

$$\Pi_n(\mathcal{A}_n) \to 0, \text{ a.s., when } n \to \infty. \tag{6}$$

The following theorem shows that (6) holds for the proposed model, with the proof of the theorem given in the appendix.

THEOREM 1. *Under Assumptions (A)-(H), our proposed model satisfies posterior consistency as defined in (6).*

4. SIMULATION STUDIES

This section investigates frequentist operating characteristics of our data compression approach with a shrinkage prior along with its competitors in high dimensional regression. In particular, we implement (3) with the Horseshoe shrinkage prior (Carvalho et al., 2010) on each of the predictor coefficients β_j, and denote it by Compressed Horseshoe (CHS). As a frequentist competitor, we implement Lasso (Tibshirani, 1996) on the full data. Additionally, we fit Lasso on randomly chosen m data points from the sample of size n, and refer to this competitor as Partial Lasso (PLasso). The ordinary Lasso on full data provides a comparison of our approach with a frequentist penalized optimizer in high dimensional regression with big n and p. On the other hand, comparison of CHS with PLasso demonstrates the inferential advantage of random compression over naive sampling of m data points out of n data points. Although the remaining section presents excellent performance of the compression 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).

In the simulation examples, we draw $n = 5000$ 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 x_i for each $i = 1, ..., n$ are simulated from $N(0, \Sigma)$, with two different constructions of Σ undertaken in simulation studies.

Scenario 1: $\Sigma = I_p$, i.e., all predictors are simulated i.i.d. We refer to this as the independent correlation structure for the predictors.

Scenario 2: $\Sigma = 0.5I_p + 0.5J_p$, where 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.

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 Lasso and PLasso are fitted with the `R` package `glmnet`.

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

$$\text{MSE} = ||\hat{\beta} - \beta^*||_2^2/p, \quad \text{MSE}_{nz} = ||\hat{\beta}_{nz} - \beta^*_{nz}||_2^2/s, \tag{7}$$

where $\hat{\beta}$ and $\hat{\beta}_{nz}$ is a point estimate for β and β_{nz}, respectively. For CHS, the point estimate is taken to be the posterior mean. Uncertainty of estimating β 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. Finally, the quantity $||X\hat{\beta} - X\beta^*||_2^2/n$ is reported to assess the predictive inference from the competitors. Notably, in all three cases (a)-(c), $s\log(p)$ is similar or larger than m, presenting

Table 1. *Mean squared prediction error*$\times 10^3$ *for all the competing models under different simulation scenarios. MSPE is computed as* $||X\hat{\beta} - X\beta^*||^2/n$ *for all the competitors.*

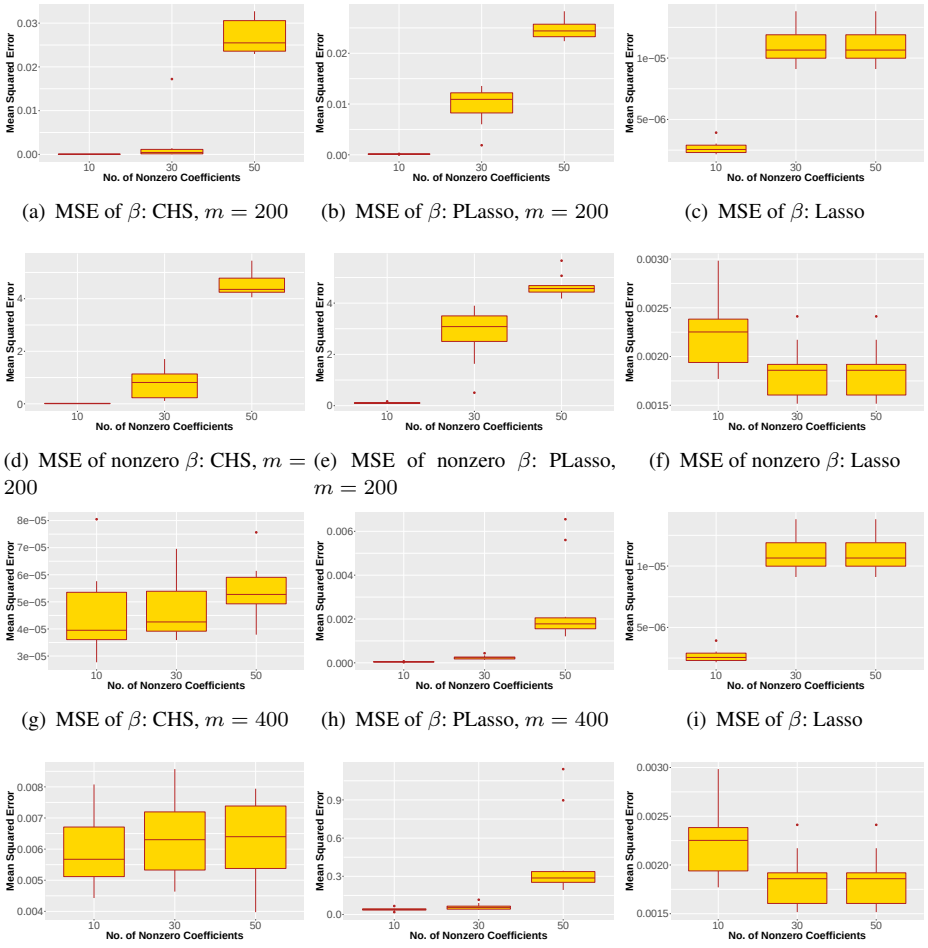
	Scenario 1, $m = 200$			Scenario 1, $m = 400$			Scenario 2, $m = 200$			Scenario 2, $m = 400$		
Sparsity	10	30	50	10	30	50	10	30	50	10	30	50
CHS	0.62	46.56	205.67	0.51	0.57	0.61	0.53	39.22	196.78	0.47	0.59	0.64
PLasso	1.95	71.97	249.70	0.62	2.28	33.19	1.36	62.89	234.63	0.58	1.75	50.49
Lasso	0.02	0.07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

challenging contexts where the theoretical guarantee (as given in Theorem 1) on estimation of β do not necessarily follow. All results presented are averaged over 20 replications.

4.1. *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, Lasso 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 PLasso, both in terms of MSE and MSE$_{nz}$. This becomes evident by comparing the performances of CHS and PLasso 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 Lasso (refer to the results under $m = 400$). As sparsity decreases and s/m becomes higher, the performance gap between CHS and PLasso 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 β, Table 1 shows notable advantage of CHS over PLasso in terms of predictive inference, especially with smaller s/m. Lasso 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 β^* 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 & Montanari, 2014; Van de Geer et al., 2014; Zhang & 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 compression when the Horseshoe prior is set on coefficients β. Table 2 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 & Montanari, 2014; Van de Geer et al., 2014; Zhang & Zhang, 2014) in Lasso 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 β: CHS, $m = 200$ (b) MSE of β: PLasso, $m = 200$ (c) MSE of β: Lasso

(d) MSE of nonzero β: CHS, $m = 200$ (e) MSE of nonzero β: PLasso, $m = 200$ (f) MSE of nonzero β: Lasso

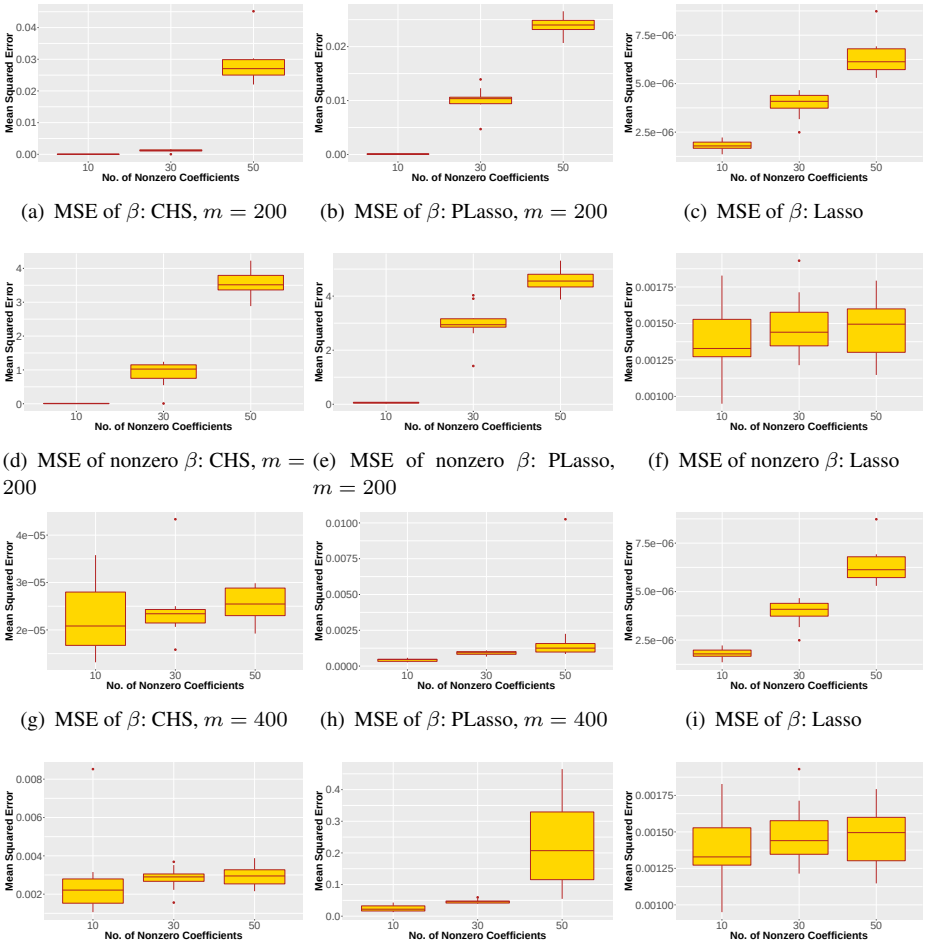
(g) MSE of β: CHS, $m = 400$ (h) MSE of β: PLasso, $m = 400$ (i) MSE of β: Lasso

(j) MSE of nonzero β: CHS, $m = 400$ (k) MSE of nonzero β: PLasso, $m = 400$ (l) MSE of nonzero β: Lasso

Fig. 1. First and third row present mean squared error (MSE) of estimating the true predictor coefficient β_0 by a point estimate of β from CHS, PLasso and Lasso for $m = 200$ and $m = 400$, respectively. Second and fourth row present mean squared error (MSE) of estimating the true nonzero coefficients in β_0 by a point estimate of the corresponding coefficients in β from CHS, PLasso and Lasso 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 β_0.

4.2. *Choice of Dimension of the Compression Matrix*

Since CHS is regarded as a computationally convenient approximation to the Horseshoe prior on the full data, the approximation accuracy is expected to increase as m approaches n. On the contrary, a higher value of m diminishes any computational gain offered by CHS. In practice, we define a model fitting statistic for CHS, given by $Mfit(m) =$



(a) MSE of β: CHS, $m = 200$ (b) MSE of β: PLasso, $m = 200$ (c) MSE of β: Lasso

(d) MSE of nonzero β: CHS, $m = 200$ (e) MSE of nonzero β: PLasso, $m = 200$ (f) MSE of nonzero β: Lasso

(g) MSE of β: CHS, $m = 400$ (h) MSE of β: PLasso, $m = 400$ (i) MSE of β: Lasso

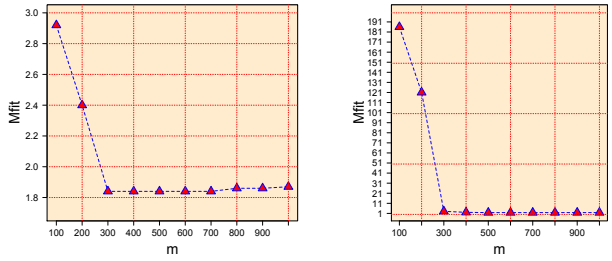
(j) MSE of nonzero β: CHS, $m = 400$ (k) MSE of nonzero β: PLasso, $m = 400$ (l) MSE of nonzero β: Lasso

Fig. 2. First and third row presenting mean squared error (MSE) of estimating the true predictor coefficient β_0 by a point estimate of β from CHS, PLasso and Lasso for $m = 200$ and $m = 400$ respectively. Second and fourth row presenting mean squared error (MSE) of estimating the true nonzero coefficients in β_0 by a point estimate of the corresponding coefficients in β from CHS, PLasso and Lasso 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 β_0.

$Mfit_1(m) + Mfit_2(m)$, where $Mfit_1(m) = \sum_{i=1}^{n}(y_i - y_{rep,i})^2$, $Mfit_2(m) = \sum_{i=1}^{n}\hat{\sigma}^2_{rep,i}$, $y_{rep,i} = \sum_{t=1}^{T} y_{rep,i,t}$ and $\hat{\sigma}^2_{rep,i} = \sum_{t=1}^{T}(y_{rep,i,t} - y_{rep,i})^2/T$. Here $y_{rep,i,t}$ is drawn from $N(x_i'\beta^{(t)}, \sigma^{2(t)})$, where $\beta^{(t)}, \sigma^{2(t)}$ are the t-th post burn-in iterates of β and σ^2 obtained from fitting (3), $t = 1, ..., T$. In the construction of $Mfit(m)$, $Mfit_1(m)$ evaluates inferential accuracy whereas $Mfit_2(m)$ indicates model fit. The metric closely mimics posterior predictive loss

Table 2. *Average coverage and average length of 95% credible intervals of β for CHS under different simulation cases.*

	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 (overall)	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 (overall)	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 (nonzero)	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 (nonzero)	5.72	5.93	5.19	5.53	5.90	5.83	5.49	6.59	4.42	5.51	5.79	5.77



(a) Model fit: $s = 10$ (b) Model fit: $s = 30$

Fig. 3. The model fitting statistic ($Mfit(m)/10^4$) values for different choices of dimension of the compression matrix. Results are presented for data generated under Scenario 1 with $n = 5000$, $p = 10000$ and with two different sparsity levels, $s = 10$ and $s = 30$.

criterion (Gelfand & Ghosh, 1998), except that the post burn-in iterates from the approximating model (3) is used in the original model, deemed necessary due to the changing dimension of the response vector with different choices of m. We propose to fit the model (3) over a grid of m values in parallel and choose the value of m after which $Mfit(m)$ values stabilize.

To assess the above algorithm for selecting m, we fit (3) for $m = 100, 200, ..., 1000$ with data generated under Scenario 1 with $s = 10$ and 30. Figure 3 shows that the model fitting statistic stabilizes at $m = 400$ and $m = 500$ respectively for $s = 10$ and $s = 30$. Interestingly, MSE_{nz} for $m = 400$ and $m = 500$ under these two simulation scenarios are 0.005 and 0.004 respectively, which are of the same order with the corresponding values obtained from the full Lasso (see Figure 1). Likewise, we find the algorithm for selecting m produces desirable inference in terms of parameter estimation in all other simulation experiments.

5. Application to Financial Stock Database

This section illustrates the performance of CHS along with its competitors for a financial data set consisting of minute by minute average log-prices of the NASDAQ stock exchange from September 10, 2018 to September 30, 2018 during trading hours. The data consists of log-prices of Apple stocks along with 3430 assets, and the aim of the data analysis is to evaluate the elasticity of the price of Apple stocks with respect to the prices of the remaining assets. This is of particular interest, since Apple, one of the biggest publicly traded companies in the world, is ubiquitous in portfolios ranging from retirement funds to small portfolios managed by individuals in the financial market. Thus accurate inference on the relationship between Apple

Table 3. *MSPE for all competitors are reported for the stock price data. We additionally report coverage and length of 95% predictive interval for CHS.*

	MSPE			Coverage	Length
Competitors	CHS	PLasso	Lasso	CHS	CHS
$m = 400$	0.01	1.60	1.36	0.99	0.03
$m = 500$	0.01	1.58	1.36	0.99	0.02

and other financial stocks allows better portfolio diversification. We employ a high dimensional linear regression model with the log-price of the Apple stock as the response and log-prices of other assets as predictors.

The data includes several assets, such as ETFs, Trust Funds, stock tracker indexes, and banks, which as expected, present a very high degree of collinearity. To avoid less desirable inference due to high collinearity, a few financial assets are removed along with assets which have very few transactions (less than 40), yielding 2014 predictors for the analysis. The data set consists of 2276 observations collected over a few days. Due to the time window of the collected data being narrow, we ignore the temporal variation of log-prices of the Apple stock.

We fit CHS with $m = 100, 200, 300, 400, 500, 600$ and observe that the model fitting statistic stabilizes after $m = 400$, with $m = 400$ and $m = 500$ yield the lowest model fitting statistic. Hence, we conclude no practical gain of fitting the compressed data model beyond $m = 500$ and present results for both $m = 400$ and $m = 500$. Inference with PLS and Lasso are also presented along with CHS. Since MSPE as described in Section 4 involves comparison between the point estimate of $X\beta$ and its true value which is not known in the real data, we formulate MSPE for the real data as $||y - X\hat{\beta}||^2/n$, where $\hat{\beta}$ is some point estimate of β. The data shows notable advantage of fitting the Horseshoe shrinkage prior over its frequentist alternatives, perhaps due to the high degree of collinearity leading to an ill conditioned $X'X$ matrix. As a result, the MSPE from the approximated inference offered by CHS demonstrates much lower value than PLS or Lasso (refer to Table 3). As expected, Lasso on the full data shows much lower MSPE than PLS. Additionally, the average coverage and length of 95% predictive interval from CHS are close to nominal, indicating satisfactory characterization of uncertainty.

6. CONCLUSION

This article presents a data compression approach in high dimensional linear regression with Gaussian scale mixture priors. The proposed approach ensures privacy of the original data by revealing little information about it to the analyst. Additionally, it leads to a massive reduction in computation for big n and p. Simulation studies show notable 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 applied to the Horseshoe 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 compression 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 our idea to high dimensional nonparametric 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).

APPENDIX

We begin by stating an important result from the random matrix theory, the proof of which is immediate following Theorem 5.31 and Corollary 5.35 of Vershynin (2010).

LEMMA A1. *Consider the $m_n \times n$ compression matrix Φ_n with each entry being drawn independently from $N(0, 1/n)$. Then, almost surely*

$$(\sqrt{n} - \sqrt{m_n} - o(\sqrt{n}))^2/n \leq e_{min}(\Phi_n \Phi_n') \leq e_{max}(\Phi_n \Phi_n') \leq (\sqrt{n} + \sqrt{m_n} + o(\sqrt{n}))^2/n, \quad \text{(A1)}$$

when both $m_n, n \to \infty$.

LEMMA A2. *There exist a sequence of test functions κ_n for testing $H_0 : \beta_n = \beta_n^*$ vs. $H_1 : \beta_n \in \mathcal{A}_n$ such that $E_{\beta_n^*}(\kappa_n) \leq \exp(-\tilde{c}_3 m_n)$, $\sup_{\beta_n \in \mathcal{A}_n} E_{\beta_n}(1 - \kappa_n) \leq \exp(-2\tilde{c}_4 m_n)$, for all large n, for some $\tilde{c}_3, \tilde{c}_4 > 0$.*

Proof. Denote $\tilde{y} = \Phi_n y$ and $\tilde{X} = \Phi_n X$. Define a sequence of test functions $\kappa_n = \max_{\xi \supset \xi^*, |\xi| \leq s_n + \tilde{s}_n} 1\{||(\tilde{X}_\xi' \tilde{X}_\xi)^{-1} \tilde{X}_\xi' \tilde{y} - \beta_{n,\xi}^*||_2 \geq \epsilon/4\}$. Let $\hat{\beta}_{n,\xi} = (\tilde{X}_\xi' \tilde{X}_\xi)^{-1} \tilde{X}_\xi' \tilde{y}$. Then

$$\begin{aligned}
E_{\beta_n^*}(\kappa_n) &\leq \sum_{\xi \supset \xi^*, |\xi| \leq s_n + \tilde{s}_n} P_{\beta_n^*}(||\hat{\beta}_{n,\xi} - \beta_{n,\xi}^*||_2 \geq \epsilon/4) = \sum_{\xi \supset \xi^*, |\xi| \leq s_n + \tilde{s}_n} P_{\beta_n^*}((\hat{\beta}_{n,\xi} - \beta_{n,\xi}^*)'(\hat{\beta}_{n,\xi} - \beta_{n,\xi}^*) \geq \epsilon^2/16) \\
&\leq \sum_{\xi \supset \xi^*, |\xi| \leq s_n + \tilde{s}_n} P_{\beta_n^*}((\hat{\beta}_{n,\xi} - \beta_{n,\xi}^*)' \tilde{X}_\xi' (\Phi_n \Phi_n')^{-1} \tilde{X}_\xi (\hat{\beta}_{n,\xi} - \beta_{n,\xi}^*) \geq \eta \epsilon^2 m_n/16 \times n/(\sqrt{m_n} + \sqrt{n} + o(\sqrt{n}))^2) \\
&\leq \sum_{\xi \supset \xi^*, |\xi| \leq s_n + \tilde{s}_n} P_{\beta_n^*}(\chi^2_{|\xi|} \geq \eta \epsilon^2 m_n/16 \times n/(\sqrt{m_n} + \sqrt{n} + o(\sqrt{n}))^2) \\
&\leq \sum_{\xi \supset \xi^*, |\xi| \leq s_n + \tilde{s}_n} P_{\beta_n^*}(\chi^2_{|\xi|} \geq (1 - \delta)\eta \epsilon^2 m_n/16) \leq \binom{p_n}{\tilde{s}_n + s_n} \exp(-2\tilde{c}_3 m_n) \leq \exp(-\tilde{c}_3 m_n),
\end{aligned}$$

for some constant $\tilde{c}_3 > 0$. where the inequality in the second line follows from two results. First, by Lemma A1, $e_{min}((\Phi_n \Phi_n')^{-1}) \geq n/(\sqrt{n} + \sqrt{m_n} + o(\sqrt{n}))^2$. Second, by assumptions (A) and (C), following the proof of Proposition 3.6 in Zhou et al. (2008), $e_{min}(\tilde{X}_\xi' \tilde{X}_\xi/m_n) \geq \eta$, for some $\eta > 0$ and for all $\xi \supset \xi^*$ such that $|\xi| \leq s_n + \tilde{s}_n$. The first inequality in the fourth line follows due to the fact that $n/(\sqrt{m_n} + \sqrt{n} + o(\sqrt{n}))^2 \to 1$ as $n \to \infty$. Hence $n/(\sqrt{m_n} + \sqrt{n} + o(\sqrt{n}))^2 \geq 1 - \delta$ for some $\delta \in (0, 1)$, for all large n. The second inequality in the fourth line in obtained by applying the Bernstein inequality (Song & Liang, 2017). The third inequality in the fourth line is obtained by the fact that $\binom{p_n}{\tilde{s}_n + s_n} \leq p_n^{\tilde{s}_n + s_n} \leq \exp((\tilde{s}_n + s_n)\log(p_n)) \leq \exp(\tilde{c}_3 m_n)$, using assumptions (C) and (E).

Consider $\zeta = \xi^* \cup \{k : |\beta_{k,n}| \geq a_n\}$. Then $\zeta \in \{\xi : \xi \supset \xi^*, |\xi| \leq s_n + \tilde{s}_n\}$. Then

$$\sup_{\beta_n \in \mathcal{A}_n} E_{\beta_n}(1 - \kappa_n) \leq \sup_{\beta_n \in \mathcal{A}_n} \{1 - P_{\beta_n}(||\hat{\beta}_{n,\zeta} - \beta_{n,\zeta}^*|| \geq \epsilon/4)\} = \sup_{\beta_n \in \mathcal{A}_n} P_{\beta_n}(||\hat{\beta}_{n,\zeta} - \beta_{n,\zeta}^*|| \leq \epsilon/4).$$

Under $\mathcal{A}_n$, $||\beta_{n,\zeta} - \beta_{n,\zeta}^*|| \geq ||\beta_n - \beta_n^*|| - ||\beta_{n,\zeta^c} - \beta_{n,\zeta^c}^*|| \geq \epsilon - a_n p_n \geq \epsilon/2$. Here the last inequality follows due to the fact that $\beta_{n,\zeta^c}^* = 0$ and for any $k \in \zeta^c$, $|\beta_{n,k}| \leq a_n$ and $a_n < \epsilon/(2p_n)$ (due to the fact

that $s_n \log(p_n)/m_n \to 0$). Using the above fact, we have

$$\sup_{\beta_n \in \mathcal{A}_n} P_{\beta_n}(||\hat{\beta}_{n,\zeta} - \beta^*_{n,\zeta}|| \leq \epsilon/4) \leq \sup_{\beta_n \in \mathcal{A}_n} P_{\beta_n}(||\hat{\beta}_{n,\zeta} - \beta_{n,\zeta}|| \geq ||\beta_{n,\zeta} - \beta^*_{n,\zeta}|| - \epsilon/4)$$
$$= \sup_{\beta_n \in \mathcal{A}_n} P_{\beta_n}(||\hat{\beta}_{n,\zeta} - \beta_{n,\zeta}|| \geq \epsilon/4)$$
$$\leq \sup_{\beta_n \in \mathcal{A}_n} P_{\beta_n}((\hat{\beta}_{n,\zeta} - \beta_{n,\zeta})' \tilde{X}'_\zeta (\Phi_n \Phi'_n)^{-1} \tilde{X}_\zeta (\hat{\beta}_{n,\zeta} - \beta_{n,\zeta}) \geq \eta\epsilon^2 m_n/16 \times n/(\sqrt{m_n} + \sqrt{n} + o(\sqrt{n}))^2)$$
$$\leq \sup_{\beta_n \in \mathcal{A}_n} P_{\beta_n}(\chi^2_{|\zeta|} \geq \eta\epsilon^2 m_n/16 \times n/(\sqrt{m_n} + \sqrt{n} + o(\sqrt{n}))^2) \leq \sup_{\beta_n \in \mathcal{A}_n} P_{\beta_n}(\chi^2_{|\zeta|} \geq (1-\delta)\eta\epsilon^2 m_n/16)$$
$$\leq \exp(-2\tilde{c}_4 m_n), \text{ for some constant } \tilde{c}_4 > 0.$$

Proof of Theorem 1:

Proof.

$$\Pi_n(\mathcal{A}_n) = \frac{\int_{\mathcal{A}_n} f(\tilde{y}|\beta_n)\pi_n(\beta_n)}{\int f(\tilde{y}|\beta_n)\pi_n(\beta_n)} = \frac{\int_{\mathcal{A}_n} \frac{f(\tilde{y}|\beta_n)}{f(\tilde{y}|\beta^*_n)}\pi_n(\beta_n)}{\int \frac{f(\tilde{y}|\beta_n)}{f(\tilde{y}|\beta^*_n)}\pi_n(\beta_n)} = \frac{\mathcal{N}_{1,n}}{\mathcal{N}_{2,n}} \leq \kappa_n + (1-\kappa_n)\frac{\mathcal{N}_{1,n}}{\mathcal{N}_{2,n}}, \tag{A2}$$

where κ_n is the sequence of tests given in Lemma A2. Note that

$$P_{\beta^*_n}(\kappa_n > \exp(-\tilde{c}_3 m_n/2)) \leq E_{\beta^*_n}(\kappa_n)\exp(\tilde{c}_3 m_n/2) \leq \exp(-\tilde{c}_3 m_n/2).$$

Therefore $\sum_{n=1}^{\infty} P_{\beta^*_n}(\kappa_n > \exp(-\tilde{c}_3 m_n/2)) < \infty$. Applying Borel-Cantelli lemma $P_{\beta^*_n}(\kappa_n > \exp(-\tilde{c}_3 m_n/2)$ infinitely often$) = 0$. Thus,

$$\kappa_n \to 0 \quad a.s. \tag{A3}$$

$$E_{\beta^*_n}((1-\kappa_n)\mathcal{N}_{1,n}) = \int (1-\kappa_n) \int_{\mathcal{A}_n} \frac{f(\tilde{y}|\beta_n)}{f(\tilde{y}|\beta^*_n)} \pi_n(\beta_n) f(\tilde{y}|\beta^*_n)$$
$$= \int_{\mathcal{A}_n} \int (1-\kappa_n) f(\tilde{y}|\beta_n)\pi_n(\beta_n) \leq \sup_{\beta_n \in \mathcal{A}_n} E_{\beta_n}(1-\kappa_n) \leq \exp(-2\tilde{c}_4 m_n).$$

Applying Borel-Cantelli lemma, $P_{\beta^*_n}((1-\kappa_n)\mathcal{N}_{1,n}\exp(m_n\tilde{c}_4) > \exp(-m_n\tilde{c}_4/2)$ infinitely often$) = 0$ so

$$\exp(m_n\tilde{c}_4)(1-\kappa_n)\mathcal{N}_{1,n} \to 0 \quad a.s.. \tag{A4}$$

Note that $\mathcal{N}_{2,n} = \int \frac{f(\tilde{y}|\beta_n)}{f(\tilde{y}|\beta^*_n)}\pi_n(\beta_n)$. Consider the set $\mathcal{H}_n = \left\{\beta_n : \frac{1}{m_n}\log\left[\frac{f(\tilde{y}|\beta^*_n)}{f(\tilde{y}|\beta_n)}\right] < \upsilon\right\}$, for $\upsilon = \tilde{c}_4/2$.

$$\exp(\tilde{c}_4 m_n)\mathcal{N}_{2,n} \geq \exp(\tilde{c}_4 m_n) \int_{\mathcal{H}_n} \exp\left(-m_n \frac{1}{m_n}\log\frac{f(\tilde{y}|\beta^*_n)}{f(\tilde{y}|\beta_n)}\right)\pi_n(\beta_n) \geq \exp((\tilde{c}_4 - \tilde{c}_4/2)m_n)\Pi_n(\mathcal{H}_n).$$

In view of (A2), (A3) and (A4), it is enough to show that $-\log(\Pi_n(\mathcal{H}_n)) = o(m_n)$.

With little algebra, we obtain

$$\frac{1}{m_n}\log\left[\frac{f(\tilde{y}|\beta^*_n)}{f(\tilde{y}|\beta_n)}\right] = \left[-(\tilde{y} - \tilde{X}\beta^*_n)'(\Phi_n\Phi'_n)^{-1}(\tilde{y} - \tilde{X}\beta^*_n) + ||\tilde{y} - \tilde{X}\beta_n||^2 - \log|\Phi_n\Phi'_n|\right]/(2m_n).$$

Notably, $-\log|\Phi_n\Phi'_n|/m_n \geq \log(n/(\sqrt{n} + \sqrt{m_n} + o(\sqrt{n}))^2) \to 0$ (by Lemma A1), as $n \to \infty$. Also,

$$P_{\beta^*_n}(\tilde{y} : (\tilde{y} - \tilde{X}\beta^*_n)'(I - (\Phi_n\Phi'_n)^{-1})(\tilde{y} - \tilde{X}\beta^*_n) > m_n/\log(n))$$
$$\leq E_{\beta^*_n}[\log(n)(\tilde{y} - \tilde{X}\beta^*_n)'(I - (\Phi_n\Phi'_n)^{-1})(\tilde{y} - \tilde{X}\beta^*_n)/m_n]$$
$$= Trace(\Phi_n\Phi'_n - I)\log(n)/m_n \leq ||\Phi_n\Phi'_n - I||_2 \log(n)/m_n \leq \log(n)/\sqrt{nm_n},$$

where the last inequality follows due to Assumption (B). Using Assumption (D), $\sum_{n=1}^{\infty} \log(n)/\sqrt{nm_n} \leq \sum_{n=1}^{\infty} \log(n)/n^{1+\tilde{\delta}} < \infty$. Hence by Borel-Cantelli lemma, $(\tilde{y} - \tilde{X}\beta_n^*)'(I - (\Phi_n\Phi_n')^{-1})(\tilde{y} - \tilde{X}\beta_n^*)/m_n \to 0$ a.s., as $n \to \infty$.

The above two results jointly imply $\mathcal{H}_n \supset \{\beta_n : -||\tilde{y} - \tilde{X}\beta_n^*||^2 + ||\tilde{y} - \tilde{X}\beta_n||^2 \leq m_n\tilde{c}_6\}$, for some constant $\tilde{c}_6 > 0$. Now use $\{\beta_n : -||\tilde{y} - \tilde{X}\beta_n^*||^2 + ||\tilde{y} - \tilde{X}\beta_n||^2 \leq m_n\tilde{c}_6\} \supset \{\beta_n : ||\beta_n - \beta_n^*||_1 < \tilde{c}_7\}$, for some constant $\tilde{c}_7 > 0$. Also, $\{\beta_n : ||\beta_n - \beta_n^*||_1 < \tilde{c}_7\} \supset \{|\beta_{j,n}| \leq \tilde{c}_7/p_n,\ \forall\, j \notin \xi^*\} \cap \{|\beta_{j,n} - \beta_{j,n}^*| \leq \tilde{c}_7/s_n\ \forall\, j \in \xi^*\}$. Now, $\pi_n(|\beta_{j,n}| \leq \tilde{c}_7/p_n,\ \forall\, j \notin \xi^*) \geq \prod_{j \notin \xi^*} \pi_n(|\beta_{j,n}| \leq a_n) \geq (1 - p_n^{-1-u})^{p_n} \to 1$, as $n \to \infty$. Here the first inequality follows as $a_n = \sqrt{s_n \log(p_n)/m_n}/p_n$ and $s_n \log(p_n)/m_n \to 0$. The second inequality follows by Assumption (G). On the other hand, $\pi_n(|\beta_{j,n} - \beta_{j,n}^*| \leq \tilde{c}_7/s_n,\ \forall\, j \in \xi^*) \geq (2\tilde{c}_7/s_n \inf_{[-M_n, M_n]} h_\mu(x))^{s_n}$, which holds for all large n as $|\beta_{j,n}^*| < M_n/2$ and $\tilde{c}_7/s_n \to 0$ as $n \to \infty$. Thus, $-\log(\pi_n(|\beta_{j,n} - \beta_{j,n}^*| \leq \tilde{c}_7/s_n,\ \forall\, j \in \xi^*)) \leq O(s_n \log(p_n)) = o(m_n)$, by Assumptions (C) and (H). Hence, $-\log(\pi_n(\mathcal{H}_n) = o(m_n)$. □

REFERENCES

ARMAGAN, A., DUNSON, D. B. & LEE, J. (2013). Generalized double Pareto shrinkage. *Statistica Sinica* **23**, 119–143.

BHATTACHARYA, A., CHAKRABORTY, A. & MALLICK, B. K. (2016). Fast sampling with gaussian scale mixture priors in high-dimensional regression. *Biometrika* , asw042.

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

CARON, F. & DOUCET, A. (2008). Sparse bayesian nonparametric regression. In *Proceedings of the 25th international conference on Machine learning*.

CARVALHO, C. M., POLSON, N. G. & SCOTT, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika* **97**, 465–480.

CASTILLO, I., SCHMIDT-HIEBER, J., VAN DER VAART, A. et al. (2015). Bayesian linear regression with sparse priors. *The Annals of Statistics* **43**, 1986–2018.

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

ELDAR, Y. C. & KUTYNIOK, G. (2012). *Compressed sensing: theory and applications*. Cambridge university press.

GELFAND, A. E. & GHOSH, S. K. (1998). Model choice: a minimum posterior predictive loss approach. *Biometrika* **85**, 1–11.

GEORGE, E. I. & MCCULLOCH, R. E. (1997). Approaches for bayesian variable selection. *Statistica sinica* , 339–373.

GOLUB, G. H. & VAN LOAN, C. F. (2012). *Matrix computations*, vol. 3. JHU press.

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

GUHANIYOGI, R. & DUNSON, D. B. (2015). Bayesian compressed regression. *Journal of the American Statistical Association* **110**, 1500–1514.

JAVANMARD, A. & MONTANARI, A. (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *The Journal of Machine Learning Research* **15**, 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**, 1487–1511.

LIU, K., KARGUPTA, H. & RYAN, J. (2005). Random projection-based multiplicative data perturbation for privacy preserving distributed data mining. *IEEE Transactions on knowledge and Data Engineering* **18**, 92–106.

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

RUE, H. (2001). Fast sampling of gaussian markov random fields. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **63**, 325–338.

SCOTT, J. G. & BERGER, J. O. (2010). Bayes and empirical-bayes multiplicity adjustment in the variable-selection problem. *The Annals of Statistics* , 2587–2619.

SONG, Q. & LIANG, F. (2017). Nearly optimal bayesian shrinkage for high dimensional regression. *arXiv preprint arXiv:1712.08964* .

SZABÓ, B., VAN DER VAART, A. W., VAN ZANTEN, J. et al. (2015). Frequentist coverage of adaptive nonparametric bayesian credible sets. *The Annals of Statistics* **43**, 1391–1428.

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* **58**, 267–288.

TING, D., FIENBERG, S. E. & TROTTINI, M. (2008). Random orthogonal matrix masking methodology for micro-data release. *International Journal of Information and Computer Security* **2**, 86–105.

VAART, A. V. D. & ZANTEN, H. V. (2011). Information rates of nonparametric gaussian process methods. *Journal of Machine Learning Research* **12**, 2095–2119.

VAN DE GEER, S., BÜHLMANN, P., RITOV, Y., DEZEURE, R. et al. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics* **42**, 1166–1202.

VAN DER VAART, A. W., VAN ZANTEN, J. H. et al. (2009). Adaptive bayesian estimation using a gaussian random field with inverse gamma bandwidth. *The Annals of Statistics* **37**, 2655–2675.

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

ZHANG, C.-H. & 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**, 217–242.

ZHAO, L. & CHEN, L. (2019). On the privacy of matrix masking-based verifiable (outsourced) computation. *IEEE Transactions on Cloud Computing* .

ZHOU, S., WASSERMAN, L. & LAFFERTY, J. D. (2008). Compressed regression. In *Advances in Neural Information Processing Systems*.

[*Received on* 2 *January* 2017. *Editorial decision on* 1 *April* 2017]