

DOI: <http://dx.doi.org/10.21123/bsj.2022.19.3.0551>

Variable Selection Using a Modified Gibbs Sampler Algorithm with Application on Rock Strength Dataset

Ghadeer J.M. Mahdi*

Othman M. Salih

Department of Mathematics, College of education for Pure sciences- ibn Al-Haitham, University of Baghdad, Iraq

*Corresponding author: mahdighadeer@gmail.com, gmaahdi@ihcoedu.uobaghdad.edu.iq

*ORCID ID: <https://orcid.org/0000-0003-4870-4034>, <https://orcid.org/0000-0002-9908-8748>

Received 11/12/2020, Accepted 31/1/2021, Published Online First 20/11/2021 Published 1/6/2022



This work is licensed under a [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/).

Abstract:

Variable selection is an essential and necessary task in the statistical modeling field. Several studies have tried to develop and standardize the process of variable selection, but it is difficult to do so. The first question a researcher needs to ask himself/herself what are the most significant variables that should be used to describe a given dataset's response. In this paper, a new method for variable selection using Gibbs sampler techniques has been developed. First, the model is defined, and the posterior distributions for all the parameters are derived. The new variable selection method is tested using four simulation datasets. The new approach is compared with some existing techniques: Ordinary Least Squared (OLS), Least Absolute Shrinkage and Selection Operator (Lasso), and Tikhonov Regularization (Ridge). The simulation studies show that the performance of our method is better than the others according to the error and the time complexity. These methods are applied to a real dataset, which is called Rock Strength Dataset. The new approach implemented using the Gibbs sampler is more powerful and effective than other approaches. All the statistical computations conducted for this paper are done using R version 4.0.3 on a single processor computer.

Keywords: Bayesian, Gibbs, Lasso, Markov chain Monte Carlo, Posterior, Ridge, Variable selection.

Introduction:

The relationship between a set of variables, x_1, x_2, \dots, x_p , and a response, y , can be expressed by the following linear regression model,

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip} + \epsilon_i, \quad (1)$$
$$\epsilon_i \sim N(0, \sigma^2) \quad \forall i = 1, \dots, n.$$

The system of n equations (Eq.1) can be expressed in matrix notation as follows,

$$y = X\beta + \epsilon, \quad \epsilon \sim MVN(0_n, \Sigma_{n \times n}) \quad (2)$$

where y is a $n \times 1$ vector, X is a $n \times (p + 1)$ matrix, β is a $(p + 1) \times 1$ vector, ϵ is a $n \times 1$ vector, 0 is a $n \times 1$ vector and Σ is a $n \times n$ matrix. Generally, selecting certain explanatory variables that can be used to describe the response variable is called feature selection (shrinkage). The feature selection is used to i) remove the unimportant variables which do not add any information; ii) reduce the computation time by shrinking the data size; iii) avoid the overfitting. To decide which variables are irrelevant is hard for high dimensional datasets. On the other hand, it is

difficult to build and interpret a model that uses all the explanatory variables. In this case, variable selection techniques can play an important role. The set of coefficients, β , can express whether the explanatory variables are important for the model or not. When the value of a coefficient is zero or very close to zero, then its corresponding variable is not significant to be chosen in the model.

Variable selection can be made using several traditional approaches. For example, Chi-square¹, ANOVA², and Pearson correlations can compute the variables' impact. Depending on the coefficient values, it can be determined whether the variable is important or not. Moreover, forward and backward selection methods are used to select the best subsets of variables by following some steps³. These methods are slow with large datasets⁴. In this paper, variables are selected based on the influence of their coefficients on the model.

In general, the set of parameters, $\{\beta_i; i = 1, \dots, n\}$, can be estimated from n of the observations using the Ordinary Least Squares (OLS) criterion.

The set of estimated parameters is denoted by $\hat{\beta}_{OLS}$ and defined as follows,

$$\begin{aligned} \hat{\beta}_{OLS} &= \operatorname{argmin}_{\beta \in \mathbb{R}^p} \|Y - X\beta\|^2 \\ &= \operatorname{argmin}_{\beta \in \mathbb{R}^p} (Y - X\beta)^T(Y - X\beta) \end{aligned} \quad (3)$$

From Eq.3, $\hat{\beta}_{OLS}$ is the value of β that gives the minimum squared norm of error between the observed value and estimated value. The first step to derive $\hat{\beta}_{OLS}$, let $h = (Y - X\beta)^T(Y - X\beta)$, and by expanding h yields:

$$\begin{aligned} h &= Y^T Y - \beta^T X^T Y - Y^T X \beta + \beta^T X^T X \beta \\ &= Y^T Y - 2\beta^T X^T Y + \beta^T X^T X \beta \end{aligned}$$

Taking the derivative to β ,

$$\frac{\partial h}{\partial \beta} = -2X^T Y + 2X^T X \beta. \text{ If } \frac{\partial h}{\partial \beta} = 0, \text{ then } X^T X \beta = X^T Y$$

Therefore,

$$\hat{\beta}_{OLS} = (X^T X)^{-1} X^T Y \quad (4)$$

OLS will have unbiased results if X and y have an approximately linear relationship. It also has a low variance if $n \gg p$. However, in real life, in many datasets such as health, business, and economy datasets, the number of explanatory variables can be much larger than the number of samples, $p \gg n$. Hence, the OLS solution is not unique. A dataset may be high variability in the estimators, which causes poor predictive and overfitting. For these types of datasets, researchers usually use Ridge and Lasso models to select variables.

This paper is organized as follows: Section 2 provides a background for Ridge and Lasso models. In section 3, Bayesian inference is discussed. Markov chain Monte Carlo and Gibbs sampler are discussed in sections 4 and 5; respectively. In section 6, a new variable selection method is applied to a simulation dataset. Real data analysis is introduced in section 7. Section 8 presents results and discussion of the variable selection for a real dataset between the new method and some commonly used methods. In the end, the conclusion is given in section 9.

Ridge and Lasso

This section reviews two variable selection (shrinkage) methods named Ridge⁵ and Lasso. Shrinkage methods can be used under some constraints depending on the size of the dataset. The regression model can be fitted using all the p variables, but the shrinkage technique improves the accuracy and stability by reducing the number of variables. The Ridge and Lasso models aim to estimate the coefficient of some variables as 0 or close to zero so that those variables can be excluded from the model. $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p)$ that minimizes

the Residual Sum Squares (RSS) is the solution to the OLS fitting procedure; i.e.,

$$RSS = \sum_{i=1}^n \left(y_i - \beta_0 - \sum_{j=1}^n \beta_j x_{ij} \right)^2$$

Similarly, Ridge regression seeks the vector $\hat{\beta}^{ridge}$ that minimizes the penalized RSS, $RSS + \lambda \sum_{j=1}^p \beta_j^2$, i.e.,

$$\begin{aligned} \hat{\beta}^{ridge} &= \operatorname{minimize} \left(\sum_{i=1}^n \left(y_i - \beta_0 - \sum_{j=1}^n \beta_j x_{ij} \right)^2 \right. \\ &\quad \left. + \lambda \sum_{j=1}^p \beta_j^2 \right), \text{ subject to } \sum_{j=1}^p \beta_j^2 \leq t \end{aligned}$$

where the value of t is the upper bound for the sum of the coefficients. The complexity parameter, λ , is greater than or equals to 0. If $\lambda = 0$, then $\hat{\beta}^{ridge} = \hat{\beta}$, as $\lambda \rightarrow \infty$, $\hat{\beta}^{ridge} \rightarrow 0_p$. And $0 < \lambda < \infty$ balances linear regression model fitting and shrinkage of the coefficients. The shrinkage penalty is small when β_1, β_2, \dots , and β_p are close to zero⁶.

Unlike OLS, ridge solutions are not unique. As a result, before the estimation, the inputs should be standardized. First, β_0 is estimated separately as $\bar{y} = \sum_{i=1}^n \frac{y_i}{n}$, and the remaining parameters can be estimated by using the data matrix X as follows,

$$\hat{\beta}^{ridge} = (X^T X + \lambda I)^{-1} X^T y, \text{ where } I \text{ is the } p \times p \text{ identity matrix.}$$

To constrain the size of OLS estimates different kinds of penalization can be considered. For example, L_1 norm can be used as penalty encompasses, so

$$\begin{aligned} \hat{\beta}^{lasso} &= \operatorname{minimize} \left(\sum_{i=1}^n \left(y_i - \beta_0 \right. \right. \\ &\quad \left. \left. - \sum_{j=1}^n \beta_j x_{ij} \right)^2 \right. \\ &\quad \left. + \lambda \sum_{j=1}^p |\beta_j| \right), \quad \lambda \geq 0 \end{aligned} \quad (5)$$

Lasso property allows excluding variables by setting their coefficients to be zero⁷. The reduced model becomes more efficient, especially when the number of variables is much larger than the number of samples, $p \gg n$.

Bayesian Inference of a Multivariate Linear Regression

Eq.2 is used to apply Gibbs sampler in the Multivariate Linear Regression (MLR). From Eq.2, it can be concluded that $y \sim MVNn(X\beta; \sigma^2 I_n)$. Therefore, the likelihood function, denoted by $L(y)$, can be expressed as follows,

$$L(y) = \prod_{i=1}^n f(y_i)$$

$$= \frac{1}{(\sqrt{\sigma^2})^n} e^{\left(\frac{-1}{2}(y-X\beta)^T(\sigma^2 I_n)^{-1}(y-X\beta)\right)}$$

$$= \frac{1}{(\sigma^2)^{\frac{n}{2}}} e^{\left(\frac{-(y-X\beta)^T(y-X\beta)}{2\sigma^2}\right)} \quad (6)$$

The prior for β is chosen to be Multivariate Normal Distribution with mean 0 and covariance matrix $c_0 I_{p+1}$; i. e., $\beta \sim MVN(0, c_0 I_{p+1})$. c_0 is usually chosen to be a large positive value that leads to a large variance⁸. The prior for σ^2 is chosen to be inverse Gamma; i. e., $\sigma^2 \sim IG(a_0, b_0)$. a_0 and b_0 are the initial values that can be any positive numbers. The conditional posterior distribution for β can be written as follows,

$$\pi(\beta|\sigma^2, D) \propto L(y) \times \prod(\sigma^2|\beta, D)$$

$$\propto \text{EXP}\left(-\frac{1}{2}\left[\frac{(y-X\beta)^T(y-X\beta)}{\sigma^2} + \frac{\beta^T\beta}{c_0}\right]\right)$$

$$\propto \text{EXP}\left(-\frac{1}{2}\left[\frac{y^T y - y^T X\beta - \beta^T X^T y + \beta^T X^T X\beta}{\sigma^2} + \frac{\beta^T\beta}{c_0}\right]\right)$$

$$\propto \text{EXP}\left(-\frac{1}{2}\left[\frac{1}{\sigma^2}(-y^T X\beta - \beta^T X y + \beta^T X^T X\beta) + \frac{\beta^T\beta}{c_0}\right]\right)$$

$$\propto \text{EXP}\left(-\frac{1}{2}\left[\frac{1}{\sigma^2}(-2\beta^T X^T y + \beta^T X^T X\beta) + \frac{\beta^T\beta}{c_0}\right]\right)$$

$$\propto \text{EXP}\left(-\frac{1}{2}\left[\frac{-2\beta^T X^T y + \beta^T X^T X\beta}{\sigma^2} + \frac{\beta^T\beta}{c_0}\right]\right)$$

$$\propto \text{EXP}\left(-\frac{1}{2}\left[-2\beta^T \frac{X^T y}{\sigma^2} + \beta^T \left(\frac{X^T X}{\sigma^2}\right)\beta + \beta^T \left(\frac{I}{c_0}\right)\beta\right]\right)$$

$$\propto \text{EXP}\left(-\frac{1}{2}\left[-2\beta^T \frac{X^T y}{\sigma^2} + \beta^T \left[\frac{X^T X}{\sigma^2} + \frac{I}{c_0}\right]\beta\right]\right).$$

Consider $b = \frac{X^T y}{\sigma^2}$ and $A = \frac{X^T X}{\sigma^2} + \frac{I}{c_0}$, so the posterior distribution for β can be written as follows,

$$\pi(\beta|\sigma^2, D) \propto f(y) \times \prod(\sigma^2)$$

$$\propto \text{EXP}\left(\left(-\frac{1}{2}[\beta^T A\beta - 2\beta^T b]\right)\right)$$

$$\propto \text{EXP}\left(-\frac{1}{2}(\beta - p)^T \phi(\beta - p)\right)$$

$$\propto \text{EXP}\left(-\frac{1}{2}(\beta - A^{-1}b)^T A(\beta - A^{-1}b)\right) \quad (7)$$

Eq.7 is a MVN density with $\mu = A^{-1}b$ and $\Sigma = A^{-1}$. Hence, the full conditional posterior for β is

$$\beta \sim MVN(A^{-1}b, A^{-1}), \text{ where } A = \frac{X^T X}{\sigma^2} + \frac{I}{c_0}$$

$$\text{and } b = \frac{X^T y}{\sigma^2}$$

The full conditional posterior distribution for σ^2 is $\pi(\sigma^2|\beta, D)$

$$\propto (\sigma^2)^{-\frac{n}{2}} \text{EXP}\left(-\frac{1}{2\sigma^2} E\right) (\sigma^2)^{-(a_0+1)} \text{EXP}\left(-\frac{b_0}{\sigma^2}\right),$$

where $E = (y - X\beta)^T (y - X\beta)$

$$\propto (\sigma^2)^{-\frac{n}{2} - (a_0+1)} \text{EXP}\left(-\frac{1}{\sigma^2}\left(\frac{E}{2} + b_0\right)\right)$$

$$\propto (\sigma^2)^{-\left(\frac{n}{2} + a_0 + 1\right)} \text{EXP}\left(-\frac{1}{\sigma^2}\left(\frac{E}{2} + b_0\right)\right)$$

The above function is the density function of the inverse gamma distribution with a shape equal to $\frac{n}{2} + a_0$ and rate equal to $\frac{E}{2} + b_0$. i.e.,

$$\sigma^2 \sim IG\left(\frac{n}{2} + a_0, \frac{E}{2} + b_0\right) \quad (8)$$

So far, the posterior distributions for β (Eq.7) and σ^2 (Eq.8) have been derived. Hence, the estimated values for β and σ^2 can be found by calculating their sample means. This can be done using Markov chain Monte Carlo (MCMC) without calculating the marginal likelihood for β and σ^2 . In the following section, a brief discussion of MCMC is given, and then a particular case from MCMC (Gibbs sampler) is explained in detail.

Markov Chain Monte Carlo

MCMC is an essential technique, and it is used frequently in many statistical applications. In many cases, it is challenging to sample from a target posterior density. Then MCMC is used⁹. There are three popular MCMC sampling techniques, such as Metropolis-Hastings, slice sampling¹⁰, and Gibbs sampling¹¹. MCMC methods are derived from a Monte Carlo (MC)¹². A chain is used to approximate samples of desired distribution in MCMC, and the approximation is generally improved after several steps have been done¹³.

Gibbs Sampler

In Bayesian inference, Gibbs sampling is commonly used by statistical inference without calculating the marginal likelihood function. A high dimensional problem can be broken down into numbers of low dimensional problems when Gibbs sampler is used. The vector of parameters should be split into several blocks, and then each block can

besampled from its conditional distribution given other blocks. That means Gibbs sampling generated posterior samples by sweeping through each block variables¹⁴. Gibbs sampling issimilar to the other MCMC algorithms that generate a chain of samples where each of them is correlated with its nearby samples. Therefore, if the independent samples are desired, the samples should be thinned to get an independent sample set⁷. Suppose there are n parameters $\theta_1, \theta_2, \dots, \theta_n$, Gibbs sampling can estimate the parameters by updating them one by one. Evaluating the joint posterior $f(\theta_1, \theta_2, \dots, \theta_n | Data)$ is the first step, and it can be done by multiplying the likelihood with the prior $f(\theta_1), f(\theta_2), \dots, f(\theta_n)$. For instance, the conditional posterior for θ_1 , $f(\theta_1 | \theta_2, \theta_3, \dots, \theta_n)$, can be found by assuming $\theta_2, \dots, \theta_n$ are fixed at current values. This process should be repeated for all the parameters $\theta_2, \theta_3, \dots, \theta_n$. Algorithm 1 summarizes Gibbs sampler steps¹⁵.

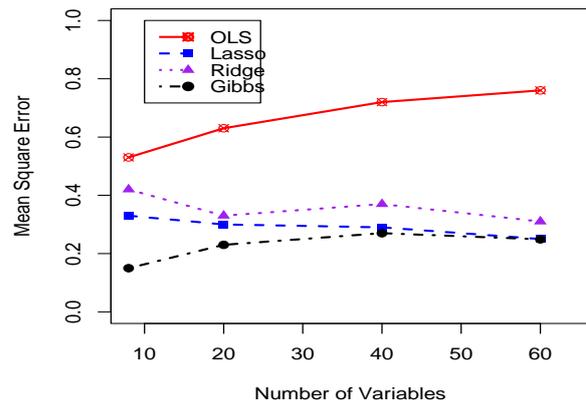
Algorithm: Gibbs Sampler

```

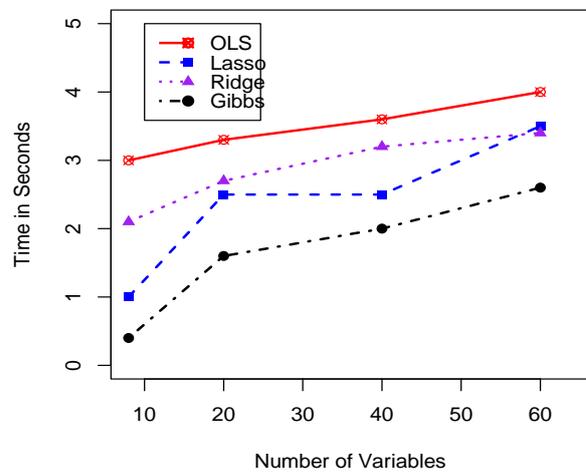
INPUT:  $\theta^{(i-1)} = (\theta_1^{(i-1)}, \theta_2^{(i-1)}, \dots, \theta_k^{(i-1)})$ 
OUTPUT:  $\theta^{(i)} = (\theta_1^{(i)}, \theta_2^{(i)}, \dots, \theta_k^{(i)})$ 
  for  $j = 1, 2, \dots, N$  do
     $\theta_1^{(j)} \sim f(\theta_1 | \theta_2^{(j-1)}, \theta_3^{(j-1)}, \dots, \theta_k^{(j-1)})$ 
     $\theta_2^{(j)} \sim f(\theta_2 | \theta_1^{(j)}, \theta_3^{(j-1)}, \dots, \theta_k^{(j-1)})$ 
     $\theta_3^{(j)} \sim f(\theta_3 | \theta_1^{(j)}, \theta_2^{(j)}, \dots, \theta_k^{(j-1)})$ 
     $\vdots$ 
     $\theta_k^{(j)} \sim f(\theta_k | \theta_1^{(j)}, \theta_2^{(j)}, \dots, \theta_{k-1}^{(j)})$ 
  end for
    
```

Simulation Studies

Four datasets with four different covariates (8, 20, 40, 60) were generated with 1000 cases. The covariates were simulated independently from a normal distribution, and then Eq.1 was used to find the responses for each dataset. The smallest dataset, 8 covariates, is discussed in detail, and the results of other datasets are summarized. Mean Squared Error (MSE) and time-complexity are represented in Fig.1. Figure 1a shows that Gibbs gives the lowest MSE in all four simulated datasets. Moreover, time consumption is checked for all datasets. Fig.1b shows that Gibbs uses less time compared to the other methods.



a. Comparing the MSE



b. Comparing the consuming time

Figure 1. Comparison among 4 methods (OLS, Lasso, Ridge, and Gibbs) in the 4 simulation datasets.

The true parameters for the first simulation dataset are: $\beta_0 = 1.1, \beta_1 = -2.2, \beta_2 = 4.3, \beta_3 = 1.2, \beta_4 = -2.2, \beta_5 = 6.7, \beta_6 = -1.3, \beta_7 = 3.3$ and $\beta_8 = 3.1$. The posterior distribution that has been derived in Eq.7 and 8 have been run with 10000 iterations. Simulated samples are thinned at every 5th sample to reduce the correlation between the samples. Both Gibbs sampler and Lasso methods are used to identify the most important variables from the 8 variables. Parameters are summarized from their corresponding posterior means, and some of them are very good estimators of the corresponding true value. In Fig.2, The samples are plotted as histograms, and the true values are marked with the blue lines. The distributions for some of the posterior samples are approximately normal. The true values of the parameters are close to the estimated parameters. The covariates

associated with $\beta_0, \beta_2, \beta_4, \beta_6$ and β_7 were selected as the most significant covariates because they were close to the true model coefficients, as shown in Table 1. However, in Lasso and Ridge methods, all the covariates were selected as important variables. Computationally, selecting all the variables as important variables is inefficient because both the error and time will increase for the large datasets.

Moreover, in Table 1, the parameters' actual values are compared with their corresponding posterior sample means. The 95% credible intervals (CI) are calculated for all the parameters. It is clear that the values that are not considered significant lie in large CI; On the other hand, all significant parameters are centered in narrow CI. This indicates that the estimation is reasonable and practical.

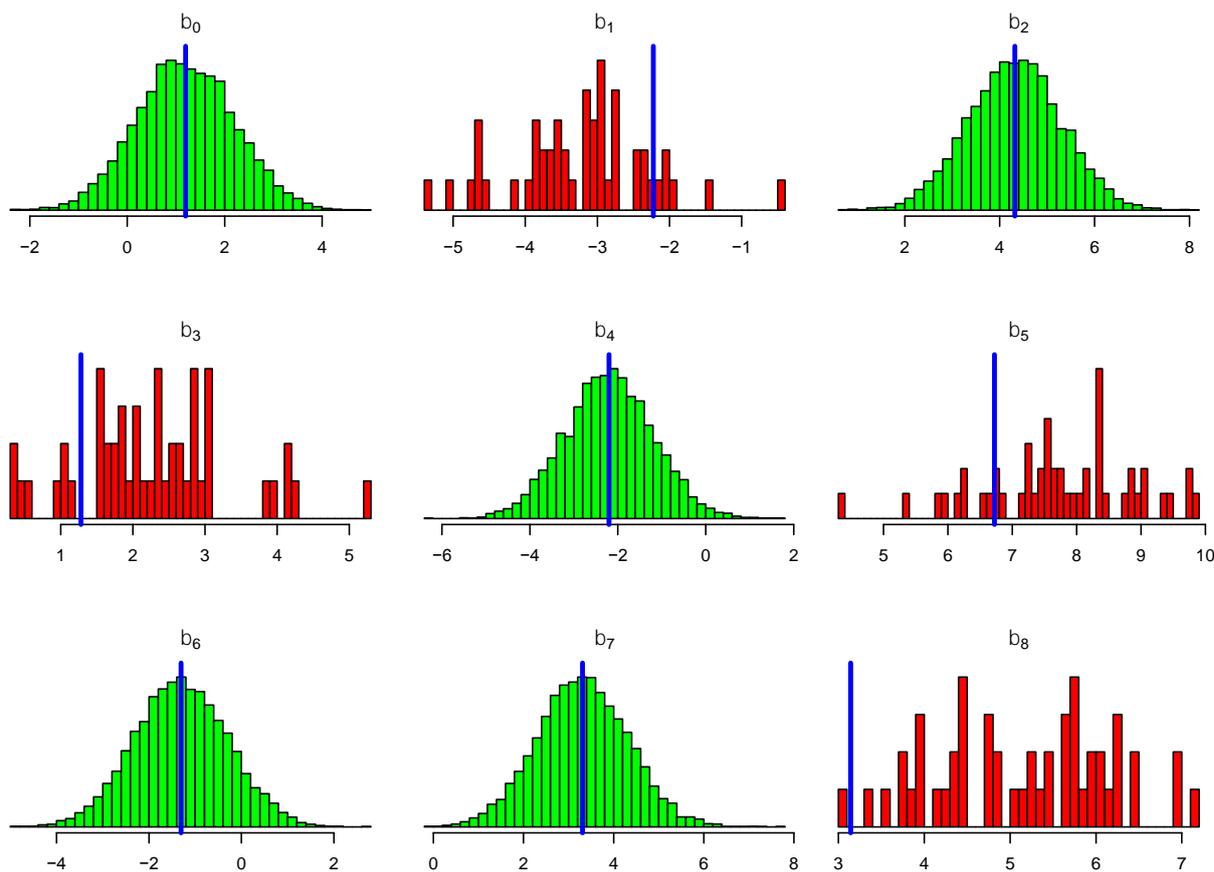


Figure 2. Posterior histograms for β_0, \dots, β_8 , blue lines denote the simulation's actual values.

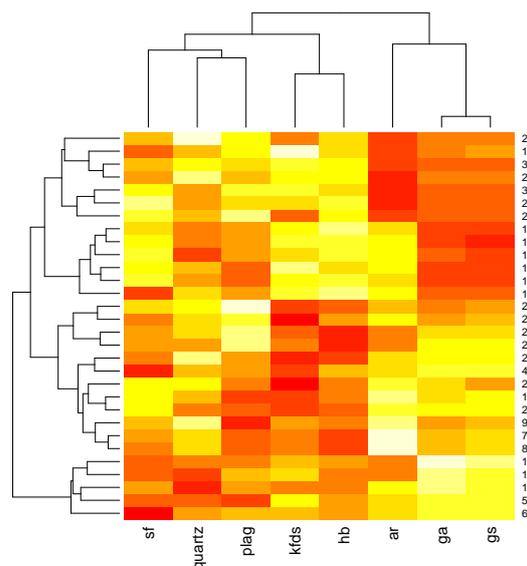
Table 1. True values, predicted values, and 95% credible intervals.

Parameter	True Value	Posterior Means	95% CI	Selected Variables (YES/NO)		
				Gibbs	Lasso	Ridge
β_0	1.1	1.003	(-0.534, 2.398)	YES	YES	YES
β_1	-2.2	-3.321	(-5.822, -1.272)	NO	YES	YES
β_2	4.3	4.221	(3.482, 5.448)	YES	YES	YES
β_3	1.2	2.322	(0.238, 4.230)	NO	YES	YES
β_4	-2.2	-2.331	(-3.238, 0.382)	YES	YES	YES
β_5	6.7	8.263	(5.239, 9.384)	NO	YES	YES
β_6	-1.3	-1.294	(-3.484, 1.823)	YES	YES	YES
β_7	3.3	3.309	(1.349, 5.282)	YES	YES	YES
β_8	3.1	4.872	(3.392, 7.849)	NO	YES	YES

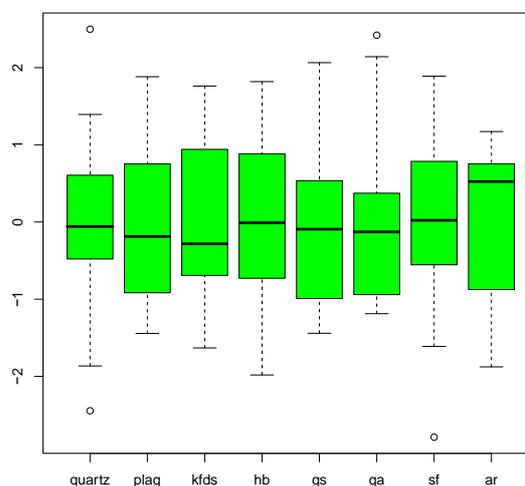
Real data Analysis

In this section, the Rock Strength Dataset (RSD) is analyzed. RSD contains information regarding the relationship between 8 predictors, which are %Quartz (**quartz**), %Plagoclase (**plag**), %K. feldspar (**kfds**), %Hornblende (**hb**), Grain size (**gs**), Grain area (**ga**), Shape Factor (**sf**), Aspect Ratio (**ar**), and the response, Uniaxial Compressive Strength (**UCS**), for 30 rock specimens. The dataset is collected from the **UCI Machine Learning Repository**.

Figure 3a shows the heatmap for the 8 predictors. The heatmap did not give us sufficient information about the data. As can be seen, the level of correlation is represented across all samples. The orange color represents the high correlation, and the low correlation is marked with yellow color. The dataset was normalized, and then predictors' boxplots are plotted. In Fig. 3b, some outliers in the dataset are realized. So, they are removed before running Gibbs and Lasso variables selection methods. The correlation matrix for the 8 predictors in the real data set (**RSD**) is given in Fig. 4. Most of the covariates are approximately normally distributed. **gs** and **ga** have a strong positive correlation, while **plag** and **kfds** have a very low correlation.



a. Heatmap for SRD variables



b. Boxplot for SRD variables

Figure 3. Heatmap and boxplot for the 8 predictors in SRD.

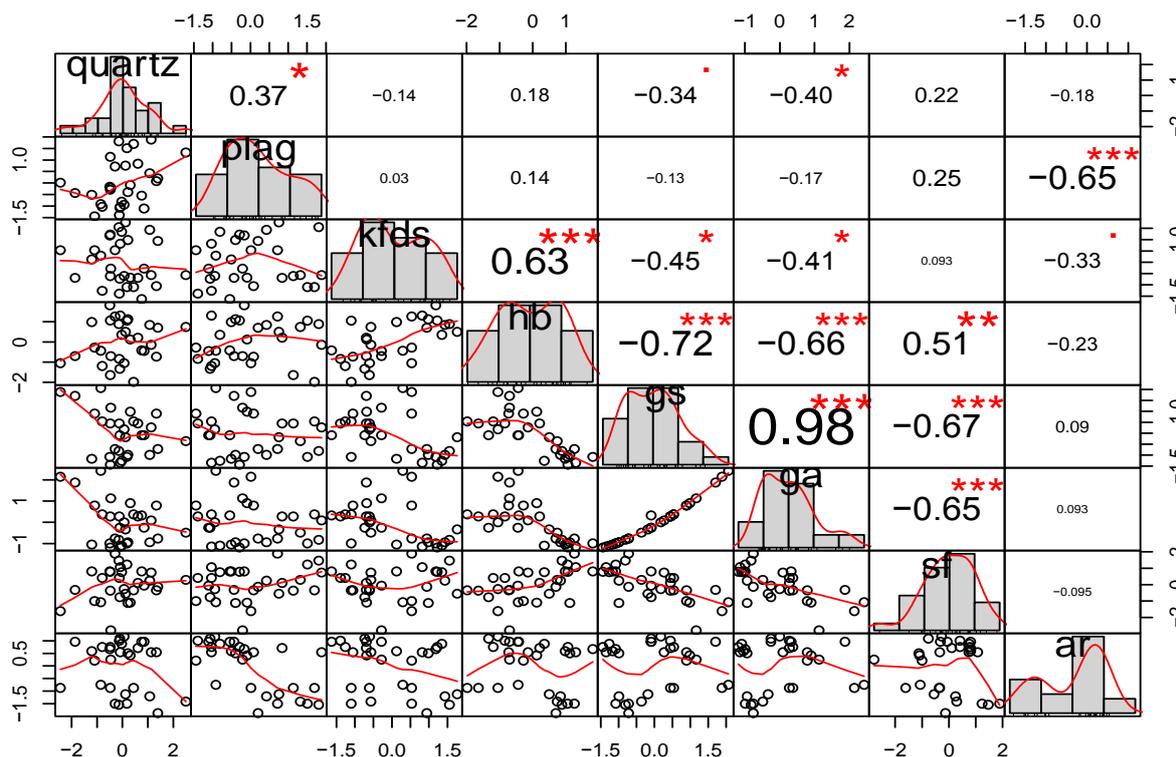


Figure 4. correlation matrix for the 8 predictors in SRD and their distributions.

Result and Discussion of the Variable Selection in RSD

Gibbs sampler has been used to select the essential variables in RSD. If all the variables are used, then the general multivariate model should be

$$y = \beta_0 + \beta_1 \text{quartz} + \beta_2 \text{plag} + \beta_3 \text{kfds} + \beta_4 \text{hb} + \beta_5 \text{gs} + \beta_6 \text{ga} + \beta_7 \text{sf} + \beta_8 \text{ar} + \epsilon$$

where $\epsilon \sim MNN(O_n, I_n)$

In the beginning, the dataset is normalized. Gibbs sampler is run for 1000 iterations to create the posterior samples of the parameters. The prediction performance is checked by using leave-one-out-cross validation (LOOCV)(10). In LOOCV, one of the observations is left out, and the model's coefficients are estimated with the rest of the observations. Since the real data has only 30 observations, the

procedure is repeated 30 times. It was found that 5 out of 8 posterior means lie inside the 95 percent credible intervals. These variables were selected as important predictors. Therefore, using the Gibbs method, the new model becomes

$$y = \beta_0 + \beta_1 \text{quartz} + \beta_3 \text{kfds} + \beta_5 \text{gs} + \beta_7 \text{sf} + \epsilon$$

where $\epsilon \sim MNN(O_n, I_n)$

OLS, Ridge, Lasso, and Gibbs selection methods were applied on the RSD, and the outputs are shown in Table 2. Gibbs method gives the smallest MSE, and takes less time compared to the other methods. Table 3. shows that the posterior means for the parameter are represented with their 95 percent CI. Gibbs selects 4 variables as essential variables: **quartz, kfds, gs, and sf**, while Lasso selects 6 variables, and Ridge selects all the 8 variables.

Table 2. Comparison between OLS, Ridge, Lasso, and Gibbs selection methods

Parameter	OLS	Ridge	Lasso	Gibbs
MSE	2.311	1.934	0.773	0.460
Time	≈8 seconds	≈6 seconds	≈3 seconds	<1 second

Table 3. Posterior means, 95% CI, and the selected variables.

Parameter	Posterior Means	95% CI	Selected Variables (YES/NO)		
			Gibbs	Lasso	Ridge
β_0	104.943	(103.533, 106.232)	YES	YES	YES
β_1 quartz	2.351	(0.375, 4.234)	YES	YES	YES
β_2 plag	-3.685	(-3.561, 13.873)	NO	NO	YES
β_3 kfds	-0.396	(-2.832, 2.134)	YES	YES	YES
β_4 hb	8.566	(-4.054, 7.184)	NO	YES	YES
β_5 gs	-3.124	(-4.334, 1.334)	YES	YES	YES
β_6 ga	-5.604	(-3.442, 12.497)	NO	YES	YES
β_7 sf	2.761	(0.476, 4.512)	YES	YES	YES
β_8 ar	9.268	(-2.442, 7.583)	NO	NO	YES

Conclusions:

A new variable selection approach using the Gibbs sampler has been discussed in this article. The posterior distributions for β and σ^2 have been derived, and the Gibbs sampler algorithm is used to sample from the corresponding distributions. The simulation datasets show that the Gibbs sampler is better than other existing methods (Lasso and Ridge, OLS). In both simulations and real datasets, the variable selection using Gibbs is faster and gives less error. As shown in SRD, the new method performs better than the other strategies by selecting only 50 percent of the variables; in contrast, Lasso and Ridge have selected 75 percent and 100 percent of the variables; respectively, with less accuracy and more time-consuming.

Authors' declaration:

- Conflicts of Interest: None.
- We hereby confirm that all the Figures and Tables in the manuscript are mine ours. Besides, the Figures and images, which are not mine ours, have been given the permission for re-publication attached with the manuscript.
- Ethical Clearance: The project was approved by the local ethical committee in University of Baghdad.

Authors' contributions statement:

Ghadeer J. M. Mahdi contributed to the design and implementation of the research, to the analysis of the results and to the writing of the manuscript. Othman M. Salih performed the computations and verified the analytical methods. All authors discussed the results and contributed to the final manuscript.

References:

1. Bahassine S, Madani A, Al-Sarem M, Kissi M. Feature selection using an improved Chi-square for Arabic text classification. JKU. 2018 May 24.
2. Surendiran B, Vadivel A. Feature selection using stepwise ANOVA discriminant analysis for mammogram mass classification. IJRTET. 2010 May;3(2):55-7.
3. Sutter JM, Kalivas JH. Comparison of forward selection, backward elimination, and generalized simulated annealing for variable selection. MICJ. 1993 Feb 1;47(1-2):60-6.
4. Pierna JA, Abbas O, Baeten V, Dardenne P. A Backward Variable Selection method for PLS regression (BVSPS). Analytica chimica acta. 2009 May 29;642(1-2):89-93.
5. Piepho HP. Ridge regression and extensions for genomewide selection in maize. Crop Science. 2009 Jul 1;49(4):1165-76.
6. Bair E, Hastie T, Paul D, Tibshirani R. Prediction by supervised principal components. JASA. 2006 Mar 1;101(473):119-37.
7. Chimisov C, Latuszynski K, Roberts G. Adapting the Gibbs sampler. arXiv preprint arXiv:1801.09299. 2018 Jan 28.
8. Mandt S, Hoffman MD, Blei DM. Stochastic gradient descent as approximate Bayesian inference. JMLR. 2017 Jan 1;18(1):4873-907.
9. Mahdi GJ. A Modified Support Vector Machine Classifiers Using Stochastic Gradient Descent with Application to Leukemia Cancer Type Dataset. BSJ. 2020 Dec 1;17(4):1255-1266. DOI: 10.21123/bsj.2020.17.4.1255.
10. Al-Sharea Z. Bayesian Model for Detection of Outliers in Linear Regression with Application to Longitudinal Data. Thesis, 2017.
11. Syring N, Hong L, Martin R. Gibbs posterior inference on value-at-risk. Scandinavian Actuarial Journal. 2019 Aug 9;2019(7):548-57.
12. Zhang Q, Mahdi G, Tinker J, Chen H. A graph-based multi-sample test for identifying pathways associated with cancer progression. Computational Biology and

- Chemistry. 2020 May 26:107285. DOI: 10.1016/j.combiolchem.2020.107285.
13. Van Ravenzwaaij D, Cassey P, Brown SD. A simple introduction to Markov Chain Monte-Carlo sampling. *PB&R*. 2018 Feb 1;25(1):143-54.
14. Mahdi GJ, Chakraborty A, Arnold ME, Rebelo AG. Efficient Bayesian modeling of large lattice data using spectral properties of Laplacian matrix. *Spatial statistics*. 2019 Mar 1;29:329-50. DOI: 10.1016/j.spasta.2019.01.003.
15. Efthymiou C, Hayes TP, Stefankovic D, Vigoda E, Yin Y. Convergence of MCMC and loopy BP in the tree uniqueness region for the hard-core model. *SIAM Journal on Computing*. 2019;48(2):581-643.

اختيار المتغيرات باستخدام خوارزمية Gibbs المطورة وتطبيقها على بيانات Rock Strength

عثمان مهدي صالح

غدير جاسم محمد مهدي

قسم الرياضيات، كلية التربية للعلوم الصرفة - ابن الهيثم، جامعة بغداد، العراق.

الخلاصة:

اختيار المتغيرات مهمة ضرورية ومطلوبة في مجال النمذجة الإحصائية. حاولت العديد من الدراسات تطوير وتوحيد طرق اختيار المتغيرات، ولكن من الصعب القيام بذلك. السؤال الأول الذي يحتاج الباحث أن يسأل نفسه عنه هو ما هو أهم المتغيرات التي يجب استخدامها لوصف الاستجابة لمجموعة بيانات معينة. في هذا العمل، تمت مناقشة طريقة جديدة في الاستدلال بايزي لاختيار المتغيرات باستخدام تقنيات عينات Gibbs. بعد تحديد النموذج، تم اشتقاق التوزيعات الخلفية لجميع المعلمات. تم اختبار طريقة الاختيار للمتغير الجديد باستخدام 4 مجاميع من البيانات. تمت مقارنة الطريقة الجديدة مع بعض الطرق المعروفة التي هي قليل مربعات الخطأ (OLS)، عامل انكماش مطلق واختيار (Lasso)، وتسوية تيكونوف (Ridge). أظهرت دراسات المحاكاة أن أداء طريقتنا أفضل من الأخرى حسب الخطأ ووقت الاستهلاك. تم تطبيق الطرق على مجموعة بيانات Rock Strength، وكانت الطريقة الجديدة التي تم تقديمها أكثر كفاءة ودقة.

الكلمات المفتاحية: اختيار المتغيرات، طريقة المربعات الصغرى، طريقة الانكماش، خوارزمية Gibbs، نظرية بايز.