Journal of Applied Statistics & Machine Learning
Vol. 3, Nos. 1-2, 2024, pp. 41-61, ISSN: 2583-2891

URL : www.blackroseindia.com/jasml

Bayesian Variable Selection with Spike-Slab prior in Random effect Model

Md Maidul Husain and Anwar Hossain*
Department of Mathematics, New Mexico Tech
**Department of Mathematics, New Mexico Tech, Socorro, NM 87801*

Received : 13 September 2024 • **Revised :** 12 October 2024;
Accepted : 20 October 2024 • **Published :** 29 December 2024

Abstract: Advances in research technologies make high-dimensional data available, and the most interesting research has been conducted on variable selection. The Bayesian variable selection is gaining interest in different fields because of its vast literature. In this paper our main objective is to investigate Bayesian variable selection (BVS) in the context of a random effect model using various prior for variance components in the spike-slab prior approach.

The spike-slab prior is viewed as a mixture distribution, where some regression coefficients concentrate around zero (spike), and the remaining coefficients have a probability of not being zero values (slab). Generally, for building a model, the normal distribution is considered a prior for regression coefficients in the spike-slab components. When the variance is unknown in normal, it could be estimated by extending the model into a hierarchical model. In a hierarchical model, different prior distributions have been proposed for the variance components. Generally, the prior distributions chosen for the variance component are uniform, inverse gamma, and half-Cauchy distribution. Both simulation and real data will be studied to investigate and evaluate how good the chosen distribution for variance components is in the random effect model for the spike-slab approach.

The application of the BVS with spike-slab prior to microarray data from ADNI (Alzheimer's Disease Neuroimaging Initiative), in a logistic regression setting (Alzheimer's Disease vs. Control) demonstrated a notable degree of dimensionality reduction. These selected genes maintain lower misclassification error percentages with higher area under the receiver operating characteristic curve(AUC-ROC) values in different machine-learning algorithms. This discovery opens up new avenues for in-depth exploration and investigation, potentially leading to the identification of biomarkers for Alzheimer's Disease(AD).

Introduction

The high-dimensional data became available in the modern era, due to modern advancements in data storage and computational power. High-dimensional data is

TO CITE THIS ARTICLE

Md Maidul Husain & Anwar Hossain(2024). Bayesian Variable Selection with Spike-Slab prior in Random effect Model, *Journal of Applied Statistics & Machine Learning,* 3(1-2), pp. 41-61.

mostly studied topic in statistics that are defined as when the observations in the data set are smaller than the total number of predictors. The most interesting problem arises in selecting a sub-set of covariates among the many covariates related to the outcome of interest. This is known as "variable selection" or "model selection" and is considered a challenging step in statistics. Statisticians are concerned about variable selection because it is hard to satisfy prediction accuracy and interpretability of the model simultaneously. When the prediction is the aim, we mainly focused on the fitting good model without regard to the number of covariates. Conversely, we try to select fewer covariates for interpretability purposes, which will provide a better unknown relationship between covariates and response. So, the variable selection provides a solution for a specific purpose, and we should not regard it as a general purpose (Rockova, 2013).

We may increase the model's precision, lessen its complexity, and save time and resources by adding a smaller number of variables to the model. There are numerous methods available for variable selection in Bayesian and classical literature.

Classical statisticians have studied variable selection extensively. Classical methods include forward selection, backward selection, and stepwise selection. These methods start with either a full or null model, and then each step variables are deleted or added to the model till getting the best model. In the forward selection approach, the variables are added gradually to the model to improve the model's fit based on predefined criteria, such as AIC, BIC, etc. On the other hand, backward elimination starts with a model that includes all potential covariates and iteratively removes the least significant covariates based on a specified criterion. Stepwise selection combines elements of forward selection and backward elimination. At each step, it considers both adding new covariates that improve the fit and removing existing covariates that no longer contribute significantly to the model. Besides, introducing the penalized term in the model, known as shrinkage methods, is another useful approach for variable selection. The most well-known shrinkage variable selection methodologies in classical statistics are the Least Absolute Shrinkage and Selection Operator (LASSO), ridge, and a combination of LASSO and ridge (elastic net).

The LASSO could be used simultaneously as a variable selection method and parameter estimation that was proposed by (Tibshirani, 1996) . Usually, LASSO adds "L1 norm" $|\beta_j|$ as a penalty term in the likelihood function. On the other hand, another shrinkage method is ridge regression, which shrinks the parameter by the size of the regression coefficients. It introduces the "L2 norm" β_j^2 in the likelihood function instead

of "L1 norm" $|\beta_j|$. Ridge regression shrinks the parameter towards zero by a constant factor (proportionally). The elastic-net method compromises between LASSO and ridge penalties (Zou & Hastie, 2005). Unlike the ridge penalty, the elastic net solution is sparse but has more non-zero parameters than the lasso penalty. Although selecting a subset of variables using forward or backward methods is computationally feasible, these methods are sensitive to the sample size. The prediction accuracy is reduced for small changes in the data set and selecting a completely different model. The penalized methods are mostly utilized for ill-conditioned data sets where the total number of observations is less than the number of covariates as well as existing multicollinearity among covariates. The main drawbacks of these methods are that the estimated parameter in LASSO can't exceed the number of observations. In contrast, the ridge can't select parameters automatically because all predictor estimates are non-zero in the model. When the data are correlated, LASSO tends to select one covariate in a group and discard the remaining covariates. Moreover, for a large number of covariates, shrinkage parameter estimation is more expensive for both LASSO and ridge.

The variable selection problem in Bayesian statistics is considered a parameter estimation where marginal posterior probability determines the inclusion of the variables into the model (O'hara & Sillanpää, 2009) . The Bayesian procedure for variable selection frequently employs the discrete choice and continuous shrinkage prior. The spike-slab prior is the discrete choice, employing a mixture distribution. In this mixing distribution, the slab is a normal distribution with a non-zero mean and variance, and the spike is a probability mass focused around zero. In the model, the slab component represents the included predictors, whereas the spike component displays the predictors that would not be included. This method was initially proposed by Mitchell and Beauchamp (Mitchell & Beauchamp, 1988) with concentration at zero and uniform diffuse slab component. The inclusion probability determines sparsity in the model; suggests Bernoulli probability with 0.5 (George & McCulloch, 1993). Later, many authors adjusted this general approach by incorporating the different structures of the variable inclusion probability, the continuous prior distribution for slab components, and the sampling strategy for fitting the model. On the other hand, the global-local (GL) shrinkage framework (Polson & Scott, 2010) includes the most popular continuous shrinkage prior. Generally, in GL shrinkage strategy, global parameters have a large mass around zero to allow more shrinkage, whereas most local parameters leave the unshrunk by considering heavy-tailed distribution.

The Horseshoe, Normal Gamma, Dirichlet-Laplace, and Horseshoe+ are the most common continuous shrinkage prior within GL framework (Carvalho *et al.*, 2010) (Brown & Griffin, 2010) (Bhattacharya *et al.*, 2015) (Bhadra *et al.*, 2017).

The most common prior in the GL technique for variable selection is the horseshoe prior, while significant covariates are estimated using the half-Cauchy distribution. But in most cases, it's difficult to determine the proper posterior distribution. In Bayesian variable selection, the literature is vast, and we will consider only spike-slab prior to this study.

The regression coefficients in the model could be fixed or random; in Bayesian analysis, fixed and random effect models can be expressed as a probability distribution. We consider the normal prior for the estimated regression coefficients in the Bayesian strategy. When the prior variance is fixed, the model is equivalent to the classical fixed effect model (O'hara & Sillanpää, 2009). For random effect, the model assumed parameters are drawn from the normal distribution with an unknown variance that would be estimated. This random effect model has advantages in tuning the parameter that will depend on the variance of the parameter. Through the hierarchical model concept, we can model this variance parameter. As a noninformative prior Inverse gamma and uniform distribution was used in modeling variance component (Browne & Draper, 2006), and Half Cauchy distribution was also used as a weakly informative prior.

In Bayesian analysis, the Spike-slab prior approach—implemented by (Mitchell & Beauchamp, 1988) into the framework of linear regression—is most commonly used to select an important subset of variables. We address the following problem in this study: "the impact of the different prior distributions of variance parameters in "spike-slab" approach for selecting significant covariates." The main objective comparing spike-slab approaches for variable selection in random effect models by considering different prior in variance component. The Kuo-Mallick (K-M) and the Stochastic Search Variable Selection (SSVS) methods will be applied in spike-slab approach for variable selection. These methods will be compared based on following criteria: degree of sparsity in the model, impact of the sample size in variable selection, ill-conditioned data (number of covariates greater than number of sample size)

Apart from these objectives, based on performance, one method will pick up specified prior variance components and will be applied in Alzheimer's disease microarray data set to identify the most significant genes. Since this data set is related

to the classification problem, all simulations will be conducted based on logistic regression.

2. Methodology

2.1. Bayesian Variable Selection (BVS)

Generally, the Bayesian approach's variable selection is straightforward based on model posterior probabilities. The model uncertainty is introduced by using the prior probabilities of each model along with the prior distribution of each parameter in the model. Baye's theorem is used to calculate the posterior probabilities. Let's consider there M competing models; the prior distribution for model m is $P(m)$, now for given data set y the posterior probability of model m is:

$$P(m|y) = \frac{P(y|m).P(m)}{\sum_{m \in M} P(y|m).P(m)} \quad m \in M$$

where, the marginal distribution of the m *th* model over all possible parameter values β_m of the model m is $P(y|m) = \int P(y|m,\beta_m)P(\beta_m|m)d\beta_m$. The best model is identified, which has the highest posterior probability. But there are some challenges in using Bayesian variable selection: prior probability, intractable likelihood and searching algorithm.

The application of variable selection is crucial for understanding the characteristics of the observed phenomena and successfully recovering any sparse underlying structures in the data. For sparse data sets "Spike-Slab" prior approach gained popularity in selecting variables with an estimation of the parameters. That method combines two probability distributions: probability mass centered at zero (referred as spike) and a continuous distribution (such as a uniform or normal distribution) that represents the slab component. Most of the coefficients in the model are concentrated around zero, while only a small subset of coefficients deviates significantly from zero. This phenomenon is known as selective shrinkage.

2.2. Framework of BVS in Logistic Model

The Bayesian variable selection framework will be described using the logistic regression model. The latent indicator variables are used for each covariate in the Bayesian framework model with "Spike-Slab" prior. The posterior summaries depend on these latent variables.

Let's consider there are p covariates $x_1, x_2, \ldots x_p$ and n individuals in the model. The response or outcome of interest y for each individual is binary (0/1); it could be disease or control. In logistic regression, we will model the probability of an outcome (i.e. $\pi = P(y = 1)$ based on the individual characteristics. The logit transformation of probability describes the linear relationship of the covariates with outcome. Mathematically the model is following,

$$logit(\pi) = \log\left(\frac{\pi}{1-\pi}\right) = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

where, π denote the probability of success event and β_i denote the regression coefficient associate with each covariates x_i in the model.

The variable selection problem for logistic regression is to find a smaller subset of the covariates which correctly classified or predicted the outcome of interest in the model. Selecting the regression coefficients β_i that is zero can be viewed as the variable selection issue. So, each regression coefficient in the model could be placed on the "spike" part (i.e.: not significant) or the "slab" part (significant). We can introduce the indicator variable I_j in the model, where I_j=1 or 0 denotes *j th* covariates presence or absence in the model. For most methods, an additional auxiliary variable called effect size $\beta_j = I_j\theta_j$ is required in the model. Now when $I_j = 1$, then $\beta_j = \theta_j$ and the interpretation is straightforward. The methods listed below vary based on β_j since the variable β_j can be interpreted in multiple ways when I_j = 0.

Since there are p covariates in the model, 2^p candidate models can be considered through indicator vector $I = (I_1, I_2, .., I_p)$ Then, the main concern focuses on the complexity or sparseness of the model required to describe the relationship between outcome and covariates. One flexible approach is to define the model's sparseness through prior probability $P(I_j = 1)$ for variable inclusion. $P(I_j = 1) = \pi\ ,\ j = 1,2, \ldots, p$

George and McCulloch (1993) suggested setting the inclusion probability to 0.5 to make all models equiprobable. Although this probability may enhance the MCMC mixing and attractive to as a null prior, the model is biased to select about half of the predictors. This is not good for the model where a small subset of the covariates is likely to be required. So the choice of $P(I_j = 1)$ value depends on the investigator's analytic approach to pick the best prior probability.

After building the model, the Markov Chain Monte Carlo (MCMC) algorithm fits the model. Using Gibbs sampler methods, the sample of regression coefficients and

indicator variables is generated from the joint posterior distribution. Fortunately, it is not required to calculate the posterior probability of all of the 2^p possible models. With the help of MCMC, it becomes convenient to identify the sub-model that exhibits promising covariates by examining their high-frequency appearances within the Gibbs sample.

2.3. *Bayesian Variable Selection Approach*

Different Bayesian variable selection approaches are mentioned in the literature. Among these methods, we can mention Gibbs variable selection (GVS), Kuo and Mallick (KM), Stochastic Search Variable Selection (SSVS), reversible jump MCMC, adaptive shrinkage with Jeffreys' prior or a Laplacian prior, and variable selection based on Zellner's g-prior (Kuo & Mallick, 1998) (George & McCulloch, 1993) (O'hara & Sillanpää, 2009) (Lesaffre & Lawson, 2012). This study considers and investigates KM and SSVS in selecting a subset of variables with different prior for variance components in the random effect model. These two approaches are differed based on how they treat θ_j, β_j and I_j. They can be readily implemented in the BUGS language since they are based on the Gibbs sampler. The following two sections are high-level descriptions of KM and SSVS methods.

2.4. *KM Approach*

Kuo and Mallick (1998) incorporated indicator variables as a parameter in the regression model to select the non-zero covariates in the model. It is a discrete process because each variable is retained or deleted from the model. In the Kuo-Mallick (KM) method, the significant regression coefficients β_j set to the slab part and the corresponding predictors β_j that has no relevance to the outcome set to the spike part (Fig: 1) Click or tap here to enter text.. The KM methods assume the independent prior for effect size $\theta_j = \beta_j\ I_j$ and indicator variables I_j, so we chose independent prior for each I_j and β_j, such that $f(I_j, \beta_j) = f(I_j).f(\beta_j)$. Now the linear relationship of covariates in the regression model is

$$\log\left(\frac{\pi}{1-\pi}\right) = \sum_{j=1}^{p} \beta_j x_j$$

This model could be considered as a discrete process where in each iteration of MCMC, predictors are either included or excluded from the model. The MCMC does

not require any tuning to fit the model. But the MCMC mixing might be poor when the prior β_j is too vague and $I_j = 0$. The regression coefficients β_j are seldom noticeable in regions where effect size θ_j has greater posterior evidence; therefore, the sampler hardly switches from $I_j = 0$ to $I_j = 1$

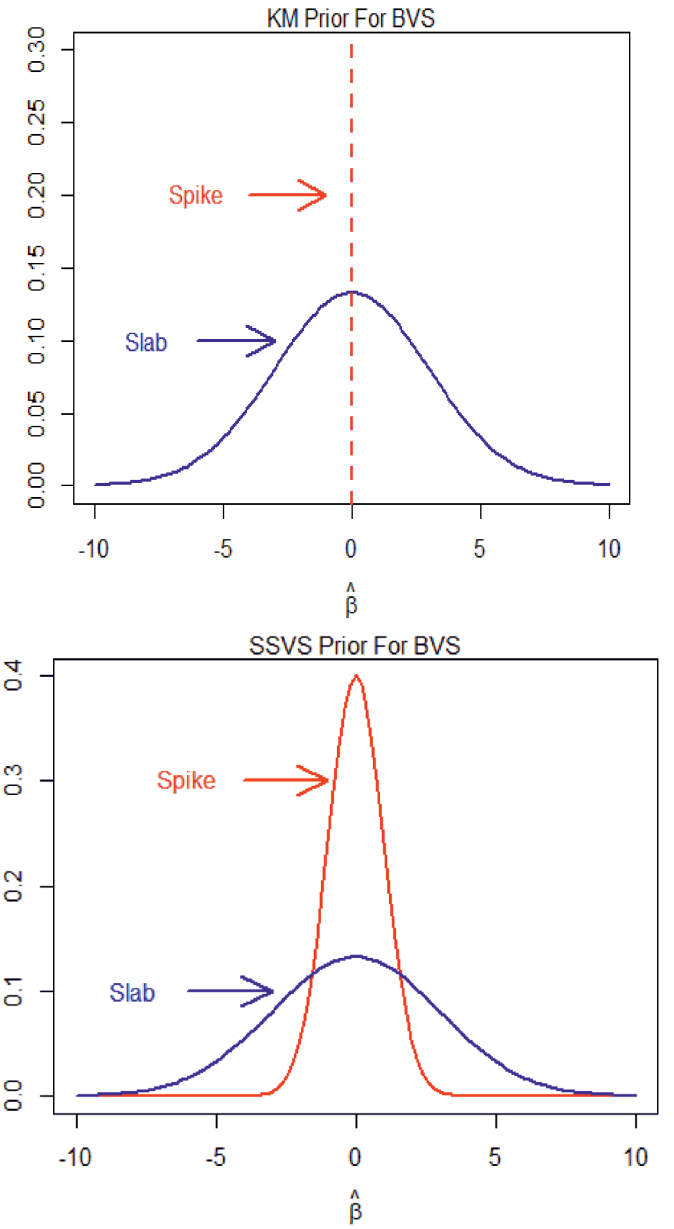

Figure 1: Bayesian Variable Selection through KM (left) and SSVS (right) approach

2.5. SSVS Approach

The stochastic search variable selection (SSVS) was proposed by George and McCulloch (1993), where the probability was applied to select a subset of variables. The subset choices in this approach are determined by latent variables derived from a hierarchical normal mixture model. Unlike the KM approach, in SSVS, the spike has a narrower distribution concentrated around zero (Fig. 1). This is considered a more realistic assumption than assuming that regression coefficients do not have any effect on the outcome. When the indicator $I_j = 0$, the β_j is drawn from the distribution concentrated around zero (spike), and the covariate has a minor effect on the outcome. On the other hand, when indicator $I_j = 1$, the β_j is drawn from the slab part, and the covariate possesses a non-zero influence on the outcome, which suggests that it has to be accounted for in the model. So, the prior distribution of β_j is affected by indicators, which violates independence assumptions of KM, i.e. $f(I_j, \beta_j) = f(I_j).f(\beta_j)$. Generally, spike-slab prior to SSVS approach is written as follows,

$$\beta_j \, I_j \sim I_j \, N(0, \sigma_j^2) + (1 - I_j) N(0, k^{-2} \sigma_j^2)$$

where σ_j^2 represents the variance of the slab part, and for large $\frac{1}{k^2}$ value with $I_j = 0$, the regression coefficients β_j stay very close to zero. The parameter tuning is challenging since $P(\beta_j | I_j = 0)$ requires a very small value while simultaneously avoiding concentration around zero. Unlike KM, the almost zero predictors are not removed from the model when $I_j = 0$, so each iteration of MCMC fits full models. This raises the computational expenses of MCMC for a large number of covariates.

3. Simulation Study

The data simulation process involves employing a logistic regression model, where the regression parameters are treated as random variables. For variable selection purposes, the KM and SSVS approach shall be applied in many settings of simulated data sets. In the Bayesian paradigm, different priors for variance components will be considered in both K-M and SSVS approaches for various settings of the simulated data. This simulation's prime objective is to evaluate the efficacy of the KM and SSVS in a sparse model. Besides, we want to gain better advantages and disadvantages of each strategy in various contexts, such as when the data is ill-conditioned $(p \geq n)$. Moreover, we try to pinpoint instances in which, for selecting a subset of variables, one approach could

be more suitable than another approach for the random effect model in the Bayesian paradigm.

We generated the covariates from the standard normal distribution for this simulation study. We varied our sample size (n), number of predictors (p), and sparsity of the data-generating model to achieve our objectives. Let X represent standard normal predictors in the model with dimension $n \times p$, where n = total sample size and p = total number of predictors. The model's sparsity is introduced by a fixed number of non-zero covariates *nzc* out of the total p covarites in the logistic regression setup. Inverse logistic transformation is now used to convert the linear predictor $\eta = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$ into probability. Then the response Y is generated from the Bernoulli distribution by using this probability

$$\pi(i) = logit^{-1}(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p) Y[i] \sim Ber(\pi(i))$$

In order to make a reliable comparison between KM and SSVS, we generated about 50 datasets for each scenario with fixed values of *n,p,* and *nzc*. The random fluctuations or outliers of data sets in a particular data set would be lessened by averaging findings across the numerous data sets.

The median probability model rule is used to select the significant covariates for both KM and SSVS approaches. We used the number of predictors as a criterion to assess the degree of sparsity attained by each methodology for evaluating each technique's efficacy. We fix the total number of non-zero covariates *nzc* in advance to compare the results with the true sparsity of the data-generating model. We repeated this fitting procedure over different datasets for fixed *nzc* and then compared the average number of predictors p identified by each method.

We can compare the accurate estimation of predictors selected by each technique and the true model sparsity. These results are presented in the column labeled P(%) in tables (Appendix A and B). In addition to evaluating the degree of sparsity, it is also important to know in identifying the true non-zero covariates in the model achieved by each method. To do this, we checked whether the covariates identified as non-zero by each method were indeed the non-zero covariates in the true model. The results are presented in the column labeled *nzc* (%) in tables (Appendix A and B). With respect to our objective, we run the following simulations plan by varying the total number of predictors p, sample size n, and the non-zero covarites *nzc* to compare the performance of both KM and SSVS with different prior for variance components.

- Degree of Sparsity: In this simulation, we fixed the total number of predictors p=20 and the sample size n = 60. The degree of sparsity imposed in the data set by selecting five different numbers of non-zero covariates nzc: 2, 5, 10, 15, 20.
- Sample Size: To examine the sample size impact on variable selection, we repeat earlier simulation plan for different sample sizes: n = 60 and n = 100
- Ill-conditioned Data: For this simulation plans, number of observations n is smaller than the number of predictors p in the model, so we keep the sample size fixed n=60 while varying the number of covariates n = 60 to n = 100.

The Bayesian variable selection methods often rely on Gibbs sampling, leading researchers to use BUGS or JAGS via the "runjags" package in R for computation. In our study, we used R and JAGS to explore the effectiveness of Bayesian variable selection in high-dimensional data, leveraging parallel processing for efficient simulations.

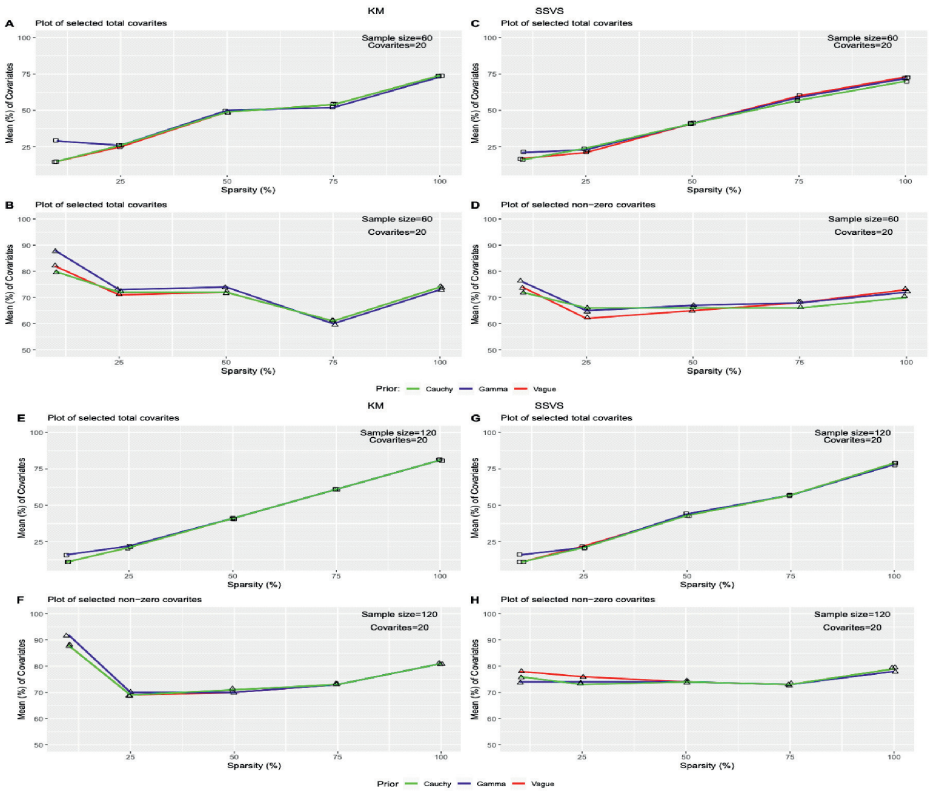

Figure 2: Comparison between KM and SSVS approach in variable selection with different prior in variance component over different levels of sparsity when sample size greater than number of covariates

These methods are compared based on the accuracy (i.e. how many true non-zero covariates are identified) over the different levels of sparsity (Fig. 2). This accuracy percentage labeled as (*nzc*(%)) in Table A.1 shows that the percentage of identified correct predictors for KM methods is higher than the SSVS approach. Comparatively, the performance KM is slightly better in higher levels of sparse model $(nzc = 10\% \text{ and } 25\,\%)$ than in less sparse model $(nzc = 75\% \text{ and } 100\,\%)$. Moreover, the gamma prior in variance components delivered better performance than the vague prior and Cauchy prior in selecting true non-zero covariates for both KM and SSVS approaches. Although the impact of the sample size is not well understood over different degree levels of sparsity for gamma and Cauchy prior, the performance gradually increased for vague prior (Fig. 2). This finding would suggest that the increased sample size has a greater impact on the posterior distribution than the prior distribution.

The simulation results for $p \geq n$ are shown in the following Fig. 3. These simulations are planned for two settings where the number of non-zero covariates in

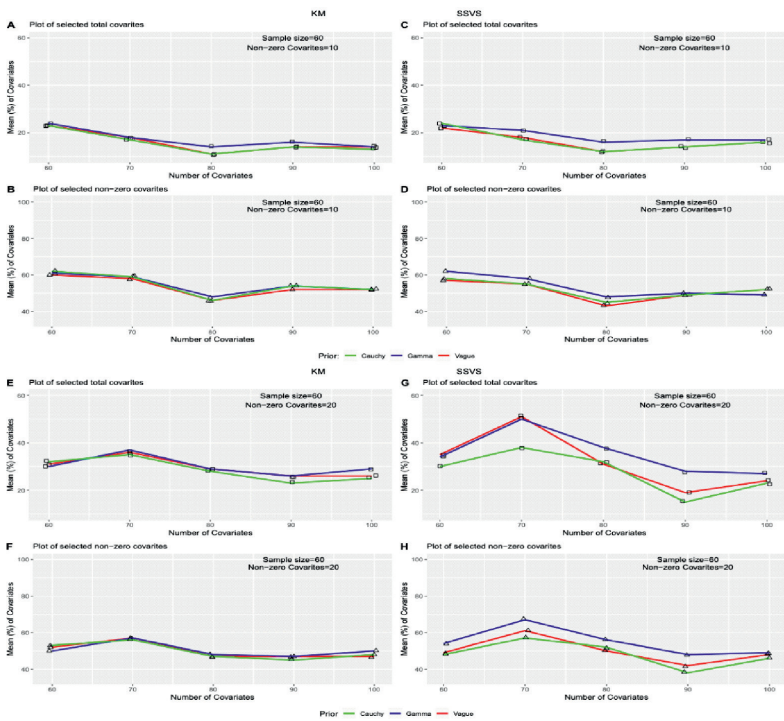

Figure 3: Comparison between KM and SSVS approach in variable selection with different prior in variance component over different levels of sparsity when number of covariates greater than sample size

the model are fixed at 10 and 20, while the total number of covariates increases from 60 to 100. The performance of the KM and SSVS is quite similar in selecting non-zero covariates for different prior settings of variance components (see column $p(\%)$) and $nzc(\%)$).

The method's performance is largely unaffected even with twice as many variables as the sample size. On the other hand, when we compare three different priors for variance, the percentage of correctly identified predictors for the gamma distribution is higher than other priors: vague and Cauchy.

In summary, from Figure 1 and 2, for a highly sparse data set, KM performs well than SSVS when $n \geq p$, whereas both approaches have similar performance for $p \geq n$. Due to the independence prior property of indicators variable and regression coefficients, the KM method is easier to implement than SSVS. Finally, we can use gamma distribution as a prior for variance components in the KM approach for variable selection from the logistic regression models. All the numerical results for the simulation are reported in the appendix.

4. REAL DATA ANALYSIS (Alzheimer's Disease Neuroimaging Initiative)

The Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni. usc. edu) is the source of the microarray data set for the disease. Michael W. Weiner, chief investigator of ADNI, launched this in 2003 as a private-public joint partnership. This database's main goal is to provide data about Alzheimer's Disease to researchers so they can better understand the condition and aid in its early identification. This data source is composed of different types of data: clinical data, imaging data (including MRI and PET images), genetic data, and biospecimen data. It is a large-scale, longitudinal study with more than 1500 participants aged 55 to 90.

Blood samples from a cohort of 811 people participating in the ADNI WGS were used for gene expression analysis. Expression profiling was performed using the Affymetrix Human Genome U219 Array from Affymetrix (Santa Clara, CA). The Robust Multi-chip (RMA) Average normalization was applied as a preprocessing step for the unprocessed expression values derived from the CEL files. Moreover, 64 samples were eliminated from the data files since they failed the quality check (QC) of the data, and from further QC steps, three questionable subjects were also removed.

The final data set contained 744 samples with 49,386 probe sets. There are four classes in the data set, and the distribution is 260 Cognitively normal (CN), 215 Early

Mild Cognitive Impairment (EMCI), 226 Late Cognitive Impairment (LMCI), and 43 Alzheimer's Disease (AD). This data set is imbalanced. In this paper, we considered only the stable participants labeled AD or CN, and the participants in the transitional state (EMCI and LMCI) were removed from the data.

4.1. Data Preprocessing

As a first step of data preprocessing, we decided to match each probe to individual genes, so we excluded certain probes from the data set that lacked associated gene annotations. The Affymetrix Human Genome 219 Plate annotation data (chip hgu219) was utilized to map the probes to ensemble IDs. Then Dropping genes in the bottom 10 percentile for over 80% of the samples. Furthermore, certain probes could be "promiscuous," which means they detect many target sequences, which could produce false findings if collapsed into individual genes. Additionally, since a number of probes may represent certain genes, combining them into a single gene may result in the loss of data. In order to ensure a more comprehensive and correct study of gene expression patterns, we decided to evaluate the probe sets as they were marked in the original dataset. We will take the median of the gene expression values when multiple probes map to the same genes.

Due to the uneven nature of the ADNI dataset, the oversampling approach was used in the second stage to balance the data for each group, namely AD and CN. The SMOTE (Synthetic Minority Oversampling Technique) is employed in our study to make a balanced dataset. This interpolation method creates samples for underrepresented classes (Chawla, 2002). The over sampling SMOTE works by joining the minority class with its five nearest neighbors in high-dimensional space. We took the difference between a sample and one of its five closest neighbors, multiplied this difference by a chance number between 0 and 1, and then added it to the sample of the minority class. Hence, the SMOTE operates within the feature space by generating sample points as random points on lines within a high-dimensional space. We used the "DMwR" package in R for implementing SMOTE. The final data set was balanced by 258 cases for each group, AD and CN, respectively.

The third step was selecting the differential expressed genes (DEGs) between AD and CN groups. To extract the DEGs from the microarray data, we utilized R's "limma" package. Within the 'limma' package, the "lmFit" function is utilized to fit a linear regression model, enabling the estimation of gene-specific effects, and "eBayes"

function is based on the empirical Bayesian method moderate the standard error over genes to identify the DEGs for two groups (Smyth, 2004). We set the cutoff value for FDR 0.01 to identify the DEGs.

4.2. *Exploratory Data Analysis*

After the first step of data preprocessing, we got the microarray data set with 16,686 genes. Further normalization is unnecessary since the ADNI microarray data has already been adjusted using log-scale transformation.

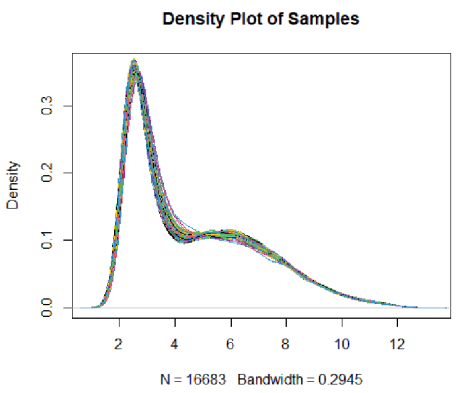

Figure 4(a): Kernal Density Plot

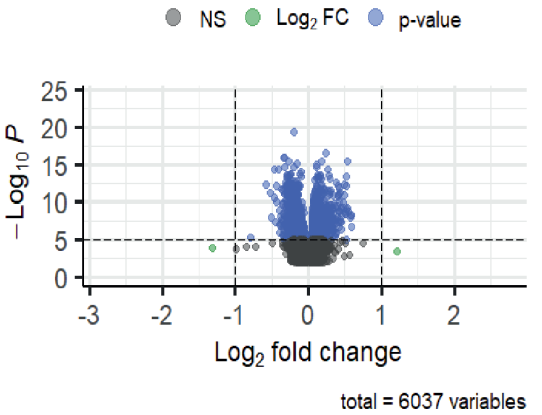

Figure 4(b): Volcano plot of DEg

To ensure the normalization of the data, Kernel Density Estimation (KDE) is employed, and a corresponding plot is provided (Fig:4(a)). The higher density corresponds to a peak point in the plot, suggesting that a certain range of gene expression values is common over the samples. The wider width of the curve indicates greater variability and may be subgroups or patterns that exist within the data. Identification of genes that caused higher variability can enhance the separation between classes (AD vs CN) while in classification.

In order to find the differentially expressed genes (DEg) from the microarray data set, we will now employ the R tool Limma (Linear Models for Microarray Data). Finding the genes with distinct expression labels for the two classes, AD and CN, is the primary goal of DEg. Since our expression data are already preprocessed, we used a linear model function "lmFit()" from the limma package to compare groups. In the design arguments of the "lmFit" function, we specified our test of hypothesis. Then the "eBays" function in the limma package is used to improve the accuracy and

stability of statistical inference. Specifically, the "eBays" function borrows information across genes to obtain reliable and stable estimates of gene-specific variances. Since the limma runs many tests simultaneously (one for each gene), we employ the "Benjamini-Hochberg" multiple correction approach to lower the false discovery rate (FDR) and get corrected p-values. The outcome provided by the limma package comprises lists of differentially expressed genes, accompanied by pertinent statistical details, including log-fold changes, p-values, and adjusted p-values. We used a volcano plot (Fig: 4(b)) to simultaneously visualize the statistical significance (P-value) and fold-changes of genes between two groups.

4.3. Feature Selection by BVS

Now we used the KM approach with a gamma prior in the variance component for selecting a small subset of genes from 1000 DEg's. Thus, the data set is divided into training and testing portions in order to get a subset of genes while controlling the proportion of AD and CN in the two parts. We repeat this procedure several times. Then average inclusion probabilities for all genes are obtained. In this context, the computation of inclusion probability entails evaluating the percentage of posterior samples where a variable is included in the model. The Kuo-Mallick approach utilize the MCMC methods for drawing samples from the posterior distribution. By drawing a sample of model parameters from the joint posterior distribution of a specific model structure, the MCMC technique explores the entire space of possible models. The inclusion probability for a variable can be determined by calculating the percentage of MCMC samples that contain the variable. The median inclusion probability rule is used to include the probability in the model. This estimate of inclusion probability is an indication of the variable's importance in the model. High inclusion probabilities are more likely to significantly impact the model's prediction, while low inclusion probabilities indicate a lesser correlation with the response variable. The BVS approach is used to choose 66 features in total, and these features were then used in Logistic Regression and Random Forest classification algorithms.

5. Classification Techniques

A classification is a form of supervised learning in which the primary objective is constructing a model using a labeled training dataset. This study's classification

approach aims to compare the effectiveness of a limited sample of genes chosen by BVS with entire DEG genes in classifying the two classes AD and CN. The number of features (genes) plays a significant role in classification. A small subset of genes decreases the analyzing complexity with time in identifying the biomarker and can reduce the model's overfitting problem. We employed two distinct classifiers, logistic regression, and random forest to perform the classification task. The data sets are divided into training data and testing data. All machine learning algorithms would be trained by using the training data set.

To assess the effectiveness of these classification algorithms, two metrics would be calculated from testing data: the proportion of misclassification errors and the area under the receiver operating characteristic curve (AUC-ROC).

The penalized logistic regression (PLR) model is utilized in our study to classify individuals with Alzheimer's disease based on their gene expression profiles. The motivation for using PLR is found in (Zhu & Hastie, 2004). The PLR can handle the situation where number of the predictors is greater than the sample size as well as can address the over-fitting issues while the logistic regression becomes unstable. By maximizing the likelihood function of the data and controlling the model complexity, PLR seeks to identify the ideal collection of genes. We used PLR in three different data sets for comparison purposes: full data set, top 1000 DEg, and genes selected by BVS. Misclassification errors from the testing data were 2.58%, 3.22%, and 9.03%, respectively, showing that while BVS selected fewer genes, these can provide further insight into potential biomarkers.

The Random Forest (RF) approach was used to build an ensemble of decision trees, which predicts clinical diagnosis by reducing overfitting through random sampling of training data. Similar to the PLR analysis, RF was applied to the same three data sets. Results indicated comparable misclassification errors across the full data set and the smaller subset of top genes, highlighting the model's robustness.

Table: Classification Performance of Logistic Regression and Random Forest over different selected sets of features

	Logistic Regression		*Random Forest*	
Features(genes) from	*Miss-classification (%)*	*AUC*	*Miss-classification (%)*	*AUC*
Full data set	2.58	0.9988	2.32	0.9985
Top DEG's	3.22	0.9878	1.56	0.9984
BVS	9.03	0.9485	2.32	0.9981

6. Conclusion

In this study, we employed Bayesian Variable Selection (BVS) with a spike-and-slab approach, incorporating appropriate priors for variance components. Through real data analysis, we demonstrated that this method effectively identified the most relevant genes from the full dataset, providing a competitive solution for handling high-dimensional data. However, caution is necessary when generalizing the results, as we only considered a limited set of prior distributions for the variance components.A key finding was that a small subset of genes maintained comparable performance to the full set when used in machine learning models. This result aligns with the known equivalence between Bayesian and frequentist regularization methods, such as LASSO, in selecting important variables (Park & Casella, 2008). Unlike frequentist approaches, which rely on cross-validation to tune hyperparameters, the Bayesian method benefits from incorporating prior knowledge, making it advantageous in variable selection.

In summary, we can say BVS with spike-slab prior is a good alternative as a dimension reduction technique that deserves further investigation in the biostatistics domain. This spike slab prior specification enables the posterior distributions of zero coefficients to concentrate around the spike while non-zero coefficients remain in the slab. The most promising subset of predictors can be identified by analyzing this separation in the posterior distribution. We described the BVS approach in the ADNI gene expression data set to select a small subset of genes. In the AD disease diagnosis from a microarray data set, there are two main challenges: accurately predicting the disease class and identifying the key genes that are responsible for AD disease. The BVS approach effectively addressed these two challenges. These chosen genes were used to distinguish between AD and CN using following classifiers: Logistic Regression and Random Forest. The analysis of a small set of genes yielded a low percentage of misclassification and a higher AUC value, indicating improved performance and accuracy in the classification task with smaller features.

References

Bhadra, A., Datta, J., Polson, N. G., & Willard, B. (2017). *The horseshoe+ estimator of ultra-sparse signals*.

Bhattacharya, A., Pati, D., Pillai, N. S., & Dunson, D. B. (2015). Dirichlet–Laplace priors for optimal shrinkage. *Journal of the American Statistical Association*, *110*(512), 1479–1490.

Brown, P. J., & Griffin, J. E. (2010). *Inference with normal-gamma prior distributions in regression problems*.

Browne, W. J., & Draper, D. (2006). *A comparison of Bayesian and likelihood-based methods for fitting multilevel models*.

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

Chawla, B. (2002). Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. *Smote: Synthetic Minority over-Sampling Technique, J. Artif. Intell. Res*, *16*, 321–357.

George, E. I., & McCulloch, R. E. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, *88*(423), 881–889.

Kuo, L., & Mallick, B. (1998). Variable selection for regression models. *Sankhyā: The Indian Journal of Statistics, Series B*, 65–81.

Lesaffre, E., & Lawson, A. B. (2012). *Bayesian biostatistics*. John Wiley & Sons.

Mitchell, T. J., & Beauchamp, J. J. (1988). Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, *83*(404), 1023–1032.

O'hara, R. B., & Sillanpää, M. J. (2009). A review of Bayesian variable selection methods: what, how and which. *Bayesian Analysis*.

Polson, N. G., & Scott, J. G. (2010). Shrink globally, act locally: Sparse Bayesian regularization and prediction. *Bayesian Statistics*, *9*(501–538), 105.

Rockova, V. (2013). *Bayesian variable selection in high-dimensional applications*.

Smyth, G. K. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology*, *3*(1).

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267–288.

Zhu, J., & Hastie, T. (2004). Classification of gene microarrays by penalized logistic regression. *Biostatistics*, *5*(3), 427–443.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *67*(2), 301–320.

Appendix A

Table A.1: Performance of KM and SSVS in variable selection for different levels of sparsity, where n= sample size, p = mean (%) of covariates selected and nzc= means (%) of non-zero covariates selected in logistic regression

Covariates = 20		*KM (n ≥ p)*					
True non-zero covariates nzc(%)		*Vague*		*Gamma*		*Cauchy*	
		p(%)	*nzc(%)*	*p(%)*	*nzc(%)*	*p(%)*	*nzc(%)*
n= 60	10.0	15.0	82.0	29.0	88.0	15.0	80.0
	25.0	25.0	71.0	26.0	73.0	26.0	72.0
	50.0	49.0	72.0	50.0	74.0	49.0	72.0
	75.0	54.0	61.0	52.0	60.0	54.0	61.0
	100.0	74.0	74.0	73.0	73.0	74.0	74.0
n=120	10.0	11.0	88.0	16.0	92.0	11.0	88.0
	25.0	21.0	69.0	22.0	70.0	21.0	69.0
	50.0	41.0	70.0	41.0	70.0	41.0	71.0
	75.0	61.0	73.0	61.0	73.0	61.0	73.0
	100.0	81.0	81.0	81.0	81.0	81.0	81.0

Covariates = 20		*SSVS (n ≥ p)*					
True non-zero covariates nzc(%)		*Vague*		*Gamma*		*Cauchy*	
		p(%)	*nzc(%)*	*p(%)*	*nzc(%)*	*p(%)*	*nzc(%)*
n= 60	10.0	17.0	74.0	21.0	76.0	16.0	72.0
	25.0	21.0	62.0	23.0	65.0	24.0	66.0
	50.0	41.0	65.0	41.0	67.0	41.0	66.0
	75.0	60.0	68.0	59.0	68.0	57.0	66.0
	100.0	73.0	73.0	72.0	72.0	70.0	70.0
n=120	10.0	11.0	78.0	16.0	74.0	11.0	76.0
	25.0	22.0	76.0	21.0	74.0	21.0	73.0
	50.0	43.0	74.0	44.0	74.0	43.0	74.0
	75.0	57.0	73.0	57.0	73.0	57.0	73.0
	100.0	79.0	79.0	78.0	78.0	79.0	79.0

Appendix B

Table A.2: Performance of KM and SSVS in variable selection when the number of covariates is greater than sample size in the sparse model, where n = sample size, p= mean (%) of covariates selected and nzc= means (%) of non-zero covariates selected in logistic regression

Covariates = 20		*KM (p≥ n)*					
True non-zero covariates nzc(%)		*Vague*		*Gamma*		*Cauchy*	
		p(%)	*nzc(%)*	*p(%)*	*nzc(%)*	*p(%)*	*nzc(%)*
Number of non-zeros covariates 10	60	23.0	60.0	24.0	61.0	23.0	62.0
	70	18.0	58.0	18.0	59.0	17.0	59.0
	80	11.0	46.0	14.0	48.0	11.0	46.0
	90	14.0	52.0	16.0	54.0	14.0	54.0
	100	14.0	52.0	14.0	52.0	13.0	52.0
Number of non-zeros covariates 20	60	31.0	52.0	30.0	50.0	32.0	53.0
	70	36.0	57.0	37.0	57.0	35.0	56.0
	80	29.0	47.0	29.0	48.0	28.0	47.0
	90	26.0	47.0	26.0	47.0	23.0	45.0
	100	26.0	47.0	29.0	50.0	25.0	48.0

Covariates = 20		SSVS (p ≥ n)					
Number of covariates		Vague		Gamma		Cauchy	
		p(%)	nzc(%)	p(%)	nzc(%)	p(%)	nzc(%)
Number of non-zeros covariates 10	60	22.0	57.0	23.0	62.0	24.0	58.0
	70	18.0	55.0	21.0	58.0	17.0	55.0
	80	12.0	43.0	16.0	48.0	12.0	45.0
	90	14.0	49.0	17.0	50.0	14.0	49.0
	100	16.32	52.0	17.0	49.0	16.0	52.0
Number of non-zeros covariates 20	60	35.0	49.0	34.0	54.0	30.0	48.0
	70	51.0	61.0	50.0	67.0	38.0	57.0
	80	31.0	50.0	38.0	56.0	32.0	52.0
	90	19.0	42.0	28.0	48.0	15.0	38.0
	100	24.0	48.0	27.0	49.0	23.0	46.0