Bayesian Analysis (2018) **13**, Number 4, pp. 1011–1035

Big Data Bayesian Linear Regression and Variable Selection by Normal-Inverse-Gamma Summation

Hang Qian*

Abstract. We introduce the normal-inverse-gamma summation operator, which combines Bayesian regression results from different data sources and leads to a simple split-and-merge algorithm for big data regressions. The summation operator is also useful for computing the marginal likelihood and facilitates Bayesian model selection methods, including Bayesian LASSO, stochastic search variable selection, Markov chain Monte Carlo model composition, etc. Observations are scanned in one pass and then the sampler iteratively combines normal-inverse-gamma distributions without reloading the data. Simulation studies demonstrate that our algorithms can efficiently handle highly correlated big data. A real-world data set on employment and wage is also analyzed.

MSC 2010 subject classifications: Primary 62E15, 62J07; secondary 68W10.

Keywords: conjugate prior, hierarchical shrinkage, MapReduce.

1 Introduction

Advances in computation and storage technologies have led to exponential growth of data volume, velocity and variety. Data scientists face the challenge of scaling up the existing Bayesian inference methods when the data are too large to be processed by a single machine. It is natural to split the big data and merge the results obtained from a plurality of data sources. Scott et al. (2016) propose the consensus Monte Carlo, in which data are partitioned among multiple machines and the data-distributed results are averaged by appropriate weights. Neiswanger et al. (2014) show that combination of subsample results obtained from parallel simulation is asymptotically consistent, and they also bound the rate of convergence. For linear regression models, Miroshnikov et al. (2015) resort to the summary statistics as the input to Markov chain Monte Carlo (MCMC) simulation. Ordonez et al. (2014) discuss the generalized sufficient statistics in stochastic search variable selection (SSVS) by George and McCulloch (1993). Ghosh and Reiter (2013) study the partitioned data for secure Bayesian model averaging.

Conjugate Bayesian linear regressions are discussed in almost every Bayesian textbook, but we study the normal-inverse-gamma (NIG) distributions from a new perspective, namely the NIG summation operator, which effectively combines the data-distributed results and facilitates the marginal likelihood computation. Our approach is applicable to Bayesian linear regressions and a variety of Bayesian hierarchical shrinkage and variable selection methods, including Bayesian ridge regressions, Bayesian least

[*] 55 Centre Street, Natick, MA 01760, USA, matlabist@gmail.com

https://doi.org/10.1214/17-BA1083

absolute shrinkage and selection operator (LASSO), SSVS, MCMC model composition (MC^3) by Madigan et al. (1995), Bayesian calibration to frequentist Akaike/Bayesian information criteria (AIC/BIC) (George and Foster, 2000), the pseudo-prior method by Carlin and Chib (1995) and a form of reversible jump MCMC (Green, 1995). We put them in a unified framework that involves two stages. First, out-of-memory data are processed by a split-and-merge NIG summation algorithm. Second, in-memory MCMC samplers (or analytic solvers) iteratively combine NIG distributions to learn the posteriors without reloading data.

In this paper, big data refer to the large number of observations (n), while the number of variables (k) is moderate in the sense that it remains feasible to store and manipulate $k \times k$ matrices in memory.[1] The conventional wisdom is that variable selection techniques are mostly ideal for large-k applications. We demonstrate that variable selection methods are also useful when several hundred variables exhibit near-perfect multicollinearity. Our algorithms can efficiently identify the promising predictors for successful out-of-sample forecast.

The remainder the paper is organized as follows. Section 2 reviews Bayesian linear regressions and Section 3 introduces the NIG summation operator and a split-and-merge algorithm for big data regressions. Section 4 applies the NIG summation operator to Bayesian variable selection methods. Section 5 is devoted to simulation studies and computational complexity analysis. Section 6 analyzes a real-world wage data set. Section 7 concludes the paper and suggests promising directions of future research.

2 Bayesian Linear Regression

Consider the multiple linear regression model:

$$Y = X\beta + \sigma\varepsilon, \tag{1}$$

where Y is a $n \times 1$ response vector, X is a $n \times k$ predictor matrix and ε is a vector of independent standard normal disturbances.

Definition 1. *The k-dimensional regression coefficients and the disturbance variance (β, σ^2) follow the distribution $NIG(\mu, \Lambda, a, b)$ if*

$$p\left(\beta, \sigma^2\right) \propto \left(\sigma^2\right)^{-\left(a+\frac{k}{2}+1\right)} e^{-\sigma^{-2}\left[b+\frac{1}{2}(\beta-\mu)'\Lambda(\beta-\mu)\right]}.$$

Also, (β, σ^2) is said to have a non-informative prior, denoted by $NIG(0_k, 0_{k\times k}, -\frac{k}{2}, 0)$, if $p(\beta, \sigma^2) \propto \sigma^{-2}$.

For notational convenience, we parameterize the NIG distribution by the precision matrix Λ. That is, $NIG(\mu, \Lambda, a, b)$ indicates that $p(\beta|\sigma^2)$ is multivariate normal with the mean μ and the precision $\sigma^{-2}\Lambda$. The covariance matrix is $\sigma^2\Lambda^{-1}$.

[1] There is a large body of literature on high-dimensional (i.e., many variables) model selection methods, see Fan and Lv (2010), Johnson and Rossell (2012), Lin et al. (2011), to name a few.

Textbook Bayesian results show that the posterior distributions are analytically tractable under the conjugate and non-informative priors. As we will frequently refer to those results, we state them as Proposition 1 and 2.

Proposition 1. *Under the conjugate prior $NIG(\mu, \Lambda, a, b)$ and n observations X, Y, the posterior distribution is given by $NIG(\overline{\mu}, \overline{\Lambda}, \overline{a}, \overline{b})$, where $\overline{\mu} = (\Lambda + X'X)^{-1}(\Lambda\mu + X'Y)$, $\overline{\Lambda} = \Lambda + X'X$, $\overline{a} = a + \frac{n}{2}$, $\overline{b} = b + \frac{1}{2}Y'Y + \frac{1}{2}\mu'\Lambda\mu - \frac{1}{2}\overline{\mu}'\overline{\Lambda}\overline{\mu}$.*

Proposition 2. *Under the non-informative prior $p(\beta, \sigma^2) \propto \sigma^{-2}$ and n observations X, Y, the posterior distribution is given by $NIG(\widetilde{\mu}, \widetilde{\Lambda}, \widetilde{a}, \widetilde{b})$, where $\widetilde{\mu} = (X'X)^{-1}X'Y$, $\widetilde{\Lambda} = X'X$, $\widetilde{a} = \frac{n-k}{2}$, $\widetilde{b} = \frac{1}{2}Y'Y - \frac{1}{2}Y'X(X'X)^{-1}X'Y$.*

If the prior hyperparameters are calibrated to $(\mu, \Lambda, a, b) = (0_k, 0_{k\times k}, -\frac{k}{2}, 0)$, then the posterior distribution in Proposition 1 reduces to the result shown in Proposition 2. The non-informative prior $p(\beta, \sigma^2) \propto \sigma^{-2}$ is not a proper NIG distribution. As far as the posterior distribution is concerned, we treat the non-informative prior as a special case of the calibrated NIG distribution, hence our notation $NIG(0_k, 0_{k\times k}, -\frac{k}{2}, 0)$.

Proposition 2 has an inverse problem: is it possible to recover observations if we are given $NIG(\widetilde{\mu}, \widetilde{\Lambda}, \widetilde{a}, \widetilde{b})$? Proposition 3 provides a pseudo-observation interpretation of the NIG distribution.

Proposition 3. *For a k-dimensional $NIG(\widetilde{\mu}, \widetilde{\Lambda}, \widetilde{a}, \widetilde{b})$ such that $\widetilde{\Lambda}$ is positive definite, $\widetilde{a} > 0$, $\widetilde{b} > 0$ and $n \equiv 2\widetilde{a} + k$ is a positive integer, let $X_1 = \Lambda^{1/2}$, $Y_1 = \Lambda^{1/2}\widetilde{\mu}$, $X_2 = 0_{(n-k)\times k}$, $Y_2 = \sqrt{\widetilde{b}/\widetilde{a}}\cdot 1_{(n-k)\times 1}$, where $\Lambda^{1/2}$ is the upper triangular Cholesky factor of $\widetilde{\Lambda}$. Then, under the non-informative prior and the pseudo observations $\binom{X_1}{X_2}, \binom{Y_1}{Y_2}$, the posterior distribution is $NIG(\widetilde{\mu}, \widetilde{\Lambda}, \widetilde{a}, \widetilde{b})$. Furthermore, if another data set X^*, Y^* yields the same posterior distribution, the data must satisfy $X^{*\prime}X^* = \widetilde{\Lambda}$, $X^{*\prime}Y^* = \widetilde{\Lambda}\widetilde{\mu}$, $Y^{*\prime}Y^* = 2\widetilde{b} + \widetilde{\mu}'\widetilde{\Lambda}\widetilde{\mu}$.*

Proposition 3 illustrates the observations extracted from the NIG distribution: X_1, Y_1 are informative on both β and σ^2, while X_2, Y_2 are not informative about β, but the non-zero Y_2 can update the distribution of σ^2. The Gaussian likelihood function $p(Y|\beta, \sigma^2) \propto e^{-\frac{1}{2}\sigma^{-2}(Y-X\beta)'(Y-X\beta)}$ can be expressed as a function of $X'X$, $X'Y$, $Y'Y$. Therefore, the pseudo observations are determined up to those sufficient statistics.

3 NIG Summation and Big Data Regression

Motivated by the pseudo-observation interpretation of the conjugate distributions, we define the NIG summation operator. The sum of two NIG distributions can be thought as the posterior distribution induced by concatenation of the observations extracted from two NIG distributions.

Definition 2. *Consider the k-dimensional $NIG(\mu_1, \Lambda_1, a_1, b_1)$ and $NIG(\mu_2, \Lambda_2, a_2, b_2)$. If a distribution $NIG(\mu, \Lambda, a, b)$ satisfies*

1. $\mu = (\Lambda_1 + \Lambda_2)^{-1}(\Lambda_1\mu_1 + \Lambda_2\mu_2)$,

2. $\Lambda = \Lambda_1 + \Lambda_2$,
3. $a = a_1 + a_2 + \frac{k}{2}$,
4. $b = b_1 + b_2 + \frac{1}{2}(\mu_1 - \mu_2)'(\Lambda_1^{-1} + \Lambda_2^{-1})^{-1}(\mu_1 - \mu_2)$,

then it is said to be the sum of two NIG distributions, denoted by

$$NIG\left(\mu, \Lambda, a, b\right) = NIG\left(\mu_1, \Lambda_1, a_1, b_1\right) + NIG\left(\mu_2, \Lambda_2, a_2, b_2\right),$$

or more compactly, $NIG(\mu, \Lambda, a, b) = \sum_{i=1}^{2} NIG(\mu_i, \Lambda_i, a_i, b_i)$.

Occasionally, it is inconvenient to invert Λ_1 or Λ_2. By completing the squares, we can rewrite Rule 4 in Definition 2 as

$$b = b_1 + b_2 + \frac{1}{2}\left(\mu_1 - \mu\right)' \Lambda_1 \left(\mu_1 - \mu\right) + \frac{1}{2}\left(\mu_2 - \mu\right)' \Lambda_2 \left(\mu_2 - \mu\right).$$

Proposition 4. *The NIG summation operator satisfies*

1. *Commutativity:*

$$\begin{aligned} &NIG(\mu_1, \Lambda_1, a_1, b_1) + NIG(\mu_2, \Lambda_2, a_2, b_2) \\ &\quad = NIG(\mu_2, \Lambda_2, a_2, b_2) + NIG(\mu_1, \Lambda_1, a_1, b_1), \end{aligned}$$

2. *Associativity:*

$$\sum_{i=1}^{3} NIG(\mu_i, \Lambda_i, a_i, b_i) = NIG(\mu_1, \Lambda_1, a_1, b_1) + \sum_{i=2}^{3} NIG(\mu_i, \Lambda_i, a_i, b_i),$$

3. *Identity element:*

$$NIG(\mu, \Lambda, a, b) + NIG\left(0_k, 0_{k\times k}, -\frac{k}{2}, 0\right) = NIG(\mu, \Lambda, a, b).$$

Proposition 4 is derived from Definition 2. Note that $NIG(0_k, 0_{k\times k}, -\frac{k}{2}, 0)$ serves as the "zero" in summation. Two use cases of the NIG summation operator are given by Proposition 5 and 6. Proofs of all propositions are provided in Supplementary Material (Qian, 2017).

Proposition 5. *Let the posterior distribution, under the prior* $NIG(\mu, \Lambda, a, b)$, *be* $NIG(\overline{\mu}, \overline{\Lambda}, \overline{a}, \overline{b})$ *by Proposition 1. Let the posterior distribution, under the non-informative prior, be* $NIG(\widetilde{\mu}, \widetilde{\Lambda}, \widetilde{a}, \widetilde{b})$ *by Proposition 2. Then we have*

$$NIG\left(\overline{\mu}, \overline{\Lambda}, \overline{a}, \overline{b}\right) = NIG\left(\mu, \Lambda, a, b\right) + NIG\left(\widetilde{\mu}, \widetilde{\Lambda}, \widetilde{a}, \widetilde{b}\right).$$

Proposition 6. *Let the posterior distribution, under the non-informative prior and full-sample observations, be $NIG(\widetilde{\mu}, \widetilde{\Lambda}, \widetilde{a}, \widetilde{b})$. Split the observations X, Y into m subsets X_i, Y_i, $i = 1, \ldots, m$. Let the subset posterior distribution, under the non-informative prior and the data X_i, Y_i, be $NIG(\widetilde{\mu}_i, \widetilde{\Lambda}_i, \widetilde{a}_i, \widetilde{b}_i)$. Then we have*

$$NIG\left(\widetilde{\mu}, \widetilde{\Lambda}, \widetilde{a}, \widetilde{b}\right) = \sum_{i=1}^{m} NIG\left(\widetilde{\mu}_i, \widetilde{\Lambda}_i, \widetilde{a}_i, \widetilde{b}_i\right).$$

An immediate consequence of Proposition 5 and 6 is a split-and-merge algorithm for big data Bayesian linear regressions.

Algorithm 1. *Consider Bayesian linear regression under the conjugate prior $NIG(\mu, \Lambda, a, b)$ and the observations X, Y, in which the sample size is so large that data cannot be stored and/or processed by a single machine. The posterior distribution $NIG(\overline{\mu}, \overline{\Lambda}, \overline{a}, \overline{b})$ shown in Proposition 1 can be obtained by the following split-and-merge algorithm.*

Step 1 *partition the big data into m subsets X_i, Y_i, $i = 1, \ldots, m$.*

Step 2 *run subset Bayesian linear regressions under the non-informative prior and obtain the subset posterior distributions $NIG(\widetilde{\mu}_i, \widetilde{\Lambda}_i, \widetilde{a}_i, \widetilde{b}_i)$, $i = 1, \ldots, m$, where $\widetilde{\mu}_i, \widetilde{\Lambda}_i, \widetilde{a}_i, \widetilde{b}_i$ are given by Proposition 2 using X_i, Y_i.*

Step 3 *gather and sum up the subset posterior distributions:*

$$NIG\left(\widetilde{\mu}, \widetilde{\Lambda}, \widetilde{a}, \widetilde{b}\right) = \sum_{i=1}^{m} NIG\left(\widetilde{\mu}_i, \widetilde{\Lambda}_i, \widetilde{a}_i, \widetilde{b}_i\right).$$

Step 4 *sum up the prior and the combined subset posterior distributions:*

$$NIG\left(\overline{\mu}, \overline{\Lambda}, \overline{a}, \overline{b}\right) = NIG\left(\mu, \Lambda, a, b\right) + NIG\left(\widetilde{\mu}, \widetilde{\Lambda}, \widetilde{a}, \widetilde{b}\right).$$

Some remarks on implementation of Algorithm 1. First, subset sizes are not necessarily the same. Second, there is no dependency between m subset regressions, which can be processed in an embarrassingly parallel fashion. Third, the NIG summation operator is associative, so Step 3 can be processed recursively by pairs. Fourth, Step 3 tolerates machine failures. If some subset regression fails to add, $\widetilde{\mu}$ reweighs the remaining regressions and the precision $\widetilde{\Lambda}$ decreases without breakdown. Fifth, commutativity and associativity justify online updating for flow data. For example, when new data X_{m+1}, Y_{m+1} come in, we can update the distribution: $NIG(\overline{\mu}, \overline{\Lambda}, \overline{a}, \overline{b}) + NIG(\widetilde{\mu}_{m+1}, \widetilde{\Lambda}_{m+1}, \widetilde{a}_{m+1}, \widetilde{b}_{m+1})$.

An application of Algorithm 1 is the big data ridge regression, which can be interpreted as Bayesian linear regression under the conjugate prior $NIG(0_k, \Lambda, a, b)$, where $\Lambda = \lambda I_{k \times k}$ and λ is the regularization parameter. The posterior distribution is

$$p\left(\beta, \sigma^2 \left| Y\right.\right) \propto \left(\sigma^2\right)^{-\left(a + \frac{n+k}{2} + 1\right)} e^{-\frac{1}{2}\sigma^{-2}\left[2b + (Y - X\beta)'(Y - X\beta) + \lambda\beta'\beta\right]}.$$

The posterior mean and mode of β is given by $\overline{\mu} = (\Lambda + X'X)^{-1}X'Y$, which is the ridge point estimator, regardless of the hyperparameter values for a and b.

4 Bayesian Variable Selection

4.1 Big Data Bayesian LASSO

The LASSO of Tibshirani (1996) is a popular shrinkage and variable selection technique. It has a Bayesian interpretation with the double exponential priors imposed on the regression coefficients. Park and Casella (2008) propose an efficient Gibbs sampler for Bayesian LASSO. Consider the model (1) with the priors:

$$p\left(\beta,\sigma^2\right)\propto\left(\sigma^2\right)^{-\left(a+\frac{k}{2}+1\right)}e^{-b\sigma^{-2}-\sum_{j=1}^{k}\lambda_j\left|\frac{\beta_j}{\sigma}\right|},\tag{2}$$

where $\lambda_1,\ldots,\lambda_k$ are regularization parameters. The posterior density is

$$p\left(\beta,\sigma^2\left|Y\right.\right)\propto\left(\sigma^2\right)^{-\left(a+\frac{n+k}{2}+1\right)}e^{-b\sigma^{-2}-\frac{1}{2}\sigma^{-2}(Y-X\beta)'(Y-X\beta)-\sum_{j=1}^{k}\lambda_j\left|\frac{\beta_j}{\sigma}\right|}.$$

The posterior mode of β coincides with the classic LASSO estimator conditional on $\sigma^2=1$.[2] The double exponential distribution has a scale mixture representation. By augmenting latent variables $\psi=(\psi_1,\ldots,\psi_k)'$ with the prior $p(\psi)=\prod_{j=1}^{k}e^{-\frac{1}{2}\lambda_j^2\psi_j}$, we have the conditional distribution $\beta,\sigma^2|\psi\sim NIG(0_k,\Lambda(\psi),a,b)$, where $\Lambda(\psi)=diag(\psi_1^{-1},\ldots,\psi_k^{-1})$. The posterior conditional distribution of ψ is also tractable:

$$p\left(\psi\left|\beta,\sigma^2,Y\right.\right)\propto\prod_{j=1}^{k}\psi_j^{-\frac{1}{2}}e^{-\frac{1}{2}\lambda_j^2\psi_j-\frac{1}{2}\sigma^{-2}\beta_j^2\psi_j^{-1}},$$

which indicates that $\psi_j^{-1}|\beta,\sigma^2,Y$ follows an inverse Gaussian distribution with parameters $|\frac{\lambda_j\sigma}{\beta_j}|$ and λ_j^2 (see Park and Casella, 2008, p. 682). Therefore, Bayesian LASSO can be handled by a Gibbs sampler that alternately samples from the inverse Gaussian distributions and the posterior NIG distributions. By Proposition 5 and 6, an efficient algorithm with one pass of big data is summarized below.

Algorithm 2. *Consider Bayesian LASSO regression (1) and (2) with the big data X,Y. We have the following Gibbs sampling algorithm:*

Step 1–3 *the same as those in Algorithm 1. After Step 3, save $NIG(\widetilde{\mu},\widetilde{\Lambda},\widetilde{a},\widetilde{b})$ only.*

Step 4 *iterate the sub-steps until (β,σ^2,ψ) draws converge to stationary distributions:*

- **4.1** *given the current draw of ψ, construct $\Lambda(\psi)=diag(\psi_1^{-1},\ldots,\psi_k^{-1})$; compute $NIG(\overline{\mu},\overline{\Lambda},\overline{a},\overline{b})=NIG(0_k,\Lambda(\psi),a,b)+NIG(\widetilde{\mu},\widetilde{\Lambda},\widetilde{a},\widetilde{b})$; generate a draw for (β,σ^2) from $NIG(\overline{\mu},\overline{\Lambda},\overline{a},\overline{b})$.*
- **4.2** *given the current draw of (β,σ^2), generate a draw for each ψ_j^{-1} from the inverse Gaussian distribution with parameters $|\frac{\lambda_j\sigma}{\beta_j}|$ and λ_j^2, $j=1,\ldots,k$.*

[2] The posterior mode matching the frequentist LASSO estimator without conditioning on σ^2 can be achieved by an alternative prior $p(\beta,\sigma^2)\propto(\sigma^2)^{-(a+\frac{k}{2}+1)}e^{-b\sigma^{-2}-\sigma^{-2}\sum_{j=1}^{k}\lambda_j|\beta_j|}$. Park and Casella (2008) note that the posterior density is concave and unimodal under (2), hence a preferable form.

4.2 Big Data SSVS

George and McCulloch (1993) develop a variable selection method in which the regression coefficients have a hierarchical mixture prior that involves small and large variance terms. The posterior mixture probabilities identify the promising subset of predictors.

Let $\gamma \equiv (\gamma_1, \ldots, \gamma_k)'$, where $\gamma_j \in \{0, 1\}$ is the variable selection indicator for the j^{th} predictor. For most applications, it is appropriate to adopt an independent prior:

$$p\left(\gamma\right) = \prod_{j=1}^{k} w_j^{\gamma_j} \left(1 - w_j\right)^{1-\gamma_j} . \tag{3}$$

The hyperparameter w_j is the prior probability of variable inclusion. To exempt a variable (e.g., the intercept term) from exclusion, we can put $w_j = 1$. Following George and Mcculloch (1997), we assume a conditionally conjugate prior for (β, σ^2):

$$\beta, \sigma^2 \left|\gamma \sim NIG\left(0_k, \Lambda\left(\gamma\right), a, b\right), \right. \tag{4}$$

where $\Lambda(\gamma) = [diag(D^{\gamma_1} d^{1-\gamma_1}, \ldots, D^{\gamma_k} d^{1-\gamma_k})]^{-1}$ and the hyperparameters satisfy $D > d > 0$. By marginalizing γ, we see that $\beta_j|\sigma^2$ follows a zero-mean Gaussian mixture distribution: one component has a large variance $\sigma^2 D$, and the other has a small variance $\sigma^2 d$. Furthermore, $\beta|\sigma^2$ has a 2^k-component mixture prior, which uses observations to assign larger posterior probabilities to the more promising components. The crux of SSVS is a Gibbs sampler that iteratively generates draws from $p(\beta, \sigma^2|\gamma, Y)$ and $p(\gamma|\beta, \sigma^2, Y)$, which circumvents the overwhelming problem of calculating the posterior probabilities for all 2^k subsets of predictors.

We note that the hierarchical prior structure in SSVS is similar to that in Bayesian LASSO. Both are Gaussian mixture models that involve latent variables (i.e., the discrete γ in SSVS and the continuous ψ in LASSO), and both contain a conjugate Bayesian linear sub-structure conditional on the latent variables. Proposition 5 and 6 suggest a big data SSVS sampler analogous to Algorithm 2.

Algorithm 3. *Given the prior (3) and (4), the likelihood (1) and the big data X, Y, we have the following SSVS Gibbs sampling algorithm:*

Step 1–3 *the same as those in Algorithm 1. After Step 3, save $NIG(\widetilde{\mu}, \widetilde{\Lambda}, \widetilde{a}, \widetilde{b})$ only.*

Step 4 *iterate the sub-steps until $(\beta, \sigma^2, \gamma)$ draws converge to stationary distributions:*

- **4.1** *given the current draw of γ, construct $\Lambda(\gamma) = [diag(D^{\gamma_1} d^{1-\gamma_1}, \ldots, D^{\gamma_k} d^{1-\gamma_k})]^{-1}$; compute $NIG(\overline{\mu}, \overline{\Lambda}, \overline{a}, \overline{b}) = NIG(0_k, \Lambda(\gamma), a, b) + NIG(\widetilde{\mu}, \widetilde{\Lambda}, \widetilde{a}, \widetilde{b})$; generate a draw for (β, σ^2) from $NIG(\overline{\mu}, \overline{\Lambda}, \overline{a}, \overline{b})$.*
- **4.2** *given the current draw of (β, σ^2), generate a draw for each γ_j from the Bernoulli distribution with the probability $\frac{w_j \phi(\beta_j; 0, \sigma^2 D)}{w_j \phi(\beta_j; 0, \sigma^2 D) + (1-w_j) \phi(\beta_j; 0, \sigma^2 d)}$, where $\phi(\cdot)$ denotes the normal density.*

4.3 Big Data MC^3

SSVS does not strictly exclude variables from the regression. The prior variance d must be small but positive, or Algorithm 3 yields a reducible Markov chain that violates the MCMC convergence conditions. MC^3 proposed by Madigan et al. (1995) and Raftery et al. (1997) can handle variable selection in the presence of degenerate distributions, which are analytically integrated out of the posterior distributions. Similar marginalization techniques are discussed in Geweke (1996) and Smith and Kohn (1996). In this subsection, we adapt MC^3 for big data applications.

Consider (4) with $d = 0$. Given a particular γ, we have the partition $\beta = \left(\begin{smallmatrix} [c]c\beta_\gamma \\ \beta_o \end{smallmatrix} \right)$, where β_γ (and β_o) represents the coefficients of the predictors included in (and excluded from) the regression. Then (4) can be decomposed as

$$p\left(\beta, \sigma^2 \left|\gamma\right.\right) = p\left(\beta_\gamma, \sigma^2 \left|\gamma\right.\right) p\left(\beta_o \left|\gamma, \beta_\gamma, \sigma^2\right.\right),$$

where $p(\beta_o|\gamma, \beta_\gamma, \sigma^2)$ is the Dirac delta with a spike at zero, and

$$\beta_\gamma, \sigma^2 \left|\gamma\right. \sim NIG\left(\mu_\gamma, \Lambda_{\gamma\gamma}, a, b\right). \tag{5}$$

By (4), the hyperparameters μ_γ is a vector of zeros and $\Lambda_{\gamma\gamma}$ equals D^{-1} times an identity matrix. In practice, $\mu_\gamma, \Lambda_{\gamma\gamma}$ can take a more general form (see below). Since β_o is essentially a vector of zeros, the model (1) is reduced to

$$Y = X_\gamma \beta_\gamma + \sigma\varepsilon, \tag{6}$$

where X_γ is a sub-matrix of X with columns selected by γ.

For variable selection, we are mostly interested in $p(\gamma|Y)$, which is proportional to the product of the prior probability $p(\gamma)$ and the marginal likelihood $p(Y|\gamma)$. Note that (5) and (6) constitute a conjugate regression with a subset of variables. Employing the NIG summation operator, we evaluate the marginal likelihood by the Bayes formula:

$$p\left(Y \left|\gamma\right.\right) = \frac{p\left(\beta_\gamma, \sigma^2 \left|\gamma\right.\right) p\left(Y \left|\beta_\gamma, \sigma^2, \gamma\right.\right)}{p\left(\beta_\gamma, \sigma^2 \left|Y, \gamma\right.\right)}, \tag{7}$$

where $p(\beta_\gamma, \sigma^2|\gamma)$ is the prior density of $NIG(\mu_\gamma, \Lambda_{\gamma\gamma}, a, b)$, while $p(\beta_\gamma, \sigma^2|Y, \gamma)$ is the posterior density of $NIG(\overline{\mu}_\gamma, \overline{\Lambda}_{\gamma\gamma}, \overline{a}_\gamma, \overline{b}_\gamma)$ obtained by Proposition 5 using the data X_γ, Y. Note that (7) holds for all β_γ and σ^2 values. We pick $\beta_\gamma = 0$ and an arbitrary σ^2 so that the likelihood function $p(Y|\beta_\gamma, \sigma^2, \gamma)$ remains a constant for all γ. The implication is that the term $p(Y|\beta_\gamma, \sigma^2, \gamma)$ can be dropped if we just evaluate (7) up to a proportionality constant.

If the sample size were small, MC^3 would be straightforward once the marginal likelihood had been evaluated by (7). However, it is too costly to reload the big data X_γ, Y in each MCMC iteration. The following dimension-reduction propositions resolve that problem.

Proposition 7. *Let $(\beta, \sigma^2) \sim NIG(\widetilde{\mu}, \widetilde{\Lambda}, \widetilde{a}, \widetilde{b})$. Let γ be the selection vector and $o = 1-\gamma$ be the complement selection vector. Partition $\beta = \left(\begin{smallmatrix} \beta_\gamma \\ \beta_o \end{smallmatrix} \right)$, $\widetilde{\mu} = \left(\begin{smallmatrix} \underline{\mu}_\gamma \\ \underline{\mu}_o \end{smallmatrix} \right)$, $\widetilde{\Lambda} = \left(\begin{smallmatrix} \underline{\Lambda}_{\gamma\gamma} & \underline{\Lambda}_{\gamma o} \\ \underline{\Lambda}_{o\gamma} & \underline{\Lambda}_{oo} \end{smallmatrix} \right)$, where*

β_γ, β_o are sub-vectors of β selected by γ, o, respectively. Assume their lengths are k_γ and k_o. Other sub-vectors/sub-matrices are defined similarly. Given $\beta_o = 0$, the conditional distribution is given by $\beta_\gamma, \sigma^2|\beta_o \sim NIG(\widetilde{\mu}_\gamma, \widetilde{\Lambda}_{\gamma\gamma}, \widetilde{a}_\gamma, \widetilde{b}_\gamma)$, where $\widetilde{\mu}_\gamma = \underline{\mu}_\gamma + \underline{\Lambda}_{\gamma\gamma}^{-1}\underline{\Lambda}_{\gamma o}\underline{\mu}_o$, $\widetilde{\Lambda}_{\gamma\gamma} = \underline{\Lambda}_{\gamma\gamma}$, $\widetilde{a}_\gamma = \widetilde{a} + \frac{k_o}{2}$, $\widetilde{b}_\gamma = \widetilde{b} + \frac{1}{2}\underline{\mu}_o'(\underline{\Lambda}_{oo} - \underline{\Lambda}_{o\gamma}\underline{\Lambda}_{\gamma\gamma}^{-1}\underline{\Lambda}_{\gamma o})\underline{\mu}_o$.

Proposition 8. *Let the k-dimensional $NIG(\widetilde{\mu}, \widetilde{\Lambda}, \widetilde{a}, \widetilde{b})$ be the posterior distribution obtained by the full data X, Y under the non-informative prior. Let the k_γ-dimensional $NIG(\widetilde{\mu}_\gamma, \widetilde{\Lambda}_{\gamma\gamma}, \widetilde{a}_\gamma, \widetilde{b}_\gamma)$ be the posterior distribution obtained by the data X_γ, Y under the non-informative prior. Then $NIG(\widetilde{\mu}_\gamma, \widetilde{\Lambda}_{\gamma\gamma}, \widetilde{a}_\gamma, \widetilde{b}_\gamma)$ is given by Proposition 7.*

Proposition 7 and 8 are useful for Bayesian linear regressions with a subset of predictors. First, by Proposition 2 we obtain $NIG(\widetilde{\mu}, \widetilde{\Lambda}, \widetilde{a}, \widetilde{b})$. Second, by Proposition 7 and 8 we recover $NIG(\widetilde{\mu}_\gamma, \widetilde{\Lambda}_{\gamma\gamma}, \widetilde{a}_\gamma, \widetilde{b}_\gamma)$. Third, by Proposition 5 we insert the prior information by NIG summation: $NIG(\overline{\mu}_\gamma, \overline{\Lambda}_{\gamma\gamma}, \overline{a}_\gamma, \overline{b}_\gamma) = NIG(\mu_\gamma, \Lambda_{\gamma\gamma}, a, b) + NIG(\widetilde{\mu}_\gamma, \widetilde{\Lambda}_{\gamma\gamma}, \widetilde{a}_\gamma, \widetilde{b}_\gamma)$. Fourth, we use (7) for evaluating $p(Y|\gamma)$ and $p(\gamma|Y)$ up to a proportionality constant.

It is seldom feasible to enumerate $p(\gamma|Y)$ for all 2^k scenarios. We resort to MC^3 by a random-walk Metropolis-Hastings sampler. Following Raftery et al. (1997), we define a neighborhood $nbd(\gamma)$ which consists of γ itself and the sets of models that select either one variable more or fewer than γ. Loosely speaking, we randomly change one element of γ as the proposal draw. We summarize the big data MC^3 algorithm in Algorithm 4.

Algorithm 4. *Given the prior (3) and (5), the model (6) and the big data X, Y, we have the following MC^3 algorithm:*

Step 1–3 *the same as those in Algorithm 1. After Step 3, save $NIG(\widetilde{\mu}, \widetilde{\Lambda}, \widetilde{a}, \widetilde{b})$ only.*

Step 4 *iterate the sub-steps until draws of γ converge to stationary distributions:*

4.1 *let the current draw be γ^*. Propose a new draw γ from $nbd(\gamma^*)$ by randomly changing one element of γ^*.*

4.2 *reduce $NIG(\widetilde{\mu}, \widetilde{\Lambda}, \widetilde{a}, \widetilde{b})$ to $NIG(\widetilde{\mu}_\gamma, \widetilde{\Lambda}_{\gamma\gamma}, \widetilde{a}_\gamma, \widetilde{b}_\gamma)$ by Proposition 7 and 8.*

4.3 *insert prior information by NIG summation:*
$NIG(\overline{\mu}_\gamma, \overline{\Lambda}_{\gamma\gamma}, \overline{a}_\gamma, \overline{b}_\gamma) = NIG(\mu_\gamma, \Lambda_{\gamma\gamma}, a, b) + NIG(\widetilde{\mu}_\gamma, \widetilde{\Lambda}_{\gamma\gamma}, \widetilde{a}_\gamma, \widetilde{b}_\gamma)$.

4.4 *evaluate the marginal likelihood up to a proportionality constant: $p(Y|\gamma) \propto \frac{p(\beta_\gamma, \sigma^2|\gamma)}{p(\beta_\gamma, \sigma^2|Y, \gamma)}$, where $p(\beta_\gamma, \sigma^2|\gamma)$ is the density of $NIG(\mu_\gamma, \Lambda_{\gamma\gamma}, a, b)$, and $p(\beta_\gamma, \sigma^2|Y, \gamma)$ is the density of $NIG(\overline{\mu}_\gamma, \overline{\Lambda}_{\gamma\gamma}, \overline{a}_\gamma, \overline{b}_\gamma)$. Both densities should be evaluated at $\beta_\gamma = 0$ and a fixed σ^2.*

4.5 *evaluate the posterior model probability up to a proportionality constant: $p(\gamma|Y) \propto p(\gamma)p(Y|\gamma)$. Meanwhile, $p(\gamma^*|Y)$ has been computed previously.*

4.6 *accept γ with the probability $\min[1, \frac{p(\gamma|Y)}{p(\gamma^*|Y)}]$. Upon acceptance, reset γ as the current model. Otherwise, γ^* remains the current model. The proportionality constants in 4.4 and 4.5 do not affect the acceptance probability.*

An interesting application of Algorithm 4 is Bayesian calibration to frequentist information criteria for model selection. The prior precision matrix $\Lambda_{\gamma\gamma}$ in (5) is not necessarily a diagonal matrix. Instead, we may adopt a g-prior (Zeller, 1986) such that $\Lambda_{\gamma\gamma} = c^{-1}X'_\gamma X_\gamma$. George and Foster (2000) show that model selection by the information criteria such as AIC and BIC corresponds to selection of maximum posterior models by calibrating the prior hyperparameters. Specifically, if the hyperparameter c and the probability w are calibrated such that $\frac{1+c}{c}[2\ln\frac{1-w}{w} + \ln(1+c)] = 2$, then the highest posterior model coincides with the model that minimizes AIC. Furthermore, if $\frac{1+c}{c}[2\ln\frac{1-w}{w} + \ln(1+c)] = \ln n$, then the highest posterior model minimizes BIC.[3] Such correspondence implies that Algorithm 4 is also a big data algorithm for frequentist model selection by the information criteria.

4.4 Other Bayesian Variable Selection Techniques

In addition to Bayesian LASSO, SSVS and MC^3, various Bayesian variable selection techniques are proposed in the literature, such as the indicator-in-regression method by Kuo and Mallick (1998), the pseudo-prior settings by Carlin and Chib (1995), reversible jump MCMC (RJMCMC) by Green (1995) and the composite space approach by Godsill (2001). We review those techniques and propose a general-purpose algorithm for big data applications.

Kuo and Mallick (1998) introduce an indicator-in-regression method. Consider (1) and decompose β such that $\beta_j = \widehat{\beta}_j\gamma_j$, $j = 1,\ldots,k$. Given a particular γ, we have the selected $\widehat{\beta}_\gamma$ and the unselected $\widehat{\beta}_o$ depending on γ_j equals to 1 or 0. The priors are given by $\widehat{\beta}_\gamma, \sigma^2|\gamma \sim NIG(\mu_\gamma, \Lambda_{\gamma\gamma}, a, b)$, $\widehat{\beta}_o|\gamma, \widehat{\beta}_\gamma, \sigma^2 \sim N(\mu_o, V_o)$, where $N(\mu_o, V_o)$ is a pseudo prior (Carlin and Chib, 1995) that has no impact on the likelihood, and μ_o, V_o are tuning parameters for efficient MCMC operation. The posterior conditional distribution $p(\widehat{\beta}_\gamma, \sigma^2|Y,\gamma)$ is the density of $NIG(\overline{\mu}_\gamma, \overline{\Lambda}_{\gamma\gamma}, \overline{a}_\gamma, \overline{b}_\gamma)$ (refer to Step 4.3 in Algorithm 4), while $p(\widehat{\beta}_o|Y,\gamma,\widehat{\beta}_\gamma,\sigma^2)$ remains the pseudo prior density.

Godsill (2001) further explores the idea of pseudo priors and demonstrates that RJMCMC, a powerful trans-dimensional model selection approach by Green (1995), can be derived from a composite model with fixed dimension. Each γ defines a model with parameters $\theta_\gamma \equiv (\beta_\gamma, \sigma^2)$. Conditional on γ, the likelihood depends on θ_γ, but not on the unused parameters denoted by $\theta_{-\gamma}$. Let $\theta = (\theta_0, \ldots, \theta_{2^k-1})$.[4] The priors are

$$p\left(\gamma,\theta\right) = p\left(\gamma\right) p\left(\theta_\gamma\left|\gamma\right.\right) p\left(\theta_{-\gamma}\left|\gamma,\theta_\gamma\right.\right),$$

where $p(\theta_\gamma|\gamma)$ is the genuine prior given by (5), while $p(\theta_{-\gamma}|\gamma,\theta_\gamma)$ is the pseudo prior. Carlin and Chib (1995) resort to the Gibbs sampler, in which the posterior conditional

[3] The calibration results hold exactly under a known σ^2, which may be replaced by an estimate $\widehat{\sigma}^2$ such as the mean squared residuals. The mean and variance of $IG(a,b)$ are $\frac{b}{a-1}$, $\frac{b^2}{(a-1)^2(a-2)}$, respectively. If the prior hyperparameters are calibrated such that b, a are large and their ratio approximately equals $\widehat{\sigma}^2$, the prior density is concentrated around $\widehat{\sigma}^2$ and the variance tends to zero. By Proposition 1 and 2, the posterior density for σ^2 is close to the prior density. Therefore, we obtain an approximation to the regression results with σ^2 fixed at $\widehat{\sigma}^2$.

[4] For notational convenience, θ_γ is indexed by the decimal representation of $k \times 1$ binary vector γ. For example, if $k = 4$, the vector $\gamma = (1,0,0,1)'$ corresponds to the decimal number 9, hence θ_9.

distributions are given by

$$\begin{aligned}&p\left(\gamma\left|Y,\theta\right.\right)\propto p\left(\gamma\right)p\left(\theta_{\gamma}\left|\gamma\right.\right)p\left(\theta_{-\gamma}\left|\gamma,\theta_{\gamma}\right.\right)p\left(Y\left|\gamma,\theta_{\gamma}\right.\right),\\&p\left(\theta\left|Y,\gamma\right.\right)=p\left(\theta_{\gamma}\left|Y,\gamma\right.\right)p\left(\theta_{-\gamma}\left|Y,\gamma,\theta_{\gamma}\right.\right),\end{aligned}$$

where $p(\theta_\gamma|Y,\gamma)$ is $NIG(\overline{\mu}_\gamma,\overline{\Lambda}_{\gamma\gamma},\overline{a}_\gamma,\overline{b}_\gamma)$ and $p(\theta_{-\gamma}|Y,\gamma,\theta_\gamma)$ remains the pseudo prior. The sampler is computationally intensive in that $p(\theta_{-\gamma}|\gamma,\theta_\gamma)$ involves parameters in 2^k-1 models. Godsill (2001) considers the Metropolis-Hastings sampler with a proposal q from the current state $(\gamma^*,\theta^*_{\gamma^*},\theta^*_{-\gamma^*})$ to a new state $(\gamma,\theta_\gamma.\theta_{-\gamma})$ such that

$$q\left(\gamma^*,\theta^*_{\gamma^*},\theta^*_{-\gamma^*}\to\gamma,\theta_\gamma.\theta_{-\gamma}\right)=q_1\left(\gamma^*\to\gamma\right)q_2\left(\theta^*_{\gamma^*}\to\theta_\gamma\right)p\left(\theta_{-\gamma}\left|\gamma,\theta_\gamma\right.\right),$$

where q_1,q_2 denote the model and parameter transition, respectively. A new state will be accepted with the probability given by $\min(1,\alpha)$, where

$$\alpha=\frac{p\left(\gamma\right)p\left(\theta_\gamma\left|\gamma\right.\right)p\left(Y\left|\gamma,\theta_\gamma\right.\right)}{p\left(\gamma^*\right)p\left(\theta^*_{\gamma^*}\left|\gamma^*\right.\right)p\left(Y\left|\gamma^*,\theta^*_{\gamma^*}\right.\right)}\cdot\frac{q_1\left(\gamma\to\gamma^*\right)q_2\left(\theta_\gamma\to\theta^*_{\gamma^*}\right)}{q_1\left(\gamma^*\to\gamma\right)q_2\left(\theta^*_{\gamma^*}\to\theta_\gamma\right)}.$$

Since the pseudo priors are canceled in α, drawing $\theta_{-\gamma}$ is conceptual and not performed in practice. This is a form of RJMCMC, in which the "dimension matching" variables, denoted by u^* and u, are chosen such that $(\theta_\gamma,u)=(u^*,\theta^*_{\gamma^*})$ with the Jacobian term being one. Also, if the model transition q_1 is a random change of an element of the current model γ^*, and the parameter transition q_2 is independent to the current parameters $\theta^*_{\gamma^*}$, with θ_γ drawn from $NIG(\overline{\mu}_\gamma,\overline{\Lambda}_{\gamma\gamma},\overline{a}_\gamma,\overline{b}_\gamma)$, then the sampler reduces to a version of MC^3, because $\frac{p(\theta_\gamma|\gamma)p(Y|\gamma,\theta_\gamma)}{q_2(\theta^*_{\gamma^*}\to\theta_\gamma)}$ is the marginal likelihood and α is determined by the Bayes factor of the two models.

To implement any of those variable selection techniques, the MCMC sampler requires at most three data-related inputs: the posterior distribution $p(\theta_\gamma|Y,\gamma)$, the likelihood function $p(Y|\theta_\gamma,\gamma)$ and the marginal likelihood $p(Y|\gamma)$. They are functions of X_γ,Y, which vary as the value of γ updates in MCMC iterations. Recall that Proposition 3 introduces pseudo observations. If they are used in MCMC simulations, then only k, instead of n, observations are cached in memory. Proposition 9 demonstrates that pseudo observations are the perfect substitute for the big data for Bayesian inference.

Proposition 9. *Consider the model (1) with the big data X,Y. Let $NIG(\widetilde{\mu},\widetilde{\Lambda},\widetilde{a},\widetilde{b})$ be the posterior distribution obtained by Proposition 2 under the non-informative prior. Let X_1,Y_1,X_2,Y_2 be the pseudo observations extracted from $NIG(\widetilde{\mu},\widetilde{\Lambda},\widetilde{a},\widetilde{b})$ by Proposition 3. Let γ be the variable selection indicator under which the regression is reduced to (6). For an arbitrary prior on $\theta_\gamma\equiv(\beta_\gamma,\sigma^2)$, the genuine and pseudo observations yield the same posterior, likelihood and marginal likelihood. That is,*

1. $p(\theta_\gamma|Y,\gamma)=p(\theta_\gamma|Y_1,Y_2,\gamma)$,
2. $p(Y|\theta_\gamma,\gamma)=p(Y_1,Y_2|\theta_\gamma,\gamma)$,
3. $p(Y|\gamma)=p(Y_1,Y_2|\gamma)$.

Though the sample size of the pseudo observations is n, only the first k observations X_1, Y_1 contain dense data, while the remaining observations are trivial: the predictors are zeros and the response values are the same. Therefore, we only need to store X_1, Y_1 as well as a scalar element of Y_2 in the computer memory. Proposition 9 leads to a generic big-data algorithm suitable for many Bayesian variable selection methods.

Algorithm 5. *Consider Bayesian variable selection for Gaussian linear regressions with big data X, Y. We have the following general-purpose algorithm:*

Step 1–3 *the same as those in Algorithm 1. After Step 3, save $NIG(\widetilde{\mu}, \widetilde{\Lambda}, \widetilde{a}, \widetilde{b})$ only.*

Step 4 *extract pseudo observations from $NIG(\widetilde{\mu}, \widetilde{\Lambda}, \widetilde{a}, \widetilde{b})$ by Proposition 3.*

Step 5 *treat pseudo observations as if they were the genuine data, and apply variable selection methods (e.g., Bayesian LASSO, SSVS, MC^3, RJMCMC, etc.) designed for in-memory computation.*

Algorithm 5 is not necessarily the most computationally efficient algorithm, but it is the simplest method that addresses the storage and computation burden induced by big data. The original data are scanned only once and then replaced by pseudo observations. The existing variable selection techniques designed for small data inputs are applicable with minimum adaption.

5 Synthetic Data Examples

In this section, we evaluate the performance of Algorithm 1 – 5 by synthetic data. We consider two regimes: 1) a correctly specified model with highly correlated predictors, and 2) a regression with non-Gaussian (skewed and leptokurtic) disturbances, which are resampled from regression residuals of real-world data used by Section 6.

5.1 Variable Selection with Highly Correlated Predictors

The data generating process (DGP) is given by (1) with $n = 10^8$, $k = 100$, $\sigma = 10$, $\beta = (1, 0.9, \dots, 0.1, 0, \dots, 0)'$. That is, among the 100 predictors, only the first 10 have non-zero coefficients. Each row of X is randomly sampled from a zero-mean multivariate normal distribution with the correlation 0.99 for all variable pairs. The DGP shows two characteristics of big data: volume (large n) and veracity (large σ and multicollinearity).

The double-floating regression data occupy 80GB disk space, and they are saved in 1000 text files; each file contains 100 thousand observations. As the full-sample data cannot be loaded into our computer memory, we resort to the split-and-merge method (Algorithm 1). We consider the following factors regarding data partition: 1) the subsample size should be small enough to allow in-memory computing of $X_i'X_i$, $X_i'Y_i$, $Y_i'Y_i$; 2) the subsample size should be large enough so that the rank of X_i equals k; 3) as we implement Algorithm 1 via MATLAB datastore and tall array objects, the subsample size should be large enough to take advantage of the matrix computing platform; 4) the

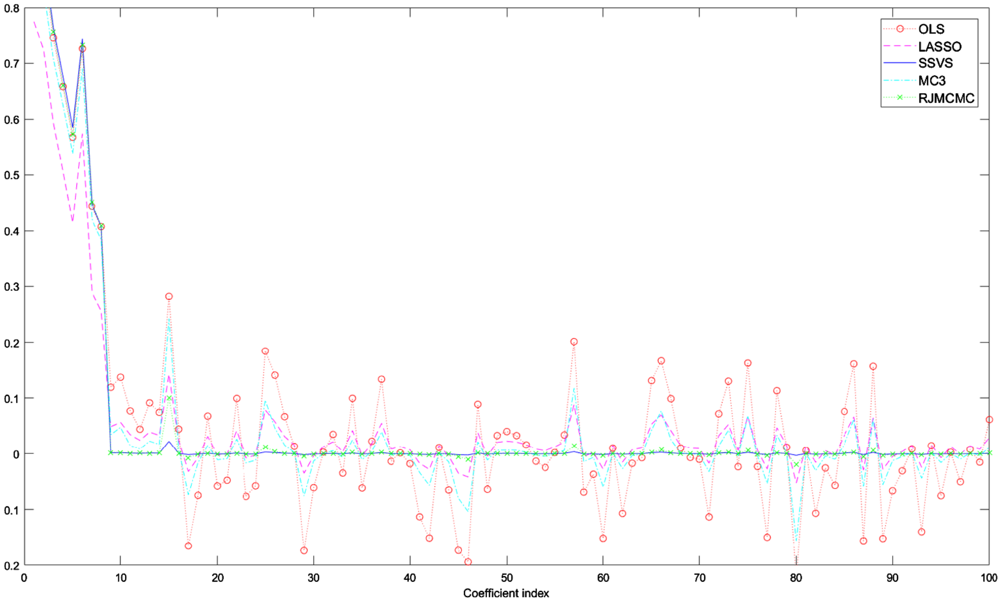

Figure 1: Posterior means of regression coefficients. Regressors have high collinearity with pairwise correlation 0.99. The first 10 of 100 predictors have non-zero coefficients: $\beta = (1, 0.9, \ldots, 0.1, 0, \ldots, 0)'$.

computing cost increases slightly with more subsets (see Section 5.3). However, when n is large, it hardly has any impact on computing speed; and 5) the data-file location matters for partition. In this example, MATLAB tall array works efficiently when it reads all observations in a text file. About 80MB data are loaded into the computer memory for the subset regressions.

Though we run Bayesian regressions, Step 3 of Algorithm 1 also produces the ordinary least squares (OLS) results. The posterior mean $\widetilde{\mu}$ equals the OLS estimator. As is shown in Figure 1, the OLS estimator is volatile due to multicollinearity.

The big data are scanned only once, and then we reuse $NIG(\widetilde{\mu}, \widetilde{\Lambda}, \widetilde{a}, \widetilde{b})$ for Bayesian LASSO, SSVS, MC^3 by Algorithm 2 – 4, and RJMCMC with pseudo observations by Algorithm 5. Figure 1 plots the posterior means of the regression coefficients $E(\beta_j|Y)$, $j = 1, \ldots, k$, which are Bayesian model averaging results. Table 1 reports the posterior probabilities $p(\gamma_j|Y)$, which indicate Bayesian model selection results, for SSVS, MC^3 and RJMCMC.

Variable selection is challenging because of high collinearity between predictors, but the big data algorithms work well. SSVS strongly favors the first 8 predictors in that $E(\gamma_j|Y) \approx 1$, and the estimated coefficients are close to the true values specified by the DGP. The true coefficients for the 9^{th} and 10^{th} predictors are 0.2 and 0.1. SSVS tends to exclude them. Given the fact that the OLS standard error is about 0.1, it is not surprising that SSVS rejects "insignificant" predictors. The estimated coefficients and

Index	SSVS	MC3	RJMCMC	Index	SSVS	MC3	RJMCMC
1	1.0000	1.0000	1.0000	91	0.0013	0.1771	0.0139
2	1.0000	1.0000	1.0000	92	0.0015	0.2132	0.0070
3	1.0000	1.0000	1.0000	93	0.0032	0.3365	0.0171
4	1.0000	1.0000	1.0000	94	0.0008	0.1938	0.0043
5	1.0000	1.0000	1.0000	95	0.0012	0.2253	0.0047
6	1.0000	1.0000	1.0000	96	0.0010	0.1738	0.0069
7	0.9757	1.0000	0.9992	97	0.0011	0.2046	0.0140
8	0.9629	0.9940	0.9810	98	0.0010	0.2054	0.0153
9	0.0014	0.3110	0.0183	99	0.0010	0.2097	0.0050
10	0.0026	0.3671	0.0174	100	0.0008	0.2015	0.0168

Table 1: Posterior probabilities $p(\gamma_j|Y)$ for the first and last 10 predictors.

probabilities for the remaining predictors are close to zero, which indicate that they are decisively excluded from the regression. RJMCMC produces similar results (the solid line and the dot-cross line largely overlap in Figure 1), though RJMCMC is slower. MC^3 calibrated to BIC produces acceptable results. The first 10 estimated coefficients are close to the OLS results, and MC^3 suppresses most spikes produced by the OLS estimator for the remaining coefficients. Bayesian LASSO estimator is also reasonable. Though the shrinkage estimator is slightly smaller compared to other estimators, LASSO effectively excludes most predictors whose true coefficients are zero by the DGP.

5.2 Regression with Skewed and Leptokurtic Disturbances

Before we analyze the real-world wage data in Section 6, we generate some synthetic data. The predictors X are copied from the real data, and the noises ε are resampled from OLS residuals using the real data. The noises have the sample skewness 2.7 and kurtosis 33.9. The synthetic response variable is constructed such that $Y = X\beta + \varepsilon$, where $\beta = (5, 4.9, \ldots, 0.1, 0, \ldots, 0)'$. That is, the first 50 of the 327 predictors have non-zero coefficients.

As is seen in Figure 2, the synthetic-data OLS estimator substantially departs from zero, and the 171^{st} and 172^{nd} coefficients are extremely large in magnitude (see Section 6 for an explanation; they correspond to Treating and Diagnosing). In contrast, Bayesian LASSO, SSVS and RJMCMC effectively shrink the estimators towards zero after the 50^{th} predictor. Meanwhile, the estimators for the first 50 coefficients are close to the true values specified by the DGP.

5.3 Computational Complexity

We measure complexity by floating point operations (flops). By convention, a floating-point addition, subtraction, multiplication, or division is counted as a flop. Solving a k-dimensional linear equation is assumed to take $\frac{2}{3}k^3$ flops.

Algorithm 1 has the complexity $O(k^2 n) + O(k^3 m)$, where n, k, m denote the number

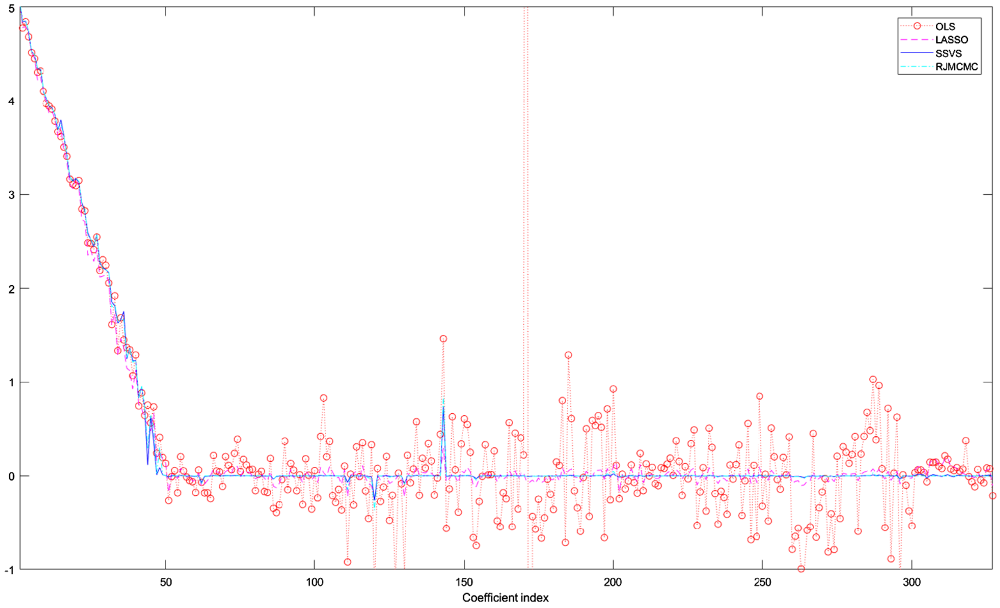

Figure 2: Posterior means of regression coefficients. Noises are skewed and leptokurtic. The first 50 of 327 predictors have non-zero coefficients: $\beta = (5, 4.9, \ldots, 0.1, 0, \ldots, 0)'$.

of observations, variables and subsets, respectively. Specifically, for the full-sample regression without data partition, the required flops are about $2k^2n + \frac{2}{3}k^3$. With data split into m subsets, flops increase by $\frac{4}{3}k^3m$, which is mild compared to $2k^2n$ when n is large.

If Algorithm 2, 3, 4 and 5 receive $NIG(\widetilde{\mu}, \widetilde{\Lambda}, \widetilde{a}, \widetilde{b})$ produced by Algorithm 1 as the input, then they only add $O(k^3r)$ flops for MCMC simulations, where r denotes the number of draws. With the NIG summation operator, the MCMC samplers will not be more computationally expensive as n increases. For big data applications, Bayesian LASSO and SSVS might run as fast as the OLS regression, because $O(k^2n)$ can be much larger than $O(k^3r)$.

To be specific, in the first synthetic data example, the OLS regression without data partition costs 2×10^{12} flops. Splitting data into 1000 subsets for Algorithm 1, we add 1.3×10^9 flops, which is less than 0.1% of the OLS computing costs. Given the outputs of Algorithm 1, we count the run-time flops for Bayesian variable selection MCMC simulations with 10^5 draws. Both Bayesian LASSO and SSVS take about 2.2×10^{11} flops. MC^3 and RJMCMC cost 1.4×10^{12} and 1.5×10^{12} flops, respectively.

6 Application

Occupational employment and wage are of interest to both researchers and the public. Despite various publications on highest paying jobs, many studies rely on survey data and self-reported incomes, which are prone to the mis-reporting problem that could induce bias or inflate the variance of the estimator. We study the Labor Condition Application (LCA) disclosure data by the Office of Foreign Labor Certification, within the Department of Labor. When a U.S. employer sponsors an H1B visa, the employer files an LCA, on behalf of the worker, that includes the job title, Standard Occupational Classification (SOC) code, rate of pay (actual wage offered for the position), and the prevailing wage for the same position in that area. Since an LCA is legal instrument prepared by employers' attorneys, we believe the job and pay information is accurate. The publicly available data contain 3 million observations for the year $2011-2016$.[5] H1B wages are typically for technical jobs that require advanced degrees. The median annual wage equals 72.4K (where K stands for thousand dollar) and the standard deviation (sd.) is 33.0K. We generate regressors by extracting 200 highest frequency (HF) words from the job titles and SOC names, as well as 100 HF employer names. We also add the year dummies, the census division and region dummies (such as New England, South Atlantic, etc.), 10 most affluent metropolitan area dummies (such as California Bay Area, New York, etc.), the employment length and part-time position dummies. There are 327 predictors in total. Details on data cleaning and processing are provided in Supplementary Material.

Table 2 illustrates the job HF words and the average wages of the subsamples that contain the key words. HF words cover occupations (e.g., Engineer, Analyst), industries (e.g., Software, Finance), experience (e.g., Senior, Assistant), positions (e.g., Manager, Lead), skills (e.g., Database, SAS), countries (e.g., USA, India), etc. Some HF words are associated with high wages, say Manager (99.3K) and Senior (97.6K), while other words such as Researcher (68.5K) and Accountant (65.8K) correspond to lower wages.

Table 3 shows the top employers and the mean wages they offer. The most generous employers are Facebook (137.6K) and Google (131.4K). The largest H1B sponsors are multinational consultancy services such as Infosys (78.9K) and Tata (68.7K). University of Michigan (67.4K) and Johns Hopkins University (65.8K) are also among the top 100 sponsors.

6.1 OLS Results

The OLS estimator is obtained from $\widetilde{\mu}$ by running Step 1–3 of Algorithm 1. Split and merge are performed by MapReduce-like (Dean and Ghemawat, 2008) MATLAB tall array operations.

Predictors with the largest positive and negative coefficients are listed in Table 4. Physician (+88.4K) and Lawyer (+55.1K) are among the highest paying jobs, while words like Preschool (-20.3K) and Assistant (-17.4K) have substantially negative impact on the rate of pay. We can also learn the joint effects of multiple regressors. For example, Resident Physician has the net effect -105.3K + 88.4K = -16.9K, which is reasonable

[5] Data were retrieved from www.foreignlaborcert.doleta.gov/performancedata.cfm on April 18, 2017. Observations include new, renewed, transferred and cap-exempt LCA.

Word	Frequency	Wage	Word	Frequency	Wage
Computer	1.2e6	76.1	Accountant	5.6e4	65.8
Analyst	9.6e5	75.4	Assistant	5.5e4	78.2
System	7.0e5	81.3	Operation	5.4e4	85.6
Engineer	6.0e5	89.0	Physician	5.1e4	160.4
Developer	5.5e5	91.4	Network	5.0e4	75.6
Software	5.4e5	92.3	Marketing	5.0e4	77.9
Program	5.0e5	69.4	Auditor	4.8e4	64.5
Application	4.2e5	92.1	Database	4.6e4	78.4
Senior	2.9e5	97.6	Physicist	4.5e4	66.0
Manager	2.7e5	99.3	Information	4.4e4	109.4
Occupational	1.9e5	76.6	Assurance	4.4e4	74.0
Technology	1.8e5	80.6	Mechanical	4.1e4	77.7
Specialist	1.6e5	68.1	Staff	4.1e4	94.9
Consultant	1.5e5	87.2	Market	3.9e4	64.3
Researcher	1.4e5	68.5	Director	3.8e4	118.3
USA	1.2e5	79.8	Professor	3.8e4	92.1
Business	1.2e5	78.4	Product	3.8e4	92.9
Lead	1.1e5	86.2	Postdoctoral	3.2e4	48.9
Associate	1.1e5	76.8	Chemist	3.0e4	59.4
Administrator	9.1e4	75.2	Surgeon	2.9e4	150.5
Tester	7.8e4	75.7	Principal	2.9e4	121.7
Scientist	7.1e4	69.9	Biologist	2.9e4	58.0
Architect	7.0e4	93.0	Industrial	2.9e4	76.9
Teacher	7.0e4	67.6	General	2.7e4	153.0
Project	6.9e4	86.7	Support	2.7e4	75.2
Finance	6.7e4	95.9	IT	2.7e4	84.7
Electronic	6.4e4	89.4	Data	2.7e4	85.1
Designer	6.3e4	75.7	Service	2.5e4	85.1
Hospital	6.1e4	95.5	Web	2.5e4	73.4
Quality	6.0e4	75.6	Therapist	2.5e4	69.3

Table 2: Highest frequency words of job titles and wage offers. The table shows the top 60 words with the corresponding frequencies and the mean annual wages (in thousand dollar) of the subsamples that contain the words.

because it is a stage of graduate medical training. More examples include Assistant Professor (-17.4K + 28.2K), Principal Economist (30.8K + 24.5K), etc.

The dummy variables are informative. As we include all the 9 census region dummies, their coefficients represent the intercept terms (i.e., baseline salaries) specific to the regions. Workers in Pacific (74.0K) have the highest income, followed by New England (70.5K), Mid Atlantic (67.9K), West South Central (67.0K), Mountain (66.1K), South Atlantic (66.0K), West North Central (65.6K), East North Central (64.7K) and East South Central (64.0K). In addition, metropolitan workers in New York (+10.6K) and Bay Area (+10.0K) earn substantially more.

Employer	Frequency	Wage	Employer	Frequency	Wage
Infosys	1.4e5	78.9	KPMG	4.4e3	73.6
Tata	6.5e4	68.7	Goldman Sachs	4.3e3	101.0
Deloitte	5.0e4	90.0	Yash	4.1e3	67.1
Capgemini	4.2e4	79.8	Mindtree	4.0e3	72.9
Wipro	4.2e4	71.3	Capital One	3.9e3	89.1
IBM	3.9e4	81.7	HTC	3.9e3	70.0
Accenture	3.3e4	78.0	KPIT	3.9e3	66.6
Tech Mahindra	2.7e4	74.4	Facebook	3.7e3	137.6
HCL	2.6e4	72.4	Ebay	3.6e3	125.8
Microsoft	2.2e4	118.8	Synechron	3.5e3	81.6
Ernst Young	1.8e4	92.2	System Soft	3.5e3	72.8
Larsen Toubro	1.8e4	66.7	BOA	3.5e3	98.4
Cognizant	1.6e4	72.9	Compunnel	3.3e3	66.7
Google	1.3e4	131.4	Randstad	3.3e3	105.8
UST Global	1.2e4	67.3	WalMart	3.2e3	116.7
Intel	1.1e4	91.7	Hitachi	2.9e3	87.3
Qualcomm	9.7e3	96.0	CVS	2.9e3	112.1
Amazon	9.6e3	111.8	Cisco	2.9e3	103.9
Dell	8.2e3	97.5	Paypal	2.9e3	125.4
Oracle	7.9e3	99.9	Salesforce	2.8e3	122.9
PWC	6.8e3	87.5	CSC	2.7e3	70.3
Apple	6.8e3	125.1	Persistent Systems	2.7e3	72.7
JP Morgan	6.5e3	103.6	HP	2.7e3	97.5
NTT	5.7e3	83.9	Itech	2.7e3	81.3
Syntel	5.4e3	75.8	Vsoft	2.6e3	65.6
Mphasis	5.2e3	73.0	Yahoo	2.6e3	108.6
Fujitsu	5.2e3	83.8	VMware	2.5e3	123.7
Mastech	5.1e3	87.6	Astir	2.5e3	66.4
Hexaware	5.0e3	70.2	Headstrong	2.5e3	79.0
Cummins	4.6e3	70.5	Ericsson	2.4e3	86.4

Table 3: Highest frequency employers and wage offers. The table shows the top 60 employers with the corresponding frequencies and their mean annual wages (in thousand dollar). The remaining high frequency employers used in the regression are SAP, Diaspark, Bloomberg, Morgan Stanley, Birlasoft, Marlabs, Broadcom, Virtusa, Verizon, University of Michigan, Reliable Software, CGI, MathWorks, Sapient, Johns Hopkins University, Avco, Marvell, LinkedIn, Management Health, Ciber, Symantec, Schlumberger, ERP, ITC, Mayo Clinic, Pyramid, Micron, Kforce, Experis, Everest, Citibank, NIH, Rite Aid, Global Foundries, Netapp, Technosoft, Cyberthink, Texas Instruments, Merrill Lynch, Credit Suisse.

Wage prediction can be performed in two steps. First, we specify the work-site region and employment terms. For example, Boston is in New England (70.5K) and one of the affluent metropolitan areas (+1.4K). For a 3-year full-time appointment, the predicted base salary is 71.9K in 2016 dollar. Second, we describe the job, say “financial analyst,

Word	Coeff	Sd.	Wage	Word	Coeff	Sd.	Wage
Physician	88.36	0.22	160.38	Resident	-105.30	0.24	57.53
Lawyer	55.07	0.27	125.70	Fellow	-39.89	0.21	56.32
Facebook	48.59	0.38	137.58	Intern	-28.27	0.34	53.83
Rite Aid	43.87	0.64	132.80	Drafter	-28.22	0.38	50.09
Dentist	42.80	0.35	126.52	Landscape	-23.82	0.53	59.87
Google	42.05	0.21	131.45	Preschool	-20.27	0.50	42.73
NIH	41.84	0.62	75.17	Food	-19.99	0.36	60.01
Petroleum	40.24	0.44	114.06	Community	-19.52	0.52	61.51
Apple	35.59	0.29	125.07	Music	-18.13	0.51	56.80
Surgeon	34.16	0.24	150.52	School	-17.44	0.35	49.62
Bloomberg	33.62	0.48	130.89	Assistant	-17.35	0.16	78.22
Director	31.91	0.13	118.30	Wholesale	-17.06	0.48	63.11
Principal	30.78	0.15	121.70	Editor	-16.41	0.44	57.65
Professor	28.18	0.24	92.10	Persistent Systems	-16.25	0.45	72.69
Mayo Clinic	27.61	0.64	102.05	Pyramid	-16.14	0.58	65.21
Microsoft	27.28	0.16	118.82	Birlasoft	-15.34	0.49	61.94
CVS	27.08	0.48	112.06	Market	-14.70	0.21	64.34
General	26.19	0.20	152.96	Graphic	-14.40	0.26	59.44
Economist	24.54	0.37	101.87	Diagnosing	-14.11	10.40	62.23
Paypal	24.52	0.45	125.42	Lab	-13.87	0.48	55.75
Ebay	24.36	0.40	125.76	UST Global	-13.70	0.22	67.27
Credit Suisse	23.85	0.78	108.86	ERS	-13.55	0.33	79.14
Morgan Stanley	23.07	0.49	111.14	Public	-13.31	0.47	58.92
Treating	22.65	10.39	62.31	Instructor	-12.32	0.23	61.37
CitiBank	22.59	0.59	113.44	Worker	-12.11	0.34	51.96
WalMart	22.50	0.41	116.67	Postdoctoral	-11.64	0.16	48.85
Pharmacist	22.01	0.24	123.92	Scholar	-11.50	0.36	51.50
Practitioner	21.35	0.40	67.34	Legal	-11.35	0.45	77.65
Amazon	21.28	0.24	111.82	KPIT	-11.22	0.38	66.61
Hydrologist	21.23	0.88	98.77	Global Foundries	-11.13	0.61	88.70

Table 4: Bayesian linear regression under the non-informative prior. The table shows the predictors with the largest positive (Column 1–4) and negative (Column 5–8) coefficients measured by posterior means (OLS estimators). Posterior standard deviations are displayed in Column 3 and 7. Mean annual wages (in thousand dollar) of the subsamples that contain the words are shown in Column 4 and 8.

senior quant developer, Fidelity", which involves HF words Finance (+13.7K), Analyst (-1.8K), Senior (+15.8K), Developer (+4.7K). Thus, our regression predicts the wage 104.3K. Since we only include 100 top employers, Fidelity does not enter our regression and the word is ignored.

A problem of the OLS regression is that HF words have heterogeneous predictive power and accuracy. Some coefficients are close to zero (by posterior mean); some nonzero coefficients are inaccurate (by large posterior sd.), and many predictors are apparently correlated. Also, there is an anomaly in Table 4: Treating has a large posterior mean 22.7K with a surge of sd. to 10.4K. Though the coefficient remains “significant” (in frequentist terminology), the unusually large sd. hints at the possibility of multicollinearity.[6] After visual inspection of the predictors, we found that Treating mostly comes from the SOC name “Health Diagnosing and Treating Practitioners”. Both Treating and Diagnosing enter the regression, and the latter has the coefficient -14.1K with sd. 10.4K. Near-perfect multicollinearity is difficult to detect when predictors are machine generated, especially in big data applications (because sd. decreases with the sample size, rendering all coefficients “significant”). As is shown below, Bayesian shrinkage and variable selection can overcome multicollinearity.

6.2 Bayesian LASSO Results

We implemented Algorithm 2 with a sequence of LASSO regularization parameters $\lambda = 500, 1000, 2000, 5000$ for HF words and employers. As for the 27 geographic and time dummy variables, they enter the regression by economic, rather than statistical, significance, and we intend not to shrink them. As the Gibbs sampler cannot proceed under $\lambda = 0$, we put a small value 0.1 for those dummy variables. Note that big data are scanned once regardless of multiple rounds of LASSO regressions under different regularization parameters, because Step 1–3 of Algorithm 2 have no reference to λ values.[7]

Table 5 demonstrates that the magnitude of shrinkage is negatively related to “t-statistics”. For example, FINANCE has a large “t-statistics”128.2. As λ increases, the LASSO estimators are stable: 14.5K, 15.2K, 15.7K and 14.7K, all of which are close to the OLS estimator 13.7. In comparison, TREATING has a small “t-statistics”2.2. Even under mild regularization, its estimator shrinks substantially, from 22.7K (OLS estimator) to 2.6K (LASSO with $\lambda = 500$). As λ increases to 1000, 2000 and 5000, its estimator quickly drops to 0.38K, 0.04K and 0.01K, respectively. Bayesian LASSO overcomes multicollinearity and effectively discards predictors with inflated variances.

Frequentist LASSO is a popular variable selection method, as L_1-penalized least squares yield corner solutions, rendering some coefficients exactly zero. Bayesian LASSO can shrink the posterior mode of weak predictors to zero, but the mode does not reveal itself from MCMC draws. Nevertheless, variable selection can be achieved by visual inspection of the posterior means. Table 5 indicates that there is a dichotomy between the strong and weak predictors, especially when λ exceeds 1000. For example, if we exclude a variable if its posterior mean is less than 0.1, most predictors are either substantially larger or substantially smaller than the threshold. Under that criterion, Bayesian LASSO selects 282, 228, 175, 105 variables when $\lambda = 500, 1000, 2000, 5000$ respectively.

[6] In this particular case, it appears that multicollinearity inflates the variances of the correlated pairs without contaminating other predictors. We removed Treating and run the regression again, the results are largely the same as those reported in Table 4.

[7] An interpretation of multiple values is an unknown regularization parameter with a uniform prior over the points in the selected grid.

Words	Coeff(t-stat)	L500	L1000	L2000	L5000	SSVS	MC^3
Software	5.70 (67.83)	5.77	5.80	5.71	5.38	6.54	5.70
USA	1.68 (8.87)	0.39	0.01	-0.01	-0.01	0.00	1.68
Finance	13.73 (128.22)	14.53	15.17	15.71	14.73	13.34	13.73
Marketing	5.54 (32.54)	4.90	4.18	2.53	0.00	5.38	5.54
Professor	28.18 (116.00)	29.56	30.13	27.58	20.94	28.36	28.13
IT	2.07 (14.21)	1.56	1.07	0.20	0.01	0.00	2.06
Sales	8.49 (48.97)	7.81	7.09	5.59	1.27	8.64	8.50
Resident	-105.30 (-447.19)	-102.95	-100.65	-96.66	-85.15	-105.51	-105.33
Secondary	1.39 (3.46)	-0.04	-0.03	-0.01	0.00	0.00	1.30
Lawyer	55.07 (202.88)	53.25	51.16	47.51	37.02	56.24	55.18
Artist	-5.50 (-14.82)	-3.31	-2.64	-1.24	-0.01	-5.13	-5.49
Family	16.30 (47.19)	13.78	11.29	5.99	0.01	16.19	16.32
Care	-3.51 (-8.82)	-1.25	-0.09	-0.02	0.00	-2.88	-3.52
Integration	5.69 (16.65)	2.94	0.59	0.02	0.00	5.96	5.69
Cost	-1.05 (-1.48)	-1.26	-1.12	-0.03	0.00	0.00	0.00
Warehouse	2.42 (5.91)	0.11	0.02	0.01	0.00	0.00	2.40
Estimator	-8.39 (-10.50)	-4.40	-0.85	-0.03	0.00	-8.91	-9.40
Treating	22.65 (2.18)	2.57	0.38	0.04	0.01	10.11	8.60
Diagnosing	-14.11 (-1.36)	1.83	0.34	0.04	0.01	0.00	0.00
Geoscientist	11.17 (12.89)	9.40	7.60	3.66	0.01	11.25	11.17
Preschool	-20.27 (-40.26)	-14.55	-8.64	-0.17	-0.01	-20.72	-20.32
Registered	-6.10 (-12.67)	-0.41	-0.02	0.00	0.00	-5.47	-6.10
IBM	0.19 (1.50)	0.12	0.08	0.03	0.01	0.00	0.17
Intel	4.71 (19.45)	3.23	2.05	0.14	0.01	4.62	4.68
Mphasis	-4.86 (-14.96)	-2.64	-0.44	-0.02	0.00	-4.42	-4.86
HTC	6.21 (16.52)	3.29	0.40	0.02	0.00	6.08	6.20
Hitachi	0.99 (2.31)	0.01	0.00	0.00	0.00	0.00	0.00
Yahoo	8.04 (17.32)	2.60	0.06	0.01	0.00	7.32	8.03
Marlabs	5.67 (11.23)	0.50	0.03	0.01	0.00	5.50	5.66
Avco	-6.76 (-12.54)	-0.82	-0.04	-0.01	0.00	-6.96	-6.81
Pyramid	-16.14 (-27.83)	-8.86	-1.44	-0.02	0.00	-16.36	-16.18
Technosoft	2.33 (3.87)	0.03	0.01	0.00	0.00	0.00	2.33
New England	70.49 (715.09)	70.15	69.98	70.15	71.06	69.34	70.48
New York	10.58 (142.95)	10.78	10.97	11.27	11.45	10.47	10.58
Two Year	3.75 (55.64)	3.65	3.57	3.46	3.10	3.84	3.75

Table 5: Bayesian linear regression with variable selection. The first column lists the predictors No. $6, 16, 26, \ldots, 326$ (due to the space limit; full results available upon request) as well as Treating and Diagnosing. In the second column, Coeff represents posterior means under the non-informative prior (OLS estimator), and "t-stat" in parenthesis are simply the ratio of the posterior mean and standard deviation. Column $3-6$ correspond to the Bayesian LASSO with regularization parameters $\lambda = 500, 1000, 2000, 5000$. Column 7 and 8 are the results for SSVS and MC^3.

6.3 SSVS and MC^3 Results

In our implementation of Algorithm 3 and 4 for SSVS and MC^3, the prior probability of selecting each HF word is assumed to be 0.5, and that of the geographic and time dummies equals 1 (so that the posterior probability still equals 1). SSVS involves the large and small variances of the mixture distribution as the tuning parameters. We experimented several pairs such as 10/0.001, 100/0.001, 1000/0.001, and the results are similar. Again, big data are scanned once under multiple pairs of tuning parameters. The Gibbs sampler generates draws for the model indicators, and we favor the most frequently visited model, which is likely to be a model with a high, if not the highest, posterior probability. SSVS selects 229 out of the 300 predictors on occupations and employers. As is seen from Table 5, our previously identified strong predictors, such as Finance, Professor, Resident and Lawyer, are also selected by SSVS and their coefficients are close to those under the LASSO regression. Unlike LASSO shrinkage on both Treating and Diagnosing, SSVS selects the former and discards the latter, which offers an alternative solution to the multicollinear anomaly in our regression.

Table 5 also reports the regression results for the most frequently visited model under MC^3 by Algorithm 4. We adopt the g-prior and hyperparameters are calibrated to the frequentist AIC criterion. See George and Foster (2000). 276 variables are selected. Coefficients of the strong predictors are close to those under SSVS.

6.4 Forecast Evaluation

Lastly, we perform out-of-sample forecast using the latest release LCA data from October 2016 to March 2017. We consider LCA cases in which job titles contain at least one of the 200 HF words and employers belong to one of the 100 HF employers. There are about 112 thousand observations for forecast evaluation. The mean absolute deviation (MAD) of the forecast by the OLS regression is 15.1K, which is reasonable as the sample sd. amounts to 33.0K. The MAD of LASSO regression is given by 14.5K, 14.0K, 13.3K, 13.6K when $\lambda = 500, 1000, 2000, 5000$ respectively. The MAD of SSVS and MC^3 are both near 15.1K. It appears that Bayesian LASSO regression with $\lambda = 2000$ works best for the current application.

7 Conclusion

The primary advantage of the NIG summation operator is the ability to merge the subset posterior distributions with data split into manageable pieces. It is also useful for Bayesian variable selection methods in which priors have mixture NIG representations, as the MCMC samplers can iteratively combine NIG distributions with a single pass of big data. Computational complexity analysis demonstrates that Bayesian variable selection algorithms with NIG summations are computationally efficient, and some MCMC samplers may run almost as fast as the OLS regression, when the sample size is large.

NIG summation can be extended to subtraction and scalar multiplication. Subtraction can be thought as taking some observations out of the NIG distribution, and scalar multiplication rescales the precision of regression data. Consider online statistical learning in which data become available in a sequential order. For example, to study the time-varying beta in the capital asset pricing model, it is common practice to employ rolling-window regressions with five or ten years of moving observations (see Ang and Chen, 2007). With NIG summation (for sequential addition of newest observations) and subtraction (for data point retirement of oldest observations), the rolling regression complexity can drop from $O(k^2n^2) + O(k^3n)$ to $O(k^3n)$. Another use case is recursive least squares with a forgetting factor $\delta \in (0, 1)$ (see Branch and Evans (2006) for a macroeconomic forecasting application), which discounts past observations at geometric rate. NIG summation and scalar multiplication like $\sum_{i=1}^{m} \delta^{m-i} NIG(\widetilde{\mu}_i, \widetilde{\Lambda}_i, \widetilde{a}_i, \widetilde{b}_i)$ lead to a weighted regression. Another direction of extending the current approach is that a collection of NIG distributions closed under summation and scalar multiplication may constitute a linear space that might have interesting theoretic properties. That will be left for future research.

Supplementary Material

Supplementary Material for Big Data Bayesian Linear Regression and Variable Selection by Normal-Inverse-Gamma Summation (DOI: 10.1214/17-BA1083SUPP; .pdf). Proofs of Proposition 1–9 and data cleaning procedures in Section 6 (in a separate document).

References

Ang, A. and Chen, J. (2007). "CAPM over the Long Run: 1926–2001." *Journal of Empirical Finance*, 14(1): 1–40. 1033

Branch, W. A. and Evans, G. W. (2006). "A Simple Recursive Forecasting Model." *Economics Letters*, 91(2): 158–166. 1033

Carlin, B. P. and Chib, S. (1995). "Bayesian Model Choice via Markov Chain Monte Carlo Methods." *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(3): 473–484. 1012, 1020

Dean, J. and Ghemawat, S. (2008). "MapReduce: Simplified Data Processing on Large Clusters." *Communications of the ACM*, 51(1): 107–113. 1026

Fan, J. and Lv, J. (2010). "A Selective Overview of Variable Selection in High Dimensional Feature Space." *Statistica Sinica*, 20(1): 101–148. 1012

George, E. I. and Foster, D. P. (2000). "Calibration and Empirical Bayes Variable Selection." *Biometrika*, 87(4): 731–747. 1012, 1020, 1032

George, E. I. and McCulloch, R. E. (1993). "Variable Selection via Gibbs Sampling." *Journal of the American Statistical Association*, 88(423): 881–889. 1011, 1017

George, E. I. and Mcculloch, R. E. (1997). "Approaches for Bayesian Variable Selection." *Statistica Sinica*, 339–374. 1017

Geweke, J. (1996). *Variable Selection and Model Comparison in Regression*. Oxford: Oxford University Press. 1018

Ghosh, J. and Reiter, J. P. (2013). "Secure Bayesian Model Averaging for Horizontally Partitioned Data." *Statistics and Computing*, 23(3): 311–322. MR3041438. 1011

Godsill, S. J. (2001). "On the Relationship between Markov Chain Monte Carlo Methods for Model Uncertainty." *Journal of Computational and Graphical Statistics*, 10(2): 230–248. 1020, 1021

Green, P. J. (1995). "Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination." *Biometrika*, 82: 711–732. 1012, 1020

Johnson, V. E. and Rossell, D. (2012). "Bayesian Model Selection in High-Dimensional Settings." *Journal of the American Statistical Association*, 107(498): 649–660. 1012

Kuo, L. and Mallick, B. (1998). "Variable Selection for Regression Models." *Sankhya: The Indian Journal of Statistics, Series B*, 60(1): 65–81. MR1717076. 1020

Lin, D., Foster, D. P., and Ungar, L. H. (2011). "VIF Regression: A Fast Regression Algorithm for Large Data." *Journal of the American Statistical Association*, 106(493): 232–247. 1012

Madigan, D., York, J., and Allard, D. (1995). "Bayesian Graphical Models for Discrete Data." *International Statistical Review*, 63(2): 215–232. 1012, 1018

Miroshnikov, A., Savel'ev, E., and Conlon, E. M. (2015). "BayesSummaryStatLM: An R package for Bayesian Linear Models for Big Data and Data Science." Manuscript: https://arxiv.org/abs/1503.00635. 1011

Neiswanger, W., Wang, C., and Xing, E. P. (2014). "Asymptotically Exact, Embarrassingly Parallel MCMC." In *Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence*, 623–632. Arlington:AUAI Press. 1011

Ordonez, C., Garcia-Alvarado, C., and Baladandayuthapani, V. (2014). "Bayesian Variable Selection in Linear Regression in One Pass for Large Datasets." *ACM Transactions on Knowledge Discovery from Data*, 9(1): 1–14. 1011

Park, T. and Casella, G. (2008). "The Bayesian Lasso." *Journal of the American Statistical Association*, 103(482): 681–686. 1016

Qian, H. (2017). "Supplementary Material for Big Data Bayesian Linear Regression and Variable Selection by Normal-Inverse-Gamma Summation." *Bayesian Analysis*. doi: https://doi.org/10.1214/17-BA1083SUPP. 1014

Raftery, A. E., Madigan, D., and Hoeting, J. A. (1997). "Bayesian Model Averaging for Linear Regression Models." *Journal of the American Statistical Association*, 92(437): 179–191. 1018, 1019

Scott, S. L., Blocker, A. W., Bonassi, F. V., Chipman, H. A., George, E. I., and McCulloch, R. E. (2016). "Bayes and Big Data: The Consensus Monte Carlo Algorithm." *International Journal of Management Science and Engineering Management*, 11: 78–88. 1011

Smith, M. and Kohn, R. (1996). "Nonparametric Regression Using Bayesian Variable Selection." *Journal of Econometrics*, 75(2): 317–343. 1018

Tibshirani, R. (1996). "Regression Shrinkage and Selection via the Lasso." *Journal of the Royal Statistical Society B*, 58: 267–288. 1016

Zeller, A. (1986). *On Assessing Prior Distributions and Bayesian Regression Analysis with g-Prior Distributions*. New York: Elsevier Science Publishers. 1020

Acknowledgments

The author would like to thank the editor, associate editor and the reviewer for the comments and insights that greatly helped improve the manuscript.