International Journal of Mathematics Trends and Technology
ISSN: 2231 – 5373/doi:10.14445/22315373/IJMTT-V68I2P504

Volume 68 Issue 2, 19-27, February, 2022

Original Article

A Comparative Study of Bayesian Stochastic Search Variable Selection Approach in Multiple Linear Regression

Christabel Nyanchama Bisonga[1], Oscar Owino Ngesa[2] and Martine Odhiambo Oleche[3]

[1]*Depertement of Mathematics, Pan African University Institute for Basic Sciences, Technology and Innovation/ Jomo Kenyatta University of Agriculture and Technology, Kenya*
[2]*Department of Mathematics and Statistics, Taita Taveta University, Taita Taveta, Kenya*
[3]*Department of Statistics and Economics, University of Nairobi, Nairobi, Kenya.*

Abstract *— The presence of insignificant predictors in models causes estimation bias and reduces prediction precision. Collinearity among predictors is a common problem that renders the design matrix unstable leading to unreliable OLS coefficient estimates. Multiple linear regression analysis in a non-regularized routine is unsatisfactory due to poor prediction as the inclusion of all variables reduces noise but increases variance and for interpretation, it becomes necessary to identify the important predictors that have a high influence on the response variable. The study implements the Bayesian Stochastic Search Variable selection (B-SSVS) algorithm in the context of multiple linear regression with the incorporation of a correlation factor prior specification to address the correlation problem which reduces the performance of the Markov chain Monte Carlo and Gibbs sampling process. Further, comparative analysis on variable selection performance with classical penalized methods Elastic Net and Least Absolute Shrinkage Selection Operator (Lasso) is done using simulated data. We found that B-SSVS with a correlation factor prior showed good performance, mixing and convergence properties based on the diagnostic tests. B-SSVS performed better in variable selection compared to Elastic Net and Lasso shrinkage methods. We also found out that Elastic Net outperforms Lasso in detecting the true predictors and has less cross-validation mean squared error.*

Keywords *— Bayesian theory, Classical penalized methods, Gibbs Sampling, Markov Chain Monte Carlo, Stochastic Search Variable selection.*

I. INTRODUCTION

Multiple linear regression model fitting is a common method used when studying the relationship between continuous dependent and independent variables because it is simple and computationally cheap. However, due to the inability to define the distribution effects of each variable, redundant variables are included which require model selection to improve the accuracy and parsimony of the final model. Finding a parsimonious model is a function of variable selection where there exist unknown subsets of predictors that are irrelevant and redundant. According to [1] statistical modeling is aimed at fitting a model with a minimized number of variables which gives a better description of the data and also results in numerical stability. [2] Terms finding of parsimonious models a variable selection problem that helps identify the explanatory variables of important correlation to the dependent variable statistically and in practice.

Reference [3] notes that selected predictors in variable selection should be diverse with minimal collinearity. Common subset selection methods are filter types that use the approach of pre-ordering stepwise search and data-driven ranking of variables causing competition bias and selecting a single model that underestimates uncertainty about quantities of interest [4]. Following [5] the single model approach for variable selection tends to select inflated coefficients and ignore uncertainty. Further failure to account for correlation among predictors tends to results in the inclusion of highly correlated redundant variables at the cost of omitting the significant ones that can improve the predictive performance [6]. Penalized or shrinkage methods such as ridge regression, least absolute shrinkage selection operator (Lasso), Elastic net rely on a parameter which there is no consensus on how to determine its suitable value making the methods unstable [7].

II. LITERATURE REVIEW

Reference [8] analyzed the use of SSVS as first introduced by George and McCulloh. Their research shows how SSVS fused ideas used in hierarchical prior designs and Gibbs sampling under data augmentation to select variables in the case of ignorable mechanisms on missing data. They outline how the binary inclusion indicator $\gamma = 1$ or 0 allowed the setup of Gibbs

based algorithm for searching model space. Their paper outlines how to set up B-SSVS informative priors (σ^2, β_γ, γ)and hyper-parameter (τ^2, c2) specification for variable selection where they point out the importance of logical augmentation when setting priors and hyper-parameters and their priors depending on the characteristic of the data set.

Reference [9] looked into two variable selection approaches; Imputation Then Select (ITS) and Simultaneously Impute And select (SIAS) in the case of incomplete data under the missing mechanisms MCAR and MAR. The proposed Bayesian SSVS procedures were done using the Gibbs Sampling algorithm on the linear regression model. In their study out of 13 available variables only 5 were selected as per the threshold MIP = 0.5 i.e. only those variables in the sample with MIP > 0.5 were considered. SIAS outperformed ITS as it generated smaller standard errors for all explanatory variables in the final model selected and compared to other models from stepwise selection and criterion-based model comparison.

Reference [10] developed the SSVS algorithm for variable selection in psychology and compared the proposed algorithm to frequentist methods such as bivariate correlation and Lasso. SSVS proved to be a more appropriate and useful method as it caters for model and parameter uncertainty. SSVS set of selected predictors explain the response variable (pain unpleasantness) more by 13% than the models from the other two methods, also SSVS includes a very significant neuron-biological predictor of pain unpleasantness consistent with pain literature, unlike the other methods that excluded it. Further, the findings show that if a predictor is selected by SSVS then it has a high likelihood of being picked by other methods i.e. SSVS was a subset of Bivariate correlation and Lasso with 91% − 99% inclusion in Lasso of predictors from SSVS versus 17% − 43% those chosen on Lasso being included in SSVS. The researcher suggests the use of narrower predictor priori if the information is available as opposed to flat uninformative priori.

Reference [11] proposed a Bayesian selection procedure for quantile regression using simple and efficient stochastic search variables and illustrated the method using simulated Boston Housing data. The approach used asymmetric Laplace distribution which allowed the use of conditional conjugacy. The algorithm was computationally fast taking 17 seconds for the complete 11, 000 samples of Boston Housing data. The QR- SSVS outperformed the models selected with frequentist methods, with frequentist methods having a higher Type 1 error rate while the 95% confidence interval for βγ obtained through QR-SSVS containing true values in most simulations.

III. DATA AND METHODS.

A. Data

To evaluate and compare the variable selection performance of the B-SSVS and other properties, we mimicked and simulated data that assumed a scenario where the dependent variable is continuous and depends on a few of the simulated predictors in the sample set such that the others are irrelevant and redundant.[15].

1. Simulation 1: Using $x_j \sim N(0,1)$; j=1,2,3,.....8 of size n- sim (N)=100 as the predictor variables with exception of $x_3 = x_2 + 0.5Z$ where Z~ N(0,1) to yield a strong correlation $\rho > 0.9$ between and to help illustrate how B-SSVS performs in case of extreme Collinearity. y_i; i = 1, 2,..., 100 the dependent variable was simulated such that $y_i = x_{i1} + 1.2x_{i2} + \varepsilon_i$ where is the error term distributed as $\varepsilon_i \sim N(0, \sigma^2)$ with $\sigma = 2.5$

2. Simulation 2: For the second simulation sample data, the independent predictors was increased to p=30. For x_j ; j=1, 2, 3,.....p =30 of size N=100, the independent variable were generated to have some correlation such that $x_j = x_j^* + Z$ where x_j^* and Z are from independent normal distribution N (0,1) to yield a correlation $(x_j, x_{-j}) = 0.5$ to help also illustrate how B-SSVS and the other two methods to be compared (Lasso and Elastic Net) perform in case of collinearity. We simulated y_i;i = 1,2,....,N = 100 the continuous dependent variable such that $y_i = \sum_{j=1}^{p} x_{ij}\beta_j + \varepsilon_i$ where $\beta_j = 2.5$ for j = 1,2....6, $\beta_j = 2.0$ for j = 7,8,.....,14, $\beta_j = 1.5$ for j = 15,16,....,25 and $\beta_j = 0$ for j = 26,27,.....30 also ε_i is the error term distributed as $\varepsilon_i \sim N(0, \sigma^2)$ with $\sigma = 1.5$.

B. Methods

a) Multiple linear regression review

Regression analysis aims to explain a response variable Y using a set of covariates/ predictor variables $X = x_1, \ldots, x_p$ assumed we have a data set (Y, X) where Y is n by 1 vector and X is n by p matrix (n is the number of observations and p is the number of predictors). In Multiple Regression it follows that $Y \in R$ is a function of variables x1…, x p. and a parameter vector β, modeled as:

$$y_i = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j + \varepsilon_i$$

$\varepsilon_i \sim N(0, \sigma^2)$, iid for i=1,……n

The parameter $\beta = (\beta 0 + \beta 1 + \ldots\ldots + \beta p)'$ are the corresponding coefficients of the predictor variable x j; j = 1,……., P estimated by Maximum Likelihood, Matrix and Least Square methods. We differentiate the minimization of the residual sum of squares, quadratic equation of parameter p+1 with respect to β to get $\hat{\beta}$:

$$\sum_{i=1}^{n} \varepsilon_i^2 = RSS(\beta) = (Y - X\beta)'(Y - X\beta)$$

$$\frac{\partial RSS}{\partial \beta} = 2X'(Y - X\beta)$$

$$X'(Y - X\beta) = 0$$

$$\therefore \hat{\beta} = (X'X)X'Y$$

$$\hat{Y} = X\hat{\beta} = X(X'X)X'Y + \varepsilon$$

b) B-SSVS

The study utilizes the stochastic search variable selection (SSVS) a Bayesian approach pioneered by [12] setting the fusion idea of using hierarchical prior designs and Gibbs sampling under data augmentation to select variables. The aim is to obtain $P(\gamma_i = 1|Y)$ which is the weight of covariate x_j being included in the model space by considering all 2^p possible models.

Let consider a sample of n observations, the data takes the form (Y, X) whose relationship is defined by the normal linear model, with $n \times 1$ vector space for Y dependent variable and $X = x_1, \ldots\ldots, x_p$ of dimension $n \times p$ predictor matrix.
We denote the canonical model $Y = \beta_0 + X\beta + \varepsilon$ with $\varepsilon_i \sim N(0, \sigma^2)\beta$ and σ being the model parameters. Precisely $Y \sim N_n(\beta_0 + X\beta, \sigma^2 I)$. We index a binary indicator $\gamma = \gamma_1, \ldots\ldots\ldots\ldots, \gamma_p$ taking values 1 or 0 for each predictor.

; If $\gamma_j = 1$ then variable x_j is included in the model otherwise if $\gamma_j = 0$ then variable x_j is excluded

We get $y_i = \beta_0 + \gamma_1\beta_1 x_{i1} + \gamma_2\beta_2 x_{i2} + \cdots + \gamma_p\beta_p x_{ip} + \varepsilon_i$.

We assume some coefficients for β_j are small likely zero or in the neighborhood of zero thus can be ignored in the model. Hence given γ we obtained the regression of the form; $Y = \beta_0 + x_\gamma\beta_\gamma + \varepsilon$ that contains predictors x_j and coefficients β_j whose γ j = 1.

An important component in B-SSVS is the specification of the prior distribution (π) because they impact the estimation of the model which also affects the parameters chosen with their bias, standard deviation, coverage rates and mean squared error [11]. Bayesian analysis requires specification of prior distributions for each parameter included in the analysis i.e. $\pi(\beta_0), \pi(\beta|\sigma^2, \gamma), \pi(\sigma^2), \pi(\gamma)$.

From the canonical model where X′X is the correlation matrix, the B-SSVS prior on the binary variable selection indicator (γ = 1 or 0) was not specifically designed to handle correlated predictors in variable selection and in their papers ([8], [11], [12] and [13]) they noted that highly correlated predictors tend to reduce MCMC convergence. Following that, we introduce the adaptive correlation factor $(X'X)^\omega$. The power $\omega \in \Re$ is used to control how the priors will smooth out correlated covariates. With $\omega > 0$ and $\omega < 0$ making the collinear predictors smooth towards or away from each other respectively as detailed in [14].

c) Hierarchical Bayesian SSVS Model

Model selection in Bayesian is based on the posterior probabilities of the model given as the product of the prior probabilities of the model parameters $\pi(\theta)$, $\theta = \beta, \sigma^2, \gamma$ and the likelihood $\prod_{i=1}^{n} f(y_i|\beta, \sigma^2)$.
To perform variable selection the following hierarchical set up is used:

$$Y|\ \beta_0, \beta,\ \sigma^2, \gamma, X \sim N(\beta_0, + X_\gamma\beta_\gamma, \sigma^2)$$

$$\pi(\beta_0) \sim (\mu, \vartheta)$$

The prior of coefficients β_j, j = 1, 2…, p given the latent binary γ_j takes the mixture of two Gaussian distribution with different variances ss presented in [15] George & McCulloch (1997).

$$\pi(\beta_j \mid \gamma_j) \sim (1-\gamma_j)N(0,\tau_j^2) + \gamma_j N(0, c_j^2\tau_j^2)$$

$$; \gamma_j = 0,\ \beta_j \sim N(0,\tau_j^2),\ \gamma_j = 1,\ \beta_j \sim N(0, c_j^2\tau_j^2)$$

; τ_j is chosen to be very small and c_j^2 large so that if $\gamma_j = 1$, β_j is included in the model and $\gamma_j = 0$, β_j tend to the neighborhood of zero thus excluded.
Model space prior $\pi(\gamma) \sim$ Bernoulli (1,α):

$$\pi(\gamma) = \prod_{j=1}^{p} \alpha^{\gamma_j}(1-\alpha)^{1-\gamma_j}$$

For that, we introduced a correlation factor matrix from [14] as $(X'X)^{\omega}$ and based on [16] $\omega = \frac{1}{2}$, a penalty factored into the prior $\pi(\gamma)$ to account correlation between covariates to obtain:

$$\pi(\gamma|\ \alpha) \propto \prod_{j=1}^{p} \alpha^{\gamma_j}(1-\alpha)^{1-\gamma_j}\sqrt{\det(X_\gamma' X_\gamma)}$$

For σ^2 a conjugate uniform Jeffrey's prior on log (σ^2), i.e., $\pi(\sigma^2) = \sigma^{-2}$ is used for this study:

$$\pi(\sigma^2|\ \gamma) \sim \mathrm{IG}\left(\frac{\nu_\gamma}{2}, \frac{\lambda_\gamma \nu_\gamma}{2}\right)$$

with $\nu_\gamma > 0$ the shape and $\lambda_\gamma \nu_\gamma > 0$ the scale being parameters of the distribution to be specified.

Parameters in the specification of prior are called hyperparameters and their definition is important in the accurate specification of the prior distribution probabilities. These include; $c_j, \tau_j, \lambda_\gamma, \nu_\gamma \mu, \vartheta$ and α. Based on previous research studies we utilized the following values for the hyperparameters; ($\nu_\gamma = \lambda_\gamma, \nu_\gamma \equiv 0$) yielding non-informative prior for σ^2, $\alpha = 0.5$ = 0.5 for γ, $\frac{\sigma_{\beta_j}}{\tau_j} \approx 1$ hence obtainτ_j, c j ≡ 10 for β and $\mu = 0, \vartheta = 2$ for β_0.
The likelihood is the probability density function that is data-driven which for the linear regression model takes the following form:

$$L(Y|\square, \sigma^2) = \prod_{i=1}^{n} f(y_i\ |\beta, \sigma^2)$$

$$= \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} \exp\left\{\frac{-1}{2\sigma^2}(Y-X\beta)'(Y-X\beta)\right\}$$

$$\propto \left(\frac{1}{\sigma^2}\right)^{\frac{n}{2}} \exp\left\{\frac{-1}{2\sigma^2}(Y-X\beta)'(Y-X\beta)\right\}$$

With observed data let call it D = (X, Y) and by Bayes theorem, the posterior distribution for any parameter θ takes the form;

$$P(\theta\ |D_n) = \frac{P(D_n|\theta)\ \pi(\theta)}{P(D_n)} = \frac{L(\theta)\pi(\theta)}{c_n}$$

$L(\theta) = \prod_{i=1}^{n} f(y_i|\beta, \sigma^2) = P(D_n|\theta)$ is the likelihood function,
$\pi(\theta)$ Is the prior of the parameters.
$C_{n=}\ P(D_n) = \int P(D_n|\theta)\ \pi(\theta)d\theta = \int L(\theta)\ \pi(\theta)d\theta$ is the normalizing constant.

d) Markov Chain Monte Carlo (MCMC) and Gibbs Sampling

Markov chain Monte Carlo procedures as outlined in [17] and [18] are a class of algorithms that samples from probability distribution in this case the posterior by making iterative steps. [19] Notes that the MCMC approach aims to create a random walk that converges to the target posterior distributions. MCMC via Gibbs sampling algorithm simulates from the target

posterior distribution to generate samples of models of higher posterior probabilities. Gibbs sequence takes$\beta_{0,}^{0}$ β^0, σ^0, γ^0, $\beta_{0,}^{1}\beta^1$, $\sigma^1, \gamma^1, \ldots\ldots\ldots\ldots\ldots\ldots\ldots, \beta_{0,.}^{M}\beta^M$, σ^M, γ^M; where M is the total number of iterations. $\beta_{0,}^{0}\beta^0$ and σ^0 was initialized to be the least-squares estimates of MLR, and γ^0 was initialized as $\gamma^0 = (1,1,\ldots,1$ of dimension p), the subsequent values of $\beta_{0,}^{k}$, β^k, $\sigma^k, \gamma^k; k = 1,2,..,M$ are obtained by successively simulating values from the stationary posterior distribution.

e) Lasso

Using the linear model $Y = \beta_0 + \mathbf{X}\beta + \varepsilon$ where Y is the response variable, X is the input variables, and ε is the error term. The Lasso minimizes the residual sum of squares based on the L1 norm:

$$\left(\sum_{i=1}^{n} y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \cdots - \beta_p x_{ip}\right)^2$$

$$\text{Subject to } \lambda\left(\sum_{j=1}^{p} |\beta_j|^2\right)^{\frac{1}{2}} < s$$

which is the penalized sum of the absolute value of the coefficients being less than a constant. $\lambda \geq 0$ controls the amount of shrinkage with some coefficients tending towards zero i.e $\beta_j \rightarrow 0$ [20]. β_0 is excluded from penalty implying that ;

$$\beta_0 = \overline{Y} = {\sum_{i=1}^{n} y}/{n}$$

$$\hat{\beta} = \text{argmin}_{\beta}\left(||Y - X\beta||^2 + \lambda \sum_{j} |\beta_j|\right)$$

$$\hat{\beta} = (X'X - \lambda I)(X'Y$$

1. Elastic Net: The method combines the L1 penalty and L2 penalty on coefficients [21]. The ridge L2 encourages the grouping effect by averaging the correlated variables and the L1 penalty promotes sparsity. It follows that the 'Elastic Net' minimizes the residual sum of squares:

$$\left(\sum_{i=1}^{n} y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \cdots - \beta_p x_{ip}\right)^2$$

$$\text{Subject to } \lambda_1 \sum_{j=1}^{p} |\beta_j| < t \text{ and } \lambda_2 \sum_{j=1}^{p} \beta_j^2 < t$$

$$\text{Letting } \alpha = \frac{\lambda_2}{\lambda_1 + \lambda_2}$$

$$\hat{\beta} = \text{argmin}_{\beta}|Y - X\beta|^2 \text{ s.t } (1-\alpha)|\beta|_1 + \alpha|\beta|^2 < t$$

where $\lambda_1, \lambda_2 \geq 0$ controls the amount of shrinkage.

IV. RESULTS AND DISCUSSION

We assessed the performance of the BSSVS with the incorporated correlation factor prior for the latent binary parameter $= 1\ or\ 0$. The summary results from simulation 1 will aid in evaluating the B-SSVS ability to select true predictors associated with the outcome and the fit of the process in the variable selection process under the defined priors, likelihood and parameters defined in the previous chapter.

The main objective in statistics is to extract information from available data while ensuring high accuracy of the parameter estimators [22]. From Table 1 below the B-SSVS has high accuracy in estimating the posterior coefficient parameter mean values at a low standard error as evident in the β_1 , β_2 coefficients whose simulation and B-SSVS estimated values are (1, 1.2) and (0.9479,1.1023) respectively. Figure 1 and Table 1 contains posterior marginal inclusion probabilities (MIP) which shows the ability of the B-SSVS technique to detect the relevant variable, based on the MIP median rule to include variables whose $P(\gamma_j = 1 \mid X,Y) > 0.5$ only the relevant and significant predictor coefficients were selected. The MIP for β_3 is tending towards

the selection criteria but these can be attributed to its high correlation to the relevant variable X2. Further, the ability of B-SSVS not to include the false positive parameter β3 signifies its power to phase out irrelevant variables that are highly correlated to outcome associated covariates similar to findings of ([23] and [24]).

Table 1. Estimation results for β from the posterior sample of the BSSVS

	Mean	Std. Err	Cred.Interval at 95%	$P(\gamma_j = 1 \mid X,Y)$	Distribution
Intercept (β_0)	0.6389	0.0242	[0.591, 0.686]	1.000	Empirical
β_1	0.9479	0.0356	[0.941, 1.022]	1.000	Empirical
β_2	1.1023	0.0208	[1.041, 1.222]	1.000	Empirical
β_3	0.4389	0.0242	[0.391, 0.486]	0.474	Empirical
β_4	0.3389	0.0242	[0.311, 0.386]	0.014	Empirical
β_5	-0.0321	0.0320	[-0.061, 0.057]	0.354	Empirical
β_6	0.4310	0.0538	[0.446, 0.657]	0.237	Empirical
β_7	-0.0021	0.0300	[-0.011, 0.037]	0.174	Empirical
β_8	1.5410	0.0538	[1.446, 1.657]	0.384	Empirical
σ^2	0.8151	0.0811	[0.672, 0.988]	1.000	Empirical

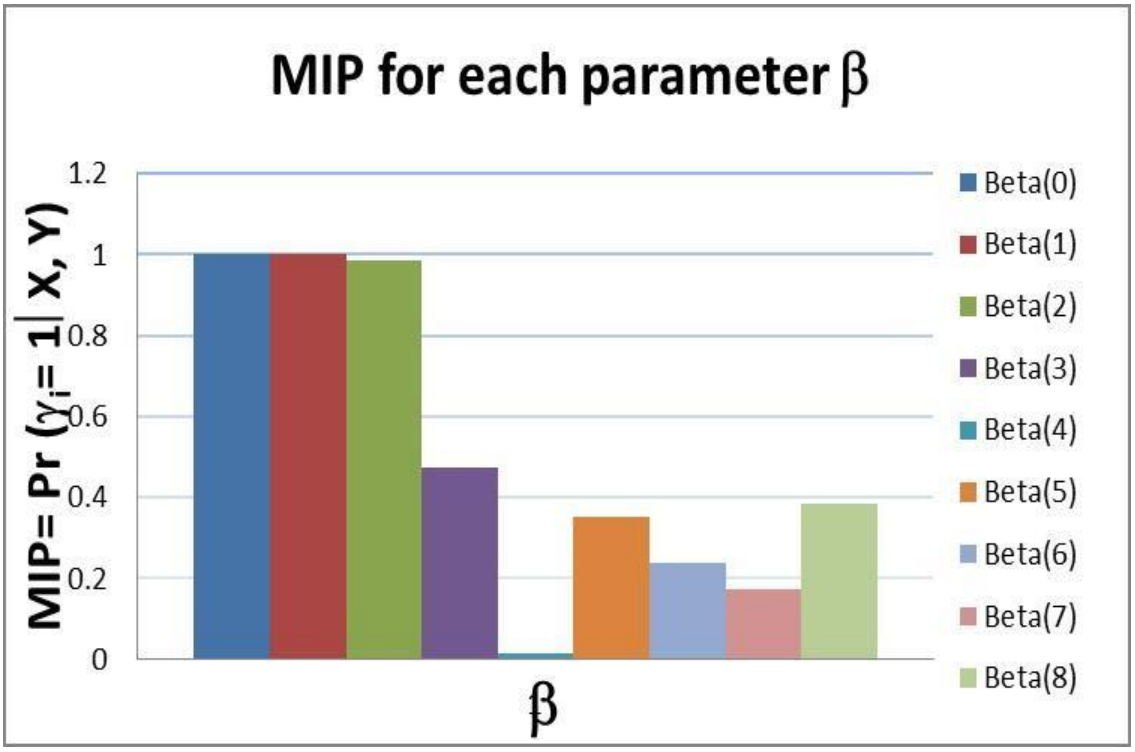

Fig. 1 Graphical presentation of marginal inclusion probability (MIP) of each β

The output for Table 2, 3 and 4 are summaries after the autocorrelations, Raftery-Lewis Diagnostics (q=0.025000, r=0.010000, s=0.950000), Geweke Diagnostics, Heidelberg Welch Diagnostics and Gelman Rubin Diagnostics met the required properties for convergence, stable and efficient MCMC sampling process.

Table 2. Estimation results for covariates from the posterior sample

Covariate	Mean	Std.dev	NSE	RNE
X1	0.993971	0.027528	0.001376	1.000000
X2	1.309894	0.070228	0.003511	1.000000
X3	-0.105704	0.055540	0.002777	1.000000
X4	-0.025452	0.030501	0.001525	1.000000
X5	0.053792	0.027939	0.001397	1.000000
X6	-0.110794	0.029613	0.001481	1.000000
X7	-0.015487	0.022197	0.001110	1.000000
X8	-0.005907	0.032155	0.001608	1.000000

Table 3. Progress summary of NSE and RNE across the Gibbs Sampling

Covariate	NSE 4%	RNE 4%	NSE 8%	RNE 8%	NSE 15%	RNE 15%
X1	0.000991	0.399020	0.000995	0.396237	0.000808	0.599878
X2	0.002778	0.396206	0.003355	0.271634	0.003086	0.321174
X3	0.003720	0.221012	0.003421	0.261266	0.003355	0.271634
X4	0.001314	0.227047	0.001116	0.314648	0.000995	0.396237
X5	0.003706	0.222615	0.003618	0.233605	0.003373	0.268729
X6	0.001272	0.242412	0.001317	0.225890	0.001201	0.271902
X7	0.002991	0.299020	0.002895	0.395237	0.002808	0.399878
X8	0.001891	0.229020	0.001095	0.266237	0.001008	0.279670

MCMC relative numerical efficiency (RNE) and numerical standard error (NSE) show the computation efficiency for the number of draws that are effective for posterior inference. Table 2 shows the mean estimates of the sampled X values with their respective NSE and RNE, across each row we see that each predictor was sampled and estimated at a very low near zero NSE and high RNE. Table 3 depicts the trend of the NSE and RNE at different % of the iteration process which clearly shows a decrease for NSE and an increase for RNE signifying an improving sampling chain as it converges to the required stationery distribution hence giving better posterior samples.

Table 4. Autocorrelation for covariates at different lags

Covariate	lag 1	lag 20	lag 50	lag 100
X1	0.494	-0.075	0.011	-0.011
X2	0.581	-0.008	-0.063	-0.014
X3	0.518	-0.094	-0.048	-0.047
X4	0.484	-0.091	-0.013	-0.116
X5	0.377	-0.080	0.024	-0.004
X6	0.425	0.097	-0.089	-0.089
X7	0.349	-0.148	-0.051	-0.018
X8	0.524	-0.082	-0.030	-0.105

Table 4 tells how much correlation exists between the MCMC draws. Autocorrelation should drop relatively quickly and decay to zero as the chain progresses. The lag k autocorrelation ρ^k is the correlation between every draw and its kth lag:

$$\rho^k = \frac{\sum_{i=1}^{n-k}(x_i - \bar{x})(x_{i+k} - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

The results, therefore, show that the B-SVS performed well as kth lag autocorrelation was smaller ask increased indicating a low degree of correlation between our draws and fast mixing. The chain requires thinning according to the lag to get independent samples applicable for posterior statistics inference.

In simulation 2, we ran an analysis using our different techniques (B-SSVS, Lasso, and Elastic Net). We looked into how the different techniques were performing in selecting relevant and active predictors that influence the dependent variable. The comparison result entails average predictors, selection power as per the true positive predictor and the tendency to type I error as per the false-positive predictor values.

Table 5. True positives (TPs) and False Positives (FPs) Predictors for each method

Method	Avg. no. of predictors	No. of TPs	No. of FPs
B-SSVS $P(\gamma_j = 1 \mid X, Y) > 0.5$	26.86	25.22	1.64
Lasso	28.44	22.34	6.10
Elastic Net	28.30	24.24	4.06

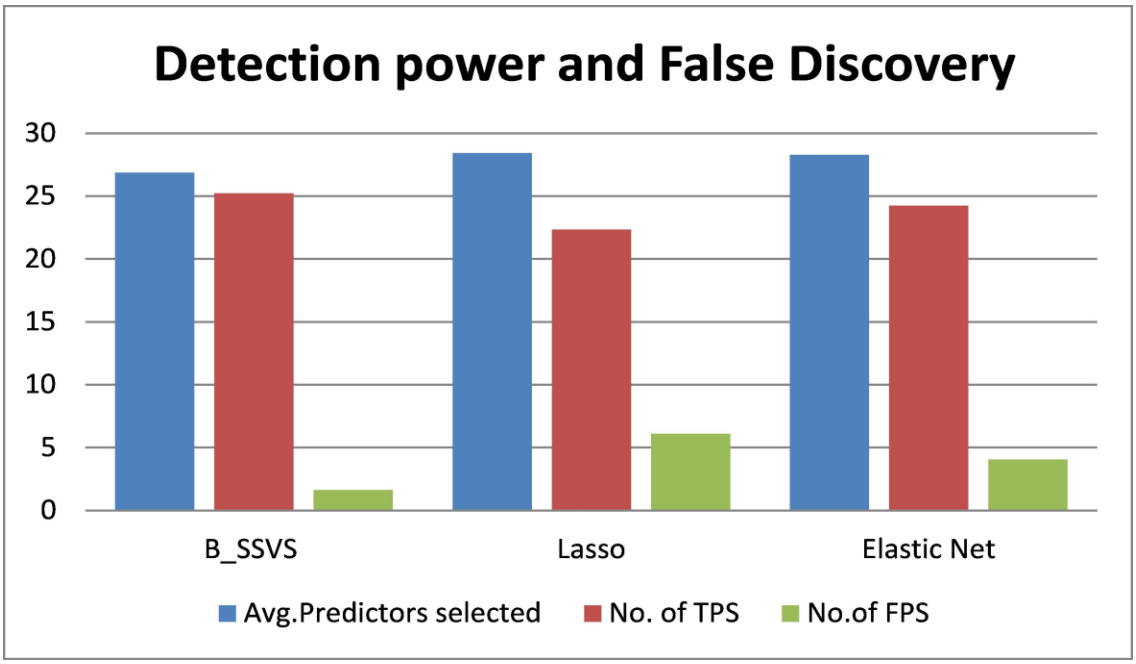

Fig. 2 Plots showing the number of average, TPs and FPs predictors selected by eachmethod

True positive predictors are those predictors in the model that have an actual and statistically significant effect on the outcome while False-positive predictors are the redundant and irrelevant ones included in the model. The value of true positives shows the selection power of a technique while the value for the false positives shows how the techniques can control false discovery rate and type I error [23]. From the simulation, 25 predictors are influencing the outcome. From Table 5 and Figure 2 B-SSVS (25.22, 1.64) can select near to accuracy the true positive predictors with minimal inclusion of the false-positive predictor followed by Elastic Net (24.34, 4.06) and lastly Lasso (22.34, 6.1)

V. CONCLUSION

With developments in technology and tools for data collection getting big data with many predictors is common and in research analysis, there is usually information asymmetry on which predictors to consider in modeling. Finding a technique that performs variable selection simultaneously rather than on individual cases was the baseline for the study to look into

B-SSVS which was compared to Elastic Net and Lasso methods. B-SSVS technique caters for correlation and uncertainty in variable selection as presence of collinearity yield 'poor' OLS estimates of the regression parameters [25]. The study has shown that B- SSVS with incorporated correlation factor prior that ensures Markov Chain is irreducible exhibits good mixing, convergence and high accuracy in selecting relevant predictors. Secondly, B-SSVS outperforms Elastic Net and Lasso, this is attributable to higher true positive predictor detection power augmented with lower false discovery hence reduced type I error, low numerical standard error with high relative numerical efficiency. On the Shrinkage methods; Elastic Net performed better than Lasso.

Due to the problems and limitations encountered in multiple regression and common classical subset selection techniques like R2, AIC, BIC, t-statistic, stepwise, forward and backward regression improvements are desirable. [26] Notes that even two-stage least squares method is less sensitive to both specification error and multiple co-linearity. B-SSVS has shown good performance in variable selection which helps improve prediction accuracy, interpretability in analysis and modeling. Elastic Net and Lasso are also averagely effective for the variable selection but do not distinctively identify true predictors related to the outcome as they tend to have a higher false-positive predictor inclusion (type I error) than the Bayesian approach. The B-SSVS with correlation prior can be extended to generalized methods where the outcome takes other forms, like binary, ordinal and count. Although, the study was based on scenarios where $p < n$ other researchers especially in genes selection studies have shown viable extension to $p > n$ scenarios.

ACKNOWLEDGMENT

This work was supported by the Pan African University.

REFERENCES

[1] Ijomah Maxwell. A., & Nwali Obinna, A. Comparative Study of Some Variable Selection Techniques In Logistic Regression. European Journal of Mathematics and Com38puter Science, (2018).
[2] Chipman, H., George, E. I., McCulloch, R. E., Clyde, M., Foster, D. P., and Stine, R. 53 A. The practical implementation of Bayesian model selection. Lecture Notes Monograph Series JSTOR , (2001) 65-134
[3] Kojima, M., and Komaki, F. Determinantal point process priors for Bayesian variable selection in linear regression. Statistica Sinica, (2016) 97-117.
[4] Raftery, A. E. Bayesian model selection in social research. Sociological methodology, (1995) 111–163.
[5] Swartz, M. D., Yu, R. K., and Shete, S. Finding factors influencing risk: comparing Bayesian stochastic search and standard variable selection methods applied to logistic regression models of cases and controls. Statistics in medicine, 27(29) (2008) 6158-6174.
[6] Kwon, D., Landi, M. T., Vannucci, M., Issaq, H. J., Prieto, D., and Pfeiffer, R. M. An efficient stochastic search for Bayesian variable selection with high-dimensional correlated predictors. Computational statistics & data analysis, 55(10) (2011) 2807-2818
[7] Wang, Y., and Witten, I. H. Pace regression (Working paper 99/12). Hamilton, New Zealand: University of Waikato, Department of Computer Science, (1999).
[8] Perrakis, K., & Ntzoufras, I. Stochastic Search Variable Selection (SSVS). Wiley. DOI: 10.1002/9781118445112.stat07829
[9] Yang, X., Belin, T. R., & Boscardin, W. J., Imputation and Variable Selection in Linear Regression Models with Missing Covariates. Biometrics, 61 (2015) 498–506. DOI: 10.1111/j.1541- 0420.2005.00317.x
[10] Bainter, S. A., McCauley, T. G., Wager, T., and Losin, E. A. Improving practices for selecting a subset of important predictors in psychology: An application to predicting pain. Advances in Methods and Practices in Psychological Science, 3 (2020) 66–80. 80. DOI:10.1177/251524591988
[11] Chen, C. W., Dunson, D. B., Reed, C., and Yu, K. Bayesian variable selection in quantile regression. Statistics and its Interface, 6 (2013) 261–274.
[12] George, E. I., and McCulloch, R. E. Variable selection via Gibbs sampling. Journal of the American Statistical Association, 88(423) (1993) 881-889.
[13] Kojima, M., & Komaki, F. Determinantal point process priors for Bayesian variable selection in linear regression. Statistica Sinica, (2016) 97-117.
[14] Krishna, A., Bondell, H. D., and Ghosh, S. K. Bayesian variable selection using an adaptive powered correlation prior. Journal of statistical planning and inference, 139(8) (2009) 2665-2674
[15] George, E. I., and McCulloch, R. E. Approaches for Bayesian variable selection. Statistica Sinica, (1997) 339–373.
[16] Yuan, M., and Lin, Y. Efficient empirical Bayes variable selection and estimation in linear models. Journal of the American Statistical Association, 100(472) (2005)1215-1225.
[17] Vats, D., and Knudson, C. Revisiting the Gelman-Rubin diagnostic. Statistical Science, 36(4): (2021) 518—529.
[18] Geweke, J., Gowrisankaran, G., and Town, R. J. Bayesian inference for hospital quality in a selection model. Econometrica, 71 (2003) 1215–1238.
[19] Cowles, M. K., and Carlin, B. P. Markov Chain Monte Carlo Convergence Diagnostics: A Comparative Review. Journal of the American Statistical Association, 91 (1996) 883–904. DOI:10.1080/01621459.1996.10476956
[20] Tibshirani, R. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological), 58 (1996) 267–288. DOI; 10.1111/j.2517- 6161.1996.tb02080.x
[21] Zou, H., and Hastie, T. Regularization and variable selection via the elastic net. Journal of the royal statistical society: series B (statistical methodology), 67(2) (2005) 301-320.
[22] Ijomah M., A., and Chris-Chinedu, J., N. Jackknife And Bootstrap Techniques In The Estimation of regression Parameters" International Journal of Mathematics Trends and Technology 65(12) (2019) 25-35
[23] Srivastava, S., and Chen, L. Comparison between the stochastic search variable selection and the least absolute shrinkage and selection operator for genome-wide association studies of rheumatoid arthritis. In BMC Proceedings 3(7) (2009) 1-7. Biomed Central.
[24] Lin, C. Y. Stochastic search variable selection for split-plot and blocked screening designs. Journal of Quality Technology, 53(1) (2021) 72-87.
[25] Martin L., W., and L.Maria Alphonse Ligori ., A Modified Least-Squares Approach to Mitigate the Effect of Collinearity in Two- Variable Regression Models, International Journal of Mathematics Trends and Technology (IJMTT), V30(1) (2016) 48-ISSN:2231-5373. www.ijmttjournal.org. Published by Seventh Sense Research Group.
[26] Qingli Pan., An Improved Two-Stage Estimator of Simultaneous Equations Models, International Journal of Mathematics Trends and Technology, 65(1) (2019) 53-56.