2022-01-14

Bayesian Variable Selection Model with Semicontinuous Response

Babatunde, Samuel

UNIVERSITY OF CALGARY

Bayesian Variable Selection Model with Semicontinuous Response

by

Samuel Ayodeji Babatunde

A THESIS
SUBMITTED TO THE FACULTY OF GRADUATE STUDIES
IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE
DEGREE OF MASTER OF SCIENCE

GRADUATE PROGRAM IN MATHEMATICS AND STATISTICS

CALGARY, ALBERTA

JANUARY, 2022

Abstract

We propose a novel Bayesian variable selection approach that identifies a set of features associated with a semicontinuous response. We used a two-part model where one of the models is a logit model that estimates the probability of zero responses while the other model is a log-normal model that estimates responses greater than zero (positive values). Stochastic Search Variable Selection (SSVS) procedure is used to randomly sample the indicator variables for variable selection which in turn searches the space of feature subsets and identifies the most promising features in the model. For the logistic model, a data augmentation approach is used to sample from the posterior density. We impose a spike-and-slab prior for the regression effects where the unselected covariates take on a prior mass at zero while the selected covariates follow a normal distribution (including the intercept and clinical covariates). Since the joint posterior density had no closed form, we employed the techniques of the Markov Chain Monte Carlo (MCMC) to sample from the posterior distribution. Simulation studies are used to assess the performance of the proposed method. We computed the average area under the receiver operating characteristic curve (AUC) to assess variable selection and compared it with competing methods. We also assessed the convergence diagnosis of our MCMC algorithm by computing the potential scale reduction factor and correlations between the marginal posterior probabilities. We finally apply our method to the coronary artery disease (CAD) data where the aim is to select important genes associated with the CAD index. This data consists of clinical covariates and gene expressions.
Keywords: Bayesian variable selection, coronary artery disease, Markov Chain Monte Carlo, Stochastic Search Variable Selection.

Acknowledgments

I give glory to God Almighty for the grace to successfully complete this work. To HIM be all the glory and honour forever.

My sincere appreciation to my brilliant and excellent supervisors, Dr. Thierry Chekouo and Dr. Tolulope Sajobi, for the opportunity to learn under their amiable supervision, I am forever grateful, I say God bless you.

To my lovely and gorgeous wife and daughter; Adetola and Gabriella, I express my profound gratitude for the sacrifice both of you paid in the course of my studies. The inconveniences, distance, to mention but a few. I love you both always.

Furthermore, thanks to my parents; Mr. and Mrs Babatunde, for their great parental, morale support and advice all through the course of studying and always. I will also like to thank my siblings and their family; The Olowus, The Ogundares, The Makindes and Nurse Babatunde Joseph (my younger brother) for their encouragement, support and care in making this dream of mine a reality. I also thank my in-laws; The Adelekes for their support and believe in me. Thanks to my friends; The Odeyemis, The Ademola, The Adekanye, The Ogundipe, The Fatokunbo also for their wonderful support.

Lastly, thanks to the Head of the Department of Mathematics and Statistics, all the Professors and Instructors, Administrative staff, Students and Colleagues. Thank you for creating a conducive and comfortable learning environment.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Background of the study

Variable selection techniques play an essential role in regression modelling, most especially in the context of high dimensional data analysis (a situation where the available observations are less than the number of covariates or risk factors related to a disease or outcome). Some of the Frequentist variable selection methods include the stepwise selection procedure (Peduzzi et al., 1980), forward (Bendel and Afifi, 1976), backward selection procedures (Derksen and Keselman, 1992) and methods based on penalized likelihood, such as the Least Absolute Shrinkage and Selection Operator (LASSO) method (Tibshirani, 1996), Ridge regression (Hoerl and Kennard, 1970), Elastic Net (Zou and Hastie, 2005), and the non-concave penalized likelihood approach (Fan and Li, 2002). In the Bayesian framework, Volinsky et al. (1997) have proposed the use of Bayesian model averaging, where a set of likely models chosen with the leaps-and-bound algorithm are fitted one at a time. More recent Bayesian approaches have been proposed (see section (2.4) for more details).

It is assumed that some sets of genes usually influence the disease state through their pathways (Li and Chekouo, 2021) and most selection methods of predicting these groups or multiple pathways is complex when outcome and nature of data involved is semicontinuous with high dimensionality. Therefore, the need to develop highly efficient and reliable method of selection due to the high dimensional nature of gene expression data. A widely used procedure for identifying genes related to survival outcomes consists of fitting univariate Cox models on each gene and selecting those that pass a threshold for significance (Rosenwald

et al., 2002). Different approaches using partial least squares (Nguyen and Rocke, 2002; Park et al., 2002) or principal components analysis (Li and Gui, 2004) have also been proposed. These methods select linear combinations of genes rather than the original variables (Sha et al., 2006).

In this study, we propose a Bayesian variable selection approach that identifies a set of important features associated with coronary artery disease (using semicontinuous response). We also explored the performance of the proposed model in comparison with some existing Bayesian and Frequentist approaches. This study used the two-part model proposed by Duan et al. (1983) where the first part is a binary model for event of having zero or positive values and the second part is a linear model. we also used a similar Stochastic Search Variable Selection (SSVS) approach introduced by George and McCulloch (1993, 1997) to select significant covariates. The SSVS procedure was used to randomly sample indicator variables for variable selection (more on this is explained in the methodology section of this study). The SSVS approach can also be seen in the context of Bayesian model averaging (Raftery et al. (1997) and Hoeting et al. (1999)). A very close method to this is that proposed by Sha et al. (2006) but they used log-normal and log-t models for the analysis of microarray data with censored outcomes and selected features independently while we used logistic model and log-normal models with application to semicontinuous response and selected features both independently and the same set of features.

1.2 Motivation

The two-part model is convenient for analysing longitudinal semicontinuous data and has inspired numerous extensions and variations (Amemiya, 1973; Belotti et al., 2015; Campbell, 2019; Gronau, 1974; Heckman, 1974, 1979; Olsen and Schafer, 2001). To the best of my knowledge, there is not yet a variable selection method for semicontinuous response. We therefore propose a Bayesian two-part model for variable selection that identifies important features and compared the performance to other existing methods. It is expected that this approach would optimise the selection process of genes that actually contribute significantly to coronary artery disease (CAD) with better prediction capabilities.

1.3 Aim and Objectives

Aim:

The aim of this study is to develop a Bayesian variable selection approach that can accurately handle high dimensional data with a semicontinuous response by identifying and selecting significant covariates or features.

The objectives of this study are:

(i.) To develop Bayesian variable selection model that can identify clinical covariates and genes associated with coronary artery disease (CAD) in the context of high dimensionality.

(ii.) To evaluate the performance of our approach using some model diagnostic metrics.

1.4 Organization of the thesis

The rest of the study is organized as follows. In Chapter 1 we present the background of study, motivation, aim and objectives. Chapter 2 is on the review of literature and approaches to modelling semicontinuous data, existing Frequentist and Bayesian methods for high dimensional data analysis as well as the challenges and findings. Chapter 3 is about the Bayesian methodological framework for our proposed method, while Chapter 4 presents the simulation studies, model diagnostics, and comparison to existing methods. Chapter 5 is on the application to coronary artery disease (CAD) data. Finally, the summary, and conclusion is given in Chapter 6.

Chapter 2

Literature Review

2.1 About Thomas Bayes and Bayesian framework

Thomas Bayes from whom the term Bayes' theorem got its name, was born in 1701 in Hertfordshire. He was an English Statistician, Philosopher and Presbyterian minister who is known for formulating the Bayes' theorem. He attended the University of Edinburgh to study logic and theology from 1719 until around 1721. Bayes never published what would have become his most famous accomplishment. His works and findings on probability theory were obtained from his notes, edited, and published after his death by Richard Price (Bellhouse, 2004).

Bayes' theorem (also known as Bayes' rule) describes the probability of an event based on prior knowledge of conditions that might be related to the event (Zalta, 2008).

Mathematically, let $\boldsymbol{\theta}$ be a parameter with probability $P(\boldsymbol{\theta})$, let y be i.i.d random samples with probability density function $P(y|\boldsymbol{\theta})$. Then, the conditional probability of $\boldsymbol{\theta}$ given y is expressed as

$$P(\boldsymbol{\theta}|y) = \frac{P(y|\boldsymbol{\theta})\,P(\boldsymbol{\theta})}{P(y)} = \frac{P(y|\boldsymbol{\theta})\,P(\boldsymbol{\theta})}{\int_{\boldsymbol{\theta}} P(y|\boldsymbol{\theta})\,P(\boldsymbol{\theta})\,d\boldsymbol{\theta}}, \tag{2.1.1}$$

Equation (2.1.1) is known as Bayes' rule for a continuous random variable, y, this is also applicable to discrete random variables with some changes to the expression of the denominator.

2.2 Review of literatures on modelling semicontinuous data

Semicontinuous data consist of mixture of zero values and continuously distributed positive values (Wang et al., 2020). Examples of semicontinuous data are: observations of total rainfall (many days don't have rainfall), household expenditures (some household spend nothing on a certain commodity), annual medical cost (a portion of the population has zero medical expense) etc. (Min and Agresti, 2002). This type of data has received lots of attention in the literature due to its occurrence under different settings and appropriate analysis to obtain accurate estimates and inferences (Yiu and Tom, 2018). Given the mixture of zero and non-zero values, it was intuitive to view the semicontinuous outcome as arising from two different stochastic processes. One process, referred to as the binary part, indicates if the outcome is zero or not, and the second referred to as the continuous part, determines the positive values conditional on the outcome being non-zero. Semicontinuous data are typically analysed using two-part models wherein the zero process and the continuous values are modelled separately using logistic regression for the binary part and log-normal for the continuous part to ensure prediction of positive values (Yiu and Tom, 2018) whereas, in this study we modelled separately and together. Jaffa et al. (2018) proposed two frameworks; the first is two-part mixed models with either correlated or non-correlated random effects in both parts and the other is based on the two-part marginal models to analyse longitudinal semicontinuous data.

Tobin (1958) proposed a censored regression model to describe household expenditures on durable goods, known as the Tobit model. Amemiya (1973), Gronau (1974), and Heckman (1974, 1979) proposed different models which are generalizations of the Tobit model, by extending it to a two-part model. Duan et al. (1983) proposed a two-part model to fit data on expenditures for medical care that uses equations to separate the modelling into two stages (the first stage is binary model while the second stage is a log-normal distribution). Jørgensen (1987); Jorgensen (1997) proposed a compound poisson exponential dispersion model for semicontinuous data. Saei et al. (1996) applied an ordinal response model that requires grouping the response outcomes into categories. Powell (1986) proposed semicontinuous parametric estimation for the Tobit model using Symmetrically Trimmed Least Squares

(STLS) estimator. Chang and Pocock (2000) applied the cumulative logit model for modelling the amount of personnel care for the elderly. This model has the capability to handle the clumping at zero and positive outcomes. Min and Agresti (2002) reviewed various methods by which "Nonnegative data with clumping at zero" can be analysed, cases where the response for the non-zero observations is continuous and when it is discrete, reviewed existing models for analyzing cross-sectional data and summarized extensions for repeated measurement analysis (e.g. longitudinal studies). Some of the models mentioned by Min and Agresti (2002) for modelling semicontinuous data includes; Tobit models, Two-part models, sample selection models, compound poisson exponential dispersion models and ordinal threshold models. They also reviewed models for Zero-Inflated data as Zero-Inflated discrete distribution models, Hurdle models, Finite mixture models and Neyman Type A distribution models. Karlsson and Laitila (2014) suggested a finite mixture of Tobit models for estimation of regression models with a censored response variable, stating the interesting features of their proposed estimator as having the potential to yield valid estimates in cases with a high degree of censoring.

2.3 Review of literatures on Frequentist variable selection model in the context of high dimensionality

The common strategies for the selection of important features include the best subset selection, stepwise (Peduzzi et al., 1980), forward (Bendel and Afifi, 1976), and backward selection procedures, which are all combinatorial (Derksen and Keselman, 1992). The backward selection can only be used when the number of features or covariates, p is less than the number of observations or sample size, n. Since the first model is considered as the model that contains all possible covariates i.e., the full model. Variables are excluded from the model according to an "exit" criterion with an exclusion threshold for p-values associated with individual coefficients. In contrast, the forward and stepwise procedures starts with a model that contains only a single covariate and the other covariates are subsequently added to the model according to an "entry" criterion of similar nature to that of backward selection. In backward selection, a backward step is performed when a variable is entered into the model. The inverse computational burden of these methods severely limits their applicability to large

datasets encountered in research.

Here, we use the work of Johnstone and Titterington (2009) to explain the statistical change or development from simple linear model to high dimensional nature. Let y_i be a response, x_i is a predictor, n is observations such that i is each individual from 1 to n, then the simple linear model can be written as

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \tag{2.3.1}$$

where $\epsilon_i \sim N(0, \sigma^2)$, the error component is independently and identically distributed normal with mean 0 and variance σ^2. The mean relationship between the response and predictor follows a straight line with intercept β_0 and coefficient of x_i as β_1. The two parameters are unknown and need to be estimated, here $n >> p$. It is very convenient to write in matrix form as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e} \tag{2.3.2}$$

where $\boldsymbol{y}$ is an $n \times 1$ vector, $\boldsymbol{X}$ is known as the design matrix with $n \times p$ dimension, $\boldsymbol{\beta}$ is the vector of parameters in model in equation (2.3.1) with $p = 2$ dimensions, while $\boldsymbol{e}$ is $n \times 1$ variance-covariance matrix which can be either homoscedastic (equal variance) or heteroscedasticity (non-equal variance) form.

The common approach to estimate equation (2.3.1) will be to use the least square or maximum likelihood approach. Using the least square approach, $\hat{\beta}$ is known as the minimizer of the sum of squares function as shown in equation (2.3.3) below;

$$\hat{\beta} = \arg\min_{\beta} \sum_i \left(y_i - \beta_0 - \beta_1 x_i\right)^2, \tag{2.3.3}$$

The matrix form in equation (2.3.2), this can be written in terms of the Euclidean or ℓ_2 norm as

$$\hat{\boldsymbol{\beta}} = \arg\min_{\beta} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2, \tag{2.3.4}$$

The distribution assumption about $\boldsymbol{e}$ implies that $\boldsymbol{y} \sim N_n\left(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}\right)$, where N_n denotes an n - variate multivariate normal distribution with mean vector $\boldsymbol{X}\boldsymbol{\beta}$, covariate matrix $\sigma^2\mathbf{I}$, where $\mathbf{I}$ is an $n \times n$ identity matrix, σ^2 is residual variance, and $\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 = \sum_i \left(y_i - \boldsymbol{x_i}^\top \boldsymbol{\beta_1}\right)^2$.

The probability function for y is then

$$P\left(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\beta}\right) = \left\{\left(2\pi\sigma^2\right)\right\}^{-\frac{n}{2}} \exp\left\{-\frac{\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2}{2\sigma^2}\right\} \tag{2.3.5}$$

The data provided $\boldsymbol{y}$ and $\boldsymbol{X}$ when viewed as a function of the parameters is now called the likelihood function and $\hat{\boldsymbol{\beta}}$ in equation (2.3.6) is called the maximum likelihood estimate denoted by MLE for short form.

$$\hat{\boldsymbol{\beta}} = \arg\max_{\beta} P\left(\boldsymbol{y}|\boldsymbol{X}\boldsymbol{\beta}\right) \tag{2.3.6}$$

It can be seen that $\hat{\boldsymbol{\beta}}$ satisfies $\boldsymbol{X}^\top\boldsymbol{X}\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^\top\boldsymbol{X}\right)^{-1}\boldsymbol{X}^\top\boldsymbol{y}$. From this it can be seen that $\boldsymbol{X}^\top\boldsymbol{X}$ is invertible i.e., the inverse form exist, then if the model is correct $\hat{\boldsymbol{\beta}}$ is said to be distributed

$$\hat{\boldsymbol{\beta}} = N_p\left(\boldsymbol{\beta}, \sigma^2\left(\boldsymbol{X}^\top\boldsymbol{X}\right)^{-1}\right) \tag{2.3.7}$$

where $\sigma^2\left(\boldsymbol{X}^\top\boldsymbol{X}\right)^{-1}$ is the variance-covariance matrix, equation (2.3.7) is also applicable for estimating multiple linear regression where $n >> p$.

It is worth noting that for the analysis of the simple linear model, $n \geq p$ must hold else $\left(\boldsymbol{X}^\top\boldsymbol{X}\right)$ is singular and the parameters would not be estimable. Also, using MLE approach when p is fixed, the asymptomatic theory becomes invalid. Now, what if $p > n$ or $p >> n$? The usual approach will be to use Regularization method which is also known as Penalized Least Squares or Penalized Maximum Likelihood. Let's start with the earliest which is the Ridge regression introduced by Hoerl and Kennard (1970) in which β can be estimated using $\hat{\boldsymbol{\beta}}_{\boldsymbol{Ridge}} = \boldsymbol{S}_{\boldsymbol{\lambda_2}}\boldsymbol{X}^\top\boldsymbol{y}$, where $\boldsymbol{S}_{\boldsymbol{\lambda_2}} = \left(\boldsymbol{X}^\top\boldsymbol{X} + \lambda_2\mathbf{I}\right)^{-1}$, λ_2 is called the ridge parameter or regularization constant. After some derivations, it was inferred that $\hat{\boldsymbol{\beta}}_{\boldsymbol{Ridge}} \sim N_p\left(\boldsymbol{S}_{\boldsymbol{\lambda_2}}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\beta},\ \sigma^2\boldsymbol{S}_{\boldsymbol{\lambda_2}}\left(\boldsymbol{X}^\top\boldsymbol{X}\right)\boldsymbol{S}_{\boldsymbol{\lambda_2}}\right)$. Johnstone and Titterington (2009) noted that the estimator $\hat{\boldsymbol{\beta}}_{\boldsymbol{Ridge}}$ is biased, but it can be calculated as λ_2 increases, bias increases but variance decreases which makes it acceptable. The matrix $\boldsymbol{X}^\top\boldsymbol{X} + \lambda_2\mathbf{I}$ of order $p \times p$ is invertible in this case but there is still a problem when data contains lots of zeros. It has been proven that if there are so many number of predictors, it is often known that only a small number of influential predictors are likely selected. Due to this disadvantage of Ridge regression, they proceeded to change the penalty function and considered what is called ℓ_0 regularization. The problem about implementing this approach of using ℓ_0 regularization is that it leads to unacceptable computational complexity. Therefore, a way out of this difficulty is to have the penalty function on the ℓ_1 norm (LASSO penalty item (Li and Xu, 2008)). The Least Absolute Shrinkage and Selection Operator (LASSO) proposed by Tibshirani (1996),

penalizes based on the maximum absolute coefficient across all outcomes for each predictor

$$\|\boldsymbol{\beta}\| = \lambda \sum_{j=1}^{p} |\beta_j|, \tag{2.3.8}$$

which can be formulated into three as shown below;

(i.) $\hat{\boldsymbol{\beta}}_{\text{Lasso}} = \arg\min_{\beta} \{\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda_1 \|\boldsymbol{\beta}\|\}$, for some λ_1

(ii.) $\hat{\boldsymbol{\beta}}_{\text{Lasso}}$ minimizes $\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2$ subject to $\|\boldsymbol{\beta}\|_1 \leq c_1(\lambda_1)$, and

(iii.) $\hat{\boldsymbol{\beta}}_{\text{Lasso}}$ minimizes $\|\boldsymbol{\beta}\|_1$ subject to $\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 \leq b_1(\lambda_1)$,

This procedure shrinks some coefficients towards zero and sets some to be exactly zero; thereby combining the favourable features of best subset selection and ridge regression, (Li and Xu, 2008). The disadvantage of the LASSO approach as pointed out by Ranstam and Cook (2018) is that the regression coefficients may not be reliably interpretable in terms of the independent risk factors as the focus is on the best-combined prediction, not on the accuracy of the estimation and interpretation of the contribution of individual variables.
Li and Xu (2008) also pointed out that the original LASSO cannot make use of more than n covariates and it is highly sensitive to high correlations among covariates. Johnstone and Titterington (2009) pointed out that the dual advantage of LASSO is its computational feasibility and generally leads to sparse solutions. However, Zou (2006) showed that for a fixed number of parameters p, LASSO in general is not consistent, also showed that the probability that zero coefficients are estimated as zero is generally less than 1. The Lasso estimate for the linear regression parameters was interpreted as a Bayesian Posterior mode estimate when the regression parameters have independent Laplace (double exponential) priors (Park and Casella, 2008). Zhao and Yu (2006) provided an almost necessary and sufficient condition for LASSO to be consistent based on the "irrepresentability conditions" on the design matrix; but the tuning parameter λ required for selection consistency can excessively shrink the non zero coefficient leading to asymptomatically biased estimates. To alleviate the major limitations of LASSO, Zhao and Yu (2006) proposed the Adaptive LASSO which makes use of data dependent weights of the form

$$P_{\lambda}(\boldsymbol{\beta}) = \lambda \sum_{j=1}^{p} \frac{|\beta_j|}{|\hat{\beta}_j^{\mathbf{I}}|\gamma'} \tag{2.3.9}$$

where $\hat{\beta}_j^{\mathbf{I}}$ is an initial estimator of β_j and $\gamma > 0$, large coefficients are shrink less and small coefficients shrinks more than in the original LASSO. For fixed p, adaptive LASSO yields selection consistency of non zero estimates of coefficients with asymptotic distribution equal to that obtained in the model with prior knowledge of the location of zero parameters.
There are other variants to the general LASSO approach, such as the Elastic Net proposed by Zou and Hastie (2005) with penalty taking the form.

$$P_\lambda(\boldsymbol{\beta}) = \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=1}^{p} |\beta_j|^2, \tag{2.3.10}$$

such that

$$\hat{\boldsymbol{\beta}}_{\text{ElasticNet}} = \arg\min_{\beta} \left\{ \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda_2 \|\boldsymbol{\beta}\|^2 + \lambda_1 \|\boldsymbol{\beta}\| \right\}, \tag{2.3.11}$$

where λ_1 and λ_2 are tuning parameters. The elastic net method includes the LASSO and ridge regression when $\lambda_1 = \lambda, \lambda_2 = 0$ or $\lambda_1 = 0, \lambda_2 = \lambda$. Also, the naive elastic net method finds an estimator using a two-stage procedure; first for each fixed λ_2 it finds the ridge regression coefficients, and then does a LASSO type shrinkage. This type of estimation incurs a double amount of shrinkage, which leads to increased bias and poor predictions. To improve the prediction performance, rescale the coefficients of the naive version of elastic net by multiplying the estimated coefficients by $(1 + \lambda_2)$ (Zou and Hastie, 2005). The relative merits of penalization regression techniques is an area of ongoing research.

2.4 Review of literatures on Bayesian variable selection model in the context of high dimensionality

A lot of research on Bayesian variable selection model has been done and ongoing. This is a very broad area of research due to so many factors like; type of response or outcome (for instance our response for this study is semicontinuous), prior choice, types of models involved, algorithm or methodology and even applicable software. It all depends on the objectives of the researcher. Let's take a look at some of the works that has been done in this area, starting with earlier work to most recent research that used the Bayesian framework for variable selection.

2.4.1 Earlier research

As early as 1993, George and McCulloch (1993) proposed a Bayesian procedure for identifying promising subsets of predictors, this procedure known as the Stochastic Search Variable Selection (SSVS) consists of the specification of a hierarchical Baye's mixture prior which used the data to assign larger posterior probability to the more promising models. The SSVS has the advantage (very fast and efficient simulation of Gibbs sampler) and more preferred because it avoids the burden of calculating the posterior probabilities of all 2^p, where p is the number of covariates in a model. In summary, the Gibbs sampler is used to search for promising models rather than computing the entire posterior (George and McCulloch, 1993). Furthermore, George and McCulloch (1997) described and compared various hierarchical mixture prior formations of variable selection uncertainty in normal linear regression models using related literatures from the works of Carlin and Chib (1995), Chipman (1996), Geweke (1996) and many more authors that has published related papers (see George and McCulloch (1997) for more details). YARDIMCI and Aydın (2002) compared Bayesian approaches such as Zellner, Occam's Window and Gibbs sampling, in terms of selecting the promising subset for the variable selection in a linear regression model, they also studied the behaviour using some classical criteria with different values of β, σ and the prior assumptions. The authors concluded with the claim that Bayesian approaches can be affected by prior changes, the prior information is influential in obtaining zero subsets and the Bayesian approaches compared to the classical behaved better with not very much difference. Sha et al. (2006) proposed a Bayesian variable selection method that allowed the identification of relevant markers by jointly assessing sets of genes using the accelerated failure time (AFT) models with log-normal and log-t distributional assumptions. Their method provided a unified procedure for the selection of relevant genes and the prediction of survivors' functions. Tüchler (2008) presented a Markov Chain Monte Carlo algorithm for both variable covariance selection in the context of logistic mixed-effects models using the SSVS approach to select explanatory variables and determine the structure of the random effects covariance matrix. Vannucci and Stingo (2010) reviewed some Bayesian variable selection methods in linear settings like regression, classification models and mixture models, they went further to propose two novel applications that integrate different sources of biological information into the analysis of experimental data and concluded that their method can handle large number

of covariates larger than the sample size, but were not specific about how large is this largeness. The method involved the evaluation of joint effect of sets of variables and the use of stochastic search techniques to explore the high dimensional variables space just like our proposed model. The application of Bayesian variables selection methods are very common in Genomics, Li and Zhang (2010) considered Bayesian variables selection in structured high-dimensional covariates spaces with applications to Genomics. They defined the situation of high dimensionality from two perspectives; one is that the dimension of the covariate space is often much higher than the number of subjects and secondly, the covariate space is highly structured, and it is desirable to incorporate this structural information into the model building process. The authors approached the Bayesian framework with the assumption that the covariates lie on an undirected graph and formulated using prior on the model space for incorporating structural information. They used simulation studies and application to DNA sequence data to illustrate their proposed method on two different graph structures. Lee et al. (2011) conducted a study on using Bayesian variable selection in Semiparametric survival model for right censored survival data sets. A shrinkage prior obtained from a scale mixture of normal and gamma distributions on the coefficients that corresponds to the predictor variables is used to handle cases when the explanatory variables are of high dimensionality. Their variable selection prior corresponds to the common lasso penalty while the likelihood function is based on the Cox proportional hazards model framework, where the cumulative baseline hazard function is modelled a priori and adaptively control the sparsity of the model. they also developed a fast MCMC algorithm with adaptive jumping rule. Then, the proposed method was evaluated via simulation studies and application to micro array data with right censored survival time and compared with other competing models just like our strategy in this study.

2.4.2 Recent research

Chekouo et al. (2015) proposed a novel Bayesian model to identify microRNAs and the regulatory networks that are associated with survival time by integrating multiple platforms into a statistical model to select the most relevant features which is one of the three groups by which data integration can be categorized according to Daemen et al. (2009). Another novel Bayesian approach developed by Chekouo et al. (2017) integrates multi-regression models

to identify a small set of biomarkers that can accurately predict time-to-event outcomes. It was pointed out that the method fully exploited the amount of available information across platforms and did not exclude any of the subjects from the analysis. Garcia-Donato and Forte (2018) proposed Bayesian methods for hypothesis testing, variable selection, and model averaging in linear models using R, the package is known as `BayesVarSel` in R environment. The disadvantage of this is that, it was built only for the context of the linear model where $n >> p$ and does not apply to high-dimensional cases. The work of Ročková and George (2018) studied the over-looked aspect of a potential penalized likelihood estimation and the use of spike-and-slab methodology for Bayesian variable selection. A new class of self-adaptive penalty functions that arises from a Bayes spike-and-slab formulation was introduced which is known as the spike-and-slab LASSO. The fully Bayes penalty is similar to the oracle performance, in the sense that it provided a viable alternative to cross-validation. A variant of the penalized likelihood approach estimates $\boldsymbol{\beta}$ is shown below, we will not dwell much on this in this study

$$\hat{\boldsymbol{\beta}} = \arg\max_{\beta \in \mathbb{R}^p} \left\{ -\frac{1}{2} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + \text{pen}_\lambda(\boldsymbol{\beta}) \right\}, \tag{2.4.1}$$

where $\text{pen}_\lambda(\boldsymbol{\beta})$ is a penalty function indexed by penalty parameter λ prioritising solutions that are suitably disciplined. The work of Gu et al. (2020) used the spike-and-slab Bayesian variable selection approach in the context of error-prone, self-reported outcomes. The performance of their approach was studied through simulation studies and application of data obtained from the Women's Health Initiative Single Nucleotide Polymorphisms (SNP) Health Association Resource, which includes extensive genotypic (>900,000 SNPs) and phenotypic data on 9,873 African American and Hispanic women. Their study showed improved sensitivity when compared to naive methods that ignores error in the self-reported outcomes, worthy of note is the conclusion that their variable selection accuracy reduced when the outcome is ascertained by error-prone self-reports but better when compared to approaches that neglect to account for the error-prone nature of self-reports. Koop and Korobilis (2020) proposed a variational Bayes algorithm for computationally efficient posterior and predictive inference in time- varying parameter (TVP) models. Their algorithm was evaluated numerically using synthetic data and its computational advantages were established using US microeconomic data. They discovered that the regression models that combined time-varying parameters with the information in many predictors have the potential to improve the prediction of

price inflation (outcome) over a number of alternative predicting models. More recent studies by Kundu et al. (2021) extended the global-local shrinkage idea to a scenario of modelling multiple response variables simultaneously. This approach explored situations of multiple outcome models like multiple regression or seemingly unrelated regression using the Bayesian method. They also went further to state the differences in their framework and that proposed by Bai and Ghosh (2018), see Kundu et al. (2021) for more details. Also, Li and Chekouo (2021) proposed a Bayesian hierarchical model for group selection problem when the group structure in known. The authors used the spike-and-slab priors for regression coefficient with the assumption that the slab component is from the family of non-local priors. They concluded that the proposed Bayesian approach using non-local prior enhances group selection performance which behaved better than those with local prior.

2.5 Markov Chain Monte Carlo (MCMC) method

The Metropolis algorithm was introduced by Metropolis et al. (1953). The algorithm was generalized by Hastings (1970), but it gained recognition after the publication of Gelfand and Smith (1990) on MCMC became widely used in statistics most especially Bayesian statistics. Robert and Casella (2011) contains more on the history of MCMC and Monte Carlo methods. A lot of research has been done using the MCMC method (Gelfand et al., 1990; Gelfand and Smith, 1990; Geman and Geman, 1984; Hastings, 1970; Tanner and Wong, 1987). MCMC method is a used for sampling from probability disributions using Markov Chains. High Performance Computing (HPC) capabilities has increased recently based on this and user-friendly software such as the different programs based on the programming language BUGS (Spiegelhalter et al., 2003), R, Python, Matlab, SAS etc. These developments boosted the use of Bayesian data analyses, particularly in genetics and ecology (Korner-Nievergelt et al., 2015). A number of modifications and extensions of MCMC methods have appeared since the 1990s.

The following illustration of MCMC methods is inspired by the work of Lesaffre and Lawson (2012), a Bayesian biostatistics textbook. Let y be the response variable, x be a random variable with probability density function $P(x|\boldsymbol{\theta})$, and $\boldsymbol{\theta}$ be the parameter vector.

Markov chain can be in two forms, namely; discrete-time Markov chain and continuous-time Markov chain depending on if the random variable is on a countable state space or continuous.

The Gibbs sampler

Gibbs sampler was introduced by Geman and Geman (1984) in the context of image processing. They assumed that the intensity of the pixels in an image have a Gibbs distribution, which is a classical distribution in statistical mechanics. Since this distribution is analytically intractable involving easily a million unknowns. Geman and Geman developed a sampling algorithm to explore the distribution. However, the Gibbs sampler became only popular in the statistical world when Gelfand and Smith (1990) showed its ability to tackle complex estimation problems in Bayesian framework. The approach is based on the property that the joint posterior $P(\theta_1, \theta_2|y)$ is completely determined by the marginal $P(\theta_2|y)$ and the conditional posterior $P(\theta_1|\theta_2, y)$ distribution. A sample from $P(\theta_1, \theta_2|y)$ is then obtained by sampling first from the marginal posterior distribution $P(\theta_2|y)$ yielding a sampled value $\tilde{\theta}_2$ and then from the conditional $P(\theta_1|\tilde{\theta}_2, y)$ yielding $\tilde{\theta}_1$ and hence $(\tilde{\theta}_1, \tilde{\theta}_2)$ is also obtained. The Gibbs sampler, on the other hand, uses the property that (under fairly general regularity conditions) a multivariate distribution is uniquely determined by its conditional distributions (Lesaffre and Lawson, 2012). The Gibbs sampler algorithm is initialized by setting starting values for the parameters, say $\theta_1^{(0)}, \theta_2^{(0)}, \theta_3^{(0)}, \cdots, \theta_p^{(0)}$ and then 'explores' the posterior distribution by generating $\theta_1^{(s)}, \theta_2^{(s)}, \cdots, \theta_p^{(s)}$, $(s = 1, 2, \cdots, S)$ in a sequential manner (S is the desired number of iterations). Therefore, given $\theta_1^{(0)}, \theta_2^{(0)}, \theta_3^{(0)}, \cdots, \theta_p^{(0)}$ at iteration s^{th}, the $(s+1)^{th}$ value for each of the parameters is generated according to the following iterative scheme:

- Sample $\theta_1^{(s+1)}$ from $P(\theta_1|\mathbf{y}, \theta_2^{(s)}, \theta_3^{(s)}, \cdots, \theta_p^{(s)})$
- Sample $\theta_2^{(s+1)}$ from $P(\theta_2|\mathbf{y}, \theta_1^{(s+1)}, \theta_3^{(s)}, \theta_4^{(s)}, \cdots, \theta_p^{(s)})$

  $\vdots$ $\qquad\qquad$ $\vdots$
- Sample $\theta_p^{(s+1)}$ from $P(\theta_p|\mathbf{y}, \theta_1^{(s+1)}, \theta_2^{(s+1)}, \cdots, \theta_{p-1}^{(s+1)})$
- Repeat the above steps from the beginning till S iterations are achieved.

The chain has the Markov property which means that given $\boldsymbol{\theta}^{(s)}, \boldsymbol{\theta}^{(s+1)}$ is independent of $\boldsymbol{\theta}^{(s-1)}, \boldsymbol{\theta}^{(s-2)}$, etc. In a probabilistic notation, $P(\boldsymbol{\theta}^{(s+1)}|\boldsymbol{\theta}^{(s)}, \boldsymbol{\theta}^{(s-1)}, \cdots, \mathbf{y}) = P(\boldsymbol{\theta}^{(s+1)}|\theta^{(s)}, \mathbf{y})$.

After a certain step, the samples from Gibbs sampler can be regarded as a sample from the joint posterior distribution of $\boldsymbol{\theta}$. Then, the first parts of the Gibbs sampler is discarded (known as **burn-in** period), a set of samples drawn from $P(\boldsymbol{\theta}|\mathbf{y})$ is used for Bayesian inference.

The Metropolis algorithm

Let density q is be the proposal density and the proposal density evaluated in $\boldsymbol{\theta}^*$ at iteration s is denoted as $q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(s)})$. Say a Markov chain is at the current state $\boldsymbol{\theta}^{(s)}$ at the s^{th} iteration when exploring the posterior distribution. The general idea behind the Metropolis algorithm goes as follows. The next state in the chain, $\boldsymbol{\theta}^*$, is sampled from a density q having its central location at $\boldsymbol{\theta}^{(s)}$. However, $\boldsymbol{\theta}^*$ is only a proposal for the new position (if it were automatically the next value we would be exploring q and not the posterior). The new position will always be accepted when it is located in an area of higher posterior mass, otherwise accepted with a certain probability. Positions with a lower posterior mass will also be visited otherwise the algorithm would be searching for the (posterior) mode and not exploring the posterior distribution. The probability of accepting the proposed position should be taken such that, after burn-in period, the Markov chain explores $P(\boldsymbol{\theta}|y)$. When the proposal density is symmetric, i.e. $q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(s)}) = q(\boldsymbol{\theta}^{(s)}|\boldsymbol{\theta}^*)$, which is known as the Metropolis algorithm. A popular choice for q is the multivariate normal distribution or multivariate t-distribution with mean equal to the current state $\boldsymbol{\theta}^{(s)}$ and a pre-specified covariance matrix $\boldsymbol{\Sigma}$. We sample $\boldsymbol{\theta}^*$ from $N(\boldsymbol{\theta}^{(s)}, \boldsymbol{\Sigma})$ at the s^{th} iteration to obtain the proposed subsequent value. Obviously, the location parameter $\boldsymbol{\theta}^{(s)}$ of the proposal density changes with s. Furthermore, like the acceptance-rejection algorithm, an appropriate decision rule is needed to either accept or reject the 'proposed' $\boldsymbol{\theta}^*$. Upon acceptance, $\boldsymbol{\theta}^*$ becomes the next value in the Markov chain, i.e. $\boldsymbol{\theta}^{(s+1)} = \boldsymbol{\theta}^*$ and proceed to $\boldsymbol{\theta}^*$. If rejected we remain at the current state $\boldsymbol{\theta}^{(s)}$, hence $\boldsymbol{\theta}^{(s+)} = \boldsymbol{\theta}^{(s)}$. The probability of accepting the proposed value depends on the posterior distribution. When the candidate lies in a region of the posterior distribution with a higher value, i.e. when $P(\boldsymbol{\theta}^*|y)/P(\boldsymbol{\theta}^{(s)}|y) > 1$, then the move will always be made and hence $\boldsymbol{\theta}^{(s+1)} = \boldsymbol{\theta}^*$. In contrast, when the candidate value lies a region of the posterior distribution with a lower value, i.e. when $P(\boldsymbol{\theta}^*|y)/P(\boldsymbol{\theta}^{(s)}|y) < 1$, then the move will be made with probability $\alpha = P(\boldsymbol{\theta}^*|y)/P(\boldsymbol{\theta}^{(s)}|y)$, and then $\boldsymbol{\theta}^{(s+1)} = \boldsymbol{\theta}^*$ with probability α. This procedure ensures that

the posterior distribution will be 'visited' more often at regions where the posterior likelihood is relatively higher.
Therefore, when the chain is at the current state $\boldsymbol{\theta}^{(s)}$, the Metropolis algorithm samples the chain values as follows:

- Sample a candidate $\boldsymbol{\theta}^*$ from the symmetric proposal density $q(\boldsymbol{\theta}^*|\boldsymbol{\theta})$, with $q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(s)})$.
- The next value $\boldsymbol{\theta}^{(s+1)}$ will be equal to
  - $\boldsymbol{\theta}^*$ with probability $\alpha(\boldsymbol{\theta}^{(s)}, \boldsymbol{\theta}^*)$ (accept proposal),
  - $\boldsymbol{\theta}^{(s)}$ otherwise (reject proposal), with

$$\alpha(\boldsymbol{\theta}^{(s)}, \boldsymbol{\theta}^*) = \min\left\{\frac{P(y|\boldsymbol{\theta}^*)P(\boldsymbol{\theta}^*)}{P(y|\boldsymbol{\theta}^{(s)})P(\boldsymbol{\theta}^{(s)})}, 1\right\}, \tag{2.5.1}$$

  where the function $\alpha(\boldsymbol{\theta}^{(s)}, \boldsymbol{\theta}^*)$ is called the probability of a move.

The Metropolis-Hastings algorithm

Recall that for the Metropolis algorithm, the proposal density $q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(s)})$ is symmetric. In some instances, it might be appropriate to choose an asymmetric proposal density. In that case, moving from $\boldsymbol{\theta}$ to $\boldsymbol{\theta}^*$ is not as easy as moving in the opposite direction. To ensure that the posterior distribution is equally accessible in all corners, the asymmetric nature of the proposal density needs to be compensated for. Hastings (1970) extended Metropolis' proposal to an asymmetric proposal density and the approach is referred to as the MH algorithm. The first step in the MH algorithm is to suggest a candidate for a move. At the s^{th} step $\boldsymbol{\theta}^*$ will be sampled from q. As in the Metropolis algorithm, one needs to decide whether to 'move' or to 'stay'. However, in order to ensure that the posterior distribution will be explored equally well in all directions, one should compensate for the fact that moving from $\boldsymbol{\theta}^{(s)}$ to $\boldsymbol{\theta}^*$ is easier or more difficult than in the opposite direction. This is done by changing the probability of making a move at the s^{th} iteration, as follows:

- Sample a candidate $\boldsymbol{\theta}^*$ from the symmetric proposal density $q(\boldsymbol{\theta}^*|\boldsymbol{\theta})$, with $\boldsymbol{\theta} = \boldsymbol{\theta}^{(s)}$.
- The next value $\boldsymbol{\theta}^{(s+1)}$ will be equal to
  - $\boldsymbol{\theta}^*$ with probability $\alpha(\boldsymbol{\theta}^{(s)}, \boldsymbol{\theta}^*)$ (accept proposal),

– $\boldsymbol{\theta}^{(s)}$ otherwise (reject proposal), with

$$\alpha(\boldsymbol{\theta}^{(s)}, \boldsymbol{\theta}^*) = \min\left\{\frac{P(\boldsymbol{\theta}^*|y)q(\boldsymbol{\theta}^{(s)}|\boldsymbol{\theta}^*)}{P(\boldsymbol{\theta}^{(s)}|y)q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(s)})}, 1\right\}, \tag{2.5.2}$$

where the function $\alpha(\boldsymbol{\theta}^{(s)}, \boldsymbol{\theta}^*)$ is called the probability of a move.

The change in acceptance probability was needed in order to ensure that the probability of moving from $\boldsymbol{\theta}$ to $\boldsymbol{\theta}^*$ is equal to the probability of the opposite move. This is called the *reversibility condition*, and the resulting chain is called a *reversible Markov chain*. There are other generalizations of the Metropolis-Hastings algorithm which includes; Independent Metropolis-Hastings, Random-Walk Metropolis-Hastings and Metropolis Within-Gibbs algorithm (Koop et al., 2007).

Chapter 3

Methodology

3.1 Introduction

In this chapter, we describe the types of models and methods used in this study. It also shows the Bayesian framework developed and MCMC algorithm as incorporated in the simulation studies and the application to coronary artery disease (CAD) data.

3.2 Model and Bayesian framework

The **Bayesian approach** applies the concept of the Bayes' rule by specifying a prior probability, $P(\boldsymbol{\theta})$ and then include the data (Likelihood), $P(y|\boldsymbol{\theta})$ to obtain updated information known as the Posterior density, $P(\boldsymbol{\theta}|y)$, which provides posterior probability on which Bayesian Inference depends. Hence, the Bayesian would express equation (2.1.1) as

$$P(\boldsymbol{\theta}|y) \propto P(y|\boldsymbol{\theta})\, P(\boldsymbol{\theta}),$$
$$\text{Posterior} \propto \text{Likelihood} \times \text{Prior}. \tag{3.2.1}$$

3.2.1 Two-part model

We employed the two-part model proposed by Duan et al. (1983) for this study. This type of model has several appealing properties, including a well-behaved likelihood function and more appropriate interpretations than the Tobit and Heckman models if the zeros are true values (Heckman, 1974, 1979; Min and Agresti, 2002; Tobin, 1958). Also, it addresses the

data in their original form, simple to fit, and relatively simple to interpret (Min and Agresti, 2002).

The first part is a binary model for event of having zero or positive values, such as the **logistic regression model**, let p_1 be the total number of features (including the intercept and clinical covariates) in the logistic regression model,

$$\text{logit}(P(y_i = 0)) = \boldsymbol{x}_{\boldsymbol{1i}}^{\top}\boldsymbol{\beta_1}, \tag{3.2.2}$$

where $\boldsymbol{x_{1i}} = (1, x_{1i2}, \cdots, x_{1ip_1})^{\top}$ is the vector of p_1 covariates associated with subject i, $i = 1, \cdots, n$ (including the intercept and clinical covariates) and $\boldsymbol{\beta_1} = (\beta_{10}, \beta_{11}, \cdots, \beta_{1p_1})^{\top}$ is the vector of parameters in the logistic regression model including the intercept and coefficients of clinical covariates.

Let p_2 be the total number of features including the intercept and clinical covariates in the linear regression model. Conditional on a positive value, the second part assumes a **log-normal distribution** or **linear model**; that is

$$\log(y_i|y_i > 0) = \boldsymbol{x}_{\boldsymbol{2i}}^{\top}\boldsymbol{\beta_2} + e_i, \tag{3.2.3}$$

where $e_i \sim N(0, \sigma^2)$ is the random component, $\boldsymbol{x_{2i}} = (1, x_{2i2}, \cdots, x_{2ip_2})^{\top}$ is the vector of p_2 covariates for subject i, $i = 1, \cdots, n_2$ such that $y_i > 0$ (including the intercept and clinical covariates), $\log(y_i|y_i > 0)$ is the log-response variable for positive values, and $\boldsymbol{\beta_2} = (\beta_{20}, \beta_{21}, \cdots, \beta_{2p_2})^{\top}$ is the vector of parameters in the linear model which includes the intercept and coefficients of clinical covariates. The vector of covariates $\boldsymbol{x_{1i}}$ and $\boldsymbol{x_{2i}}$ are the same for this study but may be assumed to be different.

Let $\mathbf{y}^*$ be $n_2 \times 1$ vector of the observed log-responses given that $y_i > 0$. Let p_c be the number of the clinical covariates, then we denote $\boldsymbol{X}^c = \boldsymbol{X}_1^c = \boldsymbol{X}_2^c$ as the $n \times p_c$ clinical covariates (e.g. age, sex, height, body mass index (bmi) etc.) in the combined, logistic, and linear models respectively. These clinical covariates are included in the design matrix for the logistic and linear models of this study. Let $\boldsymbol{X_2}$ be the $n \times p_2$ design matrix of covariates in the linear model and $\boldsymbol{X}_2^*$ be the $n_2 \times p_2$ of observed covariates (including the intercept and clinical covariates) that corresponds to the log-response of positive values, that is $\boldsymbol{X}_2^* = \{\boldsymbol{x_{2i}}, i = 1, \cdots, n, y_i > 0\}$ obtained from $\boldsymbol{X_2}$ in the linear model, then equation (3.2.2) and (3.2.3) can be expressed in matrix form as

$$P(y_i = 0) = \frac{e^{\boldsymbol{x}_{\boldsymbol{1i}}^{\top}\boldsymbol{\beta_1}}}{1 + e^{\boldsymbol{x}_{\boldsymbol{1i}}^{\top}\boldsymbol{\beta_1}}}, \quad i = 1, \cdots, n \tag{3.2.4}$$

and

$$\mathbf{y}^* = \boldsymbol{X}_{\mathbf{2}}^*\boldsymbol{\beta}_{\mathbf{2}} + \boldsymbol{e}, \tag{3.2.5}$$

where $\boldsymbol{\beta}_{\mathbf{1}}$, and $\boldsymbol{\beta}_{\mathbf{2}}$ are the $p_1 \times 1$ and $p_2 \times 1$ vectors of the regression coefficients (including the intercept and coefficients of clinical covariates) respectively and the error component follows a multivariate normal distribution, $\boldsymbol{e} \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma})$, with mean vector $\mathbf{0}$ and variance-covariance matrix, $\boldsymbol{\Sigma} = \sigma^2\mathbf{I}_{\mathbf{n_2}}$ corresponding to the residual variance in the linear model.

3.2.2 Likelihood function

The density of an observation can be written as

$$f\left(y_i\right) = \pi_i \boldsymbol{I}(y_i = 0) \times \left(1 - \pi_i\right) \boldsymbol{I}(y_i > 0) f\left(y_i | y_i > 0\right), \text{ for } y_i \geq 0, i = 1, \cdots, n.$$

where $\pi_i = P(y_i = 0)$, the probability of having zero observations, $1 - \pi_i = P(y_i > 0)$, is the probability of having positive observations, $f\left(y_i | y_i > 0\right)$ is the probability density of y_i given positive response observations, while $\boldsymbol{I}(y_i = 0)$ and $\boldsymbol{I}(y_i > 0)$ are the indicator variables for $y_i = 0$ and $y_i > 0$ respectively.

Then the joint likelihood is given by;

$$L\left(\boldsymbol{\beta_1}, \boldsymbol{\beta_2}, \sigma^2\right) = \prod_{y_i=0} P(y_i = 0) \left[\prod_{y_i>0} P(y_i > 0) f\left(y_i | y_i > 0\right)\right],$$

substituting we have;

$$\begin{aligned} L\left(\boldsymbol{\beta_1}, \boldsymbol{\beta_2}, \boldsymbol{\Sigma}\right) = P(\mathbf{y} | \boldsymbol{\beta_1}, \boldsymbol{\beta_2}, \sigma^2, \boldsymbol{X}) = \prod_{y_i=0} \left[\frac{e^{\boldsymbol{x}_{\mathbf{1}i}^\top \boldsymbol{\beta_1}}}{1 + e^{\boldsymbol{x}_{\mathbf{1}i}^\top \boldsymbol{\beta_1}}}\right] \prod_{y_i>0} \left[\frac{1}{1 + e^{\boldsymbol{x}_{\mathbf{1}i}^\top \boldsymbol{\beta_1}}}\right] \\ \times \left(\frac{1}{\sqrt{2\pi}}\right)^{n_2} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\left(\mathbf{y}^* - \boldsymbol{X}_{\mathbf{2}}^*\boldsymbol{\beta}_{\mathbf{2}}\right)^\top \boldsymbol{\Sigma}^{-1}\left(\mathbf{y}^* - \boldsymbol{X}_{\mathbf{2}}^*\boldsymbol{\beta}_{\mathbf{2}}\right)\right\}. \end{aligned} \tag{3.2.6}$$

as $P(y_i > 0) = 1/(1 + e^{\boldsymbol{x}_{\mathbf{1}i}^\top \boldsymbol{\beta_1}}) = 1 - P\left(y_i = 0\right)$, where $\boldsymbol{X} = \boldsymbol{X_1} = \boldsymbol{X_2}$ to represent the set of covariates in the logistic and linear models for the above equation in (3.2.6), which are not different in this study.

Let the total number of features, $p = p_1 = p_2$. We denote the vectors of indicator variables associated with $\boldsymbol{X_1}$ and $\boldsymbol{X_2^*}$ by $\boldsymbol{z^{(1)}} = (z_{p_c+2}^{(1)}, \cdots, z_p^{(1)})$ and $\boldsymbol{z^{(2)}} = (z_{p_c+2}^{(2)}, \cdots, z_p^{(2)})$ respectively, where p_c is the number of clinical covariates. Each time the indicator variable $z_j^{(1)}$ takes the value one, the corresponding effect β_{1j}, is selected ($j = p_c + 2, p_c + 3, \cdots, p$), otherwise β_{1j}

is not selected in the logistic model (the same is applicable to β_{2j} in the linear model). Let $\boldsymbol{X}^c = \boldsymbol{X}_1^c = \boldsymbol{X}_2^c$ be the design matrix that contains p_c clinical covariates that will not be subject to model selection and would always be constant or included in the indicator vectors $\boldsymbol{z^{(1)}}$ and $\boldsymbol{z^{(2)}}$. The superscript and subscript 1 and 2 refers to the logistic and linear model respectively.

3.2.3 Prior distribution

Given $\boldsymbol{z^{(1)}}$ and $\boldsymbol{z^{(2)}}$, the vector of indicator variables for the regression parameters in the logistic and linear models respectively, we imposed prior distributions on the parameters; $\boldsymbol{\beta_1}, \boldsymbol{\beta_2}$, $\boldsymbol{z^{(1)}}$, $\boldsymbol{z^{(2)}}$, q_1, q_2, and σ^2 as follows;

3.2.3.1 Spike-and-slab prior for regression coefficients

The spike-and-slab prior (Mitchell and Beauchamp, 1988) assume that the components of regression parameters are independent, each having a mixture of two prior distributions: one highly concentrated on zero i.e., the probability of a particular coefficient in the model to be zero (the spike) and the other is the prior distribution for the selected regression coefficients (the slab).

(i.) Given $z_j^{(1)} \sim$ Bernoulli(q_1), where the parameter q_1 is the probability of selecting important features in the logistic model, we impose a spike-and-slab prior on β_{1j} by,

$$\beta_{1j}|z_j^{(1)}, \sigma^2_{\beta_1} \sim \left(1 - z_j^{(1)}\right) \delta_0(\beta_{1j}) + z_j^{(1)} N\left(0, \sigma^2_{\beta_1}\right), \tag{3.2.7}$$

where $\delta_0(.)$ is a Dirac delta function whose value is zero everywhere except at zero and whose integral over the entire real line is one (Dirac, 1981). When feature j is not selected ($z_j^{(1)} = 0$), then $\beta_{1j} = 0$ (spike component), while the prior on β_{1j} follows a normal distribution with prior mean 0 and variance $\sigma^2_{\beta_1}$ when feature j is selected, i.e., $z_j^{(1)} = 1$ (slab component) (Fang et al., 2020). $\sigma^2_{\beta_1}$ is the prior dispersion that corresponds to the selected feature $j = p_c + 2, \cdots, p$ in the logistic model. Also, we imposed a normal distribution with mean 0 and variance $\sigma^2_{\beta_1 pc}$ as the prior on the intercept and clinical covariates $(\beta_{10}, \boldsymbol{\beta}_{1p_c})^\top$, where $\boldsymbol{\beta}_{1p_c} = (\beta_{1j},\ j = 2, \cdots, p_c + 1)$ effects of clinical covariates in the logistic model.

(ii.) Similarly, $z_j^{(2)} \sim$ Bernoulli(q_2), where the parameter q_2 is the probability of selecting important features in the linear model, then the spike-and-slab prior imposed on β_{2j} is

$$\beta_{2j}|z_j^{(2)}, \sigma^2 \sim \left(1 - z_j^{(2)}\right) \delta_0(\beta_{2j}) + z_j^{(2)} N\left(0,\, \sigma^2 s_0^2\right), \tag{3.2.8}$$

Similarly, when feature j is not selected ($z_j^{(2)} = 0$), then $\beta_{2j} = 0$ (spike component) while the prior on β_{2j} follows a normal distribution with prior mean 0 and variance $\sigma^2 s_0^2$ when feature j is selected ($z_j^{(2)} = 1$). σ^2 is the error variance and s_0^2 is prior variance that corresponds to selected feature β_{2j} in the linear model. Also, we imposed a normal distribution with mean 0 and variance $\sigma^2 s_{0pc}^2$ as the prior on the intercept and clinical covariates $(\beta_{20}, \boldsymbol{\beta}_{2p_c})^\top$, where $\boldsymbol{\beta}_{2p_c} = (\beta_{2j},\ j = 2, \cdots, p_c + 1)$ effects of clinical covariates in the linear model.

3.2.3.2 Prior distributions for other parameters

(iii.) For σ^2, let $\sigma^2 \sim IG(\nu_0/2, \nu_0\sigma_0^2/2)$, an inverse-gamma distribution with shape parameter $\nu_0/2$ and scale parameter $\nu_0\sigma_0^2/2$, where the parameters ν_0, σ_0^2 are pre-specified.

$$P(\sigma^2) = \frac{\left(\frac{\nu_0\sigma_0^2}{2}\right)^{\frac{\nu_0}{2}}}{\Gamma(\frac{\nu_0}{2})} \left(\sigma^2\right)^{-\frac{\nu_0}{2}-1} \exp\left\{-\frac{\nu_0\sigma_0^2}{2\sigma^2}\right\}. \tag{3.2.9}$$

(iv.) Prior on q_1 with the prior density denoted by $P(q_1)$: Let q_1 be the probability of selecting important features in the logistic model, and the prior of $q_1 \sim Beta(a, b)$ with parameters a and b, the prior on q_1 is expressed as

$$P(q_1) = \frac{1}{\beta(a,b)} q_1^{a-1} (1 - q_1)^{b-1};\ \ 0 \le q_1 \le 1. \tag{3.2.10}$$

where $\beta(a, b)$ is a beta function with shape parameters a, $b > 0$.

(v.) Prior on q_2 with the prior density denoted by $P(q_2)$: Similarly, let q_2 be the probability of selecting important features in the linear model, and the prior of $q_2 \sim Beta(a_2, b_2)$ with parameters a_2 and b_2. the prior on q_2 is expressed as

$$P(q_2) = \frac{1}{\beta(a_2,b_2)} q_2^{a_2-1} (1 - q_2)^{b_2-1};\ \ 0 \le q_2 \le 1. \tag{3.2.11}$$

where $\beta(a_2, b_2)$ is a beta function with shape parameters a_2, $b_2 > 0$.

(vi.) Let $P(\boldsymbol{z^{(1)}})$ be the prior density of $\boldsymbol{z^{(1)}}$, independent bernoulli distributions with probability of selecting important features q_1 defined as

$$P(\boldsymbol{z^{(1)}}) = \prod_{j=1}^{p_1} q_1^{z_j^{(1)}} (1-q_1)^{1-z_j^{(1)}} = q_1^{\sum_{j=1}^{p_1} z_j^{(1)}} \quad (1-q_1)^{\sum_{j=1}^{p_1}(1-z_j^{(1)})}, \tag{3.2.12}$$

(vii.) Also, the prior density of $\boldsymbol{z^{(2)}}$ denoted by $P(\boldsymbol{z^{(2)}})$, independent bernoulli distributions with probability of selecting important features q_2 is gievn by

$$P(\boldsymbol{z^{(2)}}) = \prod_{j=1}^{p_2} q_2^{z_j^{(2)}} (1-q_2)^{1-z_j^{(2)}} = q_2^{\sum_{j=1}^{p_2} z_j^{(2)}} \quad (1-q_2)^{\sum_{j=1}^{p_2}(1-z_j^{(2)})}, \tag{3.2.13}$$

3.2.4 Posterior distribution

Let $\mathbf{y}$ be the vector of observed response and $\boldsymbol{X} = \boldsymbol{X_1} = \boldsymbol{X_2}$. The joint posterior density of $\boldsymbol{\beta_{1z}}, \boldsymbol{\beta_{2z}},$ and σ^2 can be expressed as

$$\begin{aligned} P(\boldsymbol{\beta_{1z}}, \boldsymbol{\beta_{2z}}, \sigma^2 | \mathbf{y}, \boldsymbol{z^{(1)}}, \boldsymbol{z^{(2)}}, \boldsymbol{X}) &\propto P(\mathbf{y} | \boldsymbol{\beta_{1z}}, \boldsymbol{\beta_{2z}}, \boldsymbol{z^{(1)}}, \boldsymbol{z^{(2)}}, \sigma^2, \boldsymbol{X}) P(\boldsymbol{\beta_{1z}}, \boldsymbol{\beta_{2z}}, \sigma^2 | \boldsymbol{z^{(1)}}, \boldsymbol{z^{(2)}}) \\ &\propto P(\mathbf{y} | \boldsymbol{\beta_{1z}}, \boldsymbol{\beta_{2z}}, \boldsymbol{z^{(1)}}, \boldsymbol{z^{(2)}}, \sigma^2, \boldsymbol{X}) P(\boldsymbol{\beta_{2z}} | \sigma^2, \boldsymbol{z^{(2)}}) P(\boldsymbol{\beta_{1z}} | \boldsymbol{z^{(1)}}) P(\sigma^2 | \boldsymbol{z^{(2)}}). \end{aligned} \tag{3.2.14}$$

Since the joint posterior density in equation (3.2.14) does not have a closed or known form (see appendix A.5 for the marginal probability distribution of $\boldsymbol{\beta_{1z}}, \boldsymbol{\beta_{2z}},$ and σ^2), we proceed to obtain the full conditionals of $\boldsymbol{\beta_{1z}}, \boldsymbol{\beta_{2z}},$ and σ^2, then estimate using the Markov Chain Monte Carlo (MCMC) technique (see literature section 2.5 for more details).

3.2.4.1 The Full Conditional Density of $\boldsymbol{\beta_{2z}} | \sigma^2, \boldsymbol{z^{(2)}}, \mathbf{y}$:

The full conditional density of $\boldsymbol{\beta_{2z}} | \sigma^2, \boldsymbol{z^{(2)}}, \mathbf{y}$ is obtained as follows, (derivation related to this, is shown in appendix A.1)

$$\boldsymbol{\beta_{2z}} | \sigma^2, \boldsymbol{z^{(2)}}, \mathbf{y} \sim \text{MVN}(\boldsymbol{\mu_\beta}, \boldsymbol{\Sigma_\beta}), \tag{3.2.15}$$

where $\boldsymbol{\beta_{2z}}$ is the vector of $\left\{\beta_{2j},\, z_j^{(2)} = 1\right\}$ follows a multivariate normal distribution with mean vector $\boldsymbol{\mu_\beta}$ and variance-covariance matrix $\boldsymbol{\Sigma_\beta}$ for selected covariates (features) i.e.,

$z_j^{(2)} = 1$.

$$\boldsymbol{\Sigma_\beta} = \sigma^2 \left(\boldsymbol{X_{2z}^{*\top} X_{2z}^{*}} + \boldsymbol{\Sigma_{02z}^{-1}} \right)^{-1}$$
$$\boldsymbol{\mu_\beta} = \left(\boldsymbol{X_{2z}^{*\top} X_{2z}^{*}} + \boldsymbol{\Sigma_{02z}^{-1}} \right)^{-1} \left(\boldsymbol{X_{2z}^{*\top}} \mathbf{y}^* + \boldsymbol{\Sigma_{02z}^{-1}} \boldsymbol{\beta_{0z}} \right).$$

where $\boldsymbol{X_{2z}^{*}}$ is the matrix of selected covariates (including the intercept and clinical covariates) from the design matrix $\boldsymbol{X_2}$ when $z_j^{(2)} = 1$ and corresponds to positive log-response $\mathbf{y}^*$, $\boldsymbol{\beta_{0z}}$ is the prior mean vector of $\left\{ \beta_{2j},\, z_j^{(2)} = 1 \right\}$ and $\sigma^2 \boldsymbol{\Sigma_{02z}} = \sigma^2 \text{diag}\left(s_{0pc}^2 \boldsymbol{I}_{p_c+1}, s_0^2 \boldsymbol{I}_{p_{2z}} \right)$ is the prior variance-covariance matrix corresponding to the selected covariates (including the intercept and clinical covariates) when $z_j^{(2)} = 1$, where $p_{2z} = \#\left\{ j = p_c + 2, \cdots, p, \text{ when, } z_j^{(2)} = 1 \right\}$. Also, $\beta_{2j} = 0$ for non-selected covariates (features) i.e., $z_j^{(2)} = 0$, and $j \geq p_c + 2$. Note that, s_0^2 in equation (3.2.8) is the diagonal variance of the prior variance-covariance matrix $\boldsymbol{\Sigma_{02z}}$ when the features are selected ($z_j^{(2)} = 1$). $\sigma^2 s_0^2$ is the prior variance that corresponds to the selected feature $j = p_c + 2, \cdots, p$ in the linear model.

3.2.4.2 The Marginal Posterior Density of $\sigma^2 | \boldsymbol{z^{(2)}}, \mathbf{y}$:

To obtain the marginal posterior density $P(\sigma^2 | \boldsymbol{z^{(2)}}, \mathbf{y})$ (full conditionals), we integrate out $\boldsymbol{\beta_{2z}}$ (Appendix A.2).

$$P(\sigma^2 | \boldsymbol{z^{(2)}}, \mathbf{y}) \propto (\sigma^2)^{-\frac{\nu_n}{2} - 1} \exp\left\{ -\frac{\nu_n \sigma_n^2}{2\sigma^2} \right\}, \tag{3.2.16}$$

where

$$\boldsymbol{\Sigma_b} = \left(\boldsymbol{X_{2z}^{*\top} X_{2z}^{*}} + \boldsymbol{\Sigma_{02z}^{-1}} \right)^{-1}$$
$$\boldsymbol{\mu_b} = \left(\boldsymbol{X_{2z}^{*\top} X_{2z}^{*}} + \boldsymbol{\Sigma_{02z}^{-1}} \right)^{-1} \left(\boldsymbol{X_{2z}^{*\top}} \mathbf{y}^* + \boldsymbol{\Sigma_{02z}^{-1}} \boldsymbol{\beta_{0z}} \right)$$
$$\nu_n = \nu_0 + n_2$$
$$\sigma_n^2 = \frac{1}{\nu_n} \left(\nu_0 \sigma_0^2 + \mathbf{y}^{*\top} \mathbf{y}^* + \boldsymbol{\beta_{0z}^\top \Sigma_{02z}^{-1} \beta_{0z}} - \boldsymbol{\mu_b^\top \Sigma_b^{-1} \mu_b} \right).$$

Therefore, $\sigma^2 | \boldsymbol{z^{(2)}}, \mathbf{y} \sim \text{IG}(\frac{\nu_n}{2}, \frac{\nu_n \sigma_n^2}{2})$ or $\rho^2 = (1/\sigma^2) | \boldsymbol{z^{(2)}}, \mathbf{y} \sim \text{Gamma}(\frac{\nu_n}{2}, \frac{\nu_n \sigma_n^2}{2})$, an Inverse-Gamma $(\frac{\nu_n}{2}, \frac{\nu_n \sigma_n^2}{2})$ density.

3.2.5 Data Augmentation approach for Logistic model

Since there is no closed form of the posterior density of $\boldsymbol{\beta_{1z}} | \mathbf{y}, \boldsymbol{z_j^{(1)}}, \boldsymbol{X_1}$ and due to the logistic model involved, the usual approach will be to use a Metropolis-Hastings algorithm to sample

$\boldsymbol{\beta_{1z}}$. We used the data augmentation approach by the introduction of latent variables (Scott, 2004) and interpretation of the logistic model in terms of utilities (McFadden et al., 1973). Recall, the centered parameterization model

$$P(y_i = 0) = \frac{e^{\boldsymbol{x_{1i}^\top \beta_1}}}{1 + e^{\boldsymbol{x_{1i}^\top \beta_1}}}. \tag{3.2.17}$$

We follow the work of Tüchler (2008) and applied a two-step data augmentation to equation (3.2.17) to obtain a normal linear regression model.

Step 1: the data augmentation step removes the nonlinearity of model (3.2.17) by defining for each observation, latent utilities for choosing categories 0 and positive values (i.e. > 0) respectively as introduced by Scott (2004). Let the utilities for choosing 0 be denoted by u_i and the utilities for choosing positive values be u_{1i}, which follows a standard type I extreme value distribution and are independent of any of the model parameters as tabulated below (Tüchler, 2008).

Table 3.1: Components of the normal mixture approximation of type I extreme value error ε_i (Tüchler, 2008)

r	1	2	3	4	5	6	7	8	9	10
η_r	.004	0.040	.168	.147	.125	.101	.104	0.116	.107	.088
μ_r	5.09	3.29	1.82	1.24	.76	0.39	.04	-.31	-.67	-1.06
σ_r^2	4.5	2.02	1.1	.42	.2	.11	.08	.08	.09	.15

Then,

$$u_i = \boldsymbol{x_{1i}^\top \beta_1} + \mu_{r_i} + \varepsilon_i, \tag{3.2.18}$$

where u_i is the utilities for choosing 0 and the utilities for choosing positive values be u_{1i}, the following relationship holds, $y_i = 0$ iff $u_i > u_{1i}$ and $y_i > 0$ iff $u_i < u_{1i}$. Model (3.2.18) is linear with respect to the regression parameters. μ_{r_i} is the mean group indicator $r = 1, \cdots, 10$ for $i = 1, \cdots, n$ observations and the error is normally distributed $\varepsilon_i \sim N(\mu_{r_i}, \sigma_{r_i}^2)$ with values of mean μ_{r_i} and variance $\sigma_{r_i}^2$ in Table 3.1.

Step 2: Introduce for each error term, an indicator $r_i \in \{1, \cdots, 10\}$ (Tüchler, 2008). Approximate the type I extreme value error ε_i by the following mixture of normal distribution

$$p(\varepsilon_i) = \exp(-\varepsilon_i - e^{-\varepsilon_i}) \approx \sum_{r=1}^{10} \eta_r f_N(\varepsilon_i; \mu_r, \sigma_r^2). \tag{3.2.19}$$

such an approximation was proposed by Frühwirth-Schnatter and Wagner (2006) and applied by Fruehwirth-Schnatter and Frühwirth (2007) in the context of log-linear models. Given these indicators, the model error is normally distributed $\varepsilon_i \sim N(\mu_{r_i}, \sigma_{r_i}^2)$. Including all the utilities u_i and indicators $R_i = r_i;\ i = 1, 2, \cdots, n$, then the linear regression model (3.2.18) in matrix form is expressed as;

$$\mathbf{u} = \boldsymbol{X_1}\boldsymbol{\beta_1} + \boldsymbol{\mu} + \boldsymbol{\varepsilon}, \qquad \boldsymbol{\varepsilon} \sim N(\mathbf{0}, \boldsymbol{\Sigma_n}). \tag{3.2.20}$$

where $\underset{n\times 1}{\boldsymbol{u}} = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix}$, $\underset{n\times p_1}{\boldsymbol{X_1}} = \begin{pmatrix} 1 & x_{12} & \dots & x_{1p_1} \\ 1 & x_{22} & \dots & x_{2p_1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n2} & \dots & x_{np_1} \end{pmatrix}$, $\underset{p_1\times 1}{\boldsymbol{\beta}_1} = \begin{pmatrix} \beta_{10} \\ \beta_{11} \\ \vdots \\ \beta_{1p_1} \end{pmatrix}$, $\underset{n\times 1}{\boldsymbol{\mu}} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{pmatrix}$,

$\underset{n\times 1}{\boldsymbol{\varepsilon}} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$.

In summary, after introducing indicator variable $\boldsymbol{z^{(1)}}$, the posterior density of $\boldsymbol{\beta_{1z}}|\mathbf{u}, \boldsymbol{z^{(1)}}, \mathbf{R}$ is obtained using the data augmentation approach for the logistic model, we sampled u_i by method of generating exponential variates (equation 3.2.22) which is then used to obtain the latent indicators, $\boldsymbol{R}$. The latent indicators, $\boldsymbol{R}$, is obtained from full conditional, $P(R_i = r|u_i, \lambda_i)$ by generating probabilities that an individual i belongs to a group cluster r which is used as the group label to randomly obtain samples for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma_n}$ that is then used to sample from the full conditional density of $\boldsymbol{\beta_{1z}}|\mathbf{u}, \boldsymbol{z^{(1)}}, \mathbf{R}$ as shown in equation (3.2.25), where $r = 1, \cdots, 10$ is the group indicator in Table 3.1 for individual $i = 1, \cdots, n$ observations and $\lambda_i = \exp(\boldsymbol{x}_{\boldsymbol{1zi}}^{\top}\boldsymbol{\beta_{1z}})$ is the exponential of the linear predictor, $\boldsymbol{x_{1zi}}$ is the p_{1z} vector of selected feature for individual i, where p_{1z} is the number of selected features in the logistic model.

3.2.5.1 To Sample the Latent Utilities, $\mathbf{u}|\boldsymbol{\beta}_1$,

Let the exponential of the linear predictor, $\lambda_i = \exp(\boldsymbol{x}_{1zi}^{\top}\boldsymbol{\beta}_{1z})$, where $\boldsymbol{x}_{1zi}$ is the p_{1z} vector of selected feature for individual i, $i = 1, \cdots, n$. The errors in equation (3.2.18) are approximated by the mixture normal distributions with parameters given in Table 3.1, then conditional on λ_i, the latent utilities u_i are derived from Exponential distributions of the form,

$$\begin{aligned}
&\exp(-u_i) \sim \text{Exponential}(\lambda_i + 1), \text{ if } y_i = 0, \\
&\exp(-u_i) \sim \text{Exponential}(\lambda_i + 1) + \text{Exponential}(\lambda_i), \text{ if } y_i > 0.
\end{aligned} \tag{3.2.21}$$

Therefore, we sample the utilities by writing an algorithm for the expression (appendix A.4)

$$u_i = -\log\left(-\frac{\log(D_i)}{\lambda_i + 1} - \frac{\log(D_i^*)}{\lambda_i}\boldsymbol{I}(y_i > 0)\right), \tag{3.2.22}$$

where D_i and D_i^* are uniform(0, 1) random variables and $\boldsymbol{I}(y_i > 0)$ denotes the indicator function for positive log-response values. Equation (3.2.22) is obtained by using the method of generating exponential variates, other methods can be found in Knuth (1998) and Devroye (1986).

3.2.5.2 To Sample the Latent Indicators, $\boldsymbol{R}$ from $P(R_i = r|u_i, \lambda_i)$,

Conditional on the latent utilities u_i and the exponential of the linear predictor $\lambda_i = \exp(\boldsymbol{x}_{1zi}^{\top}\boldsymbol{\beta}_{1z})$. The Latent Indicator, $\boldsymbol{R}$ is obtained by sampling the component indicators R_i for individual i from the discrete density in Table 3.1 for $r = 1, \cdots, 10$ using equation (3.2.24) below.

The prior on the Latent Indicators $\boldsymbol{R}$,

$P(R_i = r) = \eta_r, r = 1, \cdots, 10, \ \forall\, i = 1, \cdots, n$, i.e. $P(R_i = 1) = \eta_1 = 0.004, \cdots, P(R_i = 10) = \eta_{10} = 0.088$.

The full conditional density for sampling the latent indicators, $\boldsymbol{R}$ is given by

$$\begin{aligned}
P(R_i = r|u_i, \lambda_i) &\propto P(u_i|R_i, \lambda_i)P(R_i = r) \\
&\propto \frac{1}{\sigma_r\sqrt{2\pi}} \exp\left\{-\frac{1}{2\sigma_r^2}\left(u_i - \log\lambda_i - \mu_r\right)^2\right\} \times \eta_r \\
P(R_i = r|u_i, \lambda_i) &\propto \phi(u_i; \log\lambda_i - \mu_r, \sigma_r^2) \times \eta_r.
\end{aligned} \tag{3.2.23}$$

where ϕ is the probability density function of the normal distribution with mean $\log \lambda_i + \mu_r$ and variance σ_r^2. Taking log of both sides, we have

$$\log P(R_i = r|u_i, \lambda_i) = \text{Constant} - \log \sigma_r - \frac{1}{2\sigma_r^2} \left(u_i - \log \lambda_i - \mu_r\right)^2 + \log \eta_r. \quad (3.2.24)$$

Exponent of equation (3.2.24) is then normalized to generate probabilities that an individual belongs to a cluster group $r = 1, \cdots, 10$ e.g. $(0, 1, 0, 0, 0, 0, 0, 0, 0, 0)$ gives an index or group label $r = 2$ for an individual i. This is used to sample from the posterior density of $\boldsymbol{\beta_{1z}}|\mathbf{u}, \boldsymbol{z^{(1)}}, \mathbf{R}$ as shown in equation (3.2.25) below with the values of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma_n}$ randomly chosen from Table 3.1 based on the group label assigned by the latent indicator, $\boldsymbol{R}$.

3.2.5.3 The Full Conditional Density of $\boldsymbol{\beta_{1z}}|\mathbf{u}, \boldsymbol{z^{(1)}}, \mathbf{R}$,

The full conditional density, $\boldsymbol{\beta_{1z}}|\mathbf{u}, \boldsymbol{z^{(1)}}, \mathbf{R}$ is given by (appendix A.3)

$$P(\boldsymbol{\beta_{1z}}|\mathbf{u}, \boldsymbol{z^{(1)}}, \mathbf{R}) \propto \exp\left\{-\frac{1}{2}\left(\boldsymbol{\beta_{1z}} - \boldsymbol{b_n}\right)^\top \boldsymbol{B_n^{-1}} \left(\boldsymbol{\beta_{1z}} - \boldsymbol{b_n}\right)\right\}, \quad (3.2.25)$$

where $\boldsymbol{\beta_{1z}}$ is the vector of β_{1j} when covariates are selected ($z_j^{(1)} = 1$), and $\beta_{1j} = 0$ when covariates are not selected ($z_j^{(1)} = 0$) and $j \geq p_c + 2$.

$$\begin{aligned}
\boldsymbol{\mu} &= \mu_{r_1}, \cdots, \mu_{r_n}, \\
\boldsymbol{\Sigma_n} &= \text{diag}\left\{\sigma_{r_1}^2, \sigma_{r_2}^2, \cdots, \sigma_{r_n}^2\right\}, \\
\boldsymbol{B_n} &= \left(\boldsymbol{B_{0z_1}^{-1}} + \boldsymbol{X_{1z}^\top \Sigma_n^{-1} X_{1z}}\right)^{-1}, \\
\boldsymbol{b_n} &= \boldsymbol{B_n}\left(\boldsymbol{B_{0z_1}^{-1} b_{0z1}} + \boldsymbol{X_{1z}^\top \Sigma_n^{-1}}(\mathbf{u} - \boldsymbol{\mu})\right).
\end{aligned}$$

where $\boldsymbol{b_{0z1}}$ is the prior mean vector of zeros and $\boldsymbol{B_{0z_1}} = \text{diag}\left(\sigma_{\beta_{1pc}}^2 \boldsymbol{I}_{p_c+1}, \sigma_{\beta_1}^2 \boldsymbol{I}_{p_{1z}}\right)$ is the prior variance-covariance matrix corresponding to the selected covariates (including the intercept and clinical covariates) when $z_j^{(1)} = 1$, where $p_{1z} = \#\left\{j = p_c + 2, \cdots, p, \text{ when, } z_j^{(1)} = 1\right\}$. $\boldsymbol{X_{1z}}$ is the matrix of selected covariates (including the intercept and clinical covariates) from the design matrix $\boldsymbol{X_1}$ that corresponds to $z_j^{(1)} = 1$ in the logistic model. Therefore, we sample directly from the posterior density of $\boldsymbol{\beta_{1z}}|\mathbf{u}, \boldsymbol{z^{(1)}}, \mathbf{R} \sim \mathbf{MVN}(\boldsymbol{b_n}, \boldsymbol{B_n})$ when $z_j^{(1)} = 1$, and 0 otherwise.

3.3 Variable Selection Method

Here, we obtained the full conditional densities of $\boldsymbol{z^{(1)}}$ and $\boldsymbol{z^{(2)}}$, then applied the Stochastic Search Variable Selection (SSVS) procedure. The SSVS procedure is used to randomly sample the indicator variables from its full conditional by searching the space of feature subsets and identifying the most promising features in the logistic and linear models. In this study we applied the SSVS procedure based on two assumptions. First, is to select features independently from each model (Method 1: Linear and Logistic). Secondly, we assumed that the same set of features are sampled from both models involved (Method 2: Combined model where the indicator variables $\boldsymbol{z^{(1)}} = \boldsymbol{z^{(2)}}$). The SSVS procedure is shown in the MCMC Algorithm section where $\boldsymbol{z^{(1)}}$ and $\boldsymbol{z^{(2)}}$ is updated (section 3.4).

3.3.1 Full Conditional Density of Indicator Variable, $z^{(1)}$ and $z^{(2)}$ for Method 1

Method 1: Assume that the **features/covariates are selected independently from the logistic and linear models**. To sample $\boldsymbol{z^{(1)}}$ from the logistic model we need to sample from the joint density $(\boldsymbol{z^{(1)}}, \boldsymbol{\beta_{1z}})$.

$$P(\boldsymbol{z^{(1)}}, \boldsymbol{\beta_{1z}}|\mathbf{u}, q_1) \propto P(\boldsymbol{\beta_{1z}}|\mathbf{u}, \boldsymbol{z^{(1)}})P(\boldsymbol{z^{(1)}}|\mathbf{u}, q_1),$$

where $P(\boldsymbol{z^{(1)}}|\mathbf{u}, q_1)$ is the marginal of $\boldsymbol{z^{(1)}}$ after integrating out $\boldsymbol{\beta_{1z}}$.

3.3.1.1 To Sample $z^{(1)}$ from $P(\boldsymbol{z^{(1)}}|\mathbf{u}, q_1)$,

Firstly, we sample $\boldsymbol{z^{(1)}} \sim P(\boldsymbol{z^{(1)}}|\mathbf{u}, q_1) \propto P(\mathbf{u}|z^{(1)}, \boldsymbol{\Sigma_n})P(\boldsymbol{z^{(1)}})$, after integrating out $\boldsymbol{\beta_{1z}}$. The likelihood, $P(\mathbf{u}|\boldsymbol{z^{(1)}}, \boldsymbol{\Sigma_n})$ is obtained as follows using the spike-and-slab prior defined on β_{1j} in equation(3.2.7);

$$\text{recall,} \qquad \mathbf{u} = \boldsymbol{X_1\beta_1} + \boldsymbol{\mu} + \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim N(\mathbf{0}, \boldsymbol{\Sigma_n}), \tag{3.3.1}$$

$$P(\mathbf{u}|\boldsymbol{z^{(1)}}, \boldsymbol{\Sigma_n}) = (2\pi)^{-\frac{n}{2}} |\boldsymbol{\Sigma_{bb}}|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\mathbf{u} - \boldsymbol{\mu_{bb}})^\top \boldsymbol{\Sigma_{bb}^{-1}} (\mathbf{u} - \boldsymbol{\mu_{bb}})\right\}, \tag{3.3.2}$$

where

$$\begin{aligned}
\boldsymbol{\mu_{bb}} &= E(\mathbf{u}) = \boldsymbol{X_{1z}} \boldsymbol{b_{0z_1}} + \boldsymbol{\mu}, \\
\boldsymbol{\Sigma_{bb}} &= Var(\mathbf{u}) = \boldsymbol{X_{1z}} \boldsymbol{B_{0z_1}} \boldsymbol{X_{1z}^{\top}} + \boldsymbol{\Sigma_{\mathbf{n}}}, \\
\boldsymbol{\mu} &= \mu_{r_1}, \cdots, \mu_{r_n}, \\
\boldsymbol{\Sigma_n} &= \text{diag}\left\{\sigma^2_{r_1}, \sigma^2_{r_2}, \cdots, \sigma^2_{r_n}\right\},
\end{aligned}$$

where $\boldsymbol{b_{0z1}}$ is the prior mean vector of zeros and $\boldsymbol{B_{0z_1}} = \text{diag}\left(\sigma^2_{\beta_{1pc}} \boldsymbol{I}_{p_c+1}, \sigma^2_{\beta_1} \boldsymbol{I}_{p_{1z}}\right)$ is the prior variance-covariance matrix corresponding to the selected covariates (including the intercept and clinical covariates) when $z_j^{(1)} = 1$, where $p_{1z} = \#\left\{j = p_c + 2, \cdots, p, \text{ when, } z_j^{(1)} = 1\right\}$. $\boldsymbol{X_{1z}}$ is the matrix of selected covariates (including intercept and clinical covariates) from the design matrix $\boldsymbol{X_1}$ that corresponds to $z_j^{(1)} = 1$ in the logistic model, therefore, $\mathbf{u}|\boldsymbol{z^{(1)}}, \boldsymbol{\Sigma_n} \sim \mathbf{MVN}(\boldsymbol{\mu_{bb}},\ \boldsymbol{\Sigma_{bb}})$, a multivariate normal distribution with mean $\boldsymbol{\mu_{bb}}$ and variance-covariance matrix $\boldsymbol{\Sigma_{bb}}$.

Then, we sample $\boldsymbol{z^{(1)}} \sim P(\boldsymbol{z^{(1)}}|\mathbf{u}, q_1) \propto P(\mathbf{u}|\boldsymbol{z^{(1)}}, \boldsymbol{\Sigma_n}) P(\boldsymbol{z^{(1)}})$. The logarithm of the marginal density of $\boldsymbol{z^{(1)}}$ is then given by

$$\begin{aligned}
\log P(\boldsymbol{z^{(1)}}|\mathbf{u}, q_1) &= \text{Const.} + \log P(\mathbf{u}|\boldsymbol{z^{(1)}}, \boldsymbol{\Sigma_n}) + \log P(\boldsymbol{z^{(1)}}) \\
\log P(\boldsymbol{z^{(1)}}|\mathbf{u}, q_1) &= \text{Const.} - \frac{n}{2}\log(2\pi) - \frac{1}{2}\log|\boldsymbol{\Sigma_{bb}}| - \frac{1}{2}(\mathbf{u} - \boldsymbol{\mu_{bb}})^{\top} \boldsymbol{\Sigma_{bb}^{-1}} (\mathbf{u} - \boldsymbol{\mu_{bb}}) \\
&\quad + \sum_{j=1}^{p_1} z_j^{(1)} \log q_1 + \sum_{j=1}^{p_1} (1 - z_j^{(1)}) \log(1 - q_1). \qquad (3.3.3)
\end{aligned}$$

Therefore, we sample $\boldsymbol{z^{(1)}}$ from the full conditional, $P(\boldsymbol{z^{(1)}}|\mathbf{u}, q_1)$, which is proportional to the exponent of equation (3.3.3).

3.3.1.2 To Sample q_1 from Full Conditional $P(q_1|\boldsymbol{z^{(1)}})$,

The full conditional to sample q_1 is given by

$$\begin{aligned} P(q_1|\boldsymbol{z^{(1)}}) &\propto P(\boldsymbol{z^{(1)}}|q_1)P(q_1) \qquad (3.3.4)\\ &\propto q_1^{\sum_{j=1}^{p_1} z_j^{(1)}} (1-q_1)^{\sum_{j=1}^{p_1}(1-z_j^{(1)})} \times \frac{1}{\beta(a,b)} q_1^{a-1}(1-q_1)^{b-1} \\ &\propto q_1^{\left(\sum_{j=1}^{p_1} z_j^{(1)} + a\right)-1} (1-q_1)^{\left(\sum_{j=1}^{p_1}(1-z_j^{(1)}) + b\right)-1} \\ P(q_1|\boldsymbol{z^{(1)}}) &\propto \frac{1}{\beta(a^*,b^*)} q_1^{a^*-1}(1-q_1)^{b^*-1} \qquad (3.3.5) \end{aligned}$$

Where $a^* = \sum_{j=1}^{p_1} z_j^{(1)} + a,\ \ b^* = \sum_{j=1}^{p_1}(1-z_j^{(1)}) + b.$
Therefore, we sample $q_1|\boldsymbol{z^{(1)}}$ from a $Beta(a^*, b^*)$, beta distribution with parameters a^* and b^*.

3.3.1.3 To Sample $\boldsymbol{z^{(2)}}$ from $P(\boldsymbol{z^{(2)}}|\mathbf{y}^*, \sigma^2, q_2)$,

Similarly, to sample $\boldsymbol{z^{(2)}} \sim P(\boldsymbol{z^{(2)}}|\mathbf{y}^*, \sigma^2, q_2) \propto P(\mathbf{y}^*|\boldsymbol{z^{(2)}}, \sigma^2, q_2)P(\boldsymbol{z^{(2)}})$. The full conditional density of $\boldsymbol{z^{(2)}}|\mathbf{y}^*, \sigma^2, q_2$ is obtained as follows;
recall,

$$\mathbf{y}^* = \boldsymbol{X^*_{2z}}\boldsymbol{\beta_{2z}} + \boldsymbol{e}, \qquad (3.3.6)$$

where the error component, $\boldsymbol{e} \sim \mathrm{N}(\mathbf{0}, \sigma^2\boldsymbol{I_{n_2}})$, and $\sigma^2\boldsymbol{I_{n_2}}$ is the variance-covariance matrix corresponding to the residual variance in the linear model.
Obtaining the likelihood $P(\mathbf{y}^*|\boldsymbol{\beta_{2z}}, \boldsymbol{X^*_{2z}})$ is similar to obtaining $P(\mathbf{y}^*|\boldsymbol{z^{(2)}}, \sigma^2, q_2)$ by introducing indicator variable $\boldsymbol{z^{(2)}}$ into $P(\mathbf{y}^*|\boldsymbol{\beta_{2z}}, \boldsymbol{X^*_{2z}})$. As defined earlier $\boldsymbol{X^*_{2z}}$ is the matrix of selected covariates from the design matrix $\boldsymbol{X_2}$ that corresponds to $z_j^{(2)} = 1$ and positive log response in the linear model, $\boldsymbol{\beta_{0z}}$ is the prior mean vector of zeros and $\sigma^2\boldsymbol{\Sigma_{02z}} = \sigma^2\mathrm{diag}\left(s^2_{0pc}\boldsymbol{I}_{p_c+1}, s^2_0\boldsymbol{I}_{p_{2z}}\right)$ is the prior variance-covariance matrix corresponding to the selected covariates when $z_j^{(2)} = 1$. Then using the spike-and-slab prior on β_{2j} in equation (3.2.8), the likelihood of $\mathbf{y}^*|\boldsymbol{z^{(2)}}, \sigma^2$ is obtained by,

$$P(\mathbf{y}^*|\boldsymbol{z^{(2)}}, \sigma^2) = (2\pi)^{-\frac{n_2}{2}} |\boldsymbol{\Sigma_{BB}}|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\left(\mathbf{y}^* - \boldsymbol{\mu_{BB}}\right)^\top \boldsymbol{\Sigma^{-1}_{BB}}\left(\mathbf{y}^* - \boldsymbol{\mu_{BB}}\right)\right\}, \qquad (3.3.7)$$

where

$$\boldsymbol{\Sigma_{02z}} = \text{diag}\left(s^2_{0pc}\boldsymbol{I}_{p_c+1}, s^2_0\boldsymbol{I}_{p_{2z}}\right),$$
$$\boldsymbol{\mu_{BB}} = E(\mathbf{y}^*) = \boldsymbol{X^*_{2z}}\boldsymbol{\beta_{0z}},$$
$$\boldsymbol{\Sigma_{BB}} = Var(\mathbf{y}^*) = \boldsymbol{X^*_{2z}}\sigma^2\boldsymbol{\Sigma^{-1}_{02z}}\boldsymbol{X^{*\top}_{2z}} + \sigma^2\mathbf{I_{n_2}},$$

which implies that $\mathbf{y}^*|\boldsymbol{z^{(2)}}, \sigma^2 \sim \mathbf{MVN}(\boldsymbol{\mu_{BB}},\ \boldsymbol{\Sigma_{BB}})$, a multivariate normal distribution with mean $\boldsymbol{\mu_{BB}}$ and variance-covariance matrix $\boldsymbol{\Sigma_{BB}}$ corresponding to the selected covariates when $z^{(2)}_j = 1$ and 0 otherwise in the linear model.
Then, the logarithm of the marginal density of $\boldsymbol{z^{(2)}}$ is then given by

$$\log P(\boldsymbol{z^{(2)}}|\mathbf{y}^*, \sigma^2, q_2) = \text{Const.} + \log P(\mathbf{y}^*|\boldsymbol{z^{(2)}}, \sigma^2) + \log P(\boldsymbol{z^{(2)}})$$
$$\log P(\boldsymbol{z^{(2)}}|\mathbf{y}^*, \sigma^2, q_2) = \text{Const.} - \frac{n_2}{2}\log(2\pi) - \frac{1}{2}\log|\boldsymbol{\Sigma_{BB}}| - \frac{1}{2}(\mathbf{y}^* - \boldsymbol{\mu_{BB}})^\top \boldsymbol{\Sigma^{-1}_{BB}}(\mathbf{y}^* - \boldsymbol{\mu_{BB}})$$
$$+ \sum_{j=1}^{p_2} z^{(2)}_j \log q_2 + \sum_{j=1}^{p_2}(1 - z^{(2)}_j)\log(1 - q_2). \tag{3.3.8}$$

Therefore, we sample $\boldsymbol{z^{(2)}}$ from the full conditional, $P(\boldsymbol{z^{(2)}}|\mathbf{y}^*, \sigma^2, q_2)$, which is proportional to the exponent of equation (3.3.8).

3.3.1.4 To Sample q_2 from Full Conditional $P(q_2|\boldsymbol{z^{(2)}})$,

Similarly, let q_2 be the probability parameter for selecting significant features/covariates in the linear model. The full conditional, $P(q_2|\boldsymbol{z^{(2)}}) \propto P(\boldsymbol{z^{(2)}}|q_2)P(q_2)$ is

$$P(q_2|\boldsymbol{z^{(2)}}) = \frac{1}{\beta(a^*_2, b^*_2)} q_2^{a^*_2 - 1}(1 - q_2)^{b^*_2 - 1}, \tag{3.3.9}$$

where $a^*_2 = \sum_{j=1}^{p_2} z^{(2)}_j + a_2,\ \ b^*_2 = \sum_{j=1}^{p_2}(1 - z^{(2)}_j) + b_2.$
Therefore, we sample $q_2|\boldsymbol{z^{(2)}}$ from $\sim Beta(a^*_2, b^*_2)$, beta distribution with parameters a^*_2 and b^*_2.

3.3.2 Full Conditional Density of Indicator Variable, $\boldsymbol{z^{(1)}}$ and $\boldsymbol{z^{(2)}}$ for Method 2: Combined Model such that $\boldsymbol{z^{(1)} = z^{(2)}}$

Method 2: This is also known as the **combined model** i.e., we assume that **the same features/covariates $\boldsymbol{z^{(1)} = z^{(2)}}$ are selected from the logistic and linear model**.
Let $\boldsymbol{z = z^{(1)} = z^{(2)}}$ for simplicity.

3.3.2.1 To Sample $\boldsymbol{z}$ from $P\left(\boldsymbol{z}|\mathbf{u}, \mathbf{y}^*, \sigma^2, q\right)$

The full conditional of $\boldsymbol{z}|\mathbf{u}, \mathbf{y}^*, \sigma^2, q$ combines the likelihood of the logistic and linear model $P\left(\mathbf{u}|\mathbf{y}, \boldsymbol{z}\right)$ and $P(\mathbf{y}^*|\boldsymbol{z}, \sigma^2)$ respectively and prior of $\boldsymbol{z}$, $P(\boldsymbol{z}|q)$ as shown below;

$$P\left(\boldsymbol{z}|\mathbf{u}, \mathbf{y}^*, \sigma^2, q\right) \propto P\left(\mathbf{u}|\mathbf{y}, \boldsymbol{z}\right) P(\mathbf{y}^*|\boldsymbol{z}, \sigma^2) P(\boldsymbol{z}|q) \tag{3.3.10}$$

where $\boldsymbol{z} = \boldsymbol{z^{(1)}} = \boldsymbol{z^{(2)}}$. In log form we have

$$\begin{aligned}
\log P\left(\boldsymbol{z}|\mathbf{u}, \mathbf{y}^*, \sigma^2, q\right) &= \text{Const.} + \log P\left(\mathbf{u}|\mathbf{y}, \boldsymbol{z}\right) + \log P(\mathbf{y}^*|\boldsymbol{z}, \sigma^2) + \log P\left(\boldsymbol{z}|q\right) \\
\log P\left(\boldsymbol{z}|\mathbf{u}, \mathbf{y}^*, \sigma^2, q\right) &= \text{Const.} + \log \text{MVN}(\boldsymbol{X_{1z}} \boldsymbol{b_{0z_1}} + \boldsymbol{\mu},\ \boldsymbol{X_{1z}} \boldsymbol{B_{0z_1}} \boldsymbol{X_{1z}^\top} + \boldsymbol{\Sigma_n}) \\
&\quad + \log \text{MVN}(\boldsymbol{X_{2z}^*} \boldsymbol{\beta_{0z}},\ \boldsymbol{X_{2z}^*} \sigma^2 \boldsymbol{\Sigma_{02z}^{-1}} \boldsymbol{X_{2z}^{*\top}} + \boldsymbol{\sigma^2 I_{n_2}}) \\
&\quad + \sum_{j=1}^{p_1} z_j \log q + \sum_{j=1}^{p_1} (1 - z_j) \log(1 - q).
\end{aligned} \tag{3.3.11}$$

Therefore, we sample $\boldsymbol{z}$ from the full conditional, $P\left(\boldsymbol{z}|\mathbf{u}, \mathbf{y}^*, \sigma^2, q\right)$, which is proportional to the exponent of equation (3.3.11).

3.3.2.2 To Sample q from Full Conditional $P\left(q|\boldsymbol{z}\right)$

Since the assumption is that we have the same set of covariates in the logistic and linear model, let $q = q_1 = q_2$ be the probability of selecting common or same significant features/covariates in the logistic and linear model respectively, they will have the same prior of $q \sim Beta(a, b)$, then the full conditional is also the same as that of $q_1|\boldsymbol{z^{(1)}}$, and $q_2|\boldsymbol{z^{(2)}}$ which can be expressed as

$$q|\boldsymbol{z} \sim Beta(a^*, b^*), \tag{3.3.12}$$

where $a^* = \sum_{j=1}^{p_1} z_j + a$ and $b^* = \sum_{j=1}^{p_1} (1 - z_j) + b$.
Therefore, we sample $q|\boldsymbol{z}$ from a $Beta(a^*, b^*)$, beta distribution with parameters a^* and b^*.

3.4 The MCMC Algorithm

Steps to Sample from the Full Conditionals of $\boldsymbol{\beta_2}$, $\boldsymbol{\beta_1}$, q_1, q_2, σ^2, u, R, $\boldsymbol{z^{(1)}}$, and $\boldsymbol{z^{(2)}}$.

Initialize $\sigma^{2(s)}$, $\boldsymbol{\beta_1}^{(s)}$, $q_1^{(s)}$, $q_2^{(s)}$, $\boldsymbol{z^{(1)}}^{(s)}$, $\boldsymbol{z^{(2)}}^{(s)}$, $\mathbf{R}$, we also incorporated the two Methods under study for $\boldsymbol{z^{(1)}}$ and $\boldsymbol{z^{(2)}}$, where $s = 1, 2, \cdots, S$ desired number of iterations.

For Method 1:

- **For the Linear regression model**

  i. Update σ^2: sample $\sigma^{2(s+1)} \sim \text{IG}(\frac{\nu_n}{2}, \frac{\nu_n \sigma_n^2}{2})$,

  ii. Update q_2: sample $q_2|\boldsymbol{z^{(2)}} \sim Beta(a_2^*, b_2^*)$,

  iii. Update $\boldsymbol{z^{(2)}}$**:** sample $\boldsymbol{z^{(2)}} \sim P(\boldsymbol{z^{(2)}}|\mathbf{y}^*, \sigma^2, q_2)$, i.e., from its full conditional. Using Stochastic Search Variable Selection (SSVS) method by randomly choosing either of the following:

  a. **Add/delete:** randomly pick one of the p_2 indices in $\boldsymbol{z^{(2)}}$ and switch their values from 0 to 1, or 1 to 0.

  b. **Swap:** pick independently and at random, a 0 and a 1 in $\boldsymbol{z^{(2)}}$ and switch their values.

  The proposed $\boldsymbol{z^{(2)*}}$ is accepted with probability

$$\min\left\{1, \frac{P(\boldsymbol{z^{(2)*}}|\mathbf{y}^*, \sigma^2, q_2)}{P(\boldsymbol{z^{(2)}}|\mathbf{y}^*, \sigma^2, q_2)}\right\} = \min\left\{1, \frac{P(\mathbf{y}^*|\boldsymbol{z^{(2)*}}, \sigma^2)P(\boldsymbol{z^{(2)*}})}{P(\mathbf{y}^*|\boldsymbol{z^{(2)}}, \sigma^2)P(\boldsymbol{z^{(2)}})}\right\},$$

  for the linear model.

  iv. Update $\boldsymbol{\beta_2}$: sample $\boldsymbol{\beta}_{\boldsymbol{2z}}^{\boldsymbol{(s+1)}} \sim \text{MVN}(\boldsymbol{\mu_\beta}, \boldsymbol{\Sigma_\beta})$, where $\boldsymbol{\Sigma_\beta}$ depends on $\sigma^{2(s+1)}$, and $\boldsymbol{\beta}_{\boldsymbol{2z}}^{\boldsymbol{(s+1)}}$ corresponds to the selected covariates when $z_j^{(2)} = 1$, while $\beta_{2j} = 0$ when $z_j^{(2)} = 0$ in the linear model.

- **For the Logistic regression model**

  v. Update u: sample the **Latent Utilities**, **u** from

$$u_i = -\log\left(-\frac{\log(D_i)}{\lambda_i + 1} - \frac{\log(D_i^*)}{\lambda_i}\boldsymbol{I}(y_i > 0)\right).$$

  vi. Update R: sample the **Latent Indicators**, **R** from $P(R_i = r|u_i, \lambda_i)$,

$$\log P(R_i = r|u_i, \lambda_i) = \text{Constant} - \log\sigma_r - \frac{1}{2\sigma_r^2}\left(u_i - \log\lambda_i - \mu_r\right)^2 + \log\eta_r.$$

  vii. Update q_1: sample $q_1|\boldsymbol{z^{(1)}} \sim Beta(a^*, b^*)$,

  viii. Update $\boldsymbol{z^{(1)}}$**:** sample $\boldsymbol{z^{(1)}} \sim P(\boldsymbol{z^{(1)}}|\mathbf{u}, q_1)$, i.e., from its full conditional. Using the Stochastic Search Variable Selection (SSVS) method by randomly choosing either of the following:

a. **Add/delete:** randomly pick one of the p_1 indices in $\boldsymbol{z^{(1)}}$ and switch their values from 0 to 1, or 1 to 0.

b. **Swap:** pick independently and at random, a 0 and a 1 in $\boldsymbol{z^{(1)}}$ and switch their values.

The proposed $\boldsymbol{z^{(1)*}}$ is accepted with probability

$$\min\left\{1, \frac{P(\boldsymbol{z^{(1)*}}|\mathbf{u}, q_1)}{P(\boldsymbol{z^{(1)}}|\mathbf{u}, q_1)}\right\} = \min\left\{1, \frac{P(\mathbf{u}|\mathbf{y}, \boldsymbol{z^{(1)*}})P(\boldsymbol{z^{(1)*}})}{P(\mathbf{u}|\mathbf{y}, \boldsymbol{z^{(1)}})P(\boldsymbol{z^{(1)}})}\right\},$$

for the logistic-model,

ix. Update $\boldsymbol{\beta_1}$: sample $\boldsymbol{\beta_{1z}^{(s+1)}} \sim \text{MVN}(\boldsymbol{b_n}, \boldsymbol{B_n})$ for $\boldsymbol{\beta_{1z}^{(s+1)}}$ which corresponds to the selected covariates $z_j^{(2)} = 1$, while $\beta_{1j} = 0$ when $z_j^{(2)} = 0$ in the logistic model.

- Repeat (i.) to (ix.) until the desired number of iterations is achieved.

- **For Method 2:** The **Combined Model** where it is assumed that $\boldsymbol{z^{(1)}} = \boldsymbol{z^{(2)}}$, we only need to sample $q = q_1 \sim Beta(a^*, b^*)$ and the proposal $\boldsymbol{z^*}$ is accepted with probability

$$\min\left\{1, \frac{P(\boldsymbol{z^*}|\mathbf{u}, \mathbf{y}^*, \sigma^2, q)}{P(\boldsymbol{z}|\mathbf{u}, \mathbf{y}^*, \sigma^2, q)}\right\} = \min\left\{1, \frac{P(\mathbf{u}, \mathbf{y}^*|\boldsymbol{z^*})P(\boldsymbol{z^*}|q)}{P(\mathbf{u}, \mathbf{y}^*|\boldsymbol{z})P(\boldsymbol{z}|q)}\right\}.$$

Chapter 4

Simulation Study

In this chapter, we used simulation studies to assess the performance of our proposed approach, generated data based on the various scenarios of features, p and sample sizes, n, set hyperparameters and initial values for our Markov Chain Monte Carlo (MCMC) algorithm. Furthermore, we assessed the MCMC algorithm for convergence and compared our approach to some existing variable selection approaches (Frequentist and Bayesian), followed by the discussion of the findings.

4.1 Description of simulation study data

We generated the set of covariates and 3 clinical covariates independently from $N(0,1)$, a standard normal distribution with mean 0, and variance 1 for the logistic and linear models respectively. Then chose 20 (out of p features) as important features. These data simulation was done for $n = 500$ and $n = 300$ observations (sample size) with respect to various features or covariates $p = 1000$, 500, 200, and 50. This data simulation was replicated 10 times to obtain 10 different datasets used in this study. Let $\boldsymbol{X} = \boldsymbol{X_1} = \boldsymbol{X_2}$, we expect to have high marginal posterior probabilities (MPP) ($P(z_j^{(\ell)} = 1|\boldsymbol{X}, \mathbf{y})$, $\ell = 1, 2,\ j = 1, \cdots, p$), for selected features and lower probabilities for the non selected features in the model. For the

true models, we pre-specified true values of ones and zeros for the parameters $\beta_{\ell j}$, $(\ell = 1, 2)$. True values of $\beta_{\ell j} = 1$ for selected features, $z_j^{(\ell)} = 1$ and non selected features, $\beta_{\ell j} = 0$, for $z_j^{(\ell)} = 0$ in the logistic and linear models, this is to assess the predictive performance of our proposed approach. The probability of success based on the two-part model of this study is computed by $P(y_i = 0|\boldsymbol{X_1}) = e^{\boldsymbol{x_{1i}^\top \beta_1}}/(1 + e^{\boldsymbol{x_{1i}^\top \beta_1}})$, and used to obtain the zero response data, $\mathbf{y}$ from a binomial distribution for the logistic model, the positive response data is obtained by using the linear model in equation (3.2.3), where the random component, e_i is sampled from a normal distribution with mean 0 and variance 0.25. Then, the semicontinuous response $\mathbf{y}$ is obtained by combining the zeros and positive values which mimics the real-life type of semicontinuous response data. Because our aim is to examine the performance of the proposed method (method 2: combined model) in the context of high dimensional data analysis, most of our focus will be on the cases where the number of features is larger than the sample size, i.e., $p >> n$ to monitor the performance of our approach.

4.2 Hyperparameter Setting

To run the MCMC algorithm for both the simulated data and observed data we set hyperparameter values as follows. For the spike-and-slab prior imposed on β_{1j} in the logistic regression model, we set the prior mean vector $\boldsymbol{b_{0z1}} = (0_1, \cdots, 0_p)^\top$, equal variance, $\sigma^2_{\beta_1 pc} = \sigma^2_{\beta_1} = 0.5$, identity matrix $\boldsymbol{I_{p_{1z}}}$ such that the prior variance-covariance matrix $\boldsymbol{B_{0z1}} = \text{diag}\left(\sigma^2_{\beta_{1pc}} \boldsymbol{I}_{p_c+1}, \sigma^2_{\beta_1} \boldsymbol{I}_{p_{1z}}\right)$ when $z_j^{(1)} = 1$ while the prior on $\beta_{1j} = 0$ when features are not selected, $z_j^{(1)} = 0$. Similarly, for the spike-and-slab prior imposed on β_{2j} in the linear regression model, we set the prior mean vector $\boldsymbol{\beta_{0z_2}} = (0_1, \cdots, 0_p)^\top$, prior variance $s^2_{0pc} = s^2_0 = 0.5$, σ^2 is the error variance which is unknown, and the prior variance-covariance matrix on β_{2j}, which is $\sigma^2 \boldsymbol{\Sigma_{02z}} =$

$\sigma^2 \text{diag}\left(s_{0pc}^2 \boldsymbol{I}_{p_c+1}, s_0^2 \boldsymbol{I}_{p_{2z}}\right)$ corresponding to the selected covariates when $z_j^{(2)} = 1$ and $\beta_{1j} = 0$ when features are not selected, $z_j^{(2)} = 0$. For the inverse-gamma prior on σ^2; we set $\nu_0 = 1$, $\sigma_0^2 = 0.25$, and assume a Beta distribution Beta(a, b) and Beta(a_2, b_2) on q_1 and q_2, the probability of selecting important features from the logistic and linear models respectively. The values of a, b, a_2, and b_2 are set such that there is a 10% chance or probability of selecting important features from the logistic and linear models. Lastly, the values of η_r, μ_r, and σ_r^2 in Table 3.1 is used in the data augmentation section.

4.3 Setting Initial Values for the MCMC Algorithm

We set initial values for $\boldsymbol{\beta_{1z}}, \sigma^2, q_1 = a/(a+b),\ q_2 = a_2/(a_2+b_2),$ randomly sampled $\boldsymbol{z^{(1)}}$ and $\boldsymbol{z^{(2)}}$ independently from a Binomial distribution with p trials and probability of selecting important features q_1 and q_2 i.e., Bin$(p, q_\ell), \ell = 1, 2$, with respect to the logistic and linear models. After the initial values for $z_j^{(1)}$ and $z_j^{(2)}$ indicator variable has been randomly sampled, we selected the initial covariates that correspond to these indicator variable $z_j^{(1)} = 1$ and $z_j^{(2)} = 0$, then obtained initial values of the regression effects, $\boldsymbol{X_{1z}}$ and $\boldsymbol{X_{2z}^*}$ for the logistic and linear models respectively. We ran the MCMC algorithm across ten different chains for each simulated data using 300,000 iterations, and discarded the first 30,000 samples to remove the effect of the initial values (burn-in period). The latent indicators, $\mathbf{R}$ is used to sample the latent utilities, $\mathbf{u}$. The following R packages are used for the evaluation of our proposed approach and to obtain the average AUC and standard error (SE) values of the competing models to compare with our method; `coda, AUC, mvtnorm, HDInterval, glmnet, mombf, corrplot`. Due to the complexity and high computation involved, the Advanced Research Computing (ARC) cluster of the High Performance Computing (HPC)

platform of the University of Calgary is used to run our R codes.

4.4 Competing Methods

Using the simulated data with respect to the various scenarios, we compared the proposed approach to some existing models (Table 4.1 and Table 4.2) which includes the Least Absolute Shrinkage and Selection Operator (LASSO) regression of the logistic and linear form, Elastic Net logistic and linear form, and the `mombf`: Moment and inverse moment Bayes factors, an R package introduced by Rossell et al. (2014). The `mombf` approach is only applicable to linear model, not appropriate for logistic form in high dimensional settings. For the LASSO and Elastic Net regression models we used the `glmnet()` function in the `glmnet` R package to fit the logistic and linear models by specifying a family of `binomial` and `gaussian` respectively, while the response of the logistic model was recoded to reflect binary (0 for response with zeros, $(y_i = 0)$ and 1 for positive response $(y_i > 0)$). Instead of choosing an optimal λ, we fit the LASSO and Elastic Net models for each lambda such that each time a coefficient is selected, it is classified as 1 and 0 otherwise, these resulted in vector of zeros and ones that was used to compute the inclusion probabilities. We compared these inclusion probabilities (probability of selection) with the true values of $z_j^{(\ell)},\ \ell = 1, 2,\ j = 1, \cdots, p$ to obtain the area under the receiver operating characteristic curve (AUC) values for each data simulation and averaged it to obtain the average AUC and standard error (SE) shown in Table 4.1 and Table 4.2. For this sudy, the Elastic Net penalty is controlled by $\alpha = 0.5$, and bridges the gap between LASSO regression ($\alpha = 1$) and Ridge regression ($\alpha = 0$). The tuning parameter λ controls the overall strength of the penalty (Friedman et al., 2010). Using the `mombf` approach (not suitable for logistic form), model selection and parameter estimation is based on non-

local prior. After calling the `mombf` library in R, we used the function `modelSelection()` to fit the model by specifying the appropriate prior and initial values as required by the function, then the inclusion probabilities are extracted using `fittedmodel$margpp`. These inclusion probabilities are then compared to the true values of $z_j^{(\ell)},\ \ell = 1, 2,\ j = 1, \cdots, p$ to obtain the AUC values for each data simulation and averaged to obtain the average AUC and standard error (SE) shown in Table 4.1 and Table 4.2. In `mombf`, routines are provided to evaluate prior densities, distribution functions, quantiles and modes, compute Bayes factors, marginal densities and to perform variable selection in regression set ups. The inclusion probabilities in `mombf` are refined using Rao-Blackwellization (Forte et al., 2018).

4.5 Results

Tables 4.1 and 4.2 showed the average AUC and standard error (SE) values obtained using simulation studies to assess the proposed approach (novel: BVS Combined - Semicontinuous, usual: BVS Logistic - Semicontinuous, and BVS Linear - Semicontinuous) and some existing Frequentist and Bayesian variable selection models with semicontinuous responses.

In section (4.5.1), we assessed the agreement of the results of the marginal posterior probabilities (MPP) among the 10 chains by evaluating the correlation coefficients between the marginal posterior probabilities for variable selection (Chekouo et al., 2017; Li and Chekouo, 2021; Stingo et al., 2011). After the assessment of convergence of our MCMC algorithm (see 4.5.1), we used 10 replicated data simulations, obtained the marginal posterior probabilities, $P(z_j^{(\ell)} = 1|\boldsymbol{X}, \mathbf{y})$, $\ell = 1, 2$ using one of the chains and compared to their respective true values of $z_j^{(\ell)}$ to obtain the AUC values which is averaged to obtain the average AUC, while the standard error (SE) is obtained by diving the standard deviation of AUCs by the square

root of the number of AUCs (ten). The values of the average AUC and standard error (SE) for other competing models is obtained in similar fashion based on their respective algorithm and methodology (more details in section (4.4)). The AUC values ranges from 0 to 1, high AUC values of 1 (100%) means that the model has 100% prediction accuracy while AUC values of 0 (0%) means that the model has a poor prediction accuracy (Li and Chekouo, 2021). Lower values of the standard error (SE) signify good performance of the model. It is therefore observed that our proposed approach (BVS Combined - Semicontinuous) has higher prediction accuracy when compared to other approaches. Our proposed method performed well in all scenarios, selecting all relevant important features as well as good estimates, lower standard errors and higher AUC values compared to other variable selection models as summarized in Tables 4.1 and 4.2.

Lastly, in section 4.5.2, we set a benchmark of 0.5 to signify that a feature is selected (≥ 0.5) and less than 0.5 as non-selected features for each method and scenario. The marginal posterior probabilities (MPP) plots for each method and scenario are shown in section 4.5.2 and the remaining in appendix B.3, which is the usual approach in practice since the model space is quite large and the posterior mode may be encountered only a few times in the MCMC runs (Chekouo et al., 2015; Sha et al., 2006). These plots of the marginal posterior probabilities does not include the intercept and clinical covariates because they will always be included in the model. In all scenarios, the marginal posterior probabilities (MPP) obtained using the proposed method indicate that the relevant features returned probability of inclusion values that are very close to 1 while that of the non-selected features are very close to zero. It was observed that the MPP values of the combined method behaved better than that of method 1 (as shown in section 4.5.2 and B.3).

Table 4.1: n = 300 for different features p, 20 important features. The best values in bold.

p	Model	Method	AUC (SE)
50	Combined	BVS Combined - Semicontinuous	**1.0000 (0.0000)**
50	Logistic	BVS Logistic - Semicontinuous	0.9955 (0.0029)
50	Linear	BVS Linear - Semicontinuous	**1.0000 (0.0000)**
50	Logistic	Lasso Logistic	0.9833 (0.0029)
50	Linear	Lasso Linear	0.9160 (0.0150)
50	Logistic	Elastic Net Logistic	0.9813 (0.0033)
50	Linear	Elastic Net Linear	0.9036 (0.0174)
50	Linear	**mombf Linear	0.9957 (0.0029)
200	Combined	BVS Combined - Semicontinuous	**1.0000 (0.0000)**
200	Logistic	BVS Logistic - Semicontinuous	0.9902 (0.0029)
200	Linear	BVS Linear - Semicontinuous	**1.0000 (0.0000)**
200	Logistic	Lasso Logistic	0.9777 (0.0057)
200	Linear	Lasso Linear	0.8691 (0.0153)
200	Logistic	Elastic Net Logistic	0.9765 (0.0053)
200	Linear	Elastic Net Linear	0.8591 (0.0151)
200	Linear	**mombf Linear	0.7313 (0.0505)
500	Combined	BVS Combined - Semicontinuous	**0.9990 (0.0008)**
500	Logistic	BVS Logistic - Semicontinuous	0.9653 (0.0088)
500	Linear	BVS Linear - Semicontinuous	0.9755 (0.0096)
500	Logistic	Lasso Logistic	0.9668 (0.0070)
500	Linear	Lasso Linear	0.8018 (0.0223)
500	Logistic	Elastic Net Logistic	0.9672 (0.0057)
500	Linear	Elastic Net Linear	0.8051 (0.0207)
500	Linear	**mombf Linear	0.6035 (0.0222)
1000	Combined	BVS Combined - Semicontinuous	**0.9750 (0.0168)**
1000	Logistic	BVS Logistic - Semicontinuous	0.9535 (0.0091)
1000	Linear	BVS Linear - Semicontinuous	0.9692 (0.0093)
1000	Logistic	Lasso Logistic	0.9593 (0.0090)
1000	Linear	Lasso Linear	0.7263 (0.0157)
1000	Logistic	Elastic Net Logistic	0.9541 (0.0104)
1000	Linear	Elastic Net Linear	0.7289 (0.0174)
1000	Linear	**mombf Linear	0.6307 (0.0188)

** method is not applicable to logistic

Table 4.2: n = 500 for different features p, 20 important features. The best values in bold.

p	Model	Method	AUC (SE)
50	Combined	BVS Combined - Semicontinuous	**1.0000 (0.0000)**
50	Logistic	BVS Logistic - Semicontinuous	0.9981 (0.0002)
50	Linear	BVS Linear - Semicontinuous	**1.0000 (0.0000)**
50	Logistic	Lasso Logistic	0.9970 (0.0014)
50	Linear	Lasso Linear	0.9663 (0.0084)
50	Logistic	Elastic Net Logistic	0.9963 (0.0017)
50	Linear	Elastic Net Linear	0.9642 (0.0078)
50	Linear	**mombf Linear	**1.0000 (0.0000)**
200	Combined	BVS Combined - Semicontinuous	**1.0000 (0.0000)**
200	Logistic	BVS Logistic - Semicontinuous	0.9999 (0.0001)
200	Linear	BVS Linear - Semicontinuous	**1.0000 (0.0000)**
200	Logistic	Lasso Logistic	0.9965 (0.0008)
200	Linear	Lasso Linear	0.9726 (0.0027)
200	Logistic	Elastic Net Logistic	0.9946 (0.0012)
200	Linear	Elastic Net Linear	0.9667 (0.0033)
200	Linear	**mombf Linear	**1.0000 (0.0000)**
500	Combined	BVS Combined - Semicontinuous	**1.0000 (0.0000)**
500	Logistic	BVS Logistic - Semicontinuous	0.9945 (0.0050)
500	Linear	BVS Linear - Semicontinuous	**1.0000 (0.0000)**
500	Logistic	Lasso Logistic	0.9968 (0.0007)
500	Linear	Lasso Linear	0.9541 (0.0101)
500	Logistic	Elastic Net Logistic	0.9950 (0.0009)
500	Linear	Elastic Net Linear	0.9452 (0.0109)
500	Linear	**mombf Linear	0.5957 (0.0347)
1000	Combined	BVS Combined - Semicontinuous	**1.0000 (0.0000)**
1000	Logistic	BVS Logistic - Semicontinuous	0.9948 (0.0050)
1000	Linear	BVS Linear - Semicontinuous	0.9978 (0.0007)
1000	Logistic	Lasso Logistic	0.9981 (0.0005)
1000	Linear	Lasso Linear	0.8986 (0.0173)
1000	Logistic	Elastic Net Logistic	0.9959 (0.0008)
1000	Linear	Elastic Net Linear	0.8937 (0.0157)
1000	Linear	**mombf Linear	0.6915 (0.0192)

** method is not applicable to logistic

4.5.1 Convergence Diagnostics on Simulated Data.

In this section, we performed convergence diagnostics on the proposed model using metrics like; Gelman and Rubin (1992) diagnostics statistic (boxplots) and plots of correlation matrix that showed the agreement of results across the 10 chains of MCMC algorithm (Chekouo et al., 2017; Li and Chekouo, 2021; Stingo et al., 2011). The proposed method by Gelman and Rubin (1992) used a potential scale reduction factor denoted by $\hat{R}_{gelman} = \sqrt{\frac{\hat{Var(\theta)}}{W}}$, where $\hat{Var}(\theta)$ is the estimated variance of the stationary distribution, θ is parameter of interest, and W is the mean of the variances of each chain. To interpret; values of $\hat{R}_{gelman}$ around 1 (below or above) signifies that convergence is attained while high values of $\hat{R}_{gelman}$, would suggest that one should run the chains longer to improve convergence to the stationary distribution. A limitation to this method as pointed out by Gelman and Rubin (1992) is that, multi-modal target distributions can give iterative simulation algorithms serious problems because the random walks may take a longer time to move from the region of one mode to another, then suggested that in such instances, the iterative simulation algorithm may have to be altered in order to speed up convergence by reparameterizing (Hills and Smith, 1992) or improving the jumping step in the generalised Metropolis algorithm (Green and Han, 1992). However, recent Bayesian variable selection studies (Chekouo et al., 2017; Li and Chekouo, 2021; Stingo et al., 2011) approached assessment of convergence by running the MCMC algorithm with multiple chains to create plots of correlation matrix. A correlation value of > 0.85 usually signify that convergence is achieved and there is agreement across the chains of algorithm (Chekouo et al., 2017). In this study, we also assessed the convergence of our MCMC algorithm by running the generated data through 10 different chains and obtained the correlations between the marginal posterior probabilities across these chains. The correlation plot is then used

to show the convergence $z_j^{(\ell)}$ in our MCMC algorithm. It is observed that values of the correlation as shown in the plots of correlation matrix are > 0.85 (Chekouo et al., 2017). Therefore, it has been established that our MCMC algorithm converged for all and inference can be made using any of these chains (more details in the discussion section 4.6).

Method 2: (Combined Model) with the assumption that the same features are selected from the logistic and linear model

Figure 4.1: Plot of correlation across the 10 chains in the combined model when n = 300; p = 1000

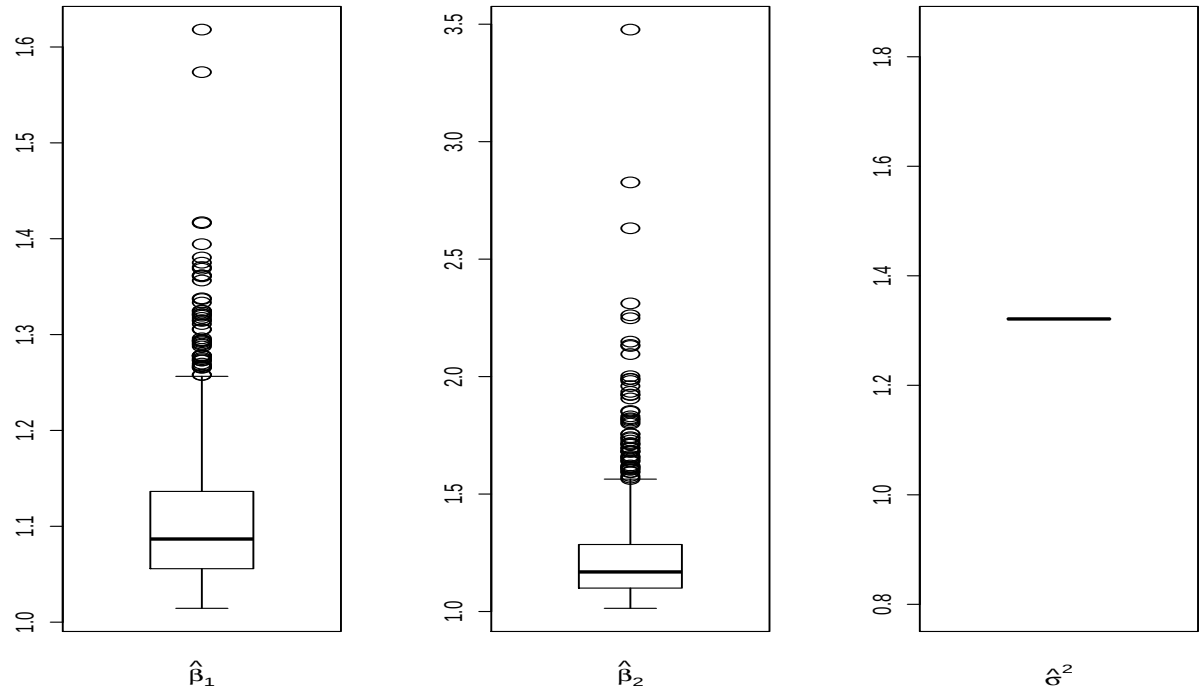


Figure 4.2: Convergence diagnostic plot of the posterior estimate samples across the 10 chains for $\hat{\beta}_1$, $\hat{\beta}_2$, and $\hat{\sigma}^2$ when n = 300; p = 1000.

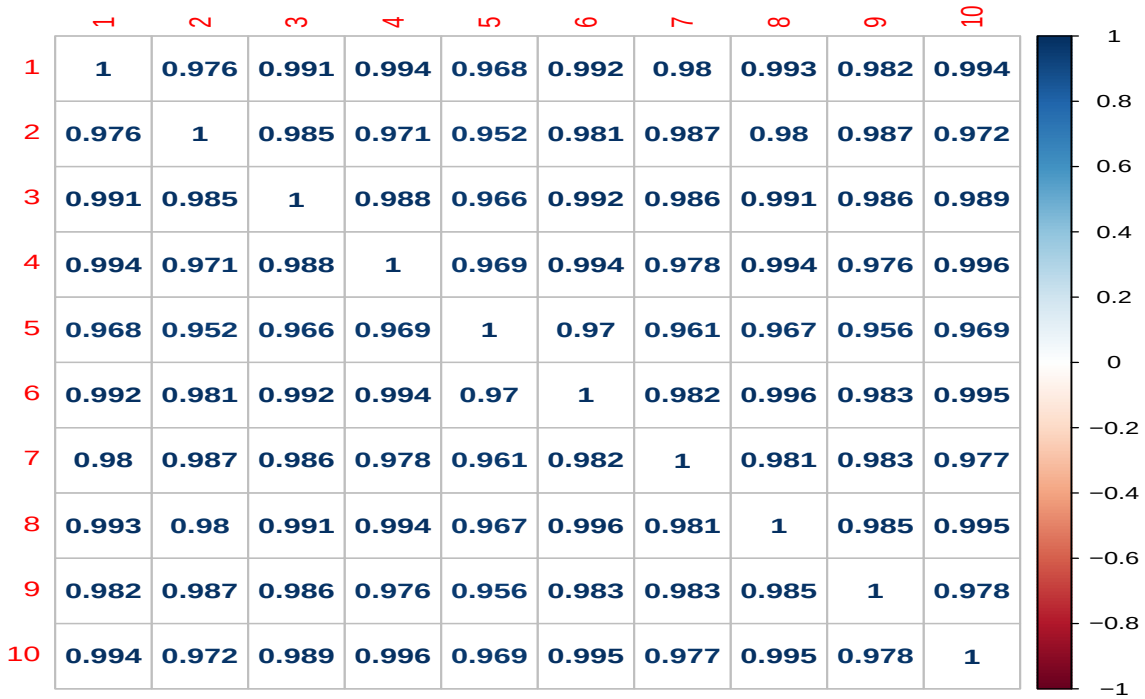


Figure 4.3: Plot of correlation across the 10 chains in the combined model when n = 500; p = 1000

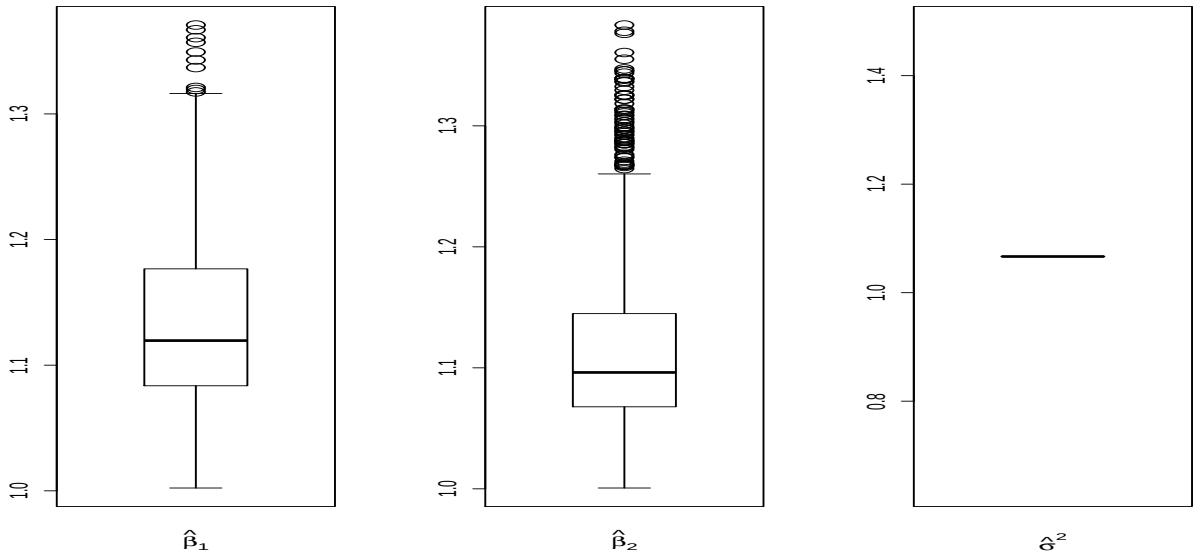


Figure 4.4: Convergence diagnostic plot of the posterior estimate samples across the 10 chains for $\hat{\beta}_1$, $\hat{\beta}_2$, and $\hat{\sigma}^2$ when n = 500; p = 1000.

Method 1: with the assumption that features are selected independently from the logistic and linear model

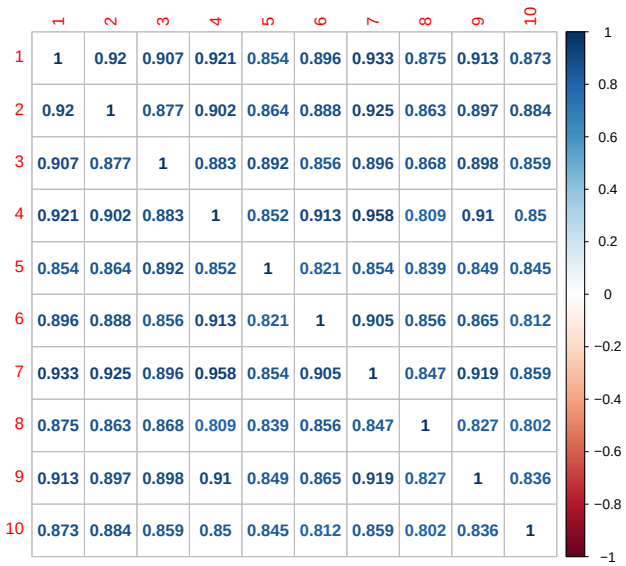


Figure 4.5: Plot of correlation across the 10 chains in the logistic model when n = 300; p = 1000

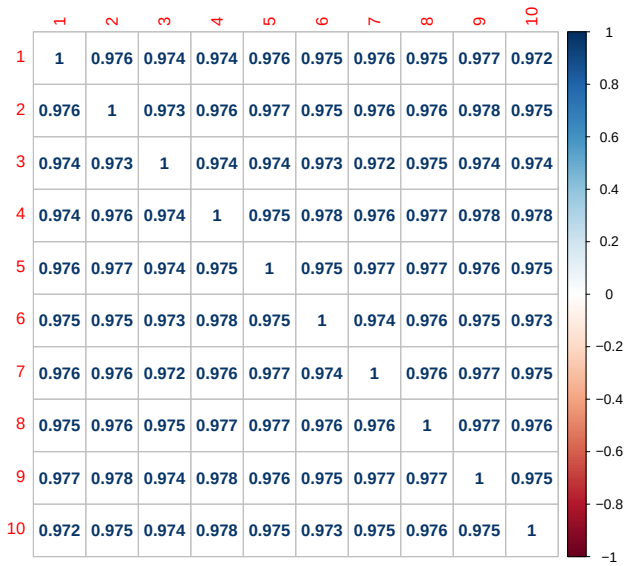


Figure 4.6: Plot of correlation across the 10 chains in the linear model when n = 300; p = 1000

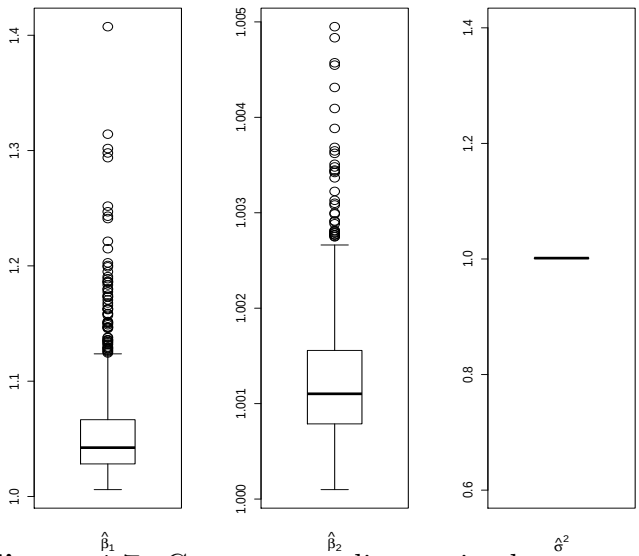


Figure 4.7: Convergence diagnostic plot of the posterior estimate samples across the 10 chains for $\hat{\beta}_1$, $\hat{\beta}_2$, and $\hat{\sigma}^2$

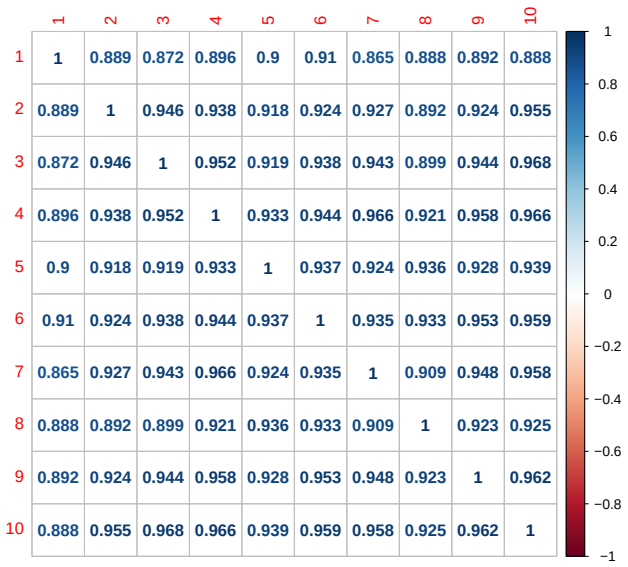


Figure 4.8: Plot of correlation across the 10 chains in the logistic model when n = 500; p = 1000

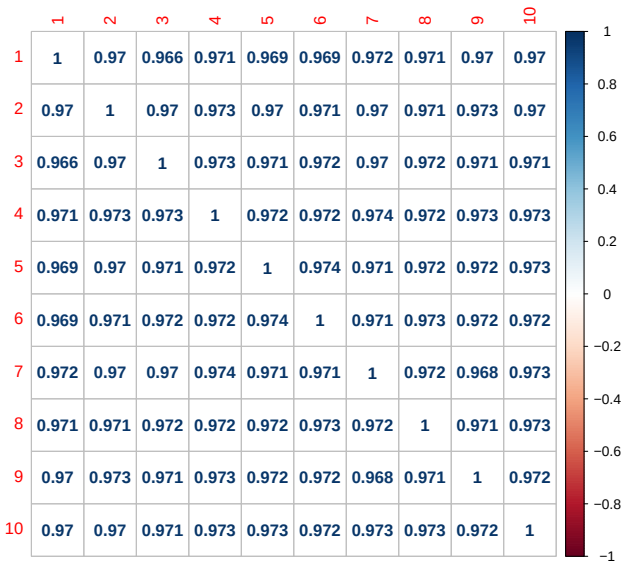


Figure 4.9: Plot of correlation across the 10 chains in the linear model when n = 500; p = 1000

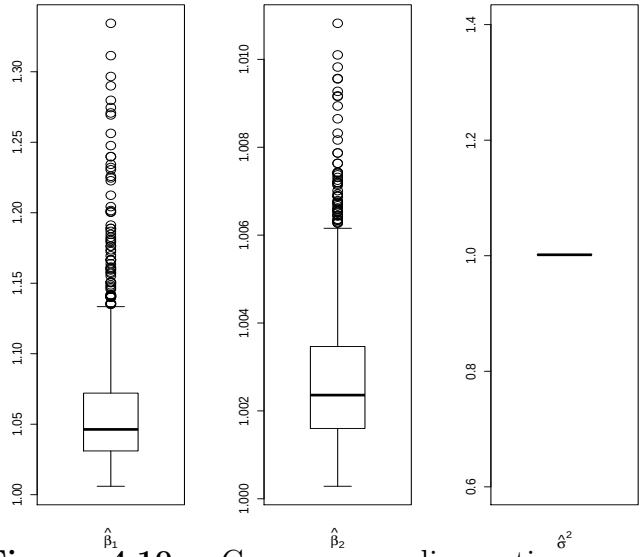


Figure 4.10: Convergence diagnostic plot of the posterior estimate samples across the 10 chains for $\hat{\beta}_1$, $\hat{\beta}_2$, and $\hat{\sigma}^2$

4.5.2 The marginal posterior probability (MPP) or probability of inclusion

Method 2: MPP using the assumption that the same set of features are selected.

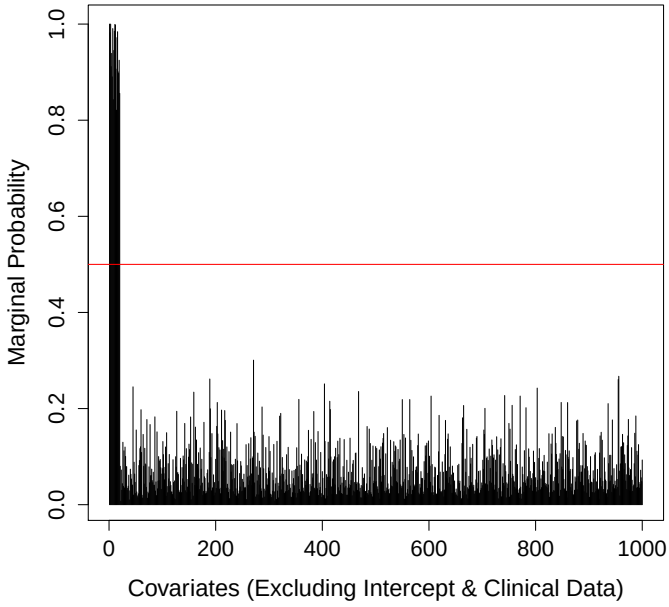


Figure 4.11: mpp when n = 300, p = 1000

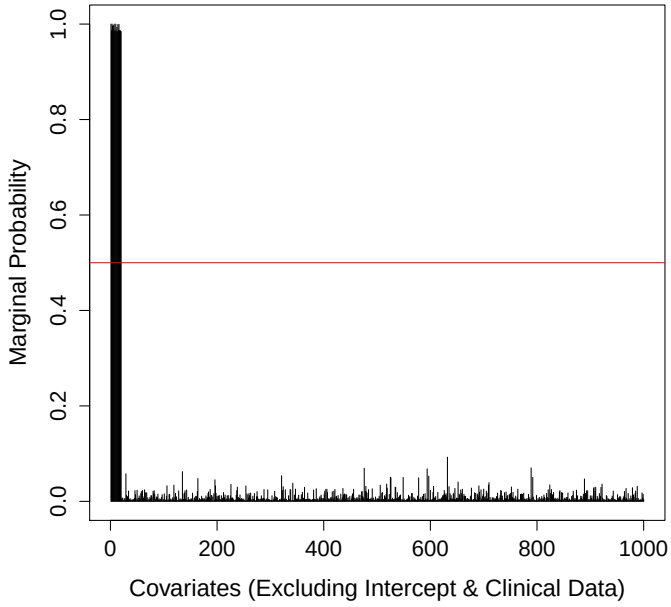


Figure 4.12: mpp when n = 500, p = 1000

Figures 4.11 and 4.12 above showed the marginal posterior probability (MPP) or inclusion probabilities obtained using the combined model (method 2) that assumed same features are selected from the logistic and linear models (i.e., $\boldsymbol{z^{(1)} = z^{(2)}}$). We observe that MPP for when $n = 500$ and $p = 1000$ returned lower probabilities of inclusion in comparison with when $n = 300$ and $p = 1000$ while both have similar MPP values that are approximately 1 for the important features. It can be inferred that the scenario under Figure 4.12 behaves better than that of Figure 4.11.

Method 1: MPP using the assumption that features are selected independently.

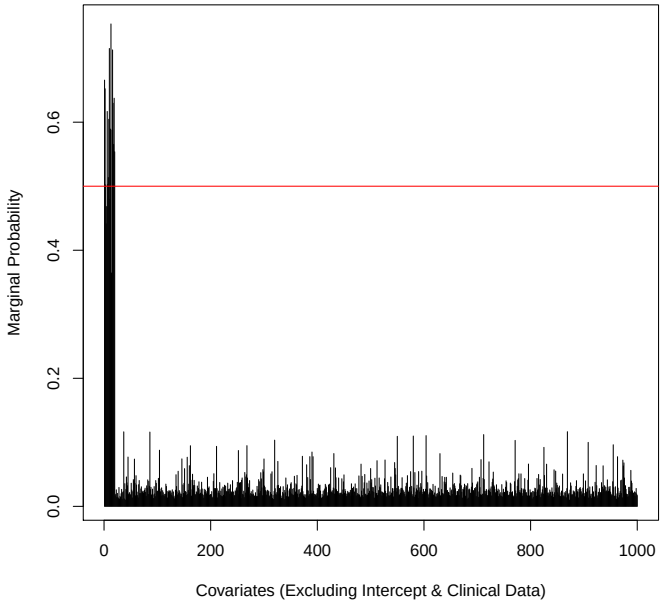


Figure 4.13: mpp when n = 300, p = 1000 for the logistic model

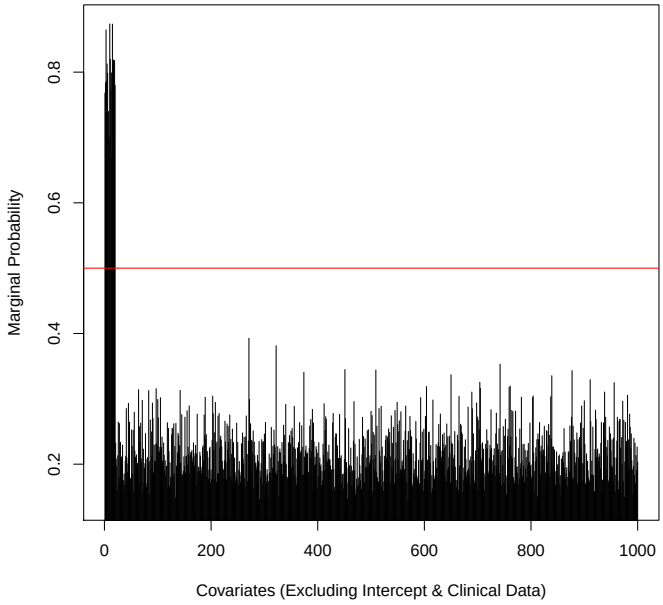


Figure 4.14: mpp when n = 300, p = 1000 for the linear model

Figures 4.13 and 4.14 above gives the marginal posterior probability (MPP) using method 1 with the assumption that features are selected independently from the logistic and linear models respectively, a scenario of when $n = 300$, $p = 1000$. It can be observed that the inclusion probabilities for the important features of the logistic model (Figure 4.14) have higher values compared to that of the linear model (Figure 4.13). However, the non-selected features for the logistic model has lower inclusion probability compared to that of the linear model.

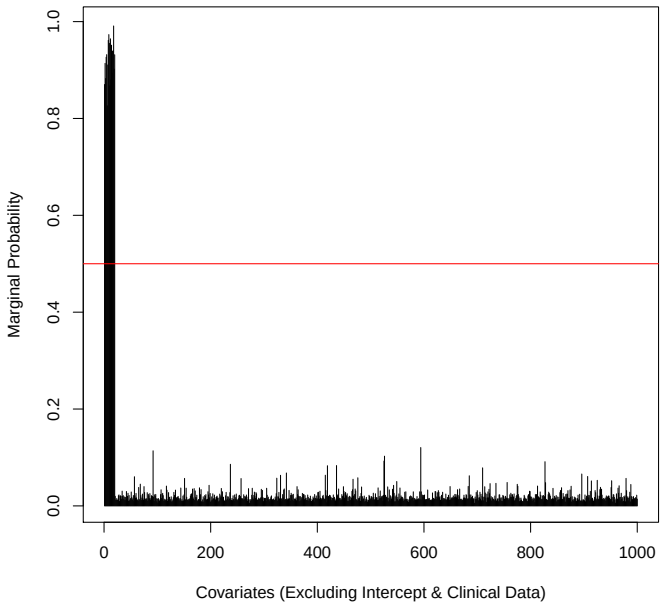


Figure 4.15: mpp when n = 500, p = 1000 for the logistic model

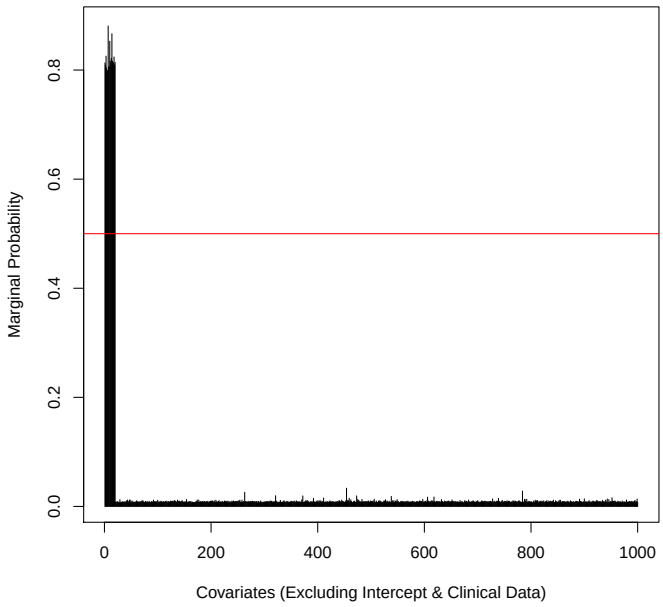


Figure 4.16: mpp when n = 500, p = 1000 for the linear model

Figures 4.15 and 4.16 above gives the marginal posterior probability (MPP) using method 1 with the assumption that features are selected independently from the logistic and linear models respectively, a scenario of when $n = 500, p = 1000$. It can be observed that the inclusion probabilities for the important features of the logistic model (Figure 4.15) have higher values compared to that of the linear model (Figure 4.16). However, the non-selected features for the linear model has lower inclusion probability compared to that of the logistic model.

4.6 Discussion

We ran the MCMC algorithm using multiple simulated dataset and obtained the average AUC and standard error, the latter was also done to obtain results for the competing models. Table 4.1 and Table 4.2 present results of the various scenarios using simulation

studies. We observed that the proposed method behaved better than other models in all scenarios of n = 300, n = 500 with different set of features $p = 1000, 500, 200$ and 50. Also, the proposed method has a higher prediction accuracy as the posterior estimates for the regression parameters are very close to their true values as shown in Table B.3 and Table B.4 in comparison to Tables B.1 and B.2. Using a threshold of 0.5, we observed that the probability of selection or inclusion of features (marginal posterior probability (MPP)) in the last column of Tables B.1, B.2, B.3, and B.4 whose values are larger than 0.5 for selected features while those non-selected features returned values less than 0.5. This is as expected and it can be interpreted by saying the selected features are the likely risk factors of an outcome or disease while the unselected features may not likely be risk factors of an outcome or disease using a threshold of 0.5 (Rosenwald et al., 2002). This is the usual approach in Bayesian variable selection methods that involves high dimensional data. Since the values of the MPP are too large to be presented on a table, the plots of the marginal posterior probabilities (MPP) is used to visualize the selected and non-selected features as shown in section 4.5.2 and appendix B.3.

Lastly, we can see that our MCMC algorithm converges based on the correlation plots in section 4.5.1 and others in appendix B.2 used to investigate the agreement across the 10 different chains for our proposed approach (combined model) and both the logistic and linear models. It was observed that the combined model converged faster compared to the other two models, most especially the logistic model. The boxplots give the estimated potential scale reduction factor, a measure used to assess the convergence of our MCMC algorithm proposed by (Gelman and Rubin, 1992). However, as it is customary in Bayesian variables selection, high correlation value of > 0.85 signify that the MCMC algorithm converged and there is agreement across the chains (Chekouo et al., 2017; Li and Chekouo, 2021; Stingo

et al., 2011). Inference can therefore be made using any of the chains.

Chapter 5

Application to coronary artery disease (CAD) data

Coronary artery disease (also called ischemic heart disease) is the most common types of cardiovascular disease in Canada. It occurs when a build-up of plaque (atherosclerosis) narrows the arteries that bring blood to the heart, limiting the amount of oxygen the heart receives (ischemia). Untreated coronary artery disease can lead to heart attack, stroke, or death (Mason and Frey, 2018).

Gene expression is the process by which genetic information encoded in DNA is used to generate gene products (e.g., proteins). A person's DNA is made up of all the genetic information needed to build an estimated 20,000 different proteins (Musunuru et al., 2017). Advances in science have allowed for the discovery and study of many biological markers (biomarkers) of gene expression (Musunuru et al., 2017). When used alone or, more commonly, in combination with known risk factors, biomarkers can help predict the possibility of developing a disease, help diagnose the presence of disease, or determine how a disease will progress in an individual patient (Musunuru et al., 2017). A phenotype refers to an observable trait. Technically speaking, even a genotype is a type of a phenotype because it is observable

(Musunuru et al., 2017).

5.1 Description of coronary artery disease (CAD) data

In this section, we applied the proposed method to coronary artery disease (CAD) data obtained from the Duke Database for Cardiovascular Disease (DDCD) which provides clinical data and gene expression data used for analysis. This data was primarily collected from individuals that took part in the cardiac catheterization study. The CATHeterization GENetics (CATHGEN) research project is a resource for the investigation of genes associated with coronary artery disease and related disorders. The project collected peripheral blood samples from consenting research subjects undergoing cardiac catheterization at Duke University Medical Center from 2001 through 2011. The data consists of gene expression data and clinical covariates. The clinical covariates consists of data such as subject ID which is a unique code assigned to participants, age at enrolment, body mass index (BMI) at enrolment, history of hypertension of participant, coronary artery disease (CAD) severity index at enrolment (response or outcome), history of diabetes of participant, history of hyperlipidemia of participant, number of diseased vessels at enrolment, race of participant defined by population stratification, sex of participant, history of smoking of participant, participant died, days from cardiac catheterization to death or last known alive, days from cardiac catheterization to subsequent myocardial infection, previous history of myocardial infarction, and case control status of the subject (Table B.6). The gene expression data is made up of 890 expressions with 48,803 probe IDs which is the identifier that refers to a set of probe pairs selected to represent expressed sequences on an array. We performed some exploratory data analysis on the clinical data and removed missing data, also replaced some few missing continuous

covariates with their mean. We also inspected the gene expression data for missing values and there was none missing, proceeded to obtain the standard deviation corresponding to each probe ID and reduced the data by retaining markers with standard deviations ≥ 1.2, then performed a univariate analysis using the two-part model independently to further reduce the data by selecting gene expressions with p-value < 0.2. After which, we combined these set of gene expression data with their respective clinical covariates using the subject ID and arrived at a dataset that is made up of 421 observation known as the sample size n, and 1,213 set of covariates that includes the clinical covariates, this is known as p according to the notation format of this study. Therefore we have $n = 421$ observations and $p = 1,213$ covariates which includes the subject ID, the response variable is coronary artery disease index denoted by CADIndex (semicontinuous response which consist of zeros and positive values), 11 clinical covariates (13 clinical covariates but Subject ID, and CADIndex excluded in Table 5.2), and mRNA gene expressions (1200).

Lastly, we applied the proposed approach to these set of data and the results are presented below.

5.2 Results

In this section, we present the results obtained using the proposed method to predict the likely risk factors and select significant genes that are associated with the coronary artery disease (CAD). After setting hyperparameter and initial values as explained in the simulation study section, we ran the MCMC algorithm with 300,000 iterations and discarded 30,000 (burn-in period) to remove the effect of initial values. The results obtained from the application of the proposed approach (Method 2: combined model) is presented below (5.2.1, 5.2.2 and 5.2.3)

while the results obtained using existing approach (Method 1) is presented in appendix B.4. The discussion section of this chapter gives more insight on the findings from the analysis and Gene Set Enrichment Analysis (GSEA) to identify over-represented gene expressions that may have been produced by the proposed method. The goal of GSEA is to determine whether members of a gene set tend to occur toward the top (or bottom) of the list, in which case the gene set is correlated with the phenotypic class distinction (Subramanian et al., 2005). The Patient characteristics and codebook of the clinical covariates is shown in Tables B.7 and B.6 respectively.

5.2.1 Convergence diagnostics on coronary artery disease (CAD) data.

Method 2: Combined method

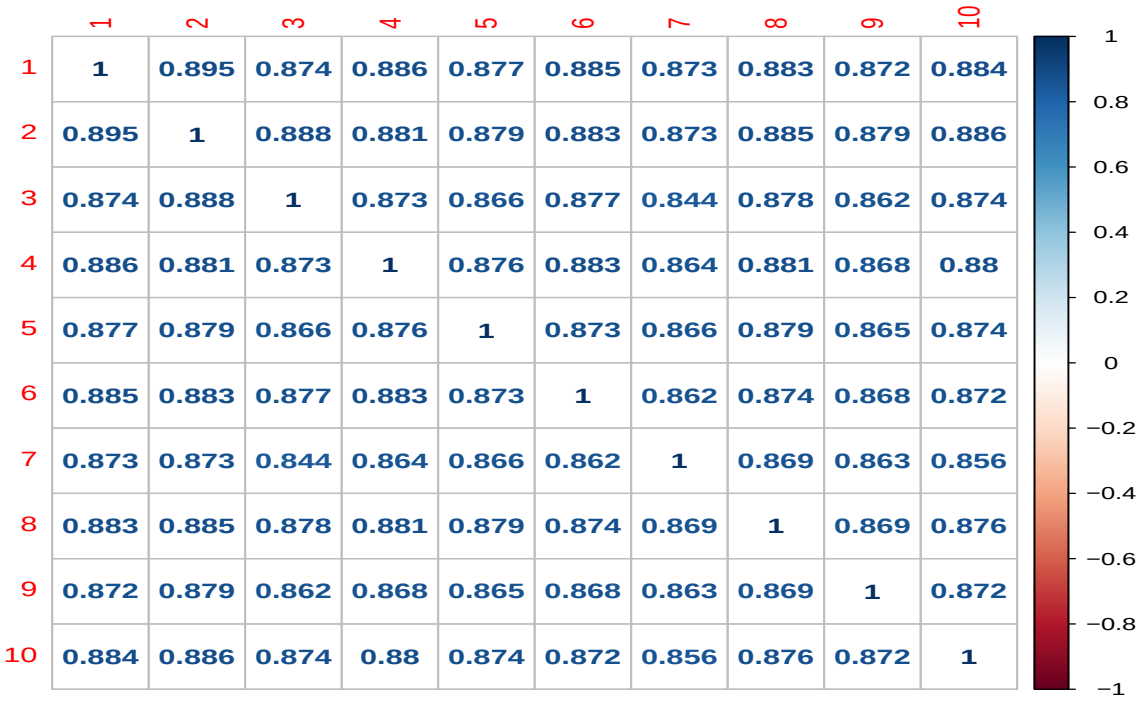


Figure 5.1: Plot of correlation across the 10 chains in the linear model using CAD data.

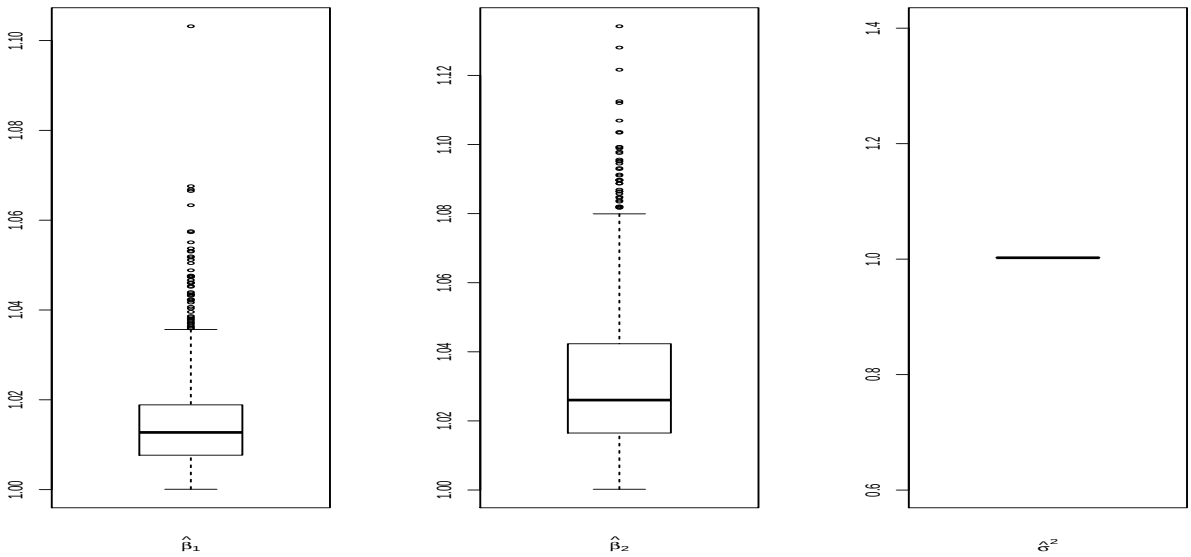


Figure 5.2: Convergence diagnostic plot of the posterior estimate samples across the 10 chains for $\hat{\beta}_1$, $\hat{\beta}_2$, and, $\hat{\sigma}^2$ using CAD data.

5.2.2 Marginal posterior probabilities (MPP) using coronary artery disease (CAD) data

Method 2: Combined method

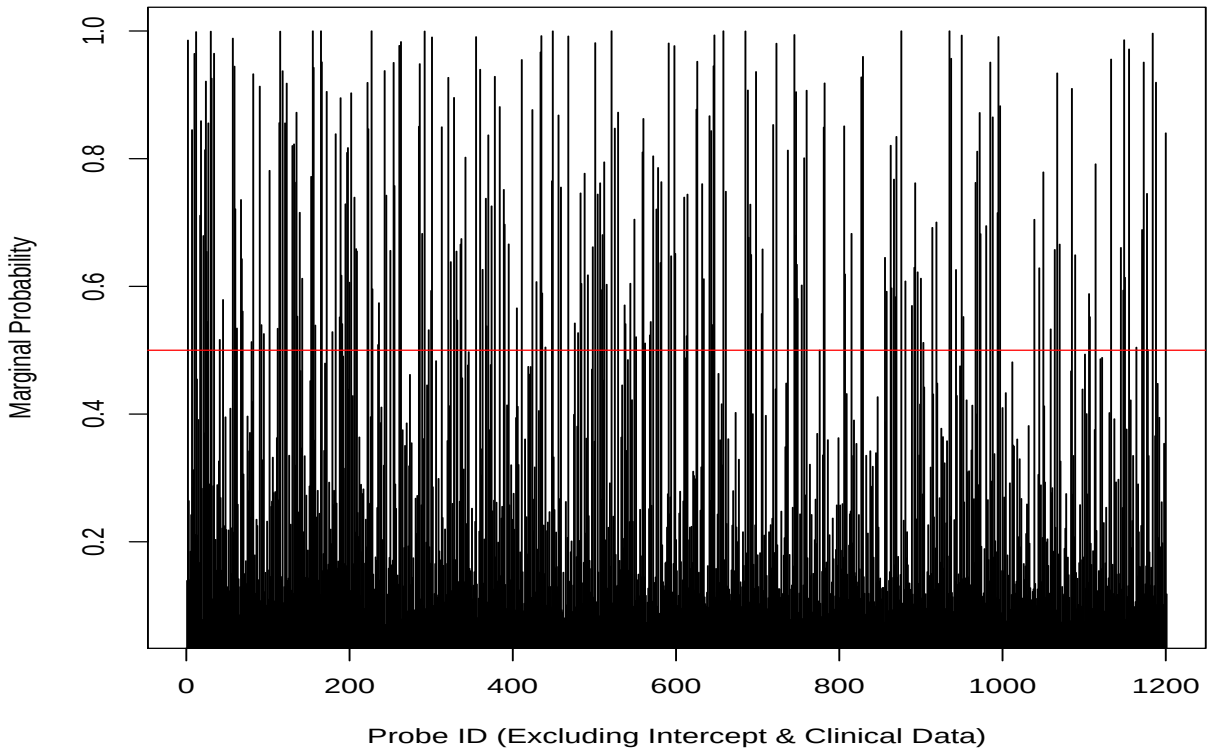


Figure 5.3: The marginal posterior probability for the logistic model using CAD data.

5.2.3 Genes associated with coronary artery disease (CAD)

Table 5.1: coronary artery disease (CAD) data: top 20 genes associated with coronary artery disease using the combined model approach (method 2: selecting the same features from the logistic and linear model, $z_j^{(1)} = z_j^{(2)}$, $j = p_c + 2, \cdots, p$, $\ell = 1, 2$).

Probe ID	**Gene Symbol**	$P(z_j^{(\ell)} = 1 \mid \boldsymbol{X}, \mathbf{y})$
ILMN_1680042	LOC649452	1.00
ILMN_1749792	SORBS1	1.00
ILMN_1792540	AR	1.00
ILMN_1893915		1.00
ILMN_1669577	PRNPIP	1.00
ILMN_1838166		1.00
ILMN_2066088	C1orf64	1.00
ILMN_2328986	SREBF1	1.00
ILMN_1664464	PTGDS	1.00
ILMN_1692056	HS3ST3A1	1.00
ILMN_1819503		1.00
ILMN_1699382	LOC647500	0.99
ILMN_1657888	SALL1	0.99
ILMN_1653980	METTL8	0.99
ILMN_1813280	LOC642222	0.99
ILMN_1800131	LOC652826	0.99
ILMN_1781027	TSNARE1	0.99
ILMN_1752843	GRM4	0.99
ILMN_1693981	SH3TC2	0.99
ILMN_1723674	NKAIN3	0.99

Based on the use of the proposed approach, we observe that about 204 important genes out of 1,200 gene expressions are associated with coronary artery disease (CAD) and top 20 of

these genes are shown in Table 5.1 above. In comparison with method 1, we inspect results from both methods to see if there are genes that are common when we use both approaches. It was observed that about 202 genes appears to be predicted as associated with coronary artery disease using both methods but the proposed approach produced a higher marginal posterior probability (MPP) values compared to method 1 and the logistic regression did not produce good values because convergence was not achieved. This is seen by carefully studying Tables 5.1 and B.5.

Table 5.2: coronary artery disease (CAD) data: the regression effects of clinical covariates associated with coronary artery disease using the combined model approach (method 2: selecting the same features from the logistic and linear model, $z_j^{(1)} = z_j^{(2)}$, $j = p_c + 2, \cdots, p$, $\ell = 1, 2$). The Posterior median (Estimate), 95% Credible Interval (C.I) and 95% Highest Posterior Density Interval (HPDI) is shown below. If the 95% C.I does not contain the null hypothesis value 0, the results are statistically significant. It can therefore be inferred that only AGE is significant in the linear model.

Combined Method			
Logistic Model	**Estimate**	**95%C.I**	**95%HPDI**
Intercept	0.0860	(-1.2834) - 1.4699	(-1.2650) - 1.4870
AGE	-0.1905	(-0.2443) - 0.1390	(-0.2439) - (-0.1387)
BMI	-0.1089	(-1.2605) - 0.9777	(-1.2617) - 0.9758
HYPERTENSION	-0.2959	(-1.6031) - 1.0046	(-1.6155) - 0.9897
DIABETES	-0.2782	(-1.5963) - 1.0520	(-1.5895) - 1.0578
HYPERCHOLESTEROLEMIA	-0.5792	(-1.8809) - 0.7598	(-1.8869) - 0.7525
RACE	-0.4436	(-1.5574) - 0.6454	(-1.5623) - 0.6393
SEX	0.6252	(-0.7238) - 1.9651	(-0.7402) - 1.9482
SMOKING	-0.4384	(-1.7794) - 0.8811	(-1.7870) - 0.8728
DEATH	-0.1414	(-1.4687) - 1.1983	(-1.4733) - 1.1916
DDEATH	-0.0833	(-1.1795) - 1.0408	(-1.1873) - 1.0324
HXMI	-0.6921	(-2.0088) - 0.6216	(-2.0142) - 0.6153
Linear Model	**Estimate**	**95%C.I**	**95%HPDI**
Intercept	0.4943	(-0.9615) - 1.9437	(-0.9327) - 1.9720
AGE	0.5720	0.5385 - 0.6076	0.5385 - 0.6076
BMI	0.5382	(-0.1635) - 1.2738	(-0.1593) - 1.2772
HYPERTENSION	0.5751	(-0.5294) - 1.6683	(-0.5338) - 1.6636
DIABETES	0.2693	(-0.8642) - 1.4159	(-0.8780) - 1.4011
HYPERCHOLESTEROLEMIA	0.2309	(-0.8684) - 1.3412	(-0.8573) - 1.3513
RACE	0.3120	(-0.3877) - 1.0176	(-0.3886) - 1.0166
SEX	-0.0103	(-1.3276) - 1.3167	(-1.3412) - 1.3016
SMOKING	0.5491	(-0.5640) - 1.6540	(-0.5700) - 1.6469
DEATH	0.3467	(-0.9124) - 1.6150	(-0.9144) - 1.6126
DDEATH	-0.6022	(-1.3822) - 0.1958	(-1.3792) - 0.1982
HXMI	0.9478	(-0.1679) - 2.0244	(-0.1685) - 2.0236

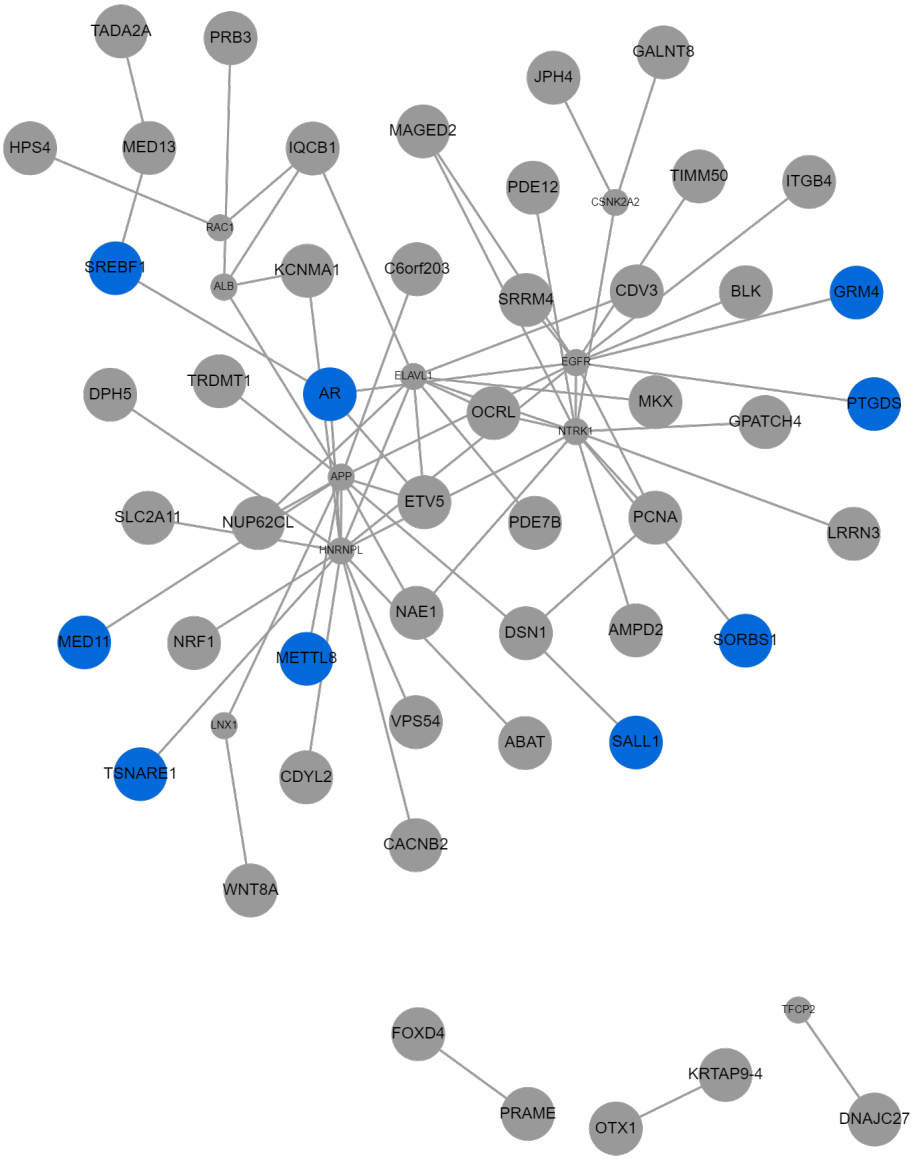


Figure 5.4: Gene regulatory network obtained from WebGestalt (Wang et al., 2013, 2017). In colour blue are part of the top 20 significant genes shown in Table 5.1 that also appear after performing Gene Set Enrichment Analysis on the set of important genes obtained from using our proposed approach. It showed that these genes interact closely with other genes based on their interactive network with other neighbouring genes.

After the application of the proposed approach to coronary artery disease (CAD) data, the set of gene symbols of all selected gene expressions are used to perform a Gene Set Enrichment Analysis to know the characteristics of this association and relationship among other genes as shown in Figure 5.4. The Gene Set Enrichment Analysis (GSEA) above is done using the

web application by Wang et al. (2017). We import the .txt file containing the gene symbol of all the selected gene expressions, then mapped to Entrez Gene ID which is National Center for Biotechnology Information (NCBI)'s database for gene-specific information. Entrez Gene includes records from genomes that have been completely sequenced, that have an active research community to contribute gene-specific information or that are scheduled for intense sequence analysis (Maglott et al., 2005). Lastly, we selected a WikiPathways (Slenter et al., 2018) to obtain the gene regulatory network.

5.3 Discussion

Figure 5.3 showed the marginal posterior probability (MPP) of selected biomarkers using a threshold of 0.5 which means that such markers of gene expression has a significant contribution to predicting the outcome which is the coronary artery disease index. Also, Figure 5.1 showed that there is agreement across the chains with correlation values > 0.85 signifying that convergence has been attained using our proposed approach on the CAD data (in line with Chekouo et al. (2017)). However, using the existing method 1 approach; the linear model converged easily using the same number of iterations as our proposed approach and same burn-in period but the logistic model may require more number of iterations to achieve convergence as shown in appendix B.4. Also, the boxplot in Figure 5.2 gives the Gelman's convergence plot (Gelman and Rubin, 1992) for posterior estimates of $\hat{\beta}_1$, $\hat{\beta}_2$, and $\hat{\sigma}^2$ using the proposed method, this showed that the algorithm converged with evidence of good mixture across the MCMC chains.

The selected genes obtained using the proposed method is shown in Table 5.1 from single nucleotide polymorphism (SNP) information. The variant G allele of the SORBS1 polymorphism

was a risk factor for hypertension. Also, The prevalence of hypertension was greater in the combined group of all subjects with the AG or GG genotype of SORBS1 than in those with the AA genotype from 40 to 80 years of age (Yamada et al., 2009). The shorter trinucleotide repeat disorders (CAG repeat) of the androgen receptor (AR) gene is associated with more severe CAD, which suggests a role for the sensitivity to androgens in the increased frequency of CAD in males (Alevizaki et al., 2003). C1orf64 closely correlated with androgen receptor (AR) expression in primary and metastatic breast tumors and C1orf64 expression was relatively higher in breast tumors with a lower grade and lobular histology (Naderi, 2017). A recent study suggested that SRARP (C1orf64) is a tumor suppressor that can be used to predict the clinical outcomes of malignant tumors (Zhang et al., 2021). sterol-regulatory element binding transcription factors (SREBFs), regulators of cholesterol metabolism, SREBF-2 and SCAP play an important role in the progression of atherosclerosis. SREBF-2 G1784C polymorphism (SREBF-2-595A/G isoforms) has been proven to be associated with early onset myocardial infarction (MI) in middle-aged male Americans (Friedlander et al., 2008). SREBF-2-595A/G isoforms and the SCAP A2386G polymorphism (SCAP-796I/V isoforms) have also been proven to be associated with prehospital sudden cardiac death in middle-aged male Finns (Fan et al., 2008). HS3ST3A1 was identified as part of the several molecular changes induced by smoking in human airway epithelium, which provided some candidate genes and microRNAs for assessing the risk of lung diseases caused by smoking (Huang et al., 2019). Rezaee et al. (2020) showed that the PTGDS gene expression levels increased significantly in the patients with vessel restenosis. SALL1 hyper-methylation has already been confirmed as the diagnostic biomarker for breast cancer and other epithelial cancers, especially for the colorectal cancer (Hill et al., 2010).

Chapter 6

Summary, and Conclusion

6.1 Summary

This study has proposed a novel Bayesian variable selection method with semicontinuous response where the nature of data is of a high dimensionality. The algorithm converged easily and has a more reliable variable selection potentials, it performs well even when sample sizes are small or larger than the number of features. We used the two-part model developed by Duan et al. (1983), then incorporated spike-and-slab prior for the regression coefficients and a normal distribution as prior on the intercept and clinical covariates parameters. Applied the Stochastic Search Variable Selection (SSVS) to randomly sample the indicator variables for variable selection. The agreement of chain behaviour is checked and marginal posterior probabilities (MPP) that passed the threshold of 0.5 is used to select important features. We also compared some existing variable selection models to the proposed approach, and it was observed that the proposed approach converged faster and behaved better than other models based on the AUC values in Tables 4.1 and 4.2 for variable selection.

Furthermore, the important risk factors selected by the proposed model has been known in previous studies to be genes associated with coronary artery disease (CAD) as explained in

the discussion section 5.3.

6.2 Conclusion

- **Consistency of our proposed method:** The findings from this study is consistent with the results from other existing Bayesian variable selection methods, the important risk factors selected by the proposed model has been known in previous studies to be genes associated with coronary artery disease (CAD) (see discussion section 5.3). It is obvious that our model behaved better than the existing models used in comparison (Tables 4.1 and 4.2).

- **Contribution to knowledge:** We have included the possibility of selecting the same and/or different sets of important features using the two-part model introduced by Duan et al. (1983) with semicontinuous response to carry out Bayesian variable selection in the context of high-dimensional data analysis. This is what most existing Bayesian variable selection methods did not mention or consider not even the classical approach. Also, the proposed method can be applied to many areas of research in health and other areas of specialisations where there are lots of zeros (also known as zero-inflated), a semicontinuous type of response data which is usually the case with most health related dataset. We have also been able to propose a method that can consistently and accurately handle high dimensional data analysis with at least one thousand ($> 1,000$) covariates using two-part model.

- **Strength of this study:** The strength of our proposed method lies in the ability to provide more accurate selection of predictors that are risk factors of outcome or disease responses (i.e., select likely genes associated with coronary artery disease) that are

semicontinuous in high dimensional data analysis. Also, the proposed method (method 2: combined model) converged faster than other Bayesian variable selection model compared in this study and can accommodate atleast a thousand number of features.

- **Limitation of the study:** One limitation that can be pointed out is that it requires a High performance computing resources if the features are higher than or equal to a thousand.

- **Future research:** There is possibility of applying this proposed method for Bayesian variable selection with other types of models for modelling nonnegative data with clumping at zero like those listed in Min and Agresti (2002), e.g using a single distribution from the exponential dispersion family to analyze semicontinuous data (Jørgensen, 1987; Jorgensen, 1997).

Bibliography

Advanced Research Computing. https://it.ucalgary.ca/research-computing-services/our-resources/high-performance-computing-hpc/advanced-research.

Alevizaki, M., Cimponeriu, A., Garofallaki, M., Sarika, H.-L., Alevizaki, C., Papamichael, C., Philippou, G., Anastasiou, E., Lekakis, J., and Mavrikakis, M. (2003). The androgen receptor gene cag polymorphism is associated with the severity of coronary artery disease in men. *Clinical endocrinology*, 59(6):749–755.

Amemiya, T. (1973). Regression analysis when the dependent variable is truncated normal. *Econometrica: Journal of the Econometric Society*, pages 997–1016.

Bai, R. and Ghosh, M. (2018). High-dimensional multivariate posterior consistency under global–local shrinkage priors. *Journal of Multivariate Analysis*, 167:157–170.

Bayes' theorem. https://en.wikipedia.org/wiki/Bayes%27_theorem.

Bellhouse, D. R. (2004). The reverend thomas bayes, frs: a biography to celebrate the tercentenary of his birth. *Statistical Science*, 19(1):3–43.

Belotti, F., Deb, P., Manning, W. G., and Norton, E. C. (2015). twopm: Two-part models. *The Stata Journal*, 15(1):3–20.

Bendel, R. B. and Afifi, A. (1976). A criterion for stepwise regression. *The American Statistician*, 30(2):85–87.

Campbell, H. (2019). Is it even rainier in north vancouver? a non-parametric rank-based test for semicontinuous longitudinal data. *Journal of Applied Statistics*, 46(7):1155–1176.

cardiac catheterization study. https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000704.v1.p1.

cardiovascular disease. https://www.canada.ca/en/public-health/services/chronic-diseases/cardiovascular-disease/six-types-cardiovascular-disease.html.

Carlin, B. P. and Chib, S. (1995). Bayesian model choice via markov chain monte carlo methods. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(3):473–484.

Chang, B.-H. and Pocock, S. (2000). Analyzing data with clumping at zero: an example demonstration. *Journal of clinical epidemiology*, 53(10):1036–1043.

Chekouo, T., Stingo, F. C., Doecke, J. D., and Do, K.-A. (2015). mirna–target gene regulatory networks: A bayesian integrative approach to biomarker selection with application to kidney cancer. *Biometrics*, 71(2):428–438.

Chekouo, T., Stingo, F. C., Doecke, J. D., and Do, K.-A. (2017). A bayesian integrative approach for multi-platform genomic data: A kidney cancer case study. *Biometrics*, 73(2):615–624.

Chipman, H. (1996). Bayesian variable selection with related predictors. *Canadian Journal of Statistics*, 24(1):17–36.

Coronary artery disease. https://www.heartandstroke.ca/heart-disease/conditions/coronary-artery-disease.

Daemen, A., Gevaert, O., Ojeda, F., Debucquoy, A., Suykens, J. A., Sempoux, C., Machiels, J.-P., Haustermans, K., and De Moor, B. (2009). A kernel-based integration of genome-wide data for clinical decision support. *Genome medicine*, 1(4):1–17.

Derksen, S. and Keselman, H. J. (1992). Backward, forward and stepwise automated subset selection algorithms: Frequency of obtaining authentic and noise variables. *British Journal of Mathematical and Statistical Psychology*, 45(2):265–282.

Devroye, L. (1986). Sample-based non-uniform random variate generation. In *Proceedings of the 18th conference on Winter simulation*, pages 260–265.

Dirac, P. A. M. (1981). *The principles of quantum mechanics*. Number 27. Oxford university press.

Duan, N., Manning, W. G., Morris, C. N., and Newhouse, J. P. (1983). A comparison of alternative models for the demand for medical care. *Journal of business & economic statistics*, 1(2):115–126.

Duke University Medical Center. https://dmpi.duke.edu/cathgen.

Fan, J. and Li, R. (2002). Variable selection for cox's proportional hazards model and frailty model. *The Annals of Statistics*, 30(1):74–99.

Fan, Y.-M., Karhunen, P. J., Levula, M., Ilveskoski, E., Mikkelsson, J., Kajander, O. A., Järvinen, O., Oksala, N., Thusberg, J., Vihinen, M., et al. (2008). Expression of sterol regulatory element-binding transcription factor (srebf) 2 and srebf cleavage-activating

protein (scap) in human atheroma and the association of their allelic variants with sudden cardiac death. *Thrombosis Journal*, 6(1):1–8.

Fang, S., Zhe, S., Lee, K.-c., Zhang, K., and Neville, J. (2020). Online bayesian sparse learning with spike and slab priors. In *2020 IEEE International Conference on Data Mining (ICDM)*, pages 142–151. IEEE.

Forte, A., Garcia-Donato, G., and Steel, M. (2018). Methods and tools for bayesian variable selection and model averaging in normal linear regression. *International Statistical Review*, 86(2):237–258.

Friedlander, Y., Schwartz, S. M., Durst, R., Meiner, V., Robertson, A. S., Erez, G., Leitersdorf, E., and Siscovick, D. S. (2008). Srebp-2 and scap isoforms and risk of early onset myocardial infarction. *Atherosclerosis*, 196(2):896–904.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1.

Fruehwirth-Schnatter, S. and Frühwirth, R. (2007). Auxiliary mixture sampling with applications to logistic models. *Computational Statistics & Data Analysis*, 51(7):3509–3528.

Frühwirth-Schnatter, S. and Wagner, H. (2006). Auxiliary mixture sampling for parameter-driven models of time series of counts with applications to state space modelling. *Biometrika*, 93(4):827–841.

Garcia-Donato, G. and Forte, A. (2018). Bayesian testing, variable selection and model averaging in linear models using r with bayesvarsel. *R J.*, 10(1):155.

Gelfand, A. E., Hills, S. E., Racine-Poon, A., and Smith, A. F. (1990). Illustration of bayesian inference in normal data models using gibbs sampling. *Journal of the American Statistical Association*, 85(412):972–985.

Gelfand, A. E. and Smith, A. F. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American statistical association*, 85(410):398–409.

Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical science*, 7(4):457–472.

Geman, S. and Geman, D. (1984). Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on pattern analysis and machine intelligence*, (6):721–741.

Generating Exponential Variates. https://en.wikipedia.org/wiki/Exponential_distribution#cite_note-14.

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

George, E. I. and McCulloch, R. E. (1997). Approaches for bayesian variable selection. *Statistica sinica*, pages 339–373.

Geweke, J. F. (1996). Bayesian inference for linear models subject to linear inequality constraints. In *Modelling and Prediction Honoring Seymour Geisser*, pages 248–263. Springer.

Green, P. J. and Han, X.-l. (1992). Metropolis methods, gaussian proposals and antithetic

variables. In *Stochastic Models, Statistical methods, and Algorithms in Image Analysis*, pages 142–164. Springer.

Gronau, R. (1974). Wage comparisons–a selectivity bias. *Journal of political Economy*, 82(6):1119–1143.

Gu, X., Tadesse, M. G., Foulkes, A. S., Ma, Y., and Balasubramanian, R. (2020). Bayesian variable selection for high dimensional predictors and self-reported outcomes. *BMC medical informatics and decision making*, 20(1):1–11.

Hastings, W. K. (1970). Monte carlo sampling methods using markov chains and their applications.

Heckman, J. (1974). Shadow prices, market wages, and labor supply. *Econometrica: journal of the econometric society*, pages 679–694.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica: Journal of the econometric society*, pages 153–161.

High Performance Computing. https://it.ucalgary.ca/research-computing-services/our-resources/high-performance-computing-hpc.

Hill, V. K., Hesson, L. B., Dansranjavin, T., Dallol, A., Bieche, I., Vacher, S., Tommasi, S., Dobbins, T., Gentle, D., Euhus, D., et al. (2010). Identification of 5 novel genes methylated in breast and other epithelial cancers. *Molecular cancer*, 9(1):1–13.

Hills, S. E. and Smith, A. F. (1992). Parameterization issues in bayesian inference. *Bayesian statistics*, 4:227–246.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67.

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999). Bayesian model averaging: a tutorial (with comments by m. clyde, david draper and ei george, and a rejoinder by the authors. *Statistical science*, 14(4):382–417.

Huang, J., Jiang, W., Tong, X., Zhang, L., Zhang, Y., and Fan, H. (2019). Identification of gene and microrna changes in response to smoking in human airway epithelium by bioinformatics analyses. *Medicine*, 98(38).

Jaffa, M. A., Gebregziabher, M., Garrett, S. M., Luttrell, D. K., Lipson, K. E., Luttrell, L. M., and Jaffa, A. A. (2018). Analysis of longitudinal semicontinuous data using marginalized two-part model. *Journal of translational medicine*, 16(1):1–15.

Johnstone, I. M. and Titterington, D. M. (2009). Statistical challenges of high-dimensional data.

Jørgensen, B. (1987). Exponential dispersion models. *Journal of the Royal Statistical Society: Series B (Methodological)*, 49(2):127–145.

Jorgensen, B. (1997). *The theory of dispersion models*. CRC Press.

Karlsson, M. and Laitila, T. (2014). Finite mixture modeling of censored regression models. *Statistical papers*, 55(3):627–642.

Knuth, D. E. (1998). Linear probing and graphs. *Algorithmica*, 22(4):561–568.

Koop, G. and Korobilis, D. (2020). Bayesian dynamic variable selection in high dimensions. *Available at SSRN 3246472*.

Koop, G., Poirier, D. J., and Tobias, J. L. (2007). *Bayesian econometric methods*. Cambridge University Press.

Korner-Nievergelt, F., Roth, T., Von Felten, S., Guélat, J., Almasi, B., and Korner-Nievergelt, P. (2015). *Bayesian data analysis in ecology using linear models with R, BUGS, and Stan*. Academic Press.

Kundu, D., Mitra, R., and Gaskins, J. T. (2021). Bayesian variable selection for multioutcome models through shared shrinkage. *Scandinavian Journal of Statistics*, 48(1):295–320.

Lee, K. H., Chakraborty, S., and Sun, J. (2011). Bayesian variable selection in semiparametric proportional hazards model for high dimensional survival data. *The International Journal of Biostatistics*, 7(1).

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

Li, F. and Zhang, N. R. (2010). Bayesian variable selection in structured high-dimensional covariate spaces with applications in genomics. *Journal of the American statistical association*, 105(491):1202–1214.

Li, H. and Gui, J. (2004). Partial cox regression analysis for high-dimensional microarray gene expression data. *Bioinformatics*, 20(suppl_1):i208–i215.

Li, W. and Chekouo, T. (2021). Bayesian group selection with non-local priors. *Computational Statistics*, pages 1–16.

Li, X. and Xu, R. (2008). *High-dimensional data analysis in cancer research*. Springer Science & Business Media.

Maglott, D., Ostell, J., Pruitt, K. D., and Tatusova, T. (2005). Entrez gene: gene-centered information at ncbi. *Nucleic acids research*, 33(suppl_1):D54–D58.

Mason, J. and Frey, N. (2018). A gene expression test to assess the likelihood of obstructive coronary artery disease. *CADTH Issues in Emerging Health Technologies*.

McFadden, D. et al. (1973). Conditional logit analysis of qualitative choice behavior.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6):1087–1092.

Min, Y. and Agresti, A. (2002). Modeling nonnegative data with clumping at zero: a survey.

Mitchell, T. J. and Beauchamp, J. J. (1988). Bayesian variable selection in linear regression. *Journal of the american statistical association*, 83(404):1023–1032.

Musunuru, K., Ingelsson, E., Fornage, M., Liu, P., Murphy, A. M., Newby, L. K., Newton-Cheh, C., Perez, M. V., Voora, D., and Woo, D. (2017). The expressed genome in cardiovascular diseases and stroke: refinement, diagnosis, and prediction: a scientific statement from the american heart association. *Circulation: Cardiovascular Genetics*, 10(4):e000037.

Naderi, A. (2017). C1orf64 is a novel androgen receptor target gene and coregulator that interacts with 14-3-3 protein in breast cancer. *Oncotarget*, 8(34):57907.

Nguyen, D. V. and Rocke, D. M. (2002). Multi-class cancer classification via partial least squares with gene expression profiles. *Bioinformatics*, 18(9):1216–1226.

Olsen, M. K. and Schafer, J. L. (2001). A two-part random-effects model for semicontinuous longitudinal data. *Journal of the American Statistical Association*, 96(454):730–745.

Park, P. J., Tian, L., and Kohane, I. S. (2002). Linking gene expression data with patient survival times using partial least squares. *Bioinformatics*, 18(suppl_1):S120–S127.

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

Peduzzi, P., Hardy, R., and Holford, T. R. (1980). A stepwise variable selection procedure for nonlinear regression models. *Biometrics*, pages 511–516.

Powell, J. L. (1986). Symmetrically trimmed least squares estimation for tobit models. *Econometrica: journal of the Econometric Society*, pages 1435–1460.

Raftery, A. E., Madigan, D., and Hoeting, J. A. (1997). Bayesian model averaging for linear regression models. *Journal of the American Statistical Association*, 92(437):179–191.

Ranstam, J. and Cook, J. (2018). Lasso regression. *Journal of British Surgery*, 105(10):1348–1348.

Rezaee, S., Kakavandi, N., Shabani, M., Khosravi, M., Hosseini-Fard, S. R., and Najafi, M. (2020). Cox and ptgds gene expression levels in pgd2 synthesis pathway are correlated with mir-520 in patients with vessel restenosis. *Endocrine, Metabolic & Immune Disorders-Drug Targets (Formerly Current Drug Targets-Immune, Endocrine & Metabolic Disorders)*, 20(9):1514–1522.

Robert, C. and Casella, G. (2011). A short history of markov chain monte carlo: Subjective recollections from incomplete data. *Statistical Science*, pages 102–115.

Ročková, V. and George, E. I. (2018). The spike-and-slab lasso. *Journal of the American Statistical Association*, 113(521):431–444.

Rosenwald, A., Wright, G., Chan, W. C., Connors, J. M., Campo, E., Fisher, R. I., Gascoyne, R. D., Muller-Hermelink, H. K., Smeland, E. B., Giltnane, J. M., et al. (2002). The use of molecular profiling to predict survival after chemotherapy for diffuse large-b-cell lymphoma. *New England Journal of Medicine*, 346(25):1937–1947.

Rossell, D., Cook, J., Telesca, D., and Roebuck, P. (2014). mombf: Moment and inverse moment bayes factors, r package version 1. 5.9.

Saei, A., Ward, J., and McGilchrist, C. (1996). Threshold models in a methadone programme evaluation. *Statistics in Medicine*, 15(20):2253–2260.

Scott, S. L. (2004). A bayesian paradigm for designing intrusion detection systems. *Computational statistics & data analysis*, 45(1):69–83.

Sha, N., Tadesse, M. G., and Vannucci, M. (2006). Bayesian variable selection for the analysis of microarray data with censored outcomes. *Bioinformatics*, 22(18):2262–2268.

Slenter, D. N., Kutmon, M., Hanspers, K., Riutta, A., Windsor, J., Nunes, N., Mélius, J., Cirillo, E., Coort, S. L., Digles, D., et al. (2018). Wikipathways: a multifaceted pathway database bridging metabolomics to other omics research. *Nucleic acids research*, 46(D1):D661–D667.

Spiegelhalter, D., Thomas, A., Best, N., and Lunn, D. (2003). Winbugs user manual.

Stingo, F. C., Chen, Y. A., Tadesse, M. G., and Vannucci, M. (2011). Incorporating biological

information into linear models: A bayesian approach to the selection of pathways and genes. *The annals of applied statistics*, 5(3).

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43):15545–15550.

Tanner, M. A. and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American statistical Association*, 82(398):528–540.

Thomas Bayes. https://en.wikipedia.org/wiki/Thomas_Bayes#cite_note-5.

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

Tobin, J. (1958). Estimation of relationships for limited dependent variables. *Econometrica: journal of the Econometric Society*, pages 24–36.

Tüchler, R. (2008). Bayesian variable selection for logistic models using auxiliary mixture sampling. *Journal of Computational and Graphical Statistics*, 17(1):76–94.

Vannucci, M. and Stingo, F. C. (2010). Bayesian models for variable selection that incorporate biological information. *Bayesian Statistics*, 9:1–20.

Volinsky, C. T., Madigan, D., Raftery, A. E., and Kronmal, R. A. (1997). Bayesian model averaging in proportional hazard models: assessing the risk of a stroke. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 46(4):433–448.

Wang, J., Duncan, D., Shi, Z., and Zhang, B. (2013). Web-based gene set analysis toolkit (webgestalt): update 2013. *Nucleic acids research*, 41(W1):W77–W83.

Wang, J., Vasaikar, S., Shi, Z., Greer, M., and Zhang, B. (2017). Webgestalt 2017: a more comprehensive, powerful, flexible and interactive gene set enrichment analysis toolkit. *Nucleic acids research*, 45(W1):W130–W137.

Wang, X., Feng, X., and Song, X. (2020). Joint analysis of semicontinuous data with latent variables. *Computational Statistics & Data Analysis*, 151:107005.

Yamada, Y., Ando, F., and Shimokata, H. (2009). Association of polymorphisms of sorbs1, gck and wisp1 with hypertension in community-dwelling japanese individuals. *Hypertension research*, 32(5):325–331.

YARDIMCI, A. and Aydın, E. (2002). Bayesian variable selection in linear regression and a comparison. *Hacettepe Journal of Mathematics and Statistics*, 31:63–76.

Yiu, S. and Tom, B. D. (2018). Two-part models with stochastic processes for modelling longitudinal semicontinuous data: Computationally efficient inference and modelling the overall marginal mean. *Statistical methods in medical research*, 27(12):3679–3695.

Zalta, E. N. (2008). Bayes' theorem.

Zhang, H., Xu, P., and Song, Y. (2021). Machine-learning-based m5c score for the prognosis diagnosis of osteosarcoma. *Journal of oncology*, 2021.

Zhao, P. and Yu, B. (2006). On model selection consistency of lasso. *The Journal of Machine Learning Research*, 7:2541–2563.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429.

Zou, H. and 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.

Appendices

Appendix A

A.1 Derivation of the full conditional density, $\beta_2|\sigma^2, \mathbf{y}$

This is the full conditional of $\beta_2|\sigma^2, \mathbf{y}$ before the introduction of vector of indicator variables $\boldsymbol{z^{(2)}}$ to the linear model.

$$
\begin{aligned}
P(\beta_2|\sigma^2,\mathbf{y}) &= \int P(\beta_1,\beta_2,\sigma^2|\mathbf{y},X)d\beta_1 \\
&= \int P(\mathbf{y}|\beta_1,\beta_2,\sigma^2,X)P(\beta_1,\beta_2,\sigma^2)d\beta_1 \\
&= \int P(\mathbf{y}|\beta_1,\beta_2,\sigma^2,X)\times P(\beta_2|\sigma^2)\times P(\sigma^2)\times P(\beta_1)d\beta_1 \\
&= \int \prod_{\substack{i=1,\\ y_i=0}}^{n}\left(\frac{e^{X_{1i}^\top\beta_1}}{1+e^{X_{1i}^\top\beta_1}}\right)\prod_{\substack{i=1,\\ y_i>0}}^{n}\left(\frac{1}{1+e^{X_{1i}^\top\beta_1}}\right)|\Sigma|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}\left[(\mathbf{y}^*-X_2^*\beta_2)^\top\right.\right. \\
&\left.\left.\times\Sigma^{-1}(\mathbf{y}^*-X_2^*\beta_2)\right]\right\}|\Sigma_0|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2\sigma^2}(\beta_2-\beta_0)^\top\Sigma_0^{-1}(\beta_2-\beta_0)\right\}P(\sigma^2)P(\beta_1)d\beta_1 \\
&= \int \prod_{\substack{i=1,\\ y_i=0}}^{n}\left(\frac{e^{X_{1i}^\top\beta_1}}{1+e^{X_{1i}^\top\beta_1}}\right)\prod_{\substack{i=1,\\ y_i>0}}^{n}\left(\frac{1}{1+e^{X_{1i}^\top\beta_1}}\right)P(\beta_1)d\beta_1|\Sigma|^{-\frac{1}{2}}|\Sigma_0|^{-\frac{1}{2}} \\
&\times\exp\left\{-\frac{1}{2}\left[(\mathbf{y}^*-X_2^*\beta_2)^\top\Sigma^{-1}(\mathbf{y}^*-X_2^*\beta_2)\right]\right\}\exp\left\{-\frac{1}{2\sigma^2}(\beta_2-\beta_0)^\top\Sigma_0^{-1}(\beta_2-\beta_0)\right\}P(\sigma^2) \\
&\propto |\Sigma|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}\left[(\mathbf{y}^*-X_2^*\beta_2)^\top\Sigma^{-1}(\mathbf{y}^*-X_2^*\beta_2)\right]\right\}|\Sigma_0|^{-\frac{1}{2}} \\
&\times\exp\left\{-\frac{1}{2\sigma^2}(\beta_2-\beta_0)^\top\Sigma_0^{-1}(\beta_2-\beta_0)\right\}\frac{\left(\frac{\nu_0\sigma_0^2}{2}\right)^{\frac{\nu_0}{2}}}{\Gamma(\frac{\nu_0}{2})}\left(\sigma^2\right)^{-\frac{\nu_0}{2}-1}\exp\left\{-\frac{\nu_0\sigma_0^2}{2\sigma^2}\right\} \\
&\propto \exp\left\{-\frac{1}{2\sigma^2}\left[(\mathbf{y}^*-X_2^*\beta_2)^\top(\mathbf{y}^*-X_2^*\beta_2)\right]\right\}\exp\left\{-\frac{1}{2\sigma^2}(\beta_2-\beta_0)^\top\Sigma_0^{-1}(\beta_2-\beta_0)\right\} \\
&\propto \exp\left\{-\frac{1}{2\sigma^2}\left(\beta_0^\top\Sigma_0^{-1}\beta_0\right)\right\}\exp\left\{-\frac{1}{2\sigma^2}\mathbf{y}^{*\top}\mathbf{y}^*\right\}\exp\left\{-\frac{1}{2}(\beta_2-\mu_\beta)^\top\Sigma_\beta^{-1}(\beta_2-\mu_\beta)\right\} \\
P(\beta_2|\sigma^2,\mathbf{y}) &\propto \exp\left\{-\frac{1}{2}(\beta_2-\mu_\beta)^\top\Sigma_\beta^{-1}(\beta_2-\mu_\beta)\right\},\ \text{where } \Sigma=\sigma^2\Sigma_0. \qquad \text{(A.1.1)}
\end{aligned}
$$

A.2 Derivation of the marginal posterior density of $\sigma^2|\mathbf{y}$

This is the full conditional of $\sigma^2|\mathbf{y}$ before the introduction of vector of indicator variables $\boldsymbol{z}^{(\mathbf{2})}$ in the linear model.

$$
\begin{aligned}
P(\sigma^2|\mathbf{y}) &= \int_{\beta_2} P(\beta_2, \sigma^2|\mathbf{y}, X) d\beta_2 \\
&= \int_{\beta_2} P(\mathbf{y}|\beta_2, \sigma^2, X) P(\beta_2, \sigma^2) d\beta_2 \\
&= \int_{\beta_2} P(\mathbf{y}|\beta_2, \sigma^2, X) \times P(\beta_2|\sigma^2) \times P(\sigma^2) d\beta_2 \\
&= P(\sigma^2) \int_{\beta_2} (2\pi)^{-\frac{n_2}{2}} \left(\sigma^2\right)^{-\frac{n_2}{2}} \exp\left\{-\frac{1}{2\sigma^2}\left[(\mathbf{y}^* - X_2^*\beta_2)^\top (\mathbf{y}^* - X_2^*\beta_2)\right]\right\} \\
&\times (2\pi)^{-\frac{p_2}{2}} \left(\sigma^2\right)^{-\frac{p_2}{2}} |\Sigma_0|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2\sigma^2}\left[(\beta_2 - \beta_0)^\top \Sigma_0^{-1} (\beta_2 - \beta_0)\right]\right\} d\beta_2 \\
&= P(\sigma^2) \int_{\beta_2} (2\pi)^{-\frac{n_2+p_2}{2}} \left(\sigma^2\right)^{-\frac{n_2+p_2}{2}} \exp\left\{-\frac{1}{2\sigma^2}\left[(\mathbf{y}^* - X_2^*\beta_2)^\top (\mathbf{y}^* - X_2^*\beta_2)\right]\right\} \\
&\times |\Sigma_0|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2\sigma^2}\left[(\beta_2 - \beta_0)^\top \Sigma_0^{-1} (\beta_2 - \beta_0)\right]\right\} d\beta_2 \\
&= P(\sigma^2) \int_{\beta_2} (2\pi)^{-\frac{n_2+p_2}{2}} \left(\sigma^2\right)^{-\frac{n_2+p_2}{2}} \exp\left\{-\frac{1}{2\sigma^2}\left[\mathbf{y}^{*\top}\mathbf{y}^* - 2\beta_2^\top X_2^{*\top}\mathbf{y}^* + \beta_2^\top X_2^{*\top} X_2^* \beta_2\right.\right. \\
&\left.\left. + \beta_2^\top \Sigma_0^{-1}\beta_2 - 2\beta_2^\top \Sigma_0^{-1}\beta_0 + \beta_0^\top \Sigma_0^{-1}\beta_0\right]\right\} d\beta_2 \\
&= P(\sigma^2) \int_{\beta_2} (2\pi)^{-\frac{n_2+p_2}{2}} \left(\sigma^2\right)^{-\frac{n_2+p_2}{2}} \exp\left\{-\frac{1}{2\sigma^2}\left[\mathbf{y}^{*\top}\mathbf{y}^* - 2\beta_2^\top \left(X_2^{*\top}\mathbf{y}^* + \Sigma_0^{-1}\beta_0\right)\right.\right. \\
&\left.\left. + \beta_2^\top \left(X_2^{*\top} X_2^* + \Sigma_0^{-1}\right)\beta_2 + \beta_0^\top \Sigma_0^{-1}\beta_0\right]\right\} d\beta_2 \\
&= P(\sigma^2) (2\pi)^{-\frac{n_2+p_2}{2}} \left(\sigma^2\right)^{-\frac{n_2+p_2}{2}} \exp\left\{-\frac{1}{2\sigma^2}\left[\mathbf{y}^{*\top}\mathbf{y}^* + \beta_0^\top \Sigma_0^{-1}\beta_0\right]\right\} \\
&\int_{\beta_2} \exp\left\{-\frac{1}{2\sigma^2}\left[\beta_2^\top \left(X_2^{*\top} X_2^* + \Sigma_0^{-1}\right)\beta_2 - 2\beta_2^\top \left(X_2^{*\top}\mathbf{y}^* + \Sigma_0^{-1}\beta_0\right)\right]\right\} d\beta_2 \\
&= P(\sigma^2) (2\pi)^{-\frac{n_2}{2}} \left(\sigma^2\right)^{-\frac{n_2}{2}} \exp\left\{-\frac{1}{2\sigma^2}\left[\mathbf{y}^{*\top}\mathbf{y}^* + \beta_0^\top \Sigma_0^{-1}\beta_0\right]\right\} \times \exp\left\{+\frac{1}{2\sigma^2}\mu_b^\top \Sigma_b^{-1}\mu_b\right\} \\
&|\Sigma_b^{-1}|^{\frac{1}{2}} \times \int_{\beta_2} (2\pi)^{-\frac{p_2}{2}} \left(\sigma^2\right)^{-\frac{p_2}{2}} |\Sigma_b^{-1}|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2\sigma^2}\left[(\beta_2 - \mu_b)^\top \Sigma_b^{-1} (\beta_2 - \mu_b)\right]\right\} d\beta_2 \\
&\propto \frac{\left(\frac{\nu_0\sigma_0^2}{2}\right)^{\frac{\nu_0}{2}}}{\Gamma(\frac{\nu_0}{2})} \left(\sigma^2\right)^{-\frac{\nu_0}{2}-1} \exp\left\{-\frac{\nu_0\sigma_0^2}{2\sigma^2}\right\} \left(\sigma^2\right)^{-\frac{n_2}{2}} \exp\left\{-\frac{1}{2\sigma^2}\left[\mathbf{y}^{*\top}\mathbf{y}^* + \beta_0^\top \Sigma_0^{-1}\beta_0 - \mu_b^\top \Sigma_b^{-1}\mu_b\right]\right\} \\
&\propto \frac{\left(\frac{\nu_0+n_2\sigma_0^2}{2}\right)^{\frac{\nu_0}{2}}}{\Gamma(\frac{\nu_0}{2})} \left(\sigma^2\right)^{-\frac{(\nu_0+n_2)}{2}-1} \exp\left\{-\frac{1}{2\sigma^2}\left[\nu_0\sigma_0^2 + \mathbf{y}^{*\top}\mathbf{y}^* + \beta_0^\top \Sigma_0^{-1}\beta_0 - \mu_b^\top \Sigma_b^{-1}\mu_b\right]\right\} \\
P(\sigma^2|\mathbf{y}) &\propto (\sigma^2)^{-\frac{\nu_n}{2}-1} \exp\left\{-\frac{\nu_n\sigma_n^2}{2\sigma^2}\right\} \qquad \text{(A.2.1)}
\end{aligned}
$$

A.3 Derivation of the posterior density of $\beta_1|\mathbf{u}, X_1, \mathbf{R}$

This is the posterior density of $\beta_1|\mathbf{u}, X_1, \mathbf{R}$ before the introduction of vector of indicator variables $\boldsymbol{z}^{(\mathbf{1})}$ in the logistic model.

$$
\begin{aligned}
P(\beta_1|\mathbf{u}, X_1, \mathbf{R}) &\propto P(\mathbf{u}|\beta_1, \mathbf{R}, X_1)P(\beta_1) \\
&\propto \left(\frac{1}{2\pi}\right)^{\frac{n}{2}} |\Sigma_n|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\mathbf{u} - X_1\beta_1 - \mu)^\top \Sigma_n^{-1}(\mathbf{u} - X_1\beta_1 - \mu)\right\} \\
&\times \left(\frac{1}{2\pi}\right)^{\frac{p_1}{2}} |B_0|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\beta_1 - b_0)^\top B_0^{-1}(\beta_1 - b_0)\right\} \\
&\propto \exp\left\{-\frac{1}{2}(\mathbf{u} - X_1\beta_1 - \mu)^\top \Sigma_n^{-1}(\mathbf{u} - X_1\beta_1 - \mu) - \frac{1}{2}(\beta_1 - b_0)^\top B_0^{-1}(\beta_1 - b_0)\right\} \\
&\propto \exp\left\{-\frac{1}{2}\left(\mathbf{u}^\top\Sigma_n^{-1}\mathbf{u} - \mathbf{u}^\top\Sigma_n^{-1}X_1\beta_1 - \mathbf{u}^\top\Sigma_n^{-1}\mu - \beta_1^\top X_1^\top\Sigma_n^{-1}\mathbf{u}\right.\right. \\
&+\beta_1^\top X_1^\top\Sigma_n^{-1}X_1\beta_1 + \beta_1^\top X_1^\top\Sigma_n^{-1}\mu - \mu^\top\Sigma_n^{-1}\mathbf{u} + \mu^\top\Sigma_n^{-1}X_1\beta_1 + \mu^\top\Sigma_n^{-1}\mu \\
&\left.\left.+\beta_1^\top B_0^{-1}\beta_1 - \beta_1^\top B_0^{-1}b_0 - b_0^\top B_0^{-1}\beta_1 + \beta_0^\top B_0^{-1}b_0\right)\right\} \\
&\propto \exp\left\{-\frac{1}{2}\left(\mathbf{u}^\top\Sigma_n^{-1}\mathbf{u} - 2\beta_1^\top X_1^\top\Sigma_n^{-1}\mathbf{u} + 2\beta_1^\top X_1^\top\Sigma_n^{-1}\mu + \beta_1^\top X_1^\top\Sigma_n^{-1}X_1\beta_1\right.\right. \\
&\left.\left.+\beta_1^\top B_0^{-1}\beta_1 - 2\beta_1^\top B_0^{-1}b_0 + b_0^\top B_0^{-1}b_0 - 2\mu^\top\Sigma_n^{-1}\mathbf{u} + \mu^\top\Sigma_n^{-1}\mu\right)\right\} \\
&\propto \exp\left\{-\frac{1}{2}\left[\beta_1^\top\left(B_0^{-1} + X_1^\top\Sigma_n^{-1}X_1\right)\beta_1 - 2\beta_1^\top\left(B_0^{-1}b_0 + X_1^\top\Sigma_n^{-1}\mathbf{u}\right.\right.\right. \\
&\left.\left.\left.-X_1^\top\Sigma_n^{-1}\mu\right)\right]\right\}\exp\left\{-\frac{1}{2}\left[\mathbf{u}^\top\Sigma_n^{-1}\mathbf{u} - 2\mu^\top\Sigma_n^{-1}\mathbf{u} + \mu^\top\Sigma_n^{-1}\mu\right]\right\} \\
&\propto \exp\left\{-\frac{1}{2}\left[\beta_1^\top\left(B_0^{-1} + X_1^\top\Sigma_n^{-1}X_1\right)\beta_1 - 2\beta_1^\top\left(B_0^{-1}b_0 + X_1^\top\Sigma_n^{-1}\mathbf{u}\right.\right.\right. \\
&\left.\left.\left.-X_1^\top\Sigma_n^{-1}\mu\right)\right]\right\}\exp\left\{-\frac{1}{2}\left(\mathbf{u} - \mu\right)^\top\Sigma_n^{-1}\left(\mathbf{u} - \mu\right)\right\}\exp\left\{\frac{1}{2}b_n^\top B_n^{-1}b_n\right\} \\
P(\beta_1|\mathbf{u}, X_1, \mathbf{R}) &\propto \exp\left\{-\frac{1}{2}\left(\beta_1 - b_n\right)^\top B_n^{-1}\left(\beta_1 - b_n\right)\right\} \qquad \text{(A.3.1)}
\end{aligned}
$$

A.4 Method of Generating Exponential Variates for the Latent Utilities, u

Since the errors in equation (3.2.20) follows a approximate normal distribution using the values for the mean and variance parameters shown in Table (3.1) and the utility is of a binary category (L = 1), we obtain

$$\begin{aligned} \exp(-u_{1i}) &\sim \text{Exponential}(\lambda_i), \quad l = 1, \cdots, L \\ \exp(-u_i) &\sim \text{Exponential}(1), \quad l = 0. \end{aligned} \tag{A.4.1}$$

Given that the categorical observation $l \in \{0, \cdots, L\}$ say $y_i = l$, the utilities u_{1i} is the maximum of all utilities

$$u_{li} = \max_{l'=0,\cdots,L} u_{l'i} \Longleftrightarrow \exp(-u_{li}) = \min_{l'=0,\cdots,L} \exp(-u_{l'i}). \tag{A.4.2}$$

Since, $\exp(-u_{li})$ is the minimum of exponentially distributed random variables, its parameter has to be equal to $1 + \lambda_i$ and for all other utilities $u_{\bar{l}i}$, $\bar{l} = 1, \cdots, L,\ \ \bar{l} \neq l$, then the following relationship holds:

$$\exp(-u_{\bar{l}i}) = \text{Exponential}(1 + \lambda_{l'i}) + \text{Exponential}(\lambda_{\bar{l}i}). \tag{A.4.3}$$

An illustration of how equation (A.4.3) becomes the expression in equation (3.2.22) can be found in the link on Generating Exponential Variates.

A.5 Marginal Distribution of $\boldsymbol{\beta_{1z}}, \boldsymbol{\beta_{2z}}$, and σ^2

To show that there was no closed form of the joint posterior distribution of $\boldsymbol{\beta_{1z}}, \boldsymbol{\beta_{2z}}$,and σ^2, we express the marginal distribution of each as follows;

The marginal probability distribution of $\boldsymbol{\beta_{1z}}$,

To obtain the marginal probability distribution of $\boldsymbol{\beta_{1z}}$, integrate out $\boldsymbol{\beta_{2z}}$ and σ^2,

$$P(\boldsymbol{\beta_{1z}}|\boldsymbol{\beta_{2z}},\sigma^2,\mathbf{y},\boldsymbol{z^{(1)}},\boldsymbol{z^{(2)}},\boldsymbol{X}) = \int_{\boldsymbol{\beta_{2z}}}\int_{\sigma^2} P(\mathbf{y}|\boldsymbol{\beta_{1z}},\boldsymbol{\beta_{2z}},\boldsymbol{z^{(1)}},\boldsymbol{z^{(2)}},\sigma^2,\boldsymbol{X})P(\boldsymbol{\beta_{1z}}|\boldsymbol{z^{(1)}}) \times P(\boldsymbol{\beta_{2z}}|\sigma^2,\boldsymbol{z^{(2)}})P(\sigma^2|\boldsymbol{z^{(2)}})d\sigma^2 d\boldsymbol{\beta_{2z}} \quad \text{(A.5.1)}$$

The marginal probability distribution of $\boldsymbol{\beta_{2z}}$,

To obtain the marginal probability distribution of $\boldsymbol{\beta_{2z}}$, integrate out $\boldsymbol{\beta_{1z}}$ and σ^2,

$$P(\boldsymbol{\beta_{2z}}|\boldsymbol{\beta_{1z}},\sigma^2,\mathbf{y},\boldsymbol{z^{(1)}},\boldsymbol{z^{(2)}},\boldsymbol{X}) = \int_{\boldsymbol{\beta_{1z}}}\int_{\sigma^2} P(\mathbf{y}|\boldsymbol{\beta_{1z}},\boldsymbol{\beta_{2z}},\boldsymbol{z^{(1)}},\boldsymbol{z^{(2)}},\sigma^2,\boldsymbol{X})P(\boldsymbol{\beta_{1z}}|\boldsymbol{z^{(1)}}) \times P(\boldsymbol{\beta_{2z}}|\sigma^2,\boldsymbol{z^{(2)}})P(\sigma^2|\boldsymbol{z^{(2)}})d\sigma^2 d\boldsymbol{\beta_{1z}} \quad \text{(A.5.2)}$$

The marginal probability distribution of σ^2,

To obtain the marginal probability distribution of σ^2, integrate out $\boldsymbol{\beta_{1z}}$ and $\boldsymbol{\beta_{2z}}$,

$$P(\sigma^2|\boldsymbol{\beta_{1z}},\boldsymbol{\beta_{2z}},\mathbf{y},\boldsymbol{z^{(1)}},\boldsymbol{z^{(2)}},\boldsymbol{X}) = \int_{\boldsymbol{\beta_{1z}}}\int_{\boldsymbol{\beta_{2z}}} P(\mathbf{y}|\boldsymbol{\beta_{1z}},\boldsymbol{\beta_{2z}},\boldsymbol{z^{(1)}},\boldsymbol{z^{(2)}},\sigma^2,\boldsymbol{X})P(\boldsymbol{\beta_{1z}}|\boldsymbol{z^{(1)}}) \times P(\boldsymbol{\beta_{2z}}|\sigma^2,\boldsymbol{z^{(2)}})P(\sigma^2|\boldsymbol{z^{(2)}})d\boldsymbol{\beta_{2z}} d\boldsymbol{\beta_{1z}} \quad \text{(A.5.3)}$$

Since, integrating equation (A.5.1), (A.5.2), and (A.5.3) cannot be performed analytically, we try to obtain the full conditional density and if it exist we can use the MCMC techniques of the Gibbs sampler but if the full conditional density does not exist we therefore apply the Metropolis-Hastings algorithm. In this study, the full conditional exist that is the reason for using a Gibbs sampler algorithm as our MCMC technique.

Appendix B

B.1 Posterior Summary for Method 1 and Method 2

Table B.1: Posterior Summary for Method 1 when n = 300, p = 500

n = 300, p = 500, with 20 important features specified out of p feature and 3 clinical covariates.							
Model	**p = 500**	**Parameter**	**True Value**	**Posterior Est.**	**95% C.I**	**95% HPDI**	**Prob. of Selection**
Logistic	Intercept	β_0	1.0000	0.5811	0.1599 - 1.0217	0.1615 - 1.0231	-
	3 Clinical	β_1	1.0000	0.8872	0.4432 - 1.4012	0.4324 - 1.3889	-
	Covariates	β_2	1.0000	0.3810	-0.0550 - 0.8247	-0.0529 - 0.8262	-
		β_3	1.0000	0.8730	0.4541 - 1.3099	0.4501 - 1.3056	-
		β_4	1.0000	0.9633	0.0000 - 1.4682	0.0000 - 1.3866	1.0000
		β_5	1.0000	0.9868	0.5533 - 1.5022	0.5360 - 1.4819	1.0000
		⋮	⋮	⋮ - ⋮	⋮ - ⋮	⋮	⋮
		β_{503}	0.0000	0.0000	0.0000 - 0.0000	0.0000 - 0.0000	0.0143
		β_{504}	0.0000	0.0000	0.0000 - 0.4704	0.0000 - 0.2814	0.0174
Linear	Intercept	β_0	1.0000	-0.8170	-1.1533 - (-0.4773)	-1.1526 - (-0.4769)	-
	3 Clinical	β_1	1.0000	0.3831	0.0992 - 0.6585	0.0983 - 0.6575	-
	Covariates	β_2	1.0000	0.2752	-0.0102 - 0.5437	-0.0103 - 0.5435	-
		β_3	1.0000	0.4228	0.1516 - 0.6835	0.1515 - 0.6835	-
		β_4	1.0000	0.3671	0.0815 - 0.6415	0.0823 - 0.6422	0.9987
		β_5	1.0000	0.5406	0.2655 - 0.8051	0.2691 - 0.8082	1.0000
		⋮	⋮	⋮ - ⋮	⋮ - ⋮	⋮	⋮
		β_{503}	0.0000	0.0000	-0.1755 - 0.0824	-0.1711 - 0.0868	0.1559
		β_{504}	0.0000	0.0000	-0.1295 - 0.1987	-0.1348 - 0.1926	0.1851
		σ^2	0.2500	0.2519	0.1949 - 0.3328	0.1895 - 0.3252	

Table B.2: Posterior Summary for Method 1 when n = 500, p = 500

n = 500, p = 500, with 20 important features specified out of p feature and 3 clinical covariates.							
Model	**p = 500**	**Parameter**	**True Value**	**Posterior Est.**	**95% C.I**	**95% HPDI**	**Prob. of Selection**
Logistic	Intercept	β_0	1.0000	0.8189	0.4641 - 1.2096	0.4517 - 1.1960	-
	3 Clinical	β_1	1.0000	0.9467	0.5510 - 1.3375	0.5440 - 1.3297	-
	Covariates	β_2	1.0000	0.6138	0.2616 - 0.9678	0.2618 - 0.9679	-
		β_3	1.0000	0.8600	0.5168 - 1.2560	0.5004 - 1.2381	-
		β_4	1.0000	1.0350	0.5071 - 1.4656	0.5037 - 1.4619	1.0000
		β_5	1.0000	0.8775	0.0000 - 1.2736	0.0000 - 1.2147	0.7573
		⋮	⋮	⋮ - ⋮	⋮ - ⋮	⋮	⋮
		β_{503}	0.0000	0.0000	0.0000 - 0.0000	0.0000 - 0.0000	0.0193
		β_{504}	0.0000	0.0000	0.0000 - 0.0000	0.0000 - 0.0000	0.0128
Linear	Intercept	β_0	1.0000	0.8665	0.7461 - 0.9862	0.7451 - 0.9851	-
	3 Clinical	β_1	1.0000	0.9180	0.8410 - 0.9936	0.8413 - 0.9939	-
	Covariates	β_2	1.0000	0.9978	0.9204 - 1.0748	0.9199 - 1.0741	-
		β_3	1.0000	0.9567	0.8899 - 1.0235	0.8903 - 1.0238	-
		β_4	1.0000	0.9819	0.9064 - 1.0572	0.9062 - 1.0569	1.0000
		β_5	1.0000	0.9176	0.8440 - 0.9918	0.8441 - 0.9918	1.0000
		⋮	⋮	⋮ - ⋮	⋮ - ⋮	⋮	⋮
		β_{503}	0.0000	0.0000	0.0000 - 0.0000	0.0000 - 0.0000	0.0224
		β_{504}	0.0000	0.0000	0.0000 - 0.0000	0.0000 - 0.0000	0.0051
		σ^2	0.2500	0.2179	0.1757 - 0.2718	0.1734 - 0.2688	

Table B.3: Posterior Summary for Method 2: Combined model when n = 300, p = 500

n = 300, p = 500, with 20 important features specified out of p feature and 3 clinical covariates.							
Model	**p = 500**	**Parameter**	**True Value**	**Posterior Est.**	**95% C.I**	**95% HPDI**	**Prob. of Selection**
Logistic	Intercept	β_0	1.0000	0.4645	0.0659 - 0.8817	0.0629 - 0.8783	-
	3 Clinical	β_1	1.0000	0.4794	0.0520 - 0.9103	0.0572 - 0.9148	-
	Covariates	β_2	1.0000	0.8084	0.3935 - 1.2483	0.3892 - 1.2435	-
		β_3	1.0000	0.6638	0.2747 - 1.0870	0.2648 - 1.0756	-
		β_4	1.0000	0.7320	0.3105 - 1.2001	0.2958 - 1.1838	1.0000
		β_5	1.0000	0.8136	0.4090 - 1.2329	0.4056 - 1.2292	1.0000
		⋮	⋮	⋮ - ⋮	⋮ - ⋮	⋮	⋮
		β_{503}	0.0000	0.0000	0.0000 - 0.0000	0.0000 - 0.0000	0.0048
		β_{504}	0.0000	0.0000	0.0000 - 0.0000	0.0000 - 0.0000	0.0023
Linear	Intercept	β_0	1.0000	0.9163	0.7654 - 1.0650	0.7671 - 1.0665	-
	3 Clinical	β_1	1.0000	0.9746	0.8756 - 1.0736	0.8755 - 1.0735	-
	Covariates	β_2	1.0000	0.9237	0.8360 - 1.0111	0.8362 - 1.0113	-
		β_3	1.0000	1.0063	0.9231 - 1.0893	0.9235 - 1.0896	-
		β_4	1.0000	1.0214	0.9217 - 1.1175	0.9221 - 1.1178	1.0000
		β_5	1.0000	0.8956	0.8041 - 0.9873	0.8044 - 0.9875	1.0000
		⋮	⋮	⋮ - ⋮	⋮ - ⋮	⋮	⋮
		β_{503}	0.0000	0.0000	0.0000 - 0.0000	0.0000 - 0.0000	0.0048
		β_{504}	0.0000	0.0000	0.0000 - 0.0000	0.0000 - 0.0000	0.0023
		σ^2	0.2500	0.2424	0.1900 - 0.3118	0.1862 - 0.3068	

Table B.4: Posterior Summary for Method 2: Combined model when n = 500, p = 500

n = 500, p = 500, with 20 important features specified out of p feature and 3 clinical covariates.							
Model	**p = 500**	**Parameter**	**True Value**	**Posterior Est.**	**95% C.I**	**95% HPDI**	**Prob. of Selection**
Logistic	Intercept	β_0	1.0000	0.8175	0.5158 - 1.1377	0.5147 - 1.1362	-
	3 Clinical	β_1	1.0000	0.9512	0.6345 - 1.2697	0.6315 - 1.2664	-
	Covariates	β_2	1.0000	0.6324	0.3443 - 0.9313	0.3474 - 0.9341	-
		β_3	1.0000	0.8679	0.5663 - 1.1840	0.5619 - 1.1793	-
		β_4	1.0000	1.0826	0.7633 - 1.4302	0.7521 - 1.4177	1.0000
		β_5	1.0000	0.9069	0.6233 - 1.2130	0.6103 - 1.1975	1.0000
		⋮	⋮	⋮ - ⋮	⋮ - ⋮	⋮	⋮
		β_{503}	0.0000	0.0000	0.0000 - 0.0000	0.0000 - 0.0000	0.0000
		β_{504}	0.0000	0.0000	0.0000 - 0.0000	0.0000 - 0.0000	0.0000
Linear	Intercept	β_0	1.0000	0.8837	0.7623 - 1.0036	0.7633 - 1.0045	-
	3 Clinical	β_1	1.0000	0.9247	0.8493 - 1.0001	0.8485 - 0.9992	-
	Covariates	β_2	1.0000	0.9969	0.9201 - 1.0741	0.9211 - 1.0750	-
		β_3	1.0000	0.9619	0.8959 - 1.0277	0.8970 - 1.0286	-
		β_4	1.0000	0.9889	0.9126 - 1.0647	0.9118 - 1.0638	1.0000
		β_5	1.0000	0.9328	0.8598 - 1.0053	0.8601 - 1.0055	1.0000
		⋮	⋮	⋮ - ⋮	⋮ - ⋮	⋮	⋮
		β_{503}	0.0000	0.0000	0.0000 - 0.0000	0.0000 - 0.0000	0.0000
		β_{504}	0.0000	0.0000	0.0000 - 0.0000	0.0000 - 0.0000	0.0000
		σ^2	0.2500	0.2494	0.2050 - 0.3055	0.2024 - 0.3020	

B.2 Other convergence diagnostics plots from simulation study

Method 2: Combined model

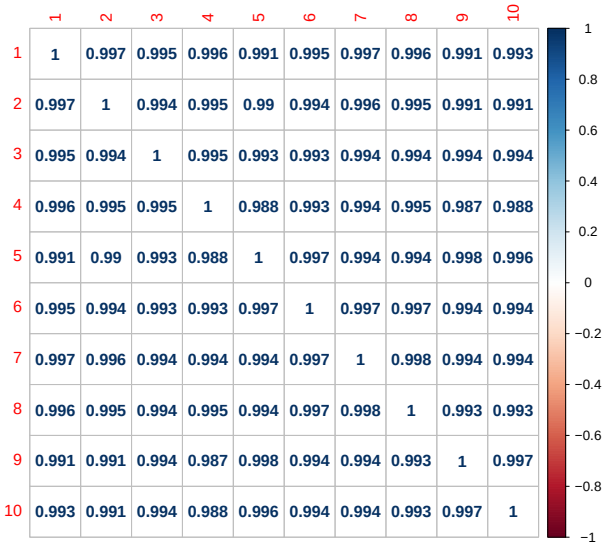


Figure B.1: Plot of correlation across the 10 chains in the combined model when n = 300; p = 500

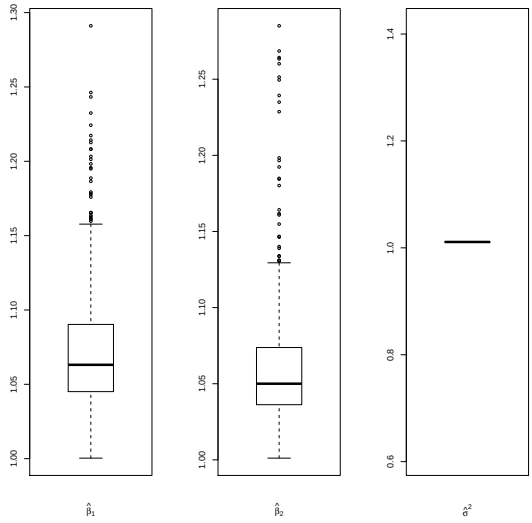


Figure B.2: Convergence diagnostic plot of the posterior estimate samples across the 10 chains for $\hat{\beta}_1$, $\hat{\beta}_2$, and $\hat{\sigma}^2$

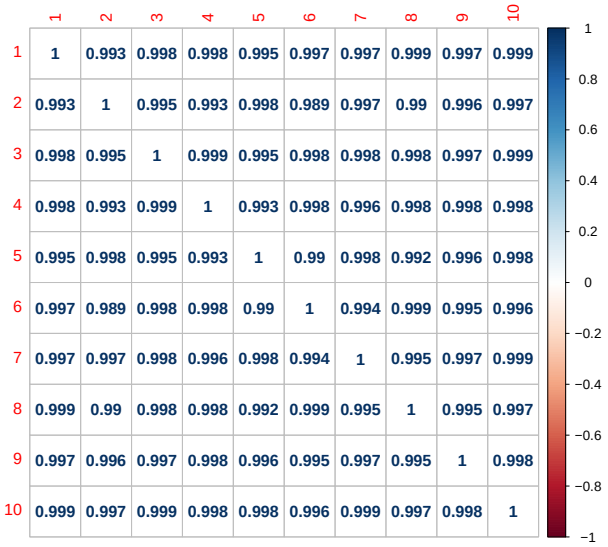


Figure B.3: Plot of correlation across the 10 chains in the combined model when n = 300; p = 200

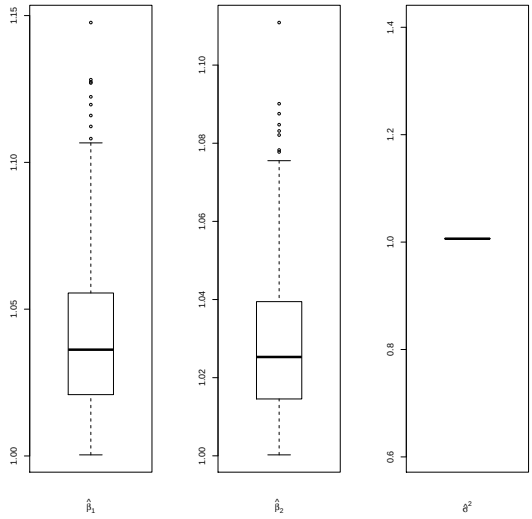


Figure B.4: Convergence diagnostic plot of the posterior estimate samples across the 10 chains for $\hat{\beta}_1$, $\hat{\beta}_2$, and $\hat{\sigma}^2$

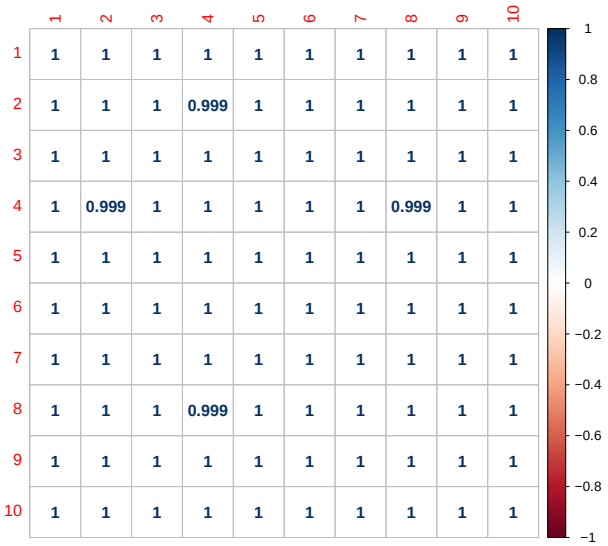


Figure B.5: Plot of correlation across the 10 chains in the combined model when n = 300; p = 50

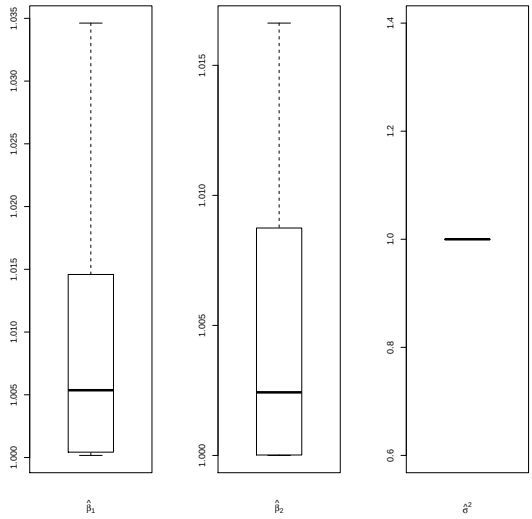


Figure B.6: Convergence diagnostic plot of the posterior estimate samples across the 10 chains for $\hat{\beta}_1$, $\hat{\beta}_2$, and $\hat{\sigma}^2$

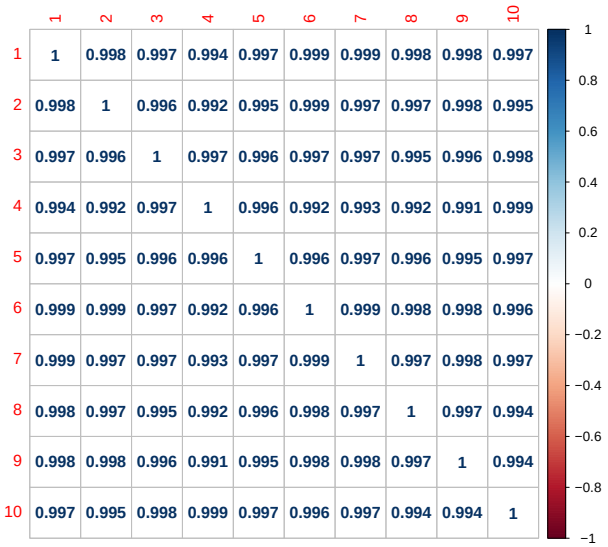


Figure B.7: Plot of correlation across the 10 chains in the combined model when n = 500; p = 500

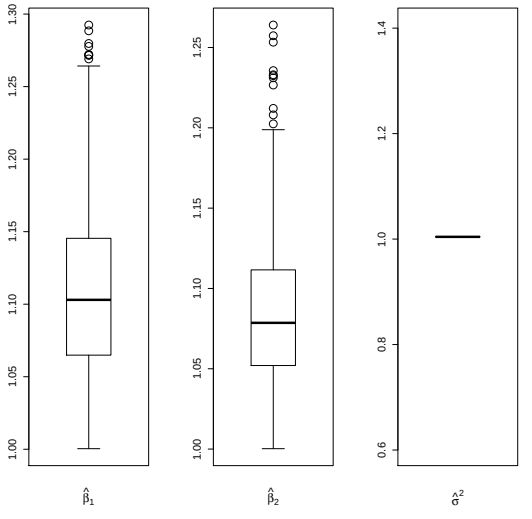


Figure B.8: Convergence diagnostic plot of the posterior estimate samples across the 10 chains for $\hat{\beta}_1$, $\hat{\beta}_2$, and $\hat{\sigma}^2$ when n = 500; p = 500

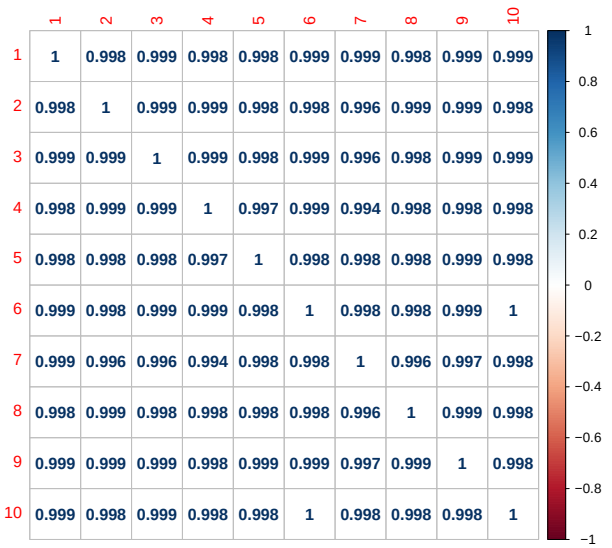


Figure B.9: Plot of correlation across the 10 chains in the combined model when n = 500; p = 200

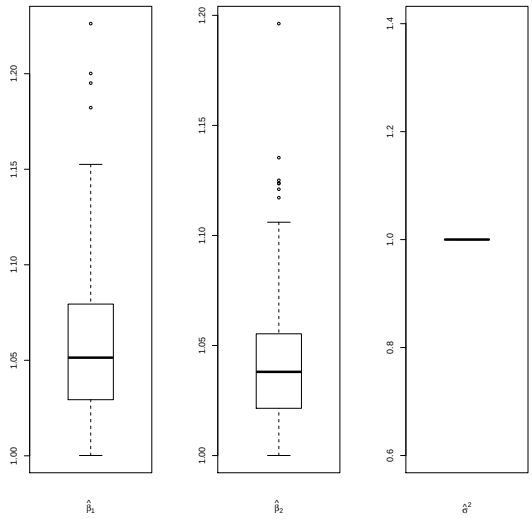


Figure B.10: Convergence diagnostic plot of the posterior estimate samples across the 10 chains for $\hat{\beta}_1$, $\hat{\beta}_2$, and $\hat{\sigma}^2$

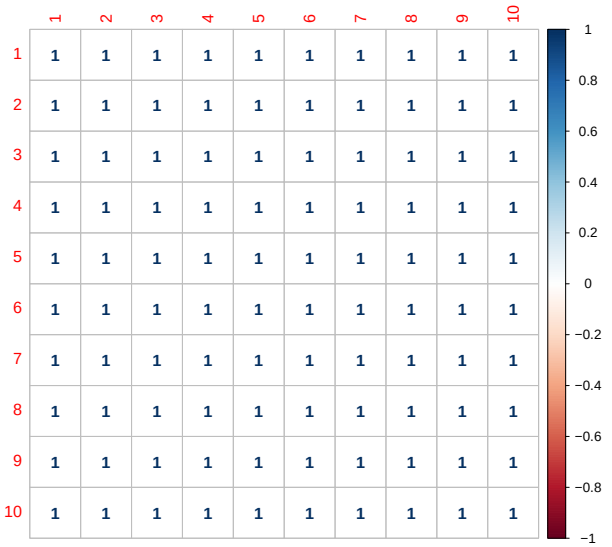


Figure B.11: Plot of correlation across the 10 chains in the combined model when n = 500; p = 50

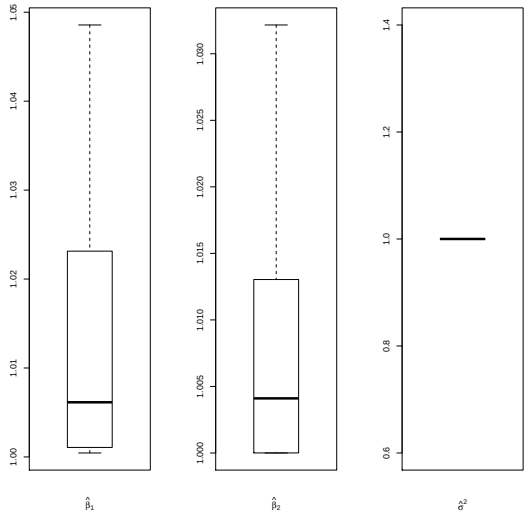
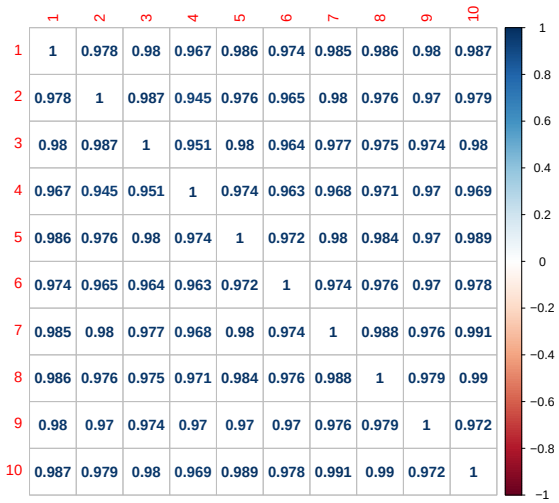


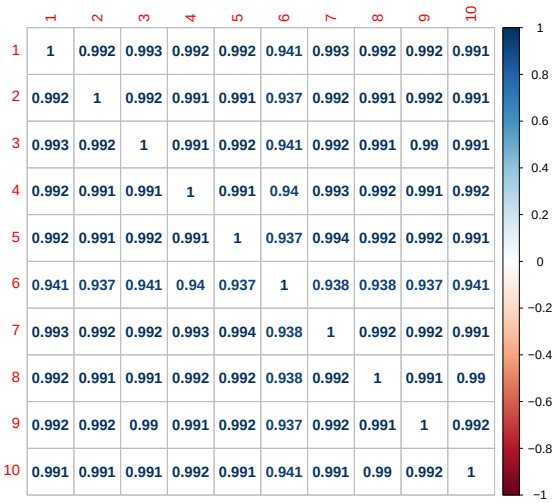
Figure B.12: Convergence diagnostic plot of the posterior estimate samples across the 10 chains for $\hat{\beta}_1$, $\hat{\beta}_2$, and $\hat{\sigma}^2$

Method 1



	1	2	3	4	5	6	7	8	9	10
1	1	0.978	0.98	0.967	0.986	0.974	0.985	0.986	0.98	0.987
2	0.978	1	0.987	0.945	0.976	0.965	0.98	0.976	0.97	0.979
3	0.98	0.987	1	0.951	0.98	0.964	0.977	0.975	0.974	0.98
4	0.967	0.945	0.951	1	0.974	0.963	0.968	0.971	0.97	0.969
5	0.986	0.976	0.98	0.974	1	0.972	0.98	0.984	0.97	0.989
6	0.974	0.965	0.964	0.963	0.972	1	0.974	0.976	0.97	0.978
7	0.985	0.98	0.977	0.968	0.98	0.974	1	0.988	0.976	0.991
8	0.986	0.976	0.975	0.971	0.984	0.976	0.988	1	0.979	0.99
9	0.98	0.97	0.974	0.97	0.97	0.97	0.976	0.979	1	0.972
10	0.987	0.979	0.98	0.969	0.989	0.978	0.991	0.99	0.972	1

Figure B.13: Plot of correlation across the 10 chains in the logistic model when n = 300; p = 500



	1	2	3	4	5	6	7	8	9	10
1	1	0.992	0.993	0.992	0.992	0.941	0.993	0.992	0.992	0.991
2	0.992	1	0.992	0.991	0.991	0.937	0.992	0.991	0.992	0.991
3	0.993	0.992	1	0.991	0.992	0.941	0.992	0.991	0.99	0.991
4	0.992	0.991	0.991	1	0.991	0.94	0.993	0.992	0.991	0.992
5	0.992	0.991	0.992	0.991	1	0.937	0.994	0.992	0.992	0.991
6	0.941	0.937	0.941	0.94	0.937	1	0.938	0.938	0.937	0.941
7	0.993	0.992	0.992	0.993	0.994	0.938	1	0.992	0.992	0.991
8	0.992	0.991	0.991	0.992	0.992	0.938	0.992	1	0.991	0.99
9	0.992	0.992	0.99	0.991	0.992	0.937	0.992	0.991	1	0.992
10	0.991	0.991	0.991	0.992	0.991	0.941	0.991	0.99	0.992	1

Figure B.14: Plot of correlation across the 10 chains in the linear model when n = 300; p = 500

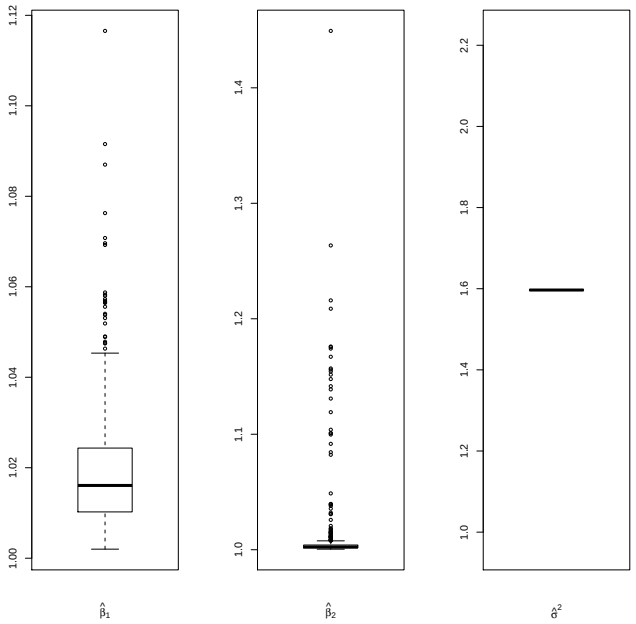


Figure B.15: Convergence diagnostic plot of the posterior estimate samples across the 10 chains for $\hat{\beta}_1$, $\hat{\beta}_2$, and $\hat{\sigma}^2$

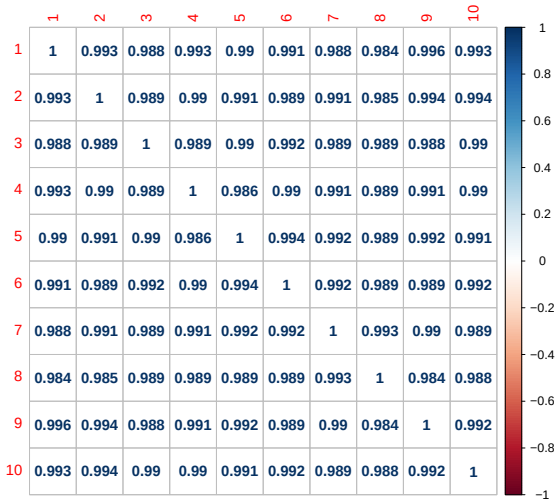


Figure B.16: Plot of correlation across the 10 chains in the logistic model when n = 300; p = 200

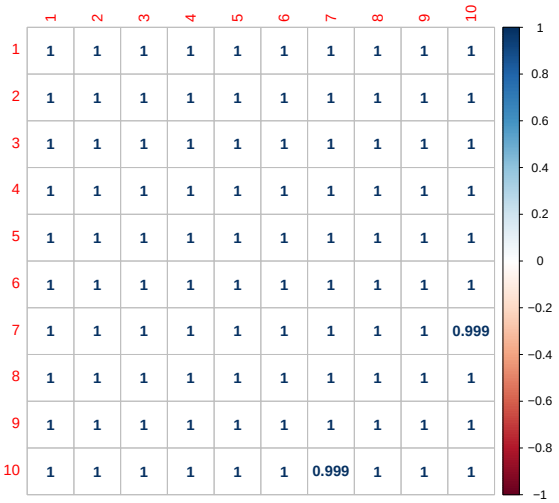


Figure B.17: Plot of correlation across the 10 chains in the linear model when n = 300; p = 200

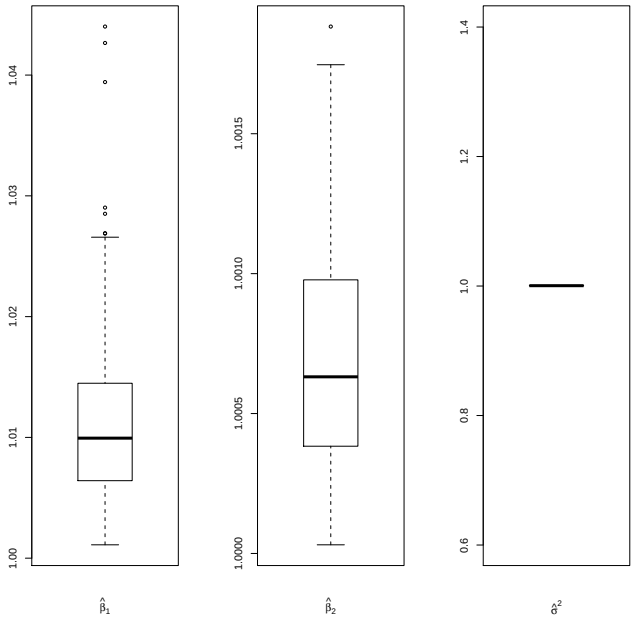


Figure B.18: Convergence diagnostic plot of the posterior estimate samples across the 10 chains for $\hat{\beta}_1$, $\hat{\beta}_2$, and $\hat{\sigma}^2$

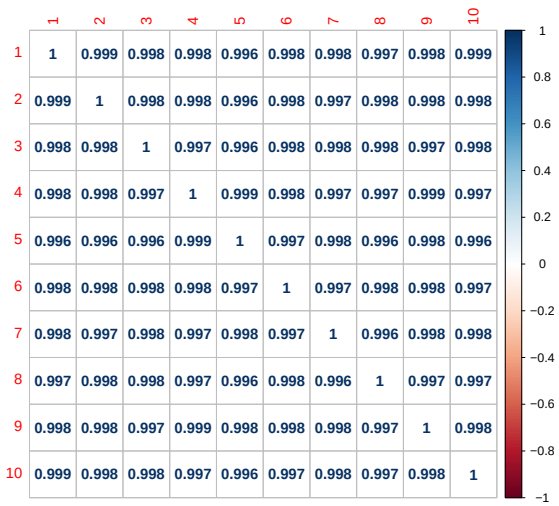


Figure B.19: Plot of correlation across the 10 chains in the logistic model when n = 300; p = 50

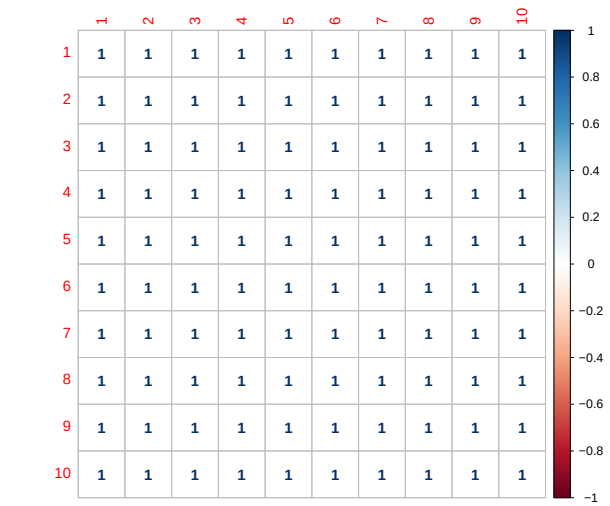


Figure B.20: Plot of correlation across the 10 chains in the linear model when n = 300; p = 50

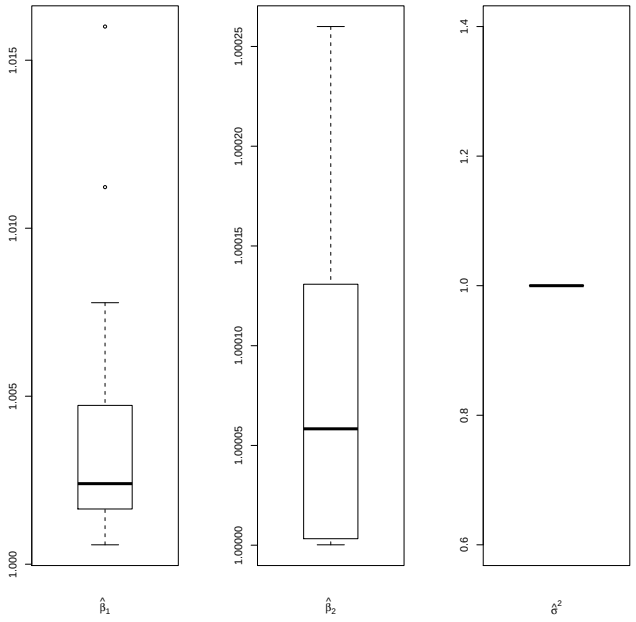
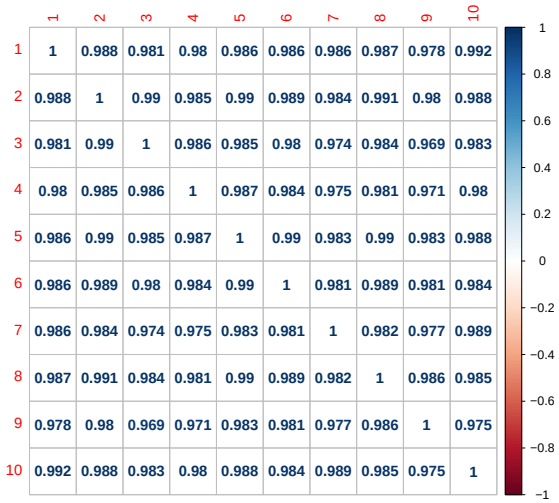
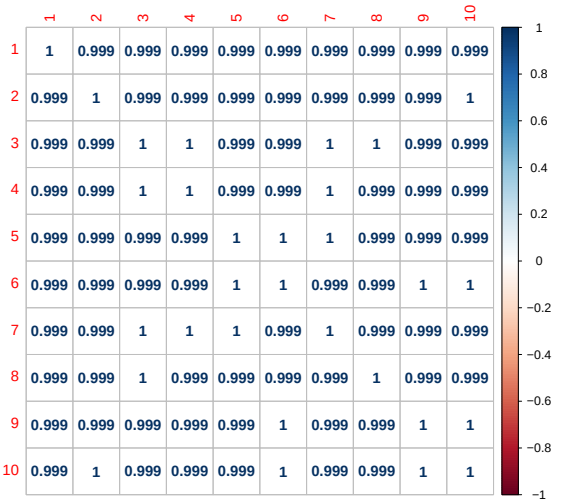


Figure B.21: Convergence diagnostic plot of the posterior estimate samples across the 10 chains for $\hat{\beta}_1$, $\hat{\beta}_2$, and $\hat{\sigma}^2$



	1	2	3	4	5	6	7	8	9	10
1	1	0.988	0.981	0.98	0.986	0.986	0.986	0.987	0.978	0.992
2	0.988	1	0.99	0.985	0.99	0.989	0.984	0.991	0.98	0.988
3	0.981	0.99	1	0.986	0.985	0.98	0.974	0.984	0.969	0.983
4	0.98	0.985	0.986	1	0.987	0.984	0.975	0.981	0.971	0.98
5	0.986	0.99	0.985	0.987	1	0.99	0.983	0.99	0.983	0.988
6	0.986	0.989	0.98	0.984	0.99	1	0.981	0.989	0.981	0.984
7	0.986	0.984	0.974	0.975	0.983	0.981	1	0.982	0.977	0.989
8	0.987	0.991	0.984	0.981	0.99	0.989	0.982	1	0.986	0.985
9	0.978	0.98	0.969	0.971	0.983	0.981	0.977	0.986	1	0.975
10	0.992	0.988	0.983	0.98	0.988	0.984	0.989	0.985	0.975	1

Figure B.22: Plot of correlation across the 10 chains in the logistic model when n = 500; p = 500



	1	2	3	4	5	6	7	8	9	10
1	1	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999
2	0.999	1	0.999	0.999	0.999	0.999	0.999	0.999	0.999	1
3	0.999	0.999	1	1	0.999	0.999	1	1	0.999	0.999
4	0.999	0.999	1	1	0.999	0.999	1	0.999	0.999	0.999
5	0.999	0.999	0.999	0.999	1	1	1	0.999	0.999	0.999
6	0.999	0.999	0.999	0.999	1	1	0.999	0.999	1	1
7	0.999	0.999	1	1	1	0.999	1	0.999	0.999	0.999
8	0.999	0.999	1	0.999	0.999	0.999	0.999	1	0.999	0.999
9	0.999	0.999	0.999	0.999	0.999	1	0.999	0.999	1	1
10	0.999	1	0.999	0.999	0.999	1	0.999	0.999	1	1

Figure B.23: Plot of correlation across the 10 chains in the linear model when n = 500; p = 500

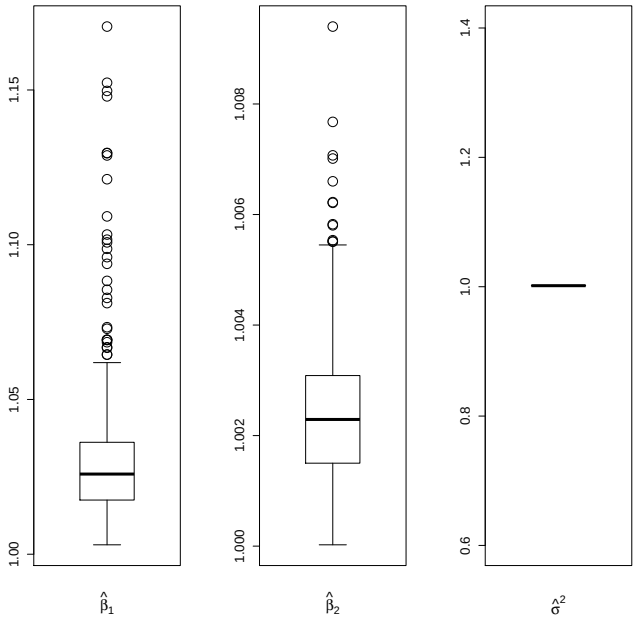


Figure B.24: Convergence diagnostic plot of the posterior estimate samples across the 10 chains for $\hat{\beta}_1$, $\hat{\beta}_2$, and $\hat{\sigma}^2$

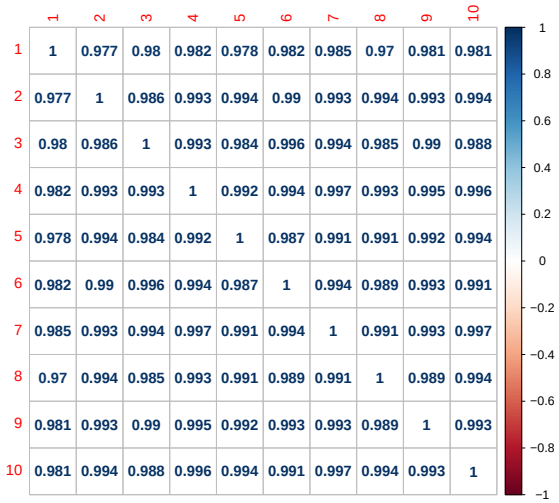


Figure B.25: Plot of correlation across the 10 chains in the logistic model when n = 500; p = 200

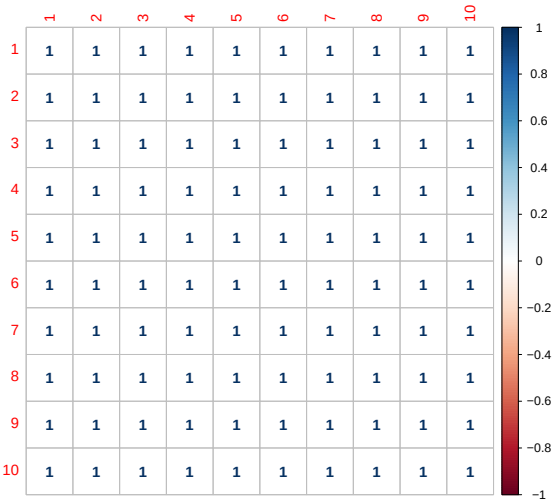


Figure B.26: Plot of correlation across the 10 chains in the linear model when n = 500; p = 200

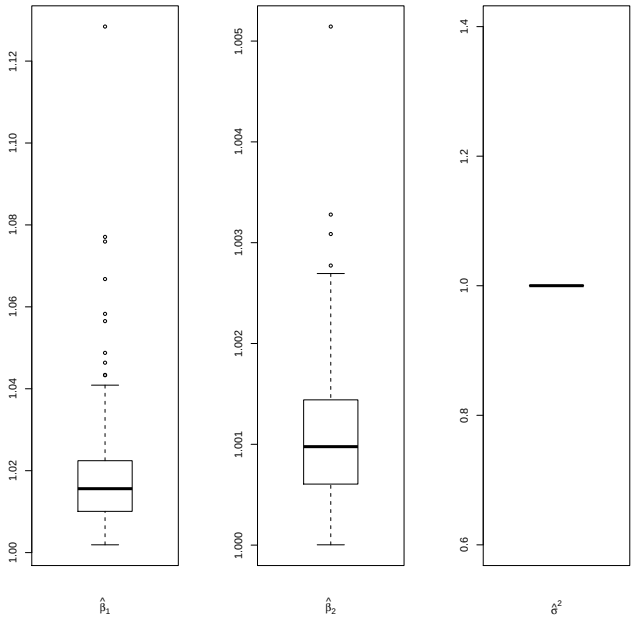


Figure B.27: Convergence diagnostic plot of the posterior estimate samples across the 10 chains for $\hat{\beta}_1$, $\hat{\beta}_2$, and $\hat{\sigma}^2$

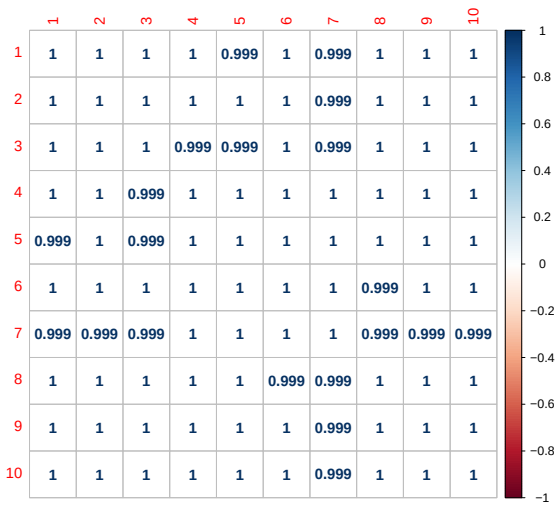


Figure B.28: Plot of correlation across the 10 chains in the logistic model when n = 500; p = 50

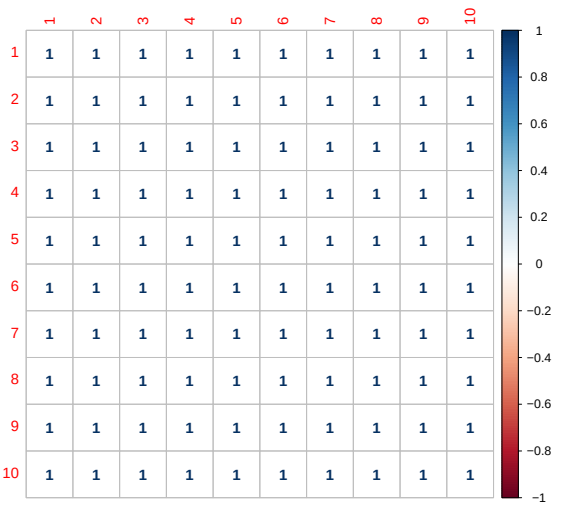


Figure B.29: Plot of correlation across the 10 chains in the linear model when n = 500; p = 50

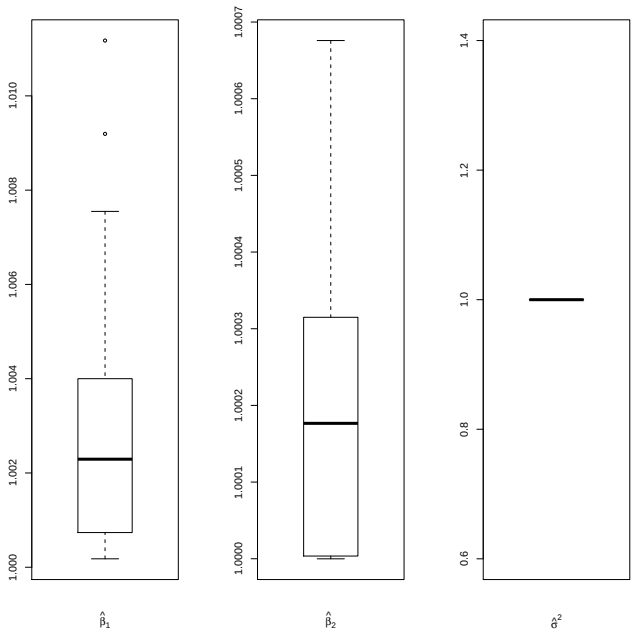


Figure B.30: Convergence diagnostic plot of the posterior estimate samples across the 10 chains for $\hat{\beta}_1$, $\hat{\beta}_2$, and $\hat{\sigma}^2$

B.3 Marginal Posterior Probability (MPP) Plots

Method 2: Combined model

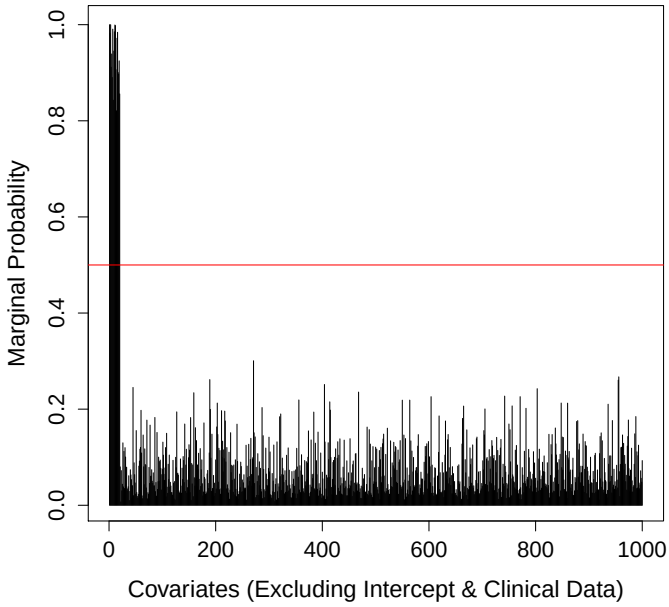


Figure B.31: mpp when n = 300, p = 1000

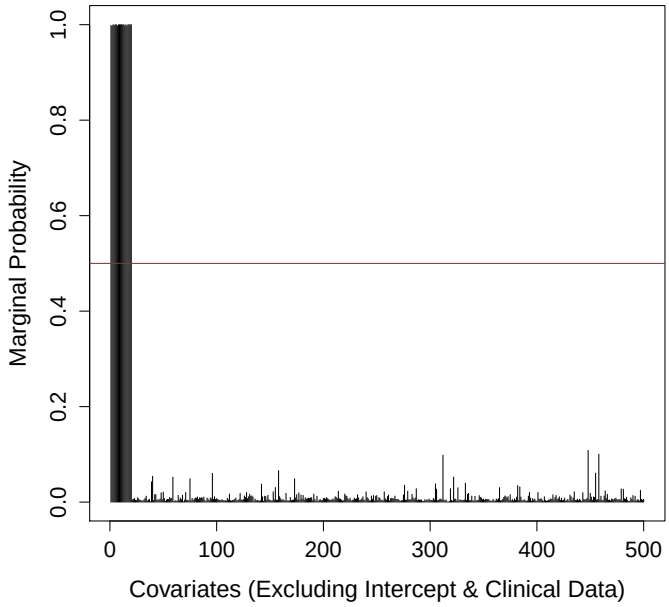


Figure B.32: mpp when n = 300, p = 500

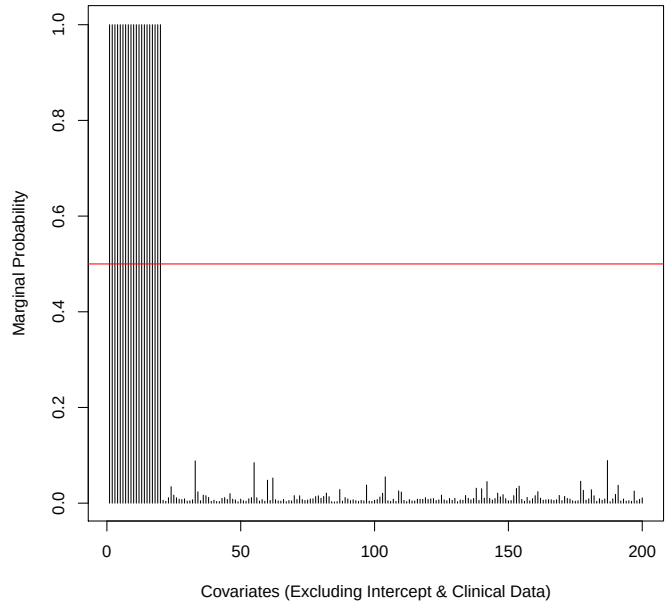


Figure B.33: mpp when n = 300, p = 200

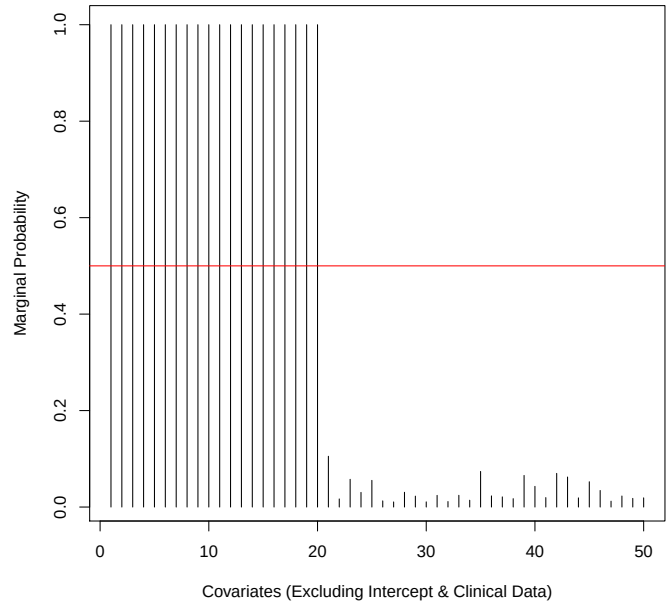


Figure B.34: mpp when n = 300, p = 50

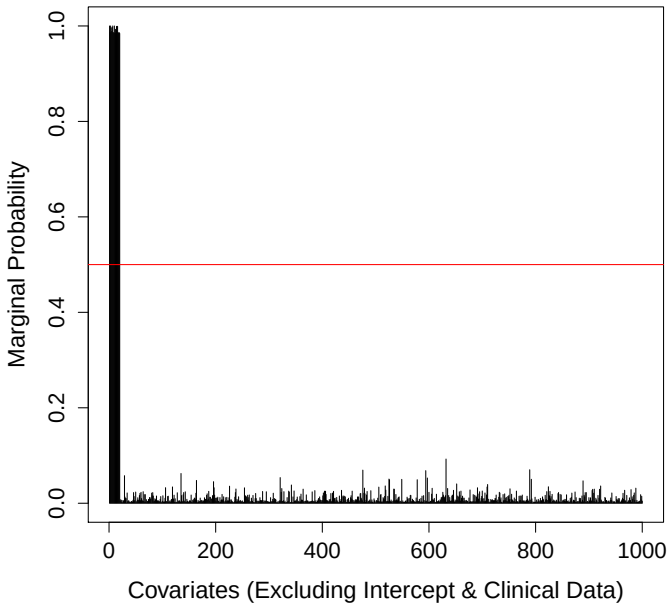


Figure B.35: mpp when n = 500, p = 1000

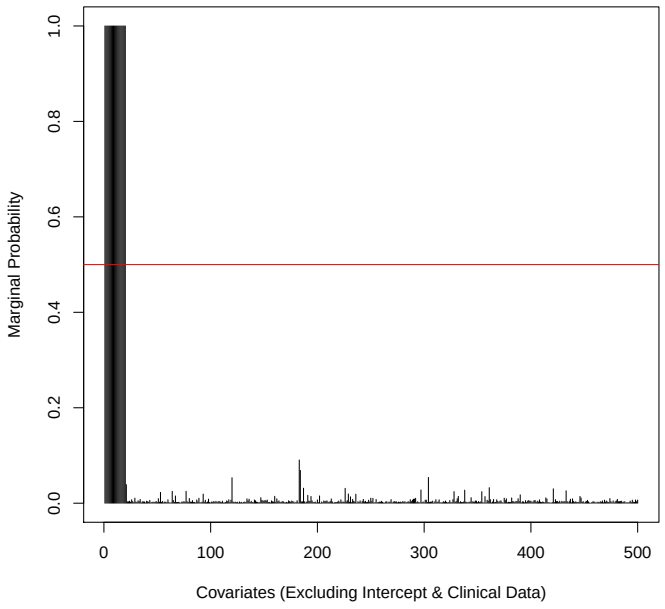


Figure B.36: mpp when n = 500, p = 500

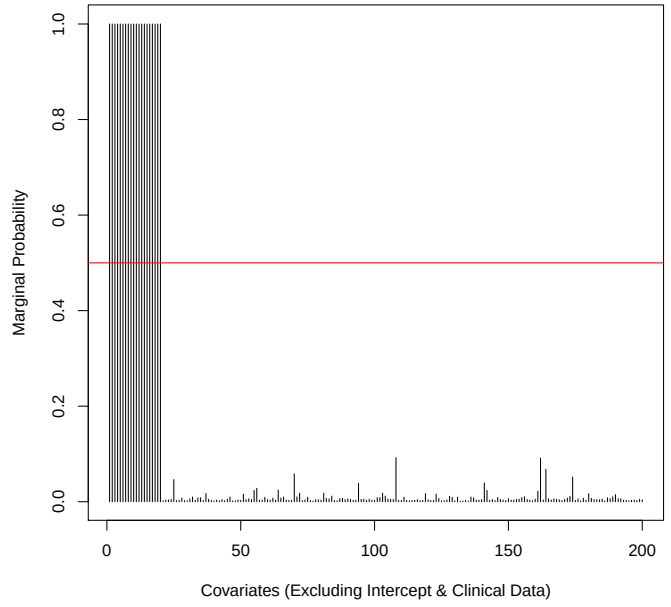


Figure B.37: mpp when n = 500, p = 200

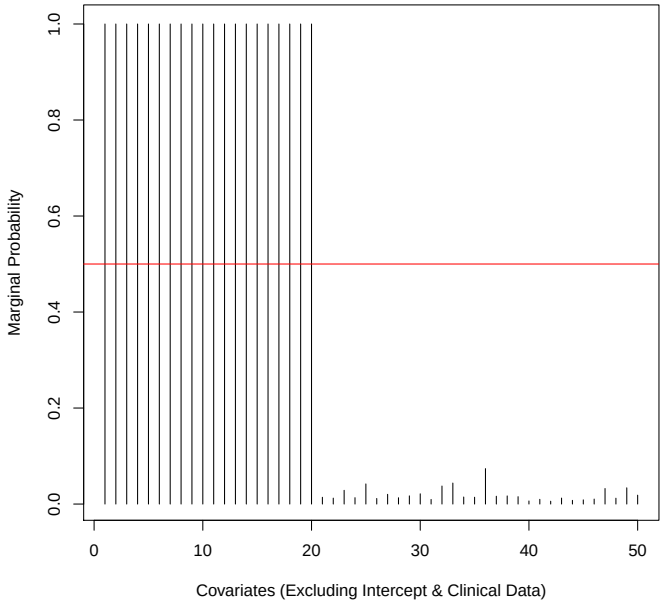


Figure B.38: mpp when n = 500, p = 50

Method 1

For Logistic model

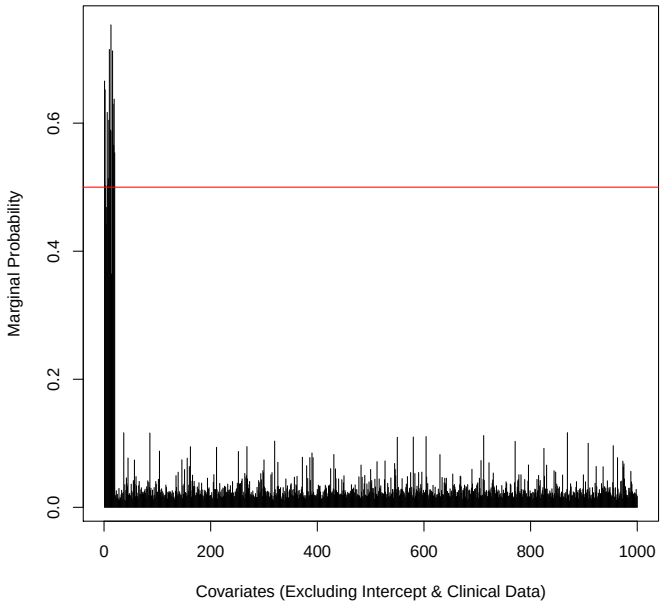


Figure B.39: mpp when n = 300, p = 1000

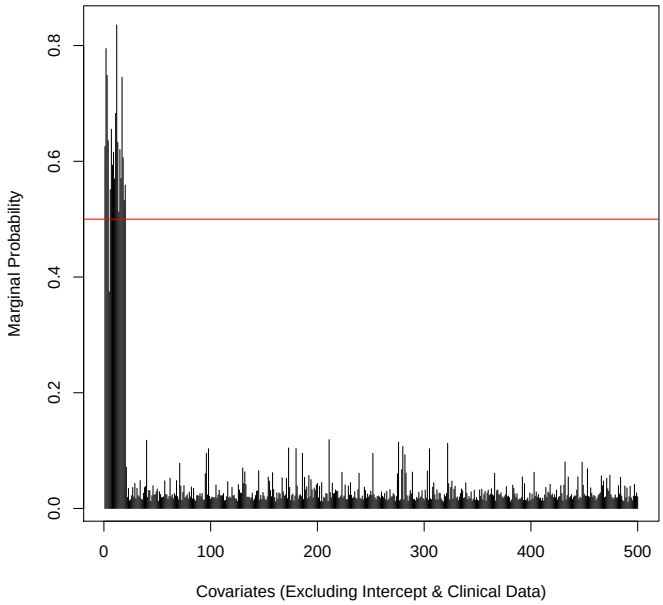


Figure B.40: mpp when n = 300, p = 500

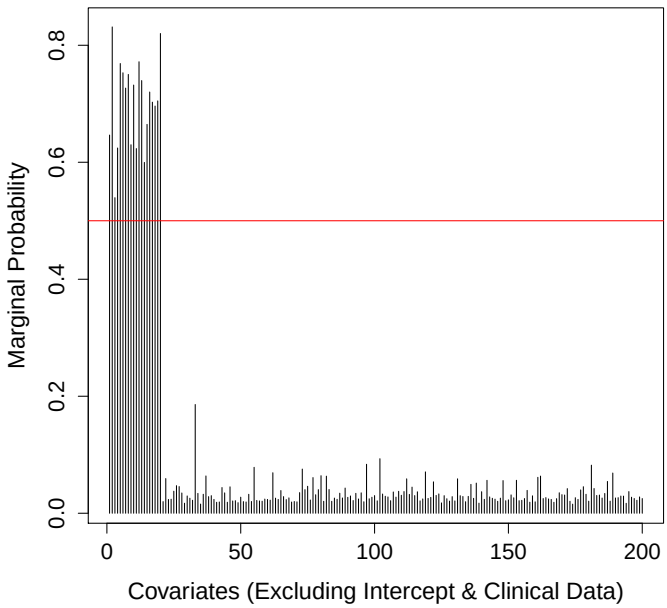


Figure B.41: mpp when n = 300, p = 200

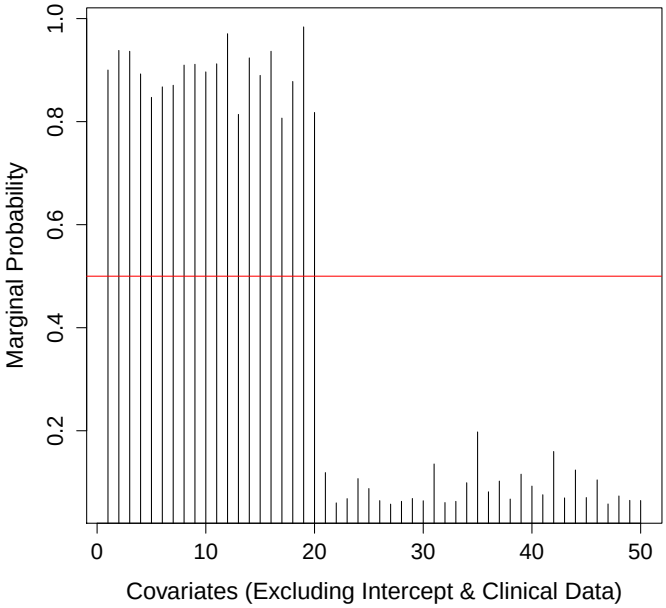


Figure B.42: mpp when n = 300, p = 50

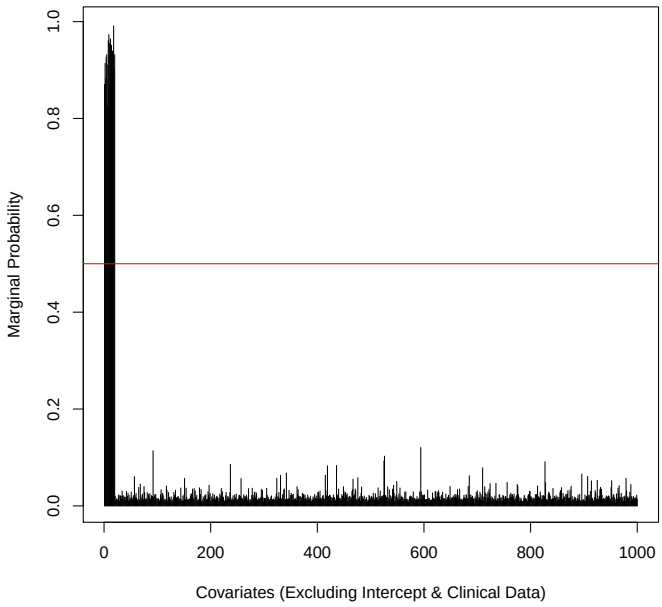


Figure B.43: mpp when n = 500, p = 1000

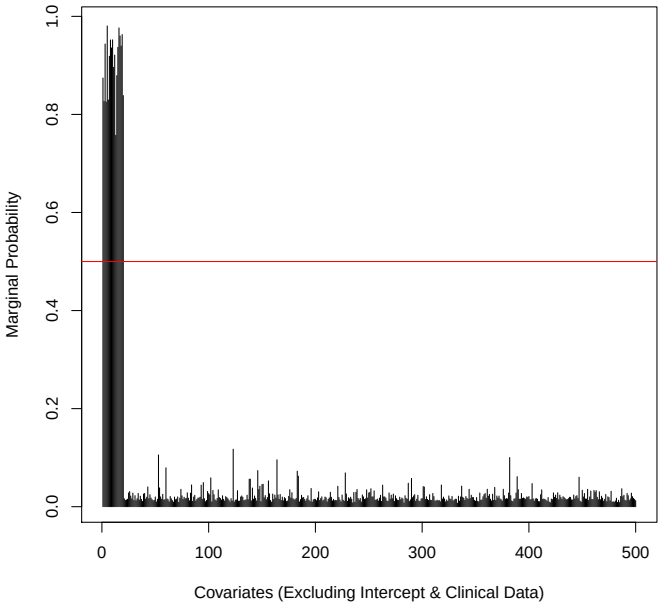


Figure B.44: mpp when n = 500, p = 500

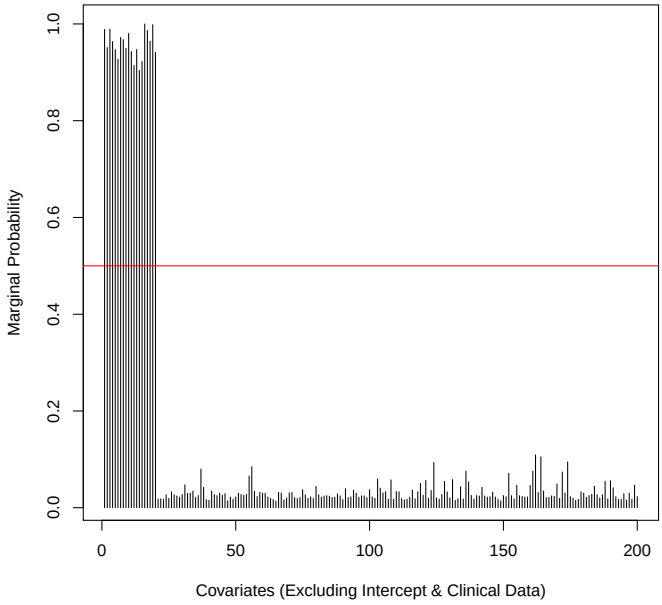


Figure B.45: mpp when n = 500, p = 200

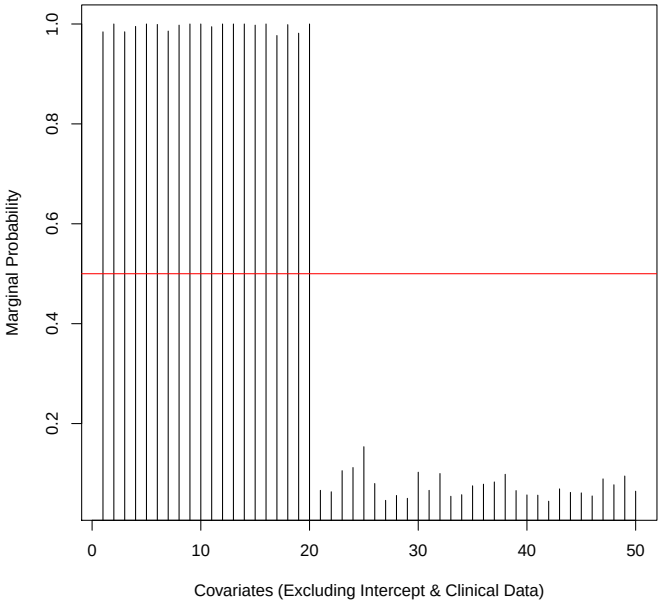


Figure B.46: mpp when n = 500, p = 50

For Linear model

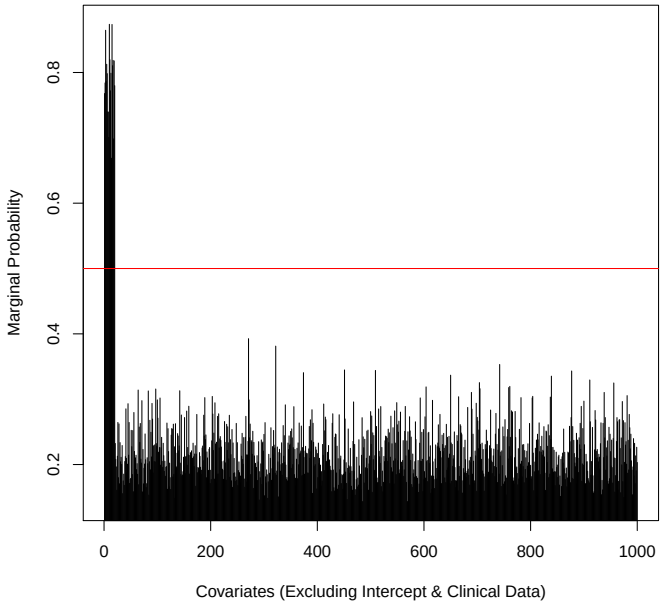


Figure B.47: mpp when n = 300, p = 1000

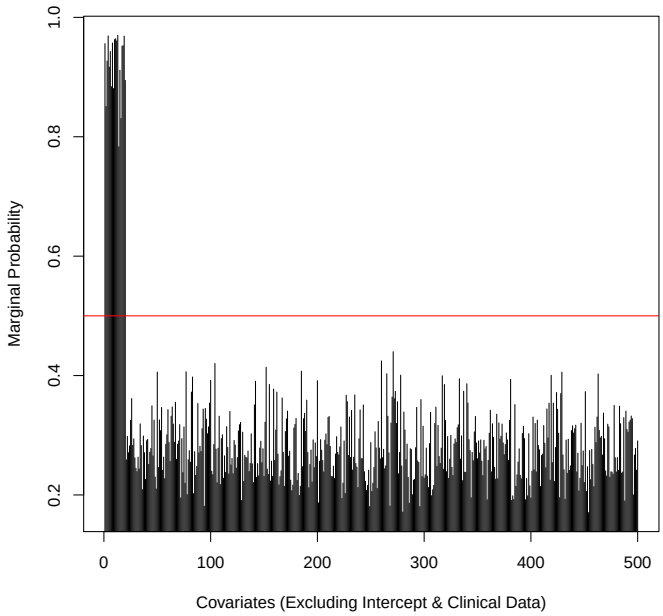


Figure B.48: mpp when n = 300, p = 500

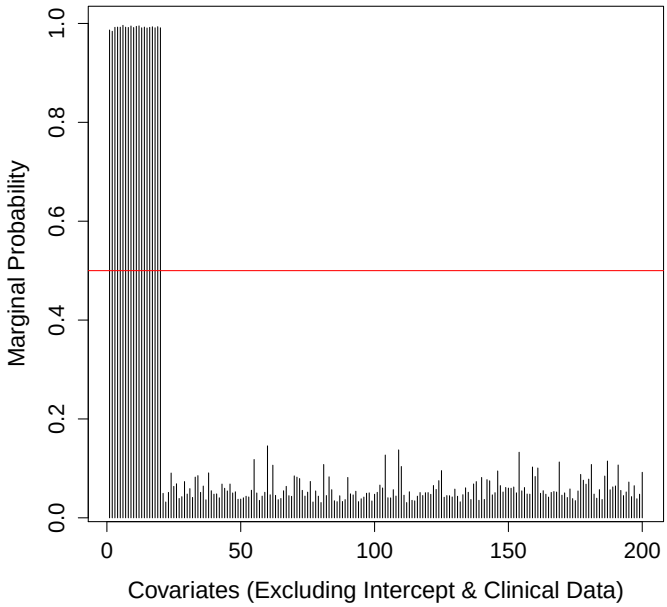


Figure B.49: mpp when n = 300, p = 200

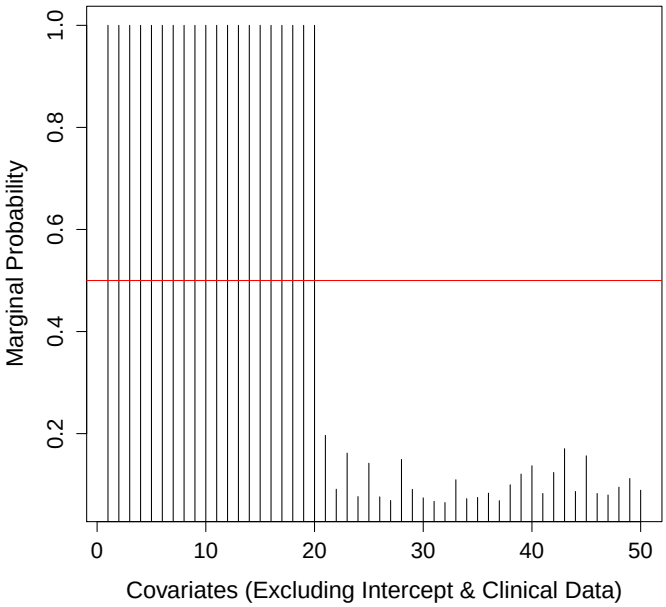


Figure B.50: mpp when n = 300, p = 50

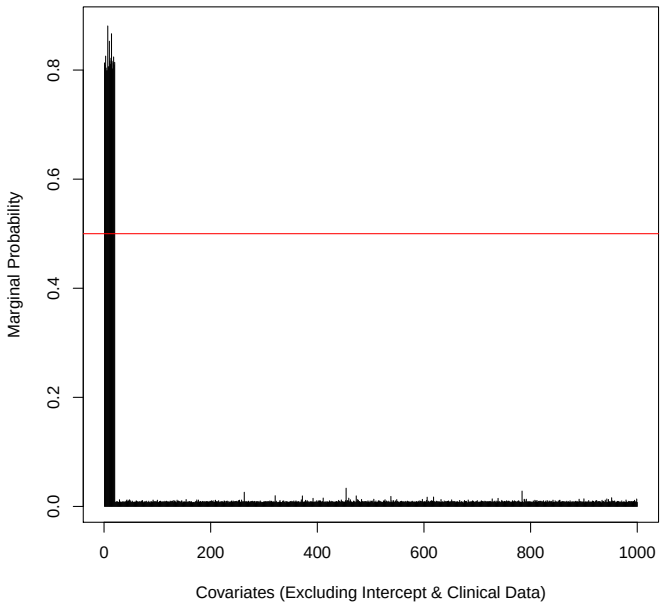


Figure B.51: mpp when n = 500, p = 1000

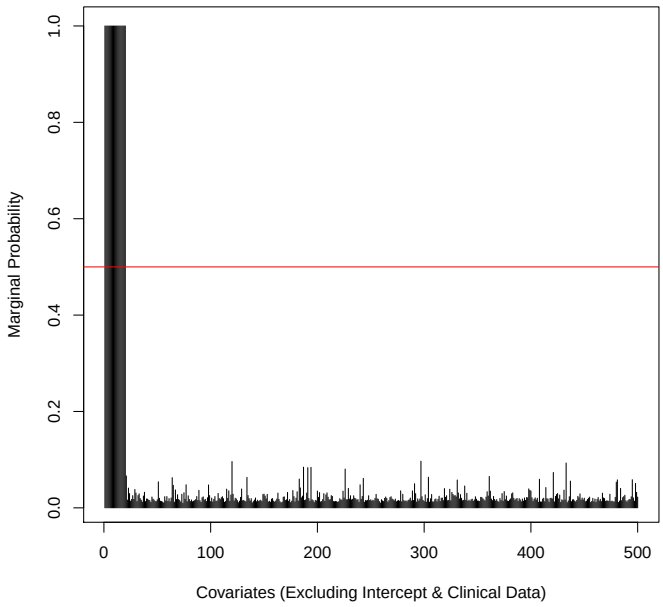


Figure B.52: mpp when n = 500, p = 500

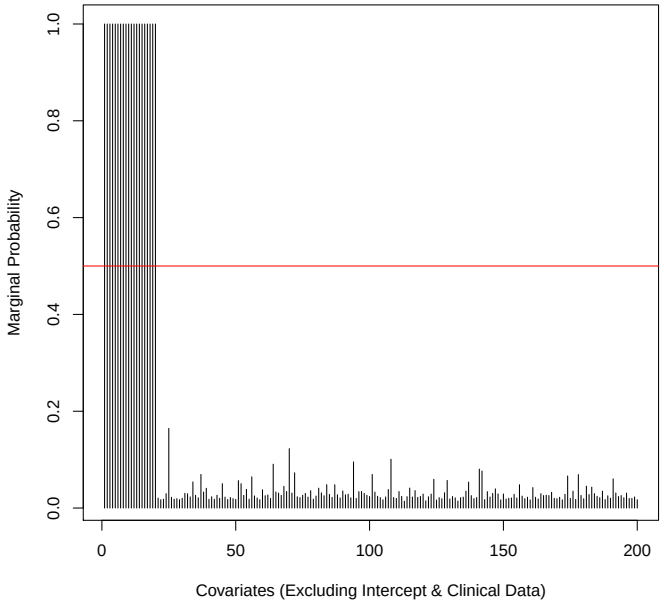


Figure B.53: mpp when n = 500, p = 200

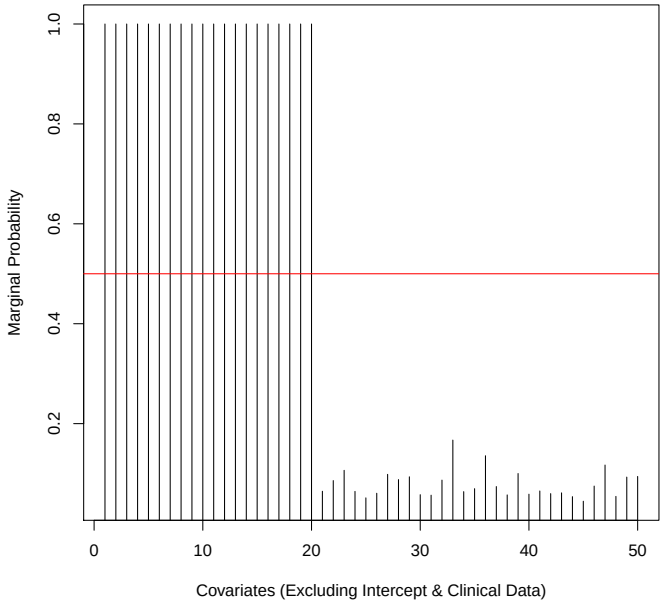
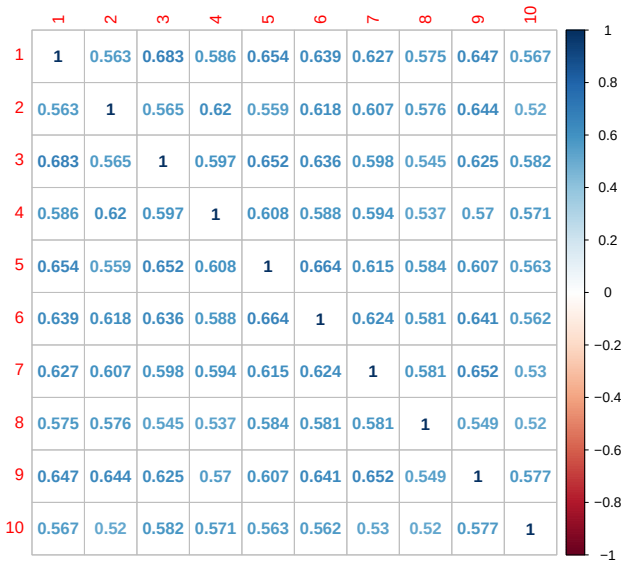


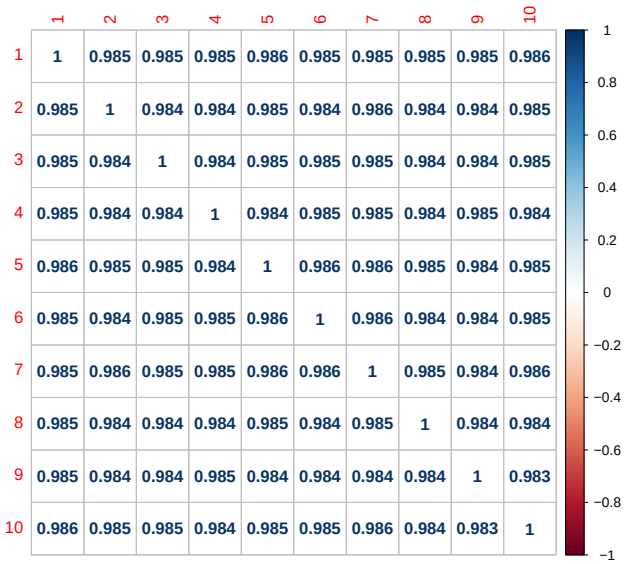
Figure B.54: mpp when n = 500, p = 50

B.4 Application of Method 1 to Coronary Artery Disease (CAD) Data



	1	2	3	4	5	6	7	8	9	10
1	1	0.563	0.683	0.586	0.654	0.639	0.627	0.575	0.647	0.567
2	0.563	1	0.565	0.62	0.559	0.618	0.607	0.576	0.644	0.52
3	0.683	0.565	1	0.597	0.652	0.636	0.598	0.545	0.625	0.582
4	0.586	0.62	0.597	1	0.608	0.588	0.594	0.537	0.57	0.571
5	0.654	0.559	0.652	0.608	1	0.664	0.615	0.584	0.607	0.563
6	0.639	0.618	0.636	0.588	0.664	1	0.624	0.581	0.641	0.562
7	0.627	0.607	0.598	0.594	0.615	0.624	1	0.581	0.652	0.53
8	0.575	0.576	0.545	0.537	0.584	0.581	0.581	1	0.549	0.52
9	0.647	0.644	0.625	0.57	0.607	0.641	0.652	0.549	1	0.577
10	0.567	0.52	0.582	0.571	0.563	0.562	0.53	0.52	0.577	1

Figure B.55: Plot of correlation across the 10 chains in the logistic model using CAD data.



	1	2	3	4	5	6	7	8	9	10
1	1	0.985	0.985	0.985	0.986	0.985	0.985	0.985	0.985	0.986
2	0.985	1	0.984	0.984	0.985	0.984	0.986	0.984	0.984	0.985
3	0.985	0.984	1	0.984	0.985	0.985	0.985	0.984	0.984	0.985
4	0.985	0.984	0.984	1	0.984	0.985	0.985	0.984	0.985	0.984
5	0.986	0.985	0.985	0.984	1	0.986	0.986	0.985	0.984	0.985
6	0.985	0.984	0.985	0.985	0.986	1	0.986	0.984	0.984	0.985
7	0.985	0.986	0.985	0.985	0.986	0.986	1	0.985	0.984	0.986
8	0.985	0.984	0.984	0.984	0.985	0.984	0.985	1	0.984	0.984
9	0.985	0.984	0.984	0.985	0.984	0.984	0.984	0.984	1	0.983
10	0.986	0.985	0.985	0.984	0.985	0.985	0.986	0.984	0.983	1

Figure B.56: Plot of correlation across the 10 chains in the linear model using CAD data.

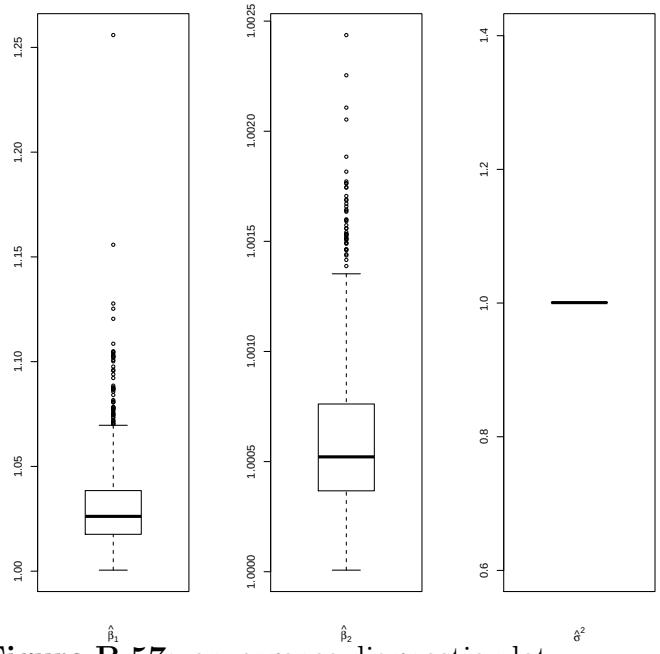


Figure B.57: onvergence diagnostic plot of the posterior estimate samples across the 10 chains for $\hat{\beta}_1$, $\hat{\beta}_2$, and, $\hat{\sigma}^2$ using CAD data.

Method 1: MPP Plot using Coronary Artery Disease (CAD) Data

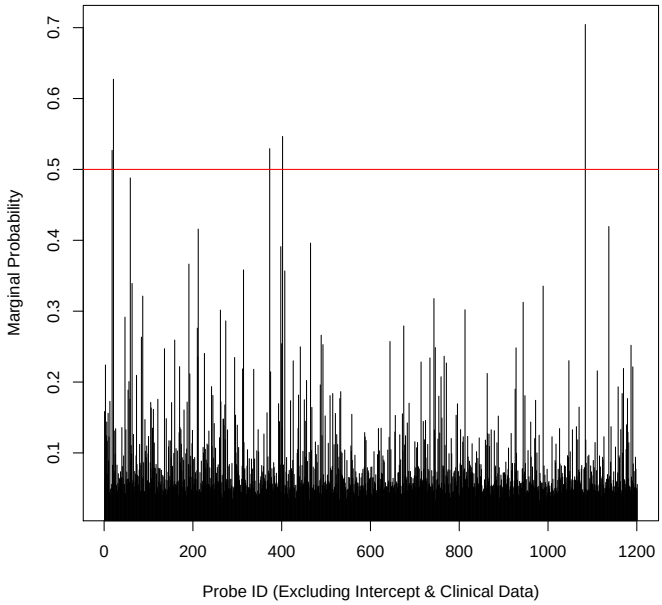


Figure B.58: The marginal posterior probability for the logistic model using CAD data.

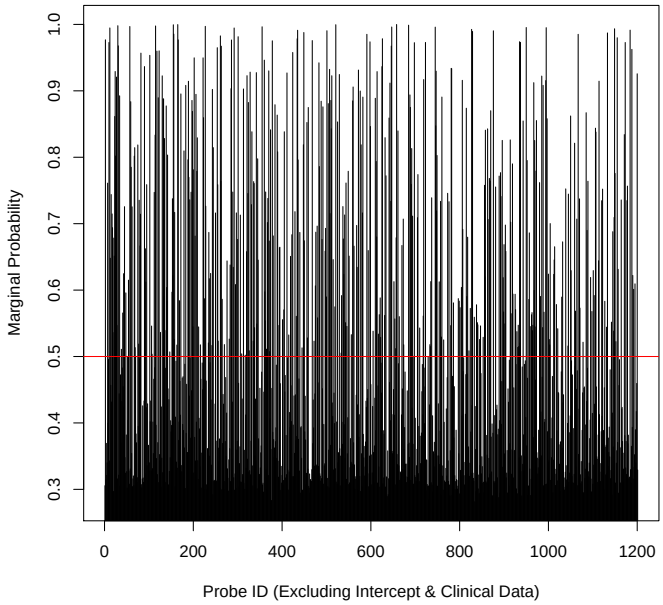


Figure B.59: The marginal posterior probability for the linear model using CAD data.

Table B.5: coronary artery disease data: top 20 genes associated with coronary artery disease using method 1 (selecting important features independently from the logistic and linear model, $j = p_c + 2, \cdots, p$, $\ell = 1, 2$).

Probe ID	**Gene Symbol**	$P(z_j^{(1)} = 1 \mid \boldsymbol{X_1}, \mathbf{y})$	$P(z_j^{(2)} = 1 \mid \boldsymbol{X_2}, \mathbf{y})$
ILMN_1749792	SORBS1	0.05	1.00
ILMN_1792540	AR	0.06	1.00
ILMN_1893915		0.05	1.00
ILMN_1669577	PRNPIP	0.05	1.00
ILMN_1838166		0.05	1.00
ILMN_2328986	SREBF1	0.11	1.00
ILMN_1664464	PTGDS	0.05	1.00
ILMN_1752843	GRM4	0.08	1.00
ILMN_1886595		0.08	1.00
ILMN_1693981	SH3TC2	0.05	1.00
ILMN_1723674	NKAIN3	0.06	1.00
ILMN_1819503		0.05	1.00
ILMN_1680042	LOC649452	0.05	0.99
ILMN_1657888	SALL1	0.04	0.99
ILMN_1813280	LOC642222	0.05	0.99
ILMN_1800131	LOC652826	0.07	0.99
ILMN_2066088	C1orf64	0.06	0.99
ILMN_1761210	LOC155100	0.05	0.99
ILMN_1714176	LOC283487	0.09	0.99

Table B.6: Clinical covariates: CATHerization GENetics (CATHEGEN) Measurement to Understand the Reclassification of Disease of Cabarrus and Kannapolis (MURDOCK) genome-wide association studies (GWAS)

Characteristics	**Data Description**	**Code**	**Type**
AGE	Age at Cath	years	Continuous
BMI	body mass index		Continuous
HYPERTENSION	hypertension by history.	1 = Y, 0 = N	Binary
CADINDEX	Index of severity of coronary artery disease at enrolment. CAD Index is set to missing: 1) When all vessel components OR numdzv is missing 2) When numdzv is not missing any of the components are present	. = missing 0 = No CAD >= 50% 19 = 1 vessel disease 50-74% 23 = > 1 vessel 50-74% OR 1 vessel >= 75% 32 = 1 vessel severe (>= 95%) 37= 2 diseases vessels 42 = 2 severe diseased (both >= 95%) 48 = 1 diseased vessel w/ proxLAD> 94% OR 2 diseased vessel w/ severe LAD 56 = 2 diseased vessels w/ severe prox LAD (>= 95%) OR 3 diseased vessels 63 = 3 diseased vessels w/ >= 1 vessel severe (>= 95%) 67 = 3 diseased vessels w/ prox LAD 74 = 3 diseased vessels w/ severe (>= 95%) prox LAD 82 = Left main disease >= 75% 100 = Left main severe (>= 95%) disease	Semicontinuous (zeros and positive values)
DIABETES	previous diagnosis of diabetes	1 = Y, 0 = N	Binary
HYPERCHOLESTEROLEMIA	previous diagnosis and/or treatment	1 = Y, 0 = N	Binary
NUMDZV	Nos. of vessels with significant occlusion (> 75%)		
RACE	race defined by population stratification	1 = Caucasian, 2 = African American, 5 = Other	Categorical (Nominal)
SEX	gender	1 = F, 0 = M	Binary
SMOKING	The patient must be smoking at least half pack of cigarettes per day to have a significant smoking history	1 = Y, 0 = N	Binary
DEATH	patient died	1 = Y, 0 = N	Binary
DDEATH	Days from a catheterization to death if dead		
DSMI	Days from cardiac catheterization to subsequent myocardial infarction		
HXMI	Previous history of myocardial infarction	1 = Y, 0 = N	Binary

Table B.7: Patients Characteristics by coronary artery disease (CAD) index data. Where 0 stands for No CAD, and 1 for positive responses of CAD.

		CAD Index		
Variable	**Overall**, N = 1,869[1]	**0**, N = 522[1]	**1**, N = 1,347[1]	**p-value**[2]
AGE	62 (54, 71)	56 (48, 65)	64 (56, 72)	<0.001
BMI	29 (25, 33)	29 (26, 36)	29 (25, 33)	0.001
HYPERTENSION	1,269 (68%)	317 (61%)	952 (71%)	<0.001
DIABETES	589 (32%)	128 (25%)	461 (34%)	<0.001
HYPERCHOLESTEROLEMIA	1,121 (60%)	234 (45%)	887 (66%)	<0.001
RACE				<0.001
1	1,332 (71%)	323 (62%)	1,009 (75%)	
2	407 (22%)	168 (32%)	239 (18%)	
5	130 (7.0%)	31 (5.9%)	99 (7.3%)	
SEX	714 (38%)	283 (54%)	431 (32%)	<0.001
SMOKING	889 (48%)	197 (38%)	692 (51%)	<0.001
DEATH	445 (24%)	83 (16%)	362 (27%)	<0.001
DDEATH	2,259 (1,808, 2,623)	2,263 (1,789, 2,643)	2,253 (1,823, 2,610)	0.30
HXMI	517 (28%)	52 (10.0%)	465 (35%)	<0.001

[1] Median (IQR) or Frequency (%)

[2] Wilcoxon rank sum test; Pearson's Chi-squared test