High-dimensional Bayesian Variable Selection with applications to Genome-wide association studies

Kannat Na Bangchang

Submitted in fulfilment of the

requirements for the Degree of Master of Science in Statistics

School of Mathematics and Statistics

College of Science and Engineering

University of Glasgow

June 2023

Abstract

Genome Wide Association studies (GWAS) are a type of experiment that aim to detect genetic variation that may be linked to a type of disease. In variable selection, a major challenge arises when the number of covariates is huge compared to the number of observations. Even if proper priors allow this to be done via Bayesian methods, with an extremely high number of covariates (i.e. many thousands or even millions) compared to the number of observations (i.e. a few hundreds), there are 2 major problems: huge computational time burdens for analysing each dataset, another is the sparsity in the number of covariates associated to the response. If data splitting is used for variable selection in the case above, this can lead to significant reduction in computational time.

GWAS typically contain many thousands of covariates (i.e. DNA variants), which makes variable selection an exceptionally computationally intensive process. Additionally, with large datasets, the MCMC sampler often becomes inefficient in terms of CPU time and shows a lack of MCMC convergence. We investigated if splitting the whole dataset into a number of small sub-datasets before running Bayesian Variable Selection (BVS) reduces the time for the MCMC sampler, improving the mixing of the Markov chain. But simultaneously, we need to investigate the impact of data splitting in terms of the properties and accuracy of the resulting model. When the data is split across columns (i.e. subsetting variables), a number of the sub-datasets may not contain the covariates associated to the response.

Hence, the covariates that are selected in each sub-dataset via using Bayesian variable selection should be finally combined to determine the final set of associated covariates. But this procedure could lead to possible biases, so we assessed how this affects the error in estimation of regression coefficients and other parameters.

Finally, we applied this technique with the real dataset that is about GWAS of heart disease from Prof.Sandosh Padmanabhan's lab at Cardiovascular Sciences at Glasgow.

Acknowledgement

Firstly, I would like to express my deepest gratitude to my supervisor Prof Dr. Mayetri Gupta for giving me the opportunity for this graduate study, and providing invaluable guidance throughout this study. Their vision, motivation, patience, and enormous knowledge have extremely inspired me. This accomplishment would not have been possible without her. I gratefully acknowledge the funding received towards my graduate study from Ministry of Higher Education, Science, Research and Innovation, Royal Thai Government. I am also thankful to Thammasat University for allowing me this greatest opportunity in life. I very much appreciate to all my teachers for all level of my education. In particular, Prof Jirawan Jitthavech and Assoc Prof Vichit Lorchirachoonkul for their guidance and support, not only in academic perspective but also living aspect. I would also like to extend my sincere thanks to my family for providing me with unfailing support and continuous encouragement throughout my years of study and through the process of researching and writing this thesis. I would not be in the position I am today without them.

Contents

Chapter 1

Introduction

1.1 Introduction

Genome Wide Association studies (GWAS) are a type of experiment that aim to detect genetic variation that may be linked to a type of disease. The main aims of GWAS are to try to determine genetic risk factors for a disease, and make predictions about who may be at risk of developing a particular disease (Bush and Moore, 2012).

One of the biggest challenges in GWAS is the extremely large potential set of variants (in millions) but a limited sample size (typically thousands) (Uffelmann et al., 2021). There are many areas for application such as estimating heritability, calculating genetic correlations, making clinical risk predictions and informing drug development. The details of GWAS and the relevant biological background are explained in Chapter 2. The general purpose of GWAS is to identify genomic sequence differences among the persons that differ phenotypically (Uffelmann et al., 2021). A genotype is the genetic makeup of an individual, a phenotype is a feature (of interest) that may be a result of the physical expression of genes.The most common sequence variations in the human genome are single-nucleotide polymorphisms (SNPs). Single nucleotide polymorphisms (SNPs) are defined as loci with alleles that differ at a single base, with the rarer allele having a frequency of at least 1 % in a random set of individuals in a population. (Keats and Sherman, 2013). The aim of GWAS is to detect SNPs that have a statistically significant association with the trait of interest. These SNPs are called genomic risk loci.

In GWAS, we evaluate the association between each genotyped marker and a phenotype of interest across a large number of individuals (Korte and Farlow, 2013). This approach was introduced about twenty years ago in human genetics (Hirschhorn and Daly, 2005) with more than 4500 published human GWAS to date (Ruth, 2020). GWAS have been applied in a range of animals and plants including mice, crops and cattle (Olsen et al., 2011).

Some traits are determined by a small number of loci with large effect sizes, which denote a simple genetic architecture. Genetic architecture describes the characteristics of genetic variation. It depends on the number of genetic variants affecting a trait, their frequencies in the population (Timpson et al., 2017). GWAS can be used in this situation. However, GWAS may be difficult for detecting complex genetic architecture (Kortre and Farlow, 2013). There are two important cases of complex genetic architecture : the first is when a trait is controlled by many rare variants and another is many common variants affecting a single phenotype.

The experimental procedure of a GWAS contains many steps, starting with the collection of DNA and phenotype information from the individuals. The information should contain disease status and demographic data. Then, genotyping of each individual is provided via using GWAS arrays, quality control is done, the imputation of untyped variants using haplotype phasing is conducted, and the statistical test for association is constructed. Moreover, sometimes a meta - analysis is conducted to combine a number of analyses to increase power. Since there is a chance to have possible biases and errors in each step, planning is important when the GWAS is set up. These steps are discussed in more detail in Chapter 2.

The performance of GWAS in identifying a true association between a SNP and trait, depend on the explanation of the phenotype variance due to the population structure. The phenotypic variance is determined by the level of two allele variants differing in their phenotype effect (the effect size) and their frequency in the sample. Both rare variants and small effect sizes may occur in GWAS (Asimit and Zeggini, 2010) . One solution to deal with a rare-variant architecture is increasing the sample size. Li et al. (2010) pointed out that increasing the sample size will improve the power to recover meaningful associations. However, increasing the sample size may not resolve all situations. Hence, one approach is to collapse several SNPs in a region into a single variable and then use this to analyse as a composite genotype (Lee et al., 2014).

One challenging aspect of the statistical analysis of GWAS is the large of number of SNPs (millions) with a comparatively small number (e.g. hundreds or thousands) of samples. A SNP is a representation of a base in the DNA sequence. The Human Genome Project SNP fact sheet (2009) states that SNPs occur in at least one percent of the population. Fitting regression models in classical statistics need the sample size to be larger than the number of variables. Hence, these methods do not directly work for GWAS studies.

Moreover, another challenge for GWAS in humans is a requirement for data on many thousand individuals to be available for detecting a large number of small effect loci (Manolio et al., 2009). There are special classes of traits for human diseases given by numerous small - effect mutations. On the other hand, loci with a medium effect size have been shown to underlie traits such as eye-colour and skin colour (Sulem et al., 2007).

Over the last decade, there are many genomic risk loci that have been found to be associated with diseases such as FTO for obesity (Lan et al., 2020) , PTPN22 for autoimmune diseases (Siminovitch, 2004) and IL - 12/IL - 23 for Crohn's diseases (Kashani and Schwartz, 2019). Moreover, GWAS may be used for supporting clinical trials for drugs targeting the relevant traits. The trait-associated genetic variants can be used as control variables in epidemiology studies to avoid confounding genetic group differences (Benjamin et al., 2012) . Further, a recent study pointed out that genomic risk prediction using genome-wide polygenic risk scores (PRSs) for coronary artery disease, atrial fibrillation, type 2 diabetes, inflammatory bowel disease and breast cancer can identify disease risk based on rare, highly penetrant mutations (Khera et al., 2018).

GWAS has many advantages as a methodology since it is a powerful tool for analysis of simple traits under additive genetic scenarios. An additive genetic model is usually employed in case-control-based GWAS (Liu et al., 2021). However, GWAS may miss uncovering the causative loci since linkage disequilibrium, preventing us from discerning SNPs. One solution is to determine the phenotype of interest by giving a score on a trait more proximal to the genetics (Benjamin et al., 2012). This technique reduces the number of loci that contribute to the trait. Thus, it leads

to increasing the power to detect them. One limitation of using GWAS is when a single causative locus has high heritability, parts of the genome are inherited together from the maternal and paternal genome, since an association there is a natural result of the linkage structure of the data, leading to the most significant SNP not being the true causative locus.

Moreover, the SNPs are often correlated because of linkage disequilibrium (Robinson, 1998). Linkage disequilibrium is a term specifying the relationship between alleles at two loci on a haplotype. There are many genes on a chromosome, these genes are jointly inherited and the coupled inheritance of genes is broken by a phenomenon called crossing over (described in Chapter 2). The correlation structure between SNPs in the genes is not considered when using a univariate regression model. Therefore, multiple regression needs to be implemented for dealing with this situation. Techniques for handling large data sets with many thousands of variables exist, such as variable subset selection in regression models, regularisation methods and Bayesian approach. These will be discussed in more detail in Chapter 3.

1.2 Motivation

There are still many challenges in the statistical modeling, and detection of associated SNPs from GWAS data. The first problem is multicollinearity, since many SNPs (covariates) are highly correlated. One possible way to deal with this is as follows. Once a variable is selected , another variable that is correlated with that variable is removed from the model. Furthermore, doing this in practice requires deciding from a data-driven manner which covariates get removed and which stay. However, in this scenario, the covariate eliminated may be the important covariate, meaning that critical information is removed from the data.

The second problem is that of large p (covariates) and relatively small n (observations) i.e., $p >> n$. Here, there is no unique solution for parameter estimation in the classical statistical framework. It means that there is also a problem for the variable subset selection, since many subsets of variables are suitable for the data. One way to deal with this situation is to find the best solution from those equally good solutions. The stochastic search algorithm could be considered via using the Bayesian framework for variable selection. This approach is also more able than classical variable selection to move between the local modes in the model space that indicate the good solutions.

Another problem is the computational time needed for the exploration of the model space using a stochastic search. Due to the large number of covariates, the estimation of parameters also consumes a huge amount of time for computation.

1.3 Objective of the study

The aim of this study is to develop, analyse and compare Bayesian variable selection (BVS) methods to assess which approach is most suitable with large p , small n data to apply to Genome Wide Association Studies (GWAS). It also investigates if techniques of data splitting and estimation based on subsets is appropriate for GWAS applications

1.4 Discussion and Comparison with some relevant articles

We now discuss two articles that are related tp split and merge Bayesian variable selection: split and merge Bayesian variable selection approach for ultrahigh dimensional regression (Song et al., 2015) and efficient genomic prediction based on whole-genome sequence data using split and merge Bayesian variable selection (Calus et al., 2016)

Although there are some points of similarity between these approaches, there are also some important contrasts. In terms of the models considered in each article, Calus et al. used logistic regression, Song et al. used the linear regression model, while my thesis contains applications to both linear regression and logistic regression models. The criteria for split in the first stage in Calus' paper is sorting based on MAF, however Song used a maximum on the marginal inclusion probability to select those covariates and for our work we used the same idea as Song' s article. The last point relates to criteria to compare the performance of the final model. Calus used cross validation on the real data set, the Mean Squared Error (MSE) was used in Songt et al., while my thesis used the length of the credible interval of the posterior mean of the regression coefficients of the associated covariates.

1.5 The novelty of our work

We propose a method for splitting the large dimensional dataset into sub-datasets for both penalised variable selection and Bayesian variable selection methods to increase computational efficiency and improve the chance of detecting associated SNPs. This step of our method also appears to improve MCMC mixing and efficiency compared to the full (unsplit data) model. For justifying this approach, we also mathematically derive results regarding the error of estimation (i.e. Expectation of SSE) in each case.

Finally, combining the selected covariates in each sub-dataset increases the chance of selecting the truly associated covariates to the response.

1.6 Structure of the thesis

The thesis starts with an overview of the molecular biology and clinical background of genome wide association studies in Chapter 2. We present a literature review of existing variable selection methods which are explained in terms of the methodology and underlying theory, described in Chapter 3. We introduce a novel splitting of datasets method, applied in the linear regression model and logistic regression model contexts, which is described in Chapter 4. In Chapter

5, we present simulation studies, describing how to simulate correlated Binomial data, the software packages used in this study, for both the linear regression models and logistic regression models. An example application of these techniques follows in Chapter 6, where the methods are applied on hyprertension GWAS dataset from the Glasgow blood pressure clinic and the results are presented. In Chapter 7, extensions of these methods for application on other large datasets is described.

Chapter 2

Molecular biology background

2.1 Introduction

Finding genetic causes underlying disease is one of the key areas of medical research, made possible by the recent advances in genotyping technology. Statistics plays a vital role in the search for genetic variants linked to diseases and statistical genetics is an area of much development. There are many specific definitions in genetics relevant to the discussion in the following chapters. This chapter will introduce basic genetic terminology .

2.2 Basic Genetic Terminology

The human genome contains of 46 chromosomes which are pairs of autosomal chromosomes (chromosomes 1 to 22) and the sex chromosomes (X and Y). Females have 2 X chromosomes and males have a combination of X and Y chromosomes. All genetic information is contained in these 46 chromosomes.

In 1928 Griffith did experiments to show how genetic information is stored (Bayrhuber et al., 1989). Moreover, Avery et al. (1944) discovered that genetic information is stored in the form of Deoxyribonucleic Acid (DNA) . Watson and Crick (1953) developed a model of DNA structure which presents it as a double helix with a sugar, phosphate backbone and nitrogenous bases on the inside. There are four bases: Adenine (A), Cytosine (C), Guanine (G) and Thymine (T). It was concluded that the sequence of the nitrogenous bases determines genetic information.

Particular sections of DNA that determine the code for proteins are called genes. Those genes are arranged in linear form on a chromosome and they are separated by non coding areas of DNA. Their position on the chromosome is called a locus. According to the Human Genome Project (2009), there are between 20,000 and 25,000 genes in the human genome. Moreover, genes can vary in size depending on the functional proteins they code for.

Proteins are built out of amino acids and the order of amino acids is defined by the order of bases in the DNA strand. The DNA is located in the cell nucleus , the information needs to be translated into proteins which are the building blocks of cells. This is done in the cells by replication of the DNA code in the form of Ribonucleic Acid (RNA) (Bayrhuber et al., 1989). The translation of the genetic code to building occurs proteins in parts of the cells called ribosomes

which are located in the cytoplasm.

Different forms of the same gene are called alleles. They initially arise if there is a change of base within a gene, creating a new base type in the population that did not exist before. Allele frequency refers to the frequency of alleles in terms of the proportion in the population.

A change in genetic material is called a mutation. Mutation can occur on different levels, on the gene level (e.g. changes of base pairs), on the chromosome level (e.g. changes of the chromosome length) and on the genome level (e.g. changes in the number of chromosomes). Mutation most commonly occurs on the gene level, when a base is changed to another base. Mutation rates at a single base are between one in ten thousand and one in one billion meioses (Bayrhuber et al., 1989). Meiosis is the process of cell division in sexually reproducing organisms, reducing the number of chromosomes in the reproductive cells called gametes (sperm in males or ovule in females) from diploid (chromosomes are arranged in pairs) to haploid (one set of chromosomes is present). There are many different causes for mutation such as exposure to the environment, through as x - ray radiation, radiation or chemical substances, or can occur spontaneously within the cell during biological processes.

Humans have two copies of the autosomal chromosomes. Sometimes the two genes on the two copies of the chromosome are identical (the allele is the same) and it is called homozygous at that locus. If the two copies are different, it is called heterozygous at that locus. The occurence of a set of alleles on a single strand of a chromosome is called the haplotype. The physical expression of genes resulting in a particular feature of the individual is called the phenotype and the genetic information that leads to the physical expression is called the genotype. An allele that determines the phenotype is called dominant when the phenotype occurs over two different alleles at a locus. An allele not leading to a physical expression is called recessive.

A mutation in a base in the DNA sequence may result in a Single Nucleotide Polymorphism (SNP). The Human Genome Project SNP fact sheet (2009) gives an overview of SNPs. Such alterations have to occur in at least one percent of the population to be classified as a SNP, meaning that the allele frequency of the rarest allele must be at least one percent. SNPs are estimated to occur every 100 to 300 bases which leads to an estimate of about 10 to 30 million SNPs in the human genome. The average mutation rate of SNP is very low about 2×10^{-8} per locus (Palmer and Cardon, 2005). SNPs are approximately equally spread throughout the whole genome.

A locus that influences a disease is called a disease susceptibility locus and a locus of known location that is used in the analysis of genetic data is referred to as a marker locus.

2.3 Hardy-Weinberg Equilibrium

Hardy (1908) and Weinberg (1908) showed that the allele frequencies at a locus in two consecutive generations will stay the same if these assumptions hold: (1) The population size is infinite, (2) there is no movement of individuals between populations, (3) there is no mutation at the locus, (4) random separation of alleles occur during gamete formation, (5) individuals within the population mate without regard to their genetic makeup and (6) individuals with a certain allele are not favoured.

These assumptions are quite stringent, but constant allele frequencies from one generation to the next, known as Hardy - Weinberg Equilibrium (HWE), is frequently observed to hold ap-

proximately. Hardy and Weinberg developed a model for the expected genotype frequencies as a function of the allele frequencies for a locus. At a di-allelic marker locus with alleles M and m, if HWE holds, the genotype frequencies can be denoted by the following:

$$p_{MM} = p_M^2, p_{Mm} = 2p_M(1 - p_M), p_{mm} = (1 - p_M)^2,$$

where p_{MM} is the frequency of genotype MM and p_M is the frequency of allele M. It should be noted that there are in fact two possible heterozygous genotypes Mm and mM, which are both represented by Mm. It is hard to determine from which haplotype m or M came, and for the analysis of genotypes this information is assumed to be of no particular importance. Hence, they are combined together into one heterozygous group, and with two alleles this gives us three genotypes.

2.4 Linkage Disequilibrium

Thomas Hunt Morgan is one of many scientists who studied the properties of chromosomes (Teare and Barrett, 2005). He discovered that there are many genes on a chromosome in those genes, recombination can occur during the formation of gametes. Genes are arranged in a linear form on a chromosome, and the further apart two genes are, the higher the chance that a crossing over occurs. That is shown in Figure 2.1. The probability of such a recombination occuring between two loci in one generation is expressed through the unit centiMorgan (cM). A genetic distance of one cM represents a probability of one percent that a crossing over between two loci will occur in one generation. The rate of recombination is not the same on each part of each chromosome, but this is a rough guideline. One cM is approximately equivalent to a distance of about one million base pairs (Mb) between two genes (National Human Genome Research Institute, 2009).

Linkage Disequilibrium (LD) is a term specifying the relationship between alleles at two loci on a haplotype. The alleles at loci that are physically closer together on the chromosomes have a higher probability of being inherited together, as the probability of a recombination event occuring between them is usually smaller than if the two loci are further apart. For testing the difference between observed and expected haplotype frequencies, LD for two loci is calculated. At two di-allelic loci A and B, where say A is a disease susceptibility locus (alleles A and a) and B represents a marker locus (alleles B and b), there are four possible haplotypes with a probability of occurrence as follows: $P(AB) = h_{AB}, P(Ab) = h_{Ab}, P(aB) = h_{aB}, P(ab) = h_{ab}$. The expected haplotype frequencies, for haplotype AB for example, can be calculated using the following model

$$h_{AB} = p_A q_B + D,$$

where D is a measure of linkage disequilibrium, and p and q represent allele frequencies at the two loci. Under equilibrium D is equal to zero, as the haplotype frequencies are equal to the product of the corresponding allele frequencies.

Conversely, D can be used as an indirect measure of distance between disease susceptibility and marker locus if constant recombination rates can be assumed on the chromosomes which

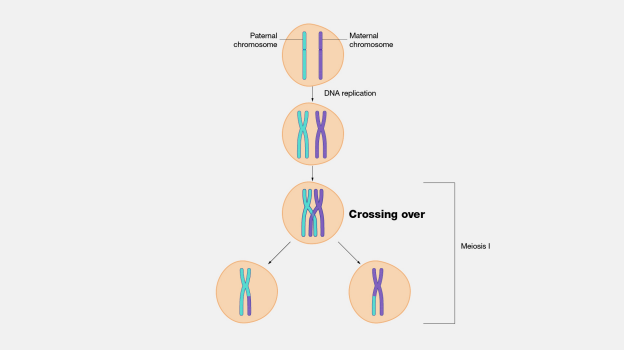


Figure 2.1: Crossing-over (National Human Genome Research Institute, 2023)

enables us to conclude that the higher the value of D, the closer two loci are together on the chromosome. The maximum LD value between two loci based on the allele frequencies at those two loci is given by

$$D \leq min(p_A(1-q_B), (1-p_A)q_B)$$

, with the lower boundary restricted by

$$D \geq max(-p_A q_B, -(1-p_A)(1-q_B))$$

.

2.5 Experimental procedure for GWAS

As mentioned in Chapter 1, genome - wide association studies (GWAS) are used to identify the association of genotypes with phenotypes via testing on the difference in the allele frequency of genetic variants in each individual (Uffelmann et al., 2021). A comprehensive overview of GWAS is presented below.

2.5.1 Study Population

The material for a GWAS experiment consists of the collection of DNA and phenotypic information from each individual within a population-based cohort. Phenotypes can be binary or continuous dependent variables. The phenotypes are tested for association with genotype. A common experimental design for GWAS is a case-control study. Cases are based on the presence of a particular phenotype while controls are based on the absence of that phenotype. Data from resources such as biological databases or cohorts with disease are often used for conducting GWAS. The large size for running a well-powered. GWAS requires significant time, cost and effort for collecting data. The bias from the data is the another point that one should be concerned about since the population should not be extreme in any characteristics. For example, when we use the information from Biobank in UK (Fry et al., 2017), these data may be biased since participants are healthier, wealthier and more educated than general population (Fry et al., 2017).

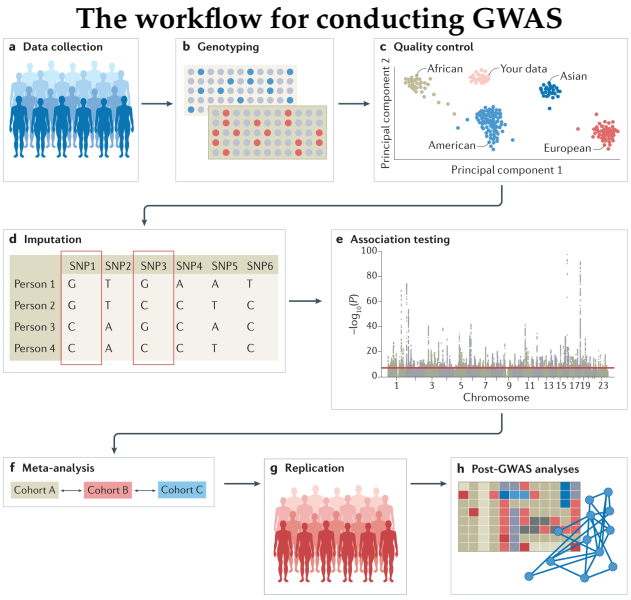


Figure 2.2: Genome - wide association studies (adapted from Uffelmann et al., 2021)

2.5.2 Steps for conducting GWAS

The first step in GWAS is data collection. We can collect data from cohorts being studied or existing ones with genotype and phenotype information in biological databases such as BioBanks. The next step is genotyping, data can be collected via using microarrays to keep common variants or using modern sequencing techniques for whole-genome sequencing (WGS). An important step is quality control: poor quality data on SNPs should be deleted, population stratification in the sample should be inspected and adjusted. The next step is imputation for missing genotypic data. Missing genotypes can be imputed via using information from matched reference populations from data bases such as TopMed (Li et al., 2009). The next step is testing for association with possible genetic variants. The key point is selecting an appropriate model. Often meta-analysis is conducted, where results from multiple cohorts are combined. The last step is post-GWAS analyses for fine mapping of variants. A classical example is silico analysis of GWAS via using information from external resources such as silico fine-mapping (Ufflemann et al., 2021). The set of steps for conducting GWAS are shown in Figure 2.2.

2.5.3 Current methods for analysis

The genotyping for individuals is a core part of the analysis. Microarrays are used for common variants but next - generation techniques are used for rare variants. Microarrays are popular since they are cheap when compared to next - generation sequencing methods. The main purpose of both approach is to detect the expression of thousands of genes simultaneously from a sample (Guo et al., 2020). However, many low-cost technologies for the new methods are becoming available. (Korte and Farlow, 2013)

The general goal of GWAS is identifying novel variant - trait associations. There are about 50,000 unique SNP - trait associations at a genome - wide significance threshold as on January 2019 (MacArthur et al., 2017). GWAS have identified risk loci for a large number of diseases and

traits such as anorexia nervosa, major depressive disorder, cancers and subtypes of cancers, type 2 diabetes, coronary artery disease, schizophrenia, inflammatory bowel disease, insomnia, body mass index (BMI) and eductional attainment (Andrews et al., 2020).

GWAS can lead to discovery of novel biological mechanisms. GWAS have implicated genes of unknown function and experimental follow-up of loci have led to the discovery of novel biological mechanisms underlying disease, for example, the role of autophagy in Crohn's disease (Henderson et al., 2012). Moreover, GWAS can be used to identify individuals at high risk of certain diseases, and thereby improve patient outcomes via early detection, prevention or treatment. For example a coding non-synonymous variant in the CFH gene explains about 50 percent of the population - attribute risk of AMD (Klein et al., 2005).

GWAS is used for multiple applications beyond gene identification. These include Mendelian randomization studies, polygenic risk scores, forensic analyses, determination of cryptic relatedness, paternity testing, clinical diagnostic genetic testing, embryonic DNA fingerprinting, determination of perinatal loss, validation of new analytic methods and quality control of next - generation sequencing data (Chaitankar et al., 2016).

2.5.4 Data analysis

Data processing is another important part of GWAS. There are many types of input data such as individual ID numbers, coded family relations, sex, phenotype information, genotypes for all variants and information on the genotyping batch. Since there are many types of input data, the analysis on the results should be particularly careful with quality control (Purcell et al., 2007). Rare or monomorphic variants should be removed, the variants not in Hardy - Weinberg equilibrium should be filtered out and imputation of missing SNPs should be considered. The matching of the phenotype and genetic data should be done accurately. There are many software tools such as PLINK that are used to analyse genetic data and conduct quality control (Purcell et al., 2007). Sample and variant quality control are performed on GWAS array data, variants that are missing are imputed using a sequenced haplotype reference (Auton et al., 2015). There are many different software tools to deal with quality control steps and imputation (Lam et al., 2020). Since the genetic data sets are very large, parallel runs are often needed. Computing clusters can distribute jobs to many computers for running. Moreover, cases and and controls should be matched by ancestry to avoid confounding - for example certain SNPs are more common in some specific groups compared to the common population. For example, in a GWAS for skin colour where cases and controls are from different regions, the yellow skin in cases would be drawn more often from the East Asian population. Ancestry in GWAS can be dealt with via using principal component analysis (Price et al., 2006): the genotypes of all individuals are used for defining clusters of individuals with similar genotypes.

Testing for association is another step in the analysis for GWAS. Linear or logistic regression models are often used to test for associations, depending on the characteristics of the phenotype. If the phenotype is continuous such as height, blood pressure, body mass index, linear regression models should be used. However, if the phenotype is binary such as the presence or absence of disease, a logistic regression model should be used instead. The dangers of model misspecification

are a negative impact on the overall goal of genome-wide association studies (GWAS) including reducing predictive power and failed to estimate the true magnitude and direction of the effects. Some covariates, for example demographic factors such as age, sex and ancestry should be included for stratification, and any confounding effect should be considered since that may reduce the statistical power for binary traits in certain samples (Pirinen et al., 2012). The individual - specific random effect term in linear or logistic models can increase statistical power for SNP discovery and control stratification at the cost of requiring greater computational resources since those terms account for genetic relationship among individuals (Zhou et al., 2018). Most researchers use a logit link function for binomially distributed case-control phenotypes in logistic regression models. Genotypes of genetic variants that are close together are not independent due to the effect of linkage disequilibrium. Hence dependence between tests should be considered when conducting a GWAS. Controlling false discovery is another important part of analysis in GWAS since testing millions of associations among individual genetic variants and a phenotype need a multiple testing threshold to avoid false positives. The International HapMap Project pointed out that since there are about one million independent common genetic variants across the human genome, the Bonferroni testing threshold should be less than 5×10^{-8}, that represents a false discovery rate of $0.05/10^6$ (Altshuler and Donnelly, 2005). However, the appropriate threshold varies depending on the population. For example, a more stringent threshold is important for populations with larger effective population sizes or if the minor allele frequency thresholds for inclusion in a GWAS are low , since low minor allele frequency variants are usually not in linkage disequilibrium with common variants. Hence adding on a multiple testing adjustment is needed. Moreover, many genetic variants have a small effect, contributing to an overall phenotype of a complex trait such as height, or type 2 diabetes.

As the last part of GWAS, meta-analysis is conducted when data from multiple cohorts are analysed together, with tools such as METAL (Willer et al., 2010) for quality control. The important considerations for Genome-wide association meta-analysis (GWAMMA) (Baselmans et al., 2019) involve using the individual cohorts following the same definitions for the data analysis plan, using harmonized phenotypes and reporting their results in a standard way. Scaling effect sizes to a standard normal distribution should be considered since phenotype measurements and their estimated absolute effect sizes cannot be compared between cohorts. Inspection at a cohort-level is done by at least two independent analysts and any issues should be resolved within the individual cohorts. In addition, meta-analysis can provide overall summary statistics.The last issue is choosing between a fixed effect model or a random effect model. A fixed effect model assumes error variances are equal across cohorts, but a random effect model tests for heterogeneity in the results. The combination of all cohorts leads to a more precise estimation of effect sizes and gives the significance of effects in GWAS via weighting each individual cohort by each sample size or by using the inverse variance technique (Willer et al., 2010).

2.5.5 Limitations of GWAS

Although GWAS is a popular technique for dealing with trait-associated variants, there are still many challenges .

The first is population stratification. There are biases especially when multiple cohorts are

used. However, this problem can also occur in relatively homogeneous populations for example, studies have uncovered population stratification and related bias in the UK Biobank mostly about 450K of the 500K individuals are white British (Abdellaoui et al., 2019). The existing methods for dealing with the effects of stratification are based on common variants. These methods contain principal component analysis, and the case of linear mixed models. However, they are insufficient if many rare variants are included in the study. The effect is even more complicated if these are demographic changes (Lawson et al.,2020). Future work needs betterapproaches for correction for population structure in GWAS and associated analyses.

The next issue is polygenicity. An extreme situation can occur when thousands of variants each have a small effect on trait and uncover underlying biological mechanisms (Watanabe et al., 2019). Rare variants of large effect may not be reported for all traits and thousands of variants are not linked to rare variants. Hence, novel techniques should be developed that account for polygenicity. High polygenicity means that individuals with the same disease may have unique genetic profiles linked with the same disease. Genetic differences are often linked to treatment sensitivity. Hence, the development of novel treatments should address this issue.

Another issue is ethical. Ethical issues related to GWAS include the use of samples and data, storage and reuse of samples and data, privacy and sharing data with individual participants. Researchers and bioethicists have pointed out that finding permissions for sample and data storage and unspecified future use is necessary. (Novembre et al., 2008). Another ethical challenge of GWAS is related to the diversity and inclusion of participants. The results from GWAS should promote health and well-being for all humans that are different by race, gender and geographical location. It implies that samples and data used for GWAS need to be representative of the global human population and the genomics workforce also need to be diverse to ensure awareness of this.

2.5.6 Gaps addressed in this dissertation

GWAS has two main challenges points in term of data analysis. The first is large dimensionality since there are many thousands of covariates (SNPs). Another point is the high correlation between SNPs, depending on genomic location. Classical variable selection methods cannot deal with either situation.

In this thesis, we investigate the use of novel Bayesian variable selection methods that can be applied to GWAS. Further we develop a novel approach to dealing with high - dimensionality in the SNP set through two - stage data - splitting. We split dataset into sub-datasets in the first stage and then combine these overall covariates that are selected in each sub-dataset (using BVS) to analyse fully via using the Bayesian Variable Selection at the second stage. These methods are described in the next two chapters.

Chapter 3

Variable Selection methods

3.1 Introduction

The standard linear regression model can be written as $\mathbf{Y} = X\boldsymbol{\beta} + \epsilon$ where $\mathbf{Y}$ is a $n \times 1$ vector of observations, $\boldsymbol{X}$ is an $n \times p$ matrix of covariates, $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown regression coefficients and σ^2 is an unknown positive scalar. The error terms $\epsilon = (\epsilon_1, \ldots, \epsilon_n)$ are assumed to be distributed independently and normally, i.e. $\epsilon \sim N(0, \sigma^2 I)$. The goal is to estimate the regression coefficient $\boldsymbol{\beta}$. In a classical statistical framework, the sample size n is larger than the number of explanatory variables p. The regression coefficient $\boldsymbol{\beta}$ can be estimated using the ordinary least squares (OLS) estimator $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'y$.

However, if there are more explanatory variables than observations, i.e. $p >> n$, the rank of $\boldsymbol{X}'\boldsymbol{X}$ is smaller than p in the data sets with $p >> n$. In this case, $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ does not exist and the OLS estimator is not unique. In this case, alternative approaches are necessary to estimate $\boldsymbol{\beta}$.

Penalised regression methods are one set of approaches to do these. Penalised regression allows to create a linear regression model that is penalised, for having too many variables in the model, by adding a constraint in the equation (Gareth et al., 2014). This is also known as shrinkage or regularisation methods. The consequence of imposing this penalty, is to reduce the coefficient values towards zero. This allows the less contributive variables to have a coefficient close to zero or equal zero.The first technique is ridge regression proposed by Hoerl and Kennard (1970), which gives a biased estimator $(\boldsymbol{X}'\boldsymbol{X} + \lambda I)^{-1}\boldsymbol{X}'y$ where I is the identity matrix and λ is a penalty parameter on the log-likelihood. Other penalised likelihood-based methods include LASSO regression introduced by Tibshirani (1996) with a penalty on the norm of the regression coefficient $\sum_{i=1}^{p} |\boldsymbol{\beta}_i|$ and the elastic net proposed by Zou and Hastie (2005) which combines ridge regression and LASSO regression.

By variable selection, namely, selecting a subset of the p variables of size p' such that $p' < n$ and the matrix inverse exists. It is a challenge to construct the model from small subsets of all variables or to choose the covariates that are associated with the response, due to the small number of covariates that are likely to be associated to the response in GWAS applications. Traditional variable selection methods like forward, backward and stepwise selection cannot be used in this situation. Due to the multicollinearity of $\boldsymbol{X}$ (when $p >> n$), it leads to the objective function not being unimodal. Hence, there are many different models that would equally fit the data set. The methods mentioned above are discussed in more detail in the following sections.

3.2 Variable Selection using penalised regression

Penalised regression can deal with high dimensional data. The general principle of this method is to estimate the regression coefficients $\boldsymbol{\beta}$ that yield the minimum value of $(Y - X\boldsymbol{\beta})'(Y - X\boldsymbol{\beta})$ under a pre-defined constraint called the penalty function. A downside to this approach that it gives biased estimates, unlike OLS.

The penalty function can be written as $p(||\boldsymbol{\beta}||)$ where $||\boldsymbol{\beta}||$ is the norm of $\boldsymbol{\beta}$. The estimate of regression coefficients ($\hat{\boldsymbol{\beta}}$) is given by

$$\hat{\boldsymbol{\beta}} = \arg\min(Y - X\boldsymbol{\beta})'(Y - X\boldsymbol{\beta})$$

under $p(||\boldsymbol{\beta}||) < t$ where t that is a constant that is to be determined. Moreover, $\hat{\boldsymbol{\beta}}$ can be rewritten as

$$\hat{\boldsymbol{\beta}} = \arg\min(Y - X\boldsymbol{\beta})'(Y - X\boldsymbol{\beta}) + \lambda p(||\boldsymbol{\beta}||)$$

where λ is a Lagrange multiplier. Different penalty functions yield different parameter estimates for penalised regression.

For the estimation of regression coefficients, the penalised maximum likelihood estimation involves adding a penalty term of log-likelihood function before maximisation. The penalised log likelihood is given by

$$\begin{aligned} M(y; \boldsymbol{\beta}, \sigma^2) &= -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}(Y - X\boldsymbol{\beta})'(Y - X\boldsymbol{\beta}) - \lambda p(||\boldsymbol{\beta}||) \\ &= -\left[\frac{n}{2}\log 2\pi + \frac{n}{2}\log\sigma^2 + \frac{1}{2\sigma^2}(Y - X\boldsymbol{\beta})'(Y - X\boldsymbol{\beta}) + \lambda p(||\boldsymbol{\beta}||)\right]. \end{aligned}$$

The value of $\boldsymbol{\beta}$ that maximises $M(y; \boldsymbol{\beta}, \sigma^2)$ yields the penalised regression biased estimator of $\boldsymbol{\beta}$. If we set

$$p(||\boldsymbol{\beta}||) = \sum_{j=1}^{p} {\boldsymbol{\beta}_j}^2,$$

it results in the ridge regression estimator. Another penalty function is

$$p(||\boldsymbol{\beta}||) = \sum_{j=1}^{p} |\boldsymbol{\beta}_j|,$$

used for LASSO, described in Section 3.2.1.

Ridge regression was introduced by Hoerl and Kennard (1970). The sum of squares of the residual can be written in the quadratic function as

$$\epsilon'\epsilon = (Y - X\boldsymbol{\beta})'(Y - X\boldsymbol{\beta}).$$

The objective of parameter estimation in regression is minimization of the residual. Under the

constraint

$$\boldsymbol{\beta}'\boldsymbol{\beta} < t,$$

constrained optimization can be done via using the Lagrange multiplier (λ), giving the resulting estimator as

$$\hat{\boldsymbol{\beta}} = (X'X - \lambda I_p)^{-1} X'y.$$

The ridge regresssion estimator can be written as

$$\hat{\boldsymbol{\beta}} = (X'X + kI_p)^{-1} X'y$$

where $k = -\lambda$ is a constant ($k > 0$). This is a biased estimator since the inverse of $X'X$ is replaced by an approximation that guarantees the existence of a solution.

Other penalised methods include LASSO regression (Tibshirani, 1996) and the elastic net (Zou and Hastie, 2005). However, these two methods do not yield closed form solutions as in ridge regression. Quadratic programming is a way to solve those solutions via varying the tuning parameter (λ). In the next section, we describe the LASSO method.

3.2.1 LASSO, Ridge regression and LARS

Least Absolute Shrinkage and Selection Operator (LASSO) is a statistical technique to deal with variable selection with high-dimensional data proposed by Tibshirani (1996). We discuss two issues for the parameter estimation in LASSO regression: first, the accuracy in the prediction and second, the interpretation of the regression coefficients.

Subset selection is a discrete process. Covariates are considered by adding them into the model one at a time (forward selection), or in backward selection, removing them from the model one at a time. LASSO is a continuous process, which reduces the effect size of regression coefficient ($\boldsymbol{\beta}$) by the nature of the constraint, if the constant t is small.

LASSO fits a penalised regression model, minimizing the cross-validation error of the log likelihood via reduction in the value of some regression coefficients and adjusting other regression coefficients to be zero. Parameter estimation from LASSO is similar to ridge regression, but the penalised log-likelihood is given by

$$l_n(\boldsymbol{\beta}) - \sum_{j=1}^{p} p(|\boldsymbol{\beta}_j|),$$

where $l_n(\boldsymbol{\beta})$ is the log likelihood given n observations and $p(|\boldsymbol{\beta}_j|)$ is the penalty function with parameter λ.
The estimator of the regression coefficients in LASSO can be derived by minimizing

$$(Y - X\boldsymbol{\beta})'(Y - X\boldsymbol{\beta})$$

under the constraint

$$\sum_{j=1}^{p} |\boldsymbol{\beta}_j| \leqslant t,$$

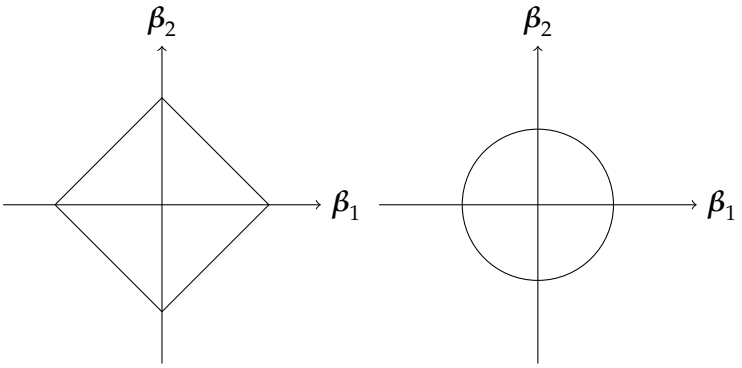


Figure 3.1: Estimation picture for the lasso (left) and ridge regression (right)

where t is a tuning parameter such that $t \geqslant 0$. A tuning parameter controls the size of shrinkage on the estimator $(\hat{\beta})$. There is no closed mathematical form for the LASSO estimator. Tibshirani (1996) proposed quadratic programming for finding the LASSO estimator. With the nature of the constraint t, if t is sufficiently small then some of the coefficients are equal to zero. Thus, the LASSO is a continuous subset selection. In term of the size in variable subset selection, t should be chosen small for the minimisation of the estimate of expected prediction error.

Next we compare ridge regression and the LASSO. With an orthonormal design matrix X, each technique can be applied in a simple way as a transformation on the least square estimate $\hat{\beta}_j$. Ridge regression and LASSO use soft-thresholding. The ridge uses a proportional shrinkage, but LASSO transforms each coefficient via a constant λ that is truncated at zero. However, in the nonorthogonal case, it is not simple to explain in words. Figure 3.1 attempts to explain this graphically. The simple case contains only two parameters (β_1 and β_2). The constraint region for ridge regression is $\beta_1^2 + \beta_2^2 \leq t$, whereas the constraint region for LASSO is $|\beta_1| + |\beta_2| \leq t$. Both methods find the first point where the elliptical contours touch the constraint region. Since the constrained region for the LASSO has corners, if the solution is at a corner then it has only one parameter (β_j) that is set to zero. Moreover, if there are more than two parameters, it implies that the constrained region has many corners. Hence, there is a high chance that many of the estimated coefficients are set to zero.

Another way to see the contraint term $|\beta_j|^q$ is through a log-prior density for β_j and this gives an equivalent of the prior distribution on the parameters in a Bayesian regression setting. If $q = 1$, it leads to the LASSO, while for $q = 2$, then it gives ridge regression. Hence, the conclusion is that ridge regression and the LASSO can be considered Bayes estimates with different priors, since the estimators of ridge regression and LASSO can be written in a general form as

$$\hat{\beta} = \arg\min(Y - X\beta)'(Y - X\beta) + \lambda \sum_{j=1}^{p} |\beta_j|^q.$$

However, there are many opportunities to choose other values for q except 1 or 2. One way is to select $q \in \{1,2\}$, which is a compromise between the LASSO and ridge regression. This was proposed by Zou and Hastie (2005) as the elastic-net penalty. The elastic net tends to select more variables than the LASSO, since the additional term distributes the weight to more variables. The

penalty function of the elastic net is given by

$$p(|\boldsymbol{\beta}_j|) = (1-\lambda)\boldsymbol{\beta}_j^2 + \lambda|\boldsymbol{\beta}_j|, \quad j = 1, \ldots, p.$$

However, the elastic net results in over-shrinkage when compared to the LASSO (Zou and Hastie, 2005). From these penalty terms, the estimated effect of most variables will be shrunk to zero, effectively excluding them from the set of relevant covariates. Under the LASSO, there is a restriction on the maximum number of variables which can be selected, which depends on the sample size n and number of variables p, i.e. $\min(n-1, p)$ (Zou and Hastie, 2005). This does not apply to the elastic net.

Efron (2004) introduced the LARS (Least Angle Regression) that is adapted from the Forward selection method. LARS is connected to the LASSO, since LARS provides the algorithm for finding the entire LASSO path. The first step is finding the covariate that is highly correlated with the response and finds the prediction via estimation of the regression coefficient using OLS, that reduces the correlation between the covariate and the residual. The procedure is repeated until all covariates are included into the model.
The algorithm for LARS consists of the following steps:

1. Standardize the covariates to have zero mean and unit norm and set $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, ..., \boldsymbol{\beta}_p = 0$.
2. Find the X_j most correlated with $\mathbf{y}$.
3. Adjust $\hat{\boldsymbol{\beta}}_j$ by increasing it from zero in the direction of its correlation with $\mathbf{y}$, stop when there is another covariate X_k that is more highly correlated than X_j with the residual $\mathbf{r} = \mathbf{y} - X\hat{\boldsymbol{\beta}}$. Calculate the residual $\mathbf{r}$ where $\mathbf{r} = \mathbf{y} - \bar{\mathbf{y}}$ and set $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, ..., \boldsymbol{\beta}_p = \mathbf{0}$.
4. Select the covariates X_j and X_k into the model and adjust the regression coefficients $\boldsymbol{\beta}_j$ and $\boldsymbol{\beta}_k$, starting at zero, until there is another covariate X_l that is as much correlated with the residual $\mathbf{r}$ when there are X_j and X_k in the model.
5. Repeat the previous process until all covariates are in the model. The estimation of regression coefficients of all are completed after $\min(n-1, p)$ steps.

From both methods above, the key part is to find the optimum value of the tuning parameter λ. One approach is to use prediction error to guide this choice. One criterion considered is $\lambda.min$, when a value of λ is chosen that gives the minimum mean cross-validated error. Another is $\lambda.1se$ that gives the largest value of λ, when the error is within 1 standard error of the minimum.

3.3 Bayesian Variable Selection

In high dimensional data sets from GWAS, typically, many covariates are not significantly associated with a given trait. Moreover, those covariates are highly correlated, leading to a multicollinearity problem. Hence the model is sparse since the coefficient of most covariates are likely yo be zero. The classical frequentist or likelihood-based variable selection via any criterion such as BIC and AIC or stepwise subset selection become infeasible when the number of variables become large (Miller, 2002). An alternative solution is Bayesian variable selection (BVS) (Cui et al.,

2010). This approach provides intuitive probabilistic interpretations and explores the model space efficiently in a stochastic way to find the model with high posterior probabilities. This approach is called stochastic search variable selection (SSVS).

There are many stochastic searching schemes have been developed such as the Gibbs variable selection, Geweke's BVS with block updates (Geweke, 1994), and the reverse jump MCMC algorithm (Green, 1995). Moreover, the application of BVS in the setting of $n << p$ has appeared in analyzing genetic data from the early 2000s. Most methods use hierarchical Bayesian modelling to combine the empirical variance with a local background variance associated with neighboring genes (Baldi and Long, 2001). BVS has been applied to GWAS data that contains millions of genetic variants or SNPs (Wakefield, 2008). We discuss BVS in more detrail in the following sections.

3.3.1 Bayesian Variable Selection in linear regression models

We start by introducing the Bayesian linear model.

Model

The linear regression model is described in the following set of equations first with non-conjugate prior distributions for $\boldsymbol{\beta}$ and σ^2.

$$y = X\boldsymbol{\beta} + \epsilon, where$$

$$\epsilon \sim N(0, \sigma^2 I_n),$$

$$\sigma^2 \sim IG(\nu/2, \nu\lambda/2), and$$

$$\boldsymbol{\beta}|\sigma^2 \sim N(\mathbf{b}, \sigma^2\mathcal{V}).$$

The prior distribution $p(\boldsymbol{\beta}|\sigma^2)$ is assumed to be a normal distribution with mean vector $\mathbf{b}$ and variance matrix $\sigma^2\mathcal{V}$. The prior of σ^2 is assumed as an Inverse Gamma distribution, with parameters ν and λ .

For Bayesian variable selection, a latent variable $\gamma_j \in 0, 1$ is introduced for each predictor X_j, where $\gamma_j = 1$ denotes that the variable X_j is included in the model and $\gamma_j = 0$ means that the variable X_j is excluded. For a prior on $\boldsymbol{\beta}_j$, George and McCulloch (1993) introduce a mixture of two normal distributions with mean zero. The first part has a small variance τ_i (favouring values of zero for $\boldsymbol{\beta}_j$), but the second part has a large variance $c_i^2\tau_i$, allowing large non-zero values, leading to

$$\boldsymbol{\beta}_j|\gamma_j \sim (1-\gamma_j)N(0, \tau_j^2) + \gamma_j N(0, c_j^2\tau_j^2), \quad c_j^2 > 1.$$

Specifying values of c_j^2 to be large makes the prior less informative. The same value $c_j^2 = c^2$ is typically used (for $j = 1, \ldots, p$), from the range $(10, 1000)$ (Smith and Kohn 1996). From the normal mixture prior and the use of the inverse gamma prior for σ^2, the linear regression model for Bayesian variable selection can now be written as:

$$y \sim N(X\boldsymbol{\beta}, \sigma^2 I_n),$$

$$\boldsymbol{\beta} \sim N(\mathbf{b}_\gamma = 0, \mathcal{V}_\gamma = D_\gamma R D_\gamma),$$

$$\sigma^2 \sim IG(\nu/2, \nu\lambda_\gamma/2), \quad \text{and}$$

$$\gamma \sim p(\gamma),$$

where $\mathbf{b}_\gamma$ is the mean vector that corresponds to the indicator variable γ, $\mathcal{V}_\gamma$ is the variance matrix that corresponds to the set of indicator variables γ, R is the prior correlation matrix of $\boldsymbol{\beta}$ and $D_\gamma = \text{diag}(a_{\gamma 1}\tau_1, \ldots, a_{\gamma p}\tau_p)$ and $a_{\gamma j} = 1$ if $\gamma_j = 0$, $a_{\gamma i} = c_i$ if $\gamma_j = 1$. Since D_γ determines the scaling of the covariance matrix, the mixture of two normal distributions is simple where $R = I_p$. (A subscript γ indicates that the parameter depends on γ). One example of choosing the prior $p(\gamma)$ is denoted by

$$p(\gamma) = \prod_{i=1}^{p} {\pi_i}^{\gamma_i}(1-\pi_i)^{1-\gamma_i},$$

where $p(\gamma_i = 1) = \pi_i$.

MCMC algorithms for Bayesian Variable Selection

The hierarchical mixture model for variable selection by George and McCulloch (1997) was extended from the non-conjugate form of the Bayesian regression model. The conjugate prior allows the exact calculation for the posterior probabilities of γ. Here, the variance σ^2 is included in the prior distribution

$$\boldsymbol{\beta}|\sigma^2, \gamma \sim N(0, \sigma^2 D_\gamma^* R_\gamma D_\gamma^*)$$

where D_γ^* is a diagonal matrix with elements denoted by $\nu_{0\gamma i}^*$, if $\gamma_i = 0$, and $\nu_{1\gamma i}^*$, if $\gamma_i = 1$. The prior of $\boldsymbol{\beta}_i$ is specified as a mixture of two normal distributions

$$\boldsymbol{\beta}_i|\gamma_i \sim (1-\gamma_i)N(0, \sigma^2\nu_{0\gamma i}^*) + \gamma_i N(0, \sigma^2\nu_{1\gamma i}^*)$$

The Gibbs sampler for variable selection in linear regression has the following steps:

1. For iteration $t = 0$, set a starting point $\gamma^{(0)}$ via using the Gray Code (Press et al., 1992).
2. For iteration $t = 1, \ldots, T$, sample a proposal $\gamma^{(*)}$ conditional on the previous iteration $\gamma^{(t-1)}$. where the posterior distribution for sampling γ is given as $p(\gamma|\boldsymbol{\beta}, \sigma)$ that is below.

Due to the conjugacy of the model, the posterior full conditional distribution of **fl** can be easily derived. At each iteration, sample γ from

$$p(\gamma|\boldsymbol{\beta}, \sigma) = |X'X|^{-1/2}|D_\gamma^* R_\gamma D_\gamma^*|^{-1/2}(\nu\lambda + S_\gamma^2)^{-(n+\nu)/2}p(\gamma)$$

where $S_\gamma^2 = Y'Y - Y'X(X'X)^{-1}X'Y$.

The advantage of using the conjugate hierarchical mixture prior is integrating out $\boldsymbol{\beta}$ and σ from the joint posterior distribution. The fast updating of γ is likely to happen in the parsimonious models. This advantage could be especially pronounced in large problems with many useless predictors (Smith and Kohn, 1996).

3.3.2 Bayesian Variable Selection in logistic regression models

Model

We now discuss the binary regression model. The standard form is discussed here, where y_i is a binary variable ($y_i \in 0, 1; i = 1, ..., n$) and has a Bernoulli distribution for a collection of n objects. We also have measurements on p covariates $x_i = (x_{i1}, ..., x_{ip})$. The parameter in the logistic model can be denoted as $g^{-1}(\nu_i)$ where g is a link function, ν_i is the linear predictor that equals $x_i\boldsymbol{\beta}$, and $\boldsymbol{\beta}$ is a $p \times 1$ column vector of regression coefficients.

Albert and Chib (1993) introduced a latent variable (y_i) which has a normal prior distribution and hence, a conjugate normal posterior distribution, where $y_i = \mathbf{X}^{'}\boldsymbol{\beta} + \epsilon_i$, the error term($\epsilon_i$) has a standard normal distribution. If $\boldsymbol{\beta}$ is specified through a prior distribution with a probit link, it leads to probit regression. A conjugate normal prior distribution is selected for $\boldsymbol{\beta}$, $p(\boldsymbol{\beta}) = N(\mathbf{b}, \mathcal{V})$, where $\mathbf{b}$ is the prior mean vector and $\mathcal{V}$ is the prior covariance matrix. Usually, a zero vector $\mathbf{b} = 0$ is chosen for the prior mean and a prior covariance matrix $\mathcal{V} = c^2 I_p$ (independent), or the g-prior $\mathcal{V} = c^2(X'X)^{-1}$.

However, the logistic regression is often more widely used over the probit model in applications to biostatistical data sets, but due to the lack of conjugacy is computationally more expensive. Holmes and Held (2006) introduced a latent variable z with the conjugate normal prior that leads to a simpler conjugate formulation for the logistic regression model. This version of the logistic regression model is discussed below.

$$y_j = \begin{cases} 1 & if\ z_j > 0 \\ 0 & \text{otherwise} \end{cases}$$

$$z_j = x_j\boldsymbol{\beta} + \epsilon_j$$

$$\epsilon_j \sim N(0, \lambda_j)$$

$$\lambda_j = (2\phi_j)^2$$

$$\phi_j \sim KS(i.i.d.)$$

$$\boldsymbol{\beta}_j \sim p(\boldsymbol{\beta})$$

The auxiliary variables ϕ_j are independent random variables from the Kolmogorov-Smirnov (KS) distribution (Devroye 1986). Andrews and Mallows (1974) proved that $2AB$ has the logistic distribution where A is Normal distribution and B is the Kolmogorov-Smirnov distribution. In this case, ϕ_j are generated from the independent of KS, then $(2\phi_j)^2$ is set as λ_j and λ_j is the variance in the Normal distribution. It leads to a normal scale mixture distribution for ϵ_j in a marginal logistic distribution. Hence, this model is equivalent to a Bayesian logistic regression model (Andrews and Mallows 1974). The prior distribution of $\boldsymbol{\beta}$ is assumed normal $N(\mathbf{b}, \mathrm{Y})$. Then, the posterior distribution of $\boldsymbol{\beta}$ is normal with mean B and covariance matrix V as the standard for Bayesian modelling (Holmes and Held 2006).

$$\boldsymbol{\beta}|z, j \sim N(B, V)$$

$$B = V(\Upsilon^{-1}\mathbf{b} + X'\lambda^{-1}z)$$

$$V = (\Upsilon^{-1} + X'\lambda^{-1}X)^{-1}$$

$$\lambda^{-1} = \text{diag}(\lambda_1^{-1}, \ldots, \lambda_n^{-1})$$

Holmes and Held (2006) extended the Bayesian logistic regression model to incorporate variable selection by including a vector of covariate indicator variables $\boldsymbol{\gamma} = (\gamma_1, ..., \gamma_p)$ where $\gamma_j \in \{0, 1\} (j = 1, \ldots, p)$ corresponds to the indicator variable in the hierarchical model for variable selection.

The hierarchical model setup for variable selection is described below:

$$y_{\gamma j} = \begin{cases} 1 & \text{if} \quad z_{\gamma j} > 0 \\ 0 & \text{otherwise} \end{cases}$$

$$z_{\gamma j} = x_{\gamma j}\boldsymbol{\beta}_\gamma + \epsilon_j$$

$$\epsilon_j \sim N(0, \lambda_j)$$

$$\lambda_j = (2\phi_j)^2$$

$$\phi_j \sim KS(i.i.d.)$$

$$\boldsymbol{\beta}_\gamma \sim N(b_\gamma, v_\gamma)$$

$$\gamma \sim p(\gamma)$$

The prior distribution $p(\gamma)$ is the product of Bernoulli distributions of the variables γ_i with prior probabilities π_i. This is given by

$$p(\gamma) = \prod_{i=1}^{p} {\pi_i}^{\gamma_i}(1 - \pi_i)^{1-\gamma_i}.$$

For γ, the Bernoulli prior is set with small prior probabilities since the expected number of selected SNPs on GWAS are small. Under the simulation studies, we choose small constant prior probabilities $\pi_i = p^*/p$ for $i = 1, ..., p$. The expected number of covariates, denoted as p^*, is set to be small, for example, three or five. However, in real data sets we do not know the exact true number of SNPs. Instead of fixing the prior probabilities, we can choose a more flexible Beta-Binomial distribution for γ, using the identity

$$p(\gamma) = \int p(\gamma|\pi)p(\pi)d\pi,$$

where $p(\gamma|\pi) = \prod_{i=1}^{p} {\pi_i}^{\gamma_i}(1 - \pi_i)^{1-\gamma_i}$ and with a hyper-prior distribution for π that is denoted by $p(\pi) = \pi^{a-1}(1 - \pi)^{b-1}/B(a, b)$ where $B(a, b)$ is a Beta function.

The prior distribution of the regression coefficient $\boldsymbol{\beta}_\gamma$ is defined for the variables for which $\gamma_i = 1$ where $(\mathbf{b}_\gamma = 0_{p_\gamma \times 1})$ and $\Upsilon = c^2 I_p$ where I_p is the identity matrix of size $p_\gamma \times p_\gamma$. The hierarchical logistic regression model gives a joint posterior distribution for $\{\boldsymbol{\beta}_\gamma, \gamma, z, \lambda\}$ that can

be written as

$$p(\boldsymbol{\beta}_\gamma, \gamma, z, \lambda | X_\gamma, y) \propto p(y|z)p(z|\lambda, \boldsymbol{\beta}_\gamma, X_\gamma)p(\boldsymbol{\beta}_\gamma|\gamma)p(\gamma)p(\lambda)$$

where $p(\lambda_i) \sim 1/4\sqrt{\lambda_i}KS(0.5\sqrt{\lambda_i})$ and $p(z|\lambda, \boldsymbol{\beta}_\gamma, X_\gamma) = N(X_\gamma \boldsymbol{\beta}_\gamma, \lambda)$.

MCMC algorithm for Bayesian Variable Selection

Since there is a high correlation between parameters in single updating which leads to slow mixing of the Markov chains, Zucknick and Richardson (2014) proposed jointly updating z, λ and $\gamma, \boldsymbol{\beta}_\gamma$ from this model.

Updating z and λ

The first update is drawn from $p(z, \lambda | \boldsymbol{\beta}, \gamma, X, y) = p(z|\boldsymbol{\beta}, \gamma, X, y)p(\lambda|z, \boldsymbol{\beta}, \gamma, X)$.

1. The inversion method can be used to draw $p(z|\boldsymbol{\beta}, \gamma, X, y)$: the steps are given below.

   (a) For $i = 1, \ldots, n$,

   $$p(z_i|\boldsymbol{\beta}, \gamma, x, y) \propto \begin{cases} \text{Logistic}(x_{\gamma i}, 1)I(z_i > 0) & \text{if } y_i = 1 \\ \text{Logistic}(x_{\gamma i}, 1)I(z_i \leq 0) & \text{if } y_i = 0 \end{cases}$$

   (b) Calculate the CDF $F(x)$ of the logistic distribution above. Sample $u_i \sim U[0, 1]$, and solve for $F(F^{-1}(u_i)) = u_i$.

2. Rejection sampling can be applied to sample $p(\lambda|z, \boldsymbol{\beta}, \gamma, x)$. Those steps consist of the following:

   (a) Sample $u_i \sim U[0, 1]$.

   (b) Sample λ_i from the candidate density

   $$g(\lambda_i) \sim GIG(0.5, 1, r_i^2) = \frac{r_i}{IG(1, |r_i|)}$$

   where IG is an Inverse Gaussian distribution with

   $$p(x) = \sqrt{\frac{|r_i|}{2\pi x^3}} \exp\left(-\frac{|r_i|(x-1)^2}{2x}\right), \qquad x \geq 0,$$

   where $r_i^2 = (z_i - x_{\gamma i}\boldsymbol{\beta}_\gamma)^2$.

   (c) If $u_i \leq \alpha(\lambda_i)$, accept λ_i where

   $$\alpha(\lambda_i) = \frac{l(r_i^2, \lambda_i)p(\lambda_i)}{Mg(\lambda_i)}$$

   with

   $$l(r_i^2, \lambda_i) = p(z_i|x_{\gamma i}, \boldsymbol{\beta}_\gamma, \lambda_i) = N(x_{\gamma i}\boldsymbol{\beta}_\gamma, \lambda_i),$$

$p(\lambda_i)$ being the Inverse Gaussian distribution. Otherwise, reject λ_i and go back to step (a).

Moreover, the use of an alternative series expansion of $KS(0.5\sqrt{\lambda_i})$ in Devroye (1986), gives

$$\alpha(\lambda_i) = N(x_{\gamma i}\boldsymbol{\beta}_i, \lambda_i)\frac{1}{4\sqrt{\lambda_i}}KS(0.5\sqrt{\lambda_i}).$$

Updating β and γ

$(\boldsymbol{\beta}_\gamma, \gamma)$ are updated jointly, using the identity

$$p(\boldsymbol{\beta}_\gamma, \gamma|z, \lambda, X) = p(\gamma|z, \lambda, X)p(\boldsymbol{\beta}_\gamma|\gamma, z, \lambda, X).$$

1. With a starting value of $\gamma = \gamma^0$, $\boldsymbol{\beta}_\gamma$ can be directly sampled from

$$N(B^*_\gamma, V^*_\gamma),$$

where $B^*_\gamma = V^*_\gamma x'_{\gamma^*}\lambda^{-1}z$ and $V^*_\gamma = (x'_{\gamma^*}\lambda^{-1}x_{\gamma^*} + v^{-1}_{\gamma^*})$.

2. Then, γ is updated using the following steps of a Metropolis-Hastings algorithm.

   (a) (Add/delete step.) At the t-th iteration, a single covariate is selected at random and the proposal distribution is given by

$$q(\gamma^*_j) = \begin{cases} 1 & \text{if } \gamma_j = 0 \\ 0 & \text{if } \gamma_j = 1 \end{cases}$$

   (b) The acceptance probability for updating γ is given by

$$\alpha(\gamma) = \min\left(1, \frac{|V_{\gamma^*}|^{1/2}|v_\gamma|^{1/2}\exp(0.5B'_{\gamma^*}V^{-1}_{\gamma^*}B_{\gamma^*})}{|V_\gamma|^{1/2}|v^*_\gamma|^{1/2}\ \exp(0.5B'_\gamma V^{-1}_\gamma B_\gamma)}\frac{(1-\pi_i)}{\pi_i}\right),$$

   where π_j is the prior probability that γ_j takes the value 1.

   (c) Set

$$\gamma^t = \begin{cases} \gamma^* & \text{with probability } \alpha(\gamma) \\ \gamma^{t-1} & \text{otherwise} \end{cases}$$

There are two packages in R statistical software utilized in our study. We use the *bvsflex* package on R-forge for the variable selection in logistic regression models. (http://bvsflex.r-forge.r-project.org) (Zucknick, 2013). Moreover, the *BayesVarSel* is the package for variable selection in Linear regression models (Gonzalo and Anabel, 2018).

The proposed methods and simulation results are presented in the next chapter (Chapter 4).

Chapter 4

Variable selection with data splitting

4.1 Introduction

In variable selection, a major challenge arises when the number of covariates is huge compared to the number of observations. Even if proper priors allow this to be done via Bayesian methods, with an extremely high number of covariates (i.e. many thousands or even millions) compared to the number of observations (i.e. a few hundreds), there are 2 major problems: huge computational time burdens for analysing each dataset, another is the sparsity in the number of covariates associated to the response. If data splitting is used for variable selection in the case above, this can lead to significant reduction in computational time.

GWAS typically contain many thousands of covariates, which makes variable selection an exceptionally computationally intensive process. Additionally, with large datasets, the MCMC sampler often becomes inefficient in terms of CPU time and shows a lack of MCMC convergence. We wished to investigate if splitting the whole dataset into a number of small sub-datasets before running BVS reduces the time for the MCMC sampler, improving the mixing of the Markov chain. But simultaneously, we need to investigate the impact of data splitting in terms of the properties and accuracy of the resulting model. When the data is split column-wise, a number of the sub-datasets may not contain the covariates associated to the response.

Hence, the covariates that are selected in each sub-dataset should be finally combined to determine the final set of associated covariates. But this procedure could lead to possible biases, the quantity is $E(SSE)$, so we need to assess how this affects the error is estimation of regression coefficients and other parameters.

4.2 The data splitting procedure

For simplicity, we start with linear regression models. The model is $y = X\boldsymbol{\beta} + \epsilon$, $\epsilon \sim N(0, \sigma^2 I)$, where y is an n dimensional vector of observations, X is $n \times p$ dimensional matrix with known constants and ϵ is a vector of errors. Since sum of squares of errors (SSE) is the measure of the difference between responses and predicted value, $E(SSE)$ is the expected value of SSE that is the average of the SSE. Hence, $E(SSE)$ can explain the error in the overall. However, $E(SSE)$ in the

unsplit data is given by

$$E(SSE) = (n-p)\sigma^2.$$

For the first stage, the full design matrix X is partitioned as $X_1^*, X_2^*, ..., X_k^*$ where k is the number of splits into sub-datasets.

Each of X^* can be written in matrix form as the following.

$$X_1^* = \begin{bmatrix} x_{1,1} & x_{1,2} & x_{1,3} & \dots & x_{1,p_1} \\ x_{2,1} & x_{2,2} & x_{2,3} & \dots & x_{2,p_1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & x_{n,3} & \dots & x_{n,p_1} \end{bmatrix}$$

$$X_2^* = \begin{bmatrix} x_{1,p_1+1} & x_{1,p_1+2} & x_{1,p_1+3} & \dots & x_{1,p_1+p_2} \\ x_{2,p_1+1} & x_{2,p_1+2} & x_{2,p_1+3} & \dots & x_{2,p_1+p_2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{n,p_1+1} & x_{n,p_1+2} & x_{n,p_1+3} & \dots & x_{n,p_1+p_2} \end{bmatrix}$$

$$\vdots$$

$$X_k^* = \begin{bmatrix} x_{1,p_{k-1}+1} & x_{1,p_{k-1}+2} & x_{1,p_{k-1}+3} & \dots & x_{1,p_1+p_2+...+p_k} \\ x_{2,p_{k-1}+1} & x_{2,p_{k-1}+2} & x_{2,p_{k-1}+3} & \dots & x_{2,p_1+p_2+...+p_k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{n,p_{k-1}+1} & x_{n,p_{k-1}+2} & x_{n,p_{k-1}+3} & \dots & x_{n,p_1+p_2+...+p_k} \end{bmatrix}$$

where $p_1 + p_2 + ... + p_k = p$, $p_1 = p_2 = ... = p_k$, p being the total number of covariates in the whole data set and $p_1, p_2, ..., p_k$ are the number of covariates in each of k sub-datasets.

Fitting the full model by ordinary least squares yields an estimated parameter

$$\hat{\boldsymbol{\beta}} = (X'X)^{-1}X'y \sim N(\boldsymbol{\beta}, \sigma^2(X'X)^{-1}).$$

The predicted value of $\mathbf{y}$ based on the fitted model is

$$\hat{\mathbf{y}} = X\hat{\boldsymbol{\beta}}$$

and the error of prediction is

$$\mathbf{y} - \hat{\mathbf{y}}.$$

A common summary of the predictive ability of the fitted model is the unconditional mean squared error (MSE). Hence, we will start with the evaluation and comparison of sum squared error (SSE) between the full and split data models.

Since we need to split the whole dataset into k equally sized sub-datasets, we rewrite the data set in the following form. First we write the response in the form of a $nk \times 1$ vector $\tilde{y}$, where $\tilde{y}$ is the vector y repeated k times. Similarly, $\tilde{X}$ is a block matrix made of $nk \times p$ blocks and $\hat{\boldsymbol{\beta}}^*$ is

the matrix of regression coefficients in each sub-dataset that has $p \times 1$ elements. Mathematically, we write,

$$\tilde{Y} = \begin{bmatrix} \mathbf{y} \\ \mathbf{y} \\ \vdots \\ \mathbf{y} \end{bmatrix}, \hat{\boldsymbol{\beta}}^* = \begin{bmatrix} \hat{\boldsymbol{\beta}}_1^* \\ \hat{\boldsymbol{\beta}}_2^* \\ \vdots \\ \hat{\boldsymbol{\beta}}_k^* \end{bmatrix} \text{ and } \tilde{X} = \begin{bmatrix} \tilde{X}_1 & \tilde{0} & \dots & \tilde{0} \\ \tilde{0} & \tilde{X}_2 & \dots & \tilde{0} \\ \vdots & \vdots & \vdots & \vdots \\ \tilde{0} & \tilde{0} & \dots & \tilde{X}_k \end{bmatrix}, \tag{4.1}$$

where $\tilde{X}_1, \tilde{X}_2, ..., \tilde{X}_k$ are the matrices of covariates corresponding to each sub-dataset and $\hat{\boldsymbol{\beta}}_1^*, \hat{\boldsymbol{\beta}}_2^*, ..., \hat{\boldsymbol{\beta}}_k^*$ the regression coefficient vectors estimated from each corresponding sub-dataset.

By definition,

$$\begin{aligned}
SSE &= (\tilde{Y} - \tilde{X}\hat{\boldsymbol{\beta}}^*)'(\tilde{Y} - \tilde{X}\hat{\boldsymbol{\beta}}^*) \\
&= \begin{bmatrix} \mathbf{y} - \tilde{X}_1\hat{\boldsymbol{\beta}}_1^* & \mathbf{y} - \tilde{X}_2\hat{\boldsymbol{\beta}}_2^* & \dots \mathbf{y} - \tilde{X}_k\hat{\boldsymbol{\beta}}_k^* \end{bmatrix} \begin{bmatrix} \mathbf{y} - \tilde{X}_1\hat{\boldsymbol{\beta}}_1^* \\ \mathbf{y} - \tilde{X}_2\hat{\boldsymbol{\beta}}_2^* \\ \vdots \\ \mathbf{y} - \tilde{X}_k\hat{\boldsymbol{\beta}}_k^* \end{bmatrix} \\
&= (\mathbf{y} - \tilde{X}_1\hat{\boldsymbol{\beta}}_1^*)'(\mathbf{y} - \tilde{X}_1\hat{\boldsymbol{\beta}}_1^*) + (\mathbf{y} - \tilde{X}_2\hat{\boldsymbol{\beta}}_2^*)'(\mathbf{y} - \tilde{X}_2\hat{\boldsymbol{\beta}}_2^*) + ... + (\mathbf{y} - \tilde{X}_k\hat{\boldsymbol{\beta}}_k^*)'(\mathbf{y} - \tilde{X}_k\hat{\boldsymbol{\beta}}_k^*), \\
&\text{(where } \hat{\boldsymbol{\beta}}_1^* = (\tilde{X_1}'\tilde{X_1})^{-1}\tilde{X_1}'\mathbf{y}, \hat{\boldsymbol{\beta}}_2^* = (\tilde{X_2}'\tilde{X_2})^{-1}\tilde{X_2}'\mathbf{y}, ..., \hat{\boldsymbol{\beta}}_k^* = (\tilde{X_k}'\tilde{X_k})^{-1}\tilde{X_k}'\mathbf{y}.) \\
&= \sum_{j=1}^{k} (\mathbf{y} - \tilde{X}_j\hat{\boldsymbol{\beta}}_j^*)'(\mathbf{y} - \tilde{X}_j\hat{\boldsymbol{\beta}}_j^*) \\
&= \sum_{j=1}^{k} (\mathbf{y} - \tilde{X}_j(\tilde{X}_j'\tilde{X}_j)^{-1}\tilde{X}_j'\mathbf{y})'(\mathbf{y} - \tilde{X}_j(\tilde{X}_j'\tilde{X}_j)^{-1}\tilde{X}_j'\mathbf{y}).
\end{aligned}$$

Now, let $P_j = \tilde{X}_j(\tilde{X}_j'\tilde{X}_j)^{-1}\tilde{X}_j'$, the j^{th} projection matrix. Hence

$$\begin{aligned}
SSE &= \sum_{j=1}^{k} (\mathbf{y} - P_j\mathbf{y})'(\mathbf{y} - P_j\mathbf{y}) \\
&= \sum_{j=1}^{k} \mathbf{y}'(I - P_j)(I - P_j)\mathbf{y} \\
&= \sum_{j=1}^{k} \mathbf{y}'(I - P_j)\mathbf{y}
\end{aligned}$$

(since P_j is idempotent, I being the identity matrix of dimension n).

Therefore,

$$E(SSE) = \sum_{j=1}^{k} E[\mathbf{y}'(I - P_j)\mathbf{y}].$$

This can now be written as follows.

$$
\begin{aligned}
E(SSE) &= \sum_{j=1}^{k}\{\text{trace}[(I-P_j)Var(\mathbf{y})] + E(\mathbf{y}')(I-P_j)E(\mathbf{y})\}\text{(following Prop. 3.22 page 79 Bingham and Fry, 2010)} \\
&= \sum_{j=1}^{k}\{\text{trace}[(I-\tilde{X}_j(\tilde{X}_j'\tilde{X}_j)^{-1}\tilde{X}_j')]\sigma^2 I + (\boldsymbol{\beta}'X')(I-P_j)(X\boldsymbol{\beta})\} \\
&= \sum_{j=1}^{k}\{[\text{trace}(I) - \text{trace}(\tilde{X}_j(\tilde{X}_j'\tilde{X}_j)^{-1}\tilde{X}_j')]\sigma^2 I + (\boldsymbol{\beta}'X')(I-P_j)(X\boldsymbol{\beta})\} \\
&= \sum_{j=1}^{k}\{[n - \text{trace}((\tilde{X}_j'\tilde{X}_j)^{-1}\tilde{X}_j'\tilde{X}_j]\sigma^2 I + (\boldsymbol{\beta}'X')(I-P_j)(X\boldsymbol{\beta})\}
\end{aligned}
$$

(using the property that trace(AB) =trace(BA) where A is $m \times n$ and B is $n \times m$.)

$$
= \sum_{j=1}^{k}\{[(n-p_j)\sigma^2] + (\boldsymbol{\beta}'X')(I-P_j)(X\boldsymbol{\beta})\},
$$

where p_j is the number of covariates in each split sub - data set, I is the $n \times n$ identity matrix and I_j^* is the $p_j \times p_j$ identity matrix.

In order to check the effect of variation in n, k, p on the expected value of SSE, we conducted a simulation study. We also compared the bias of the parameter estimation via using covariates from combining results from each sub-dataset compared to the whole dataset. These simulations and results are described in the next section.

4.3 Simulation studies

4.3.1 General Settings

We simulated 3 datasets 20, 50, and 100 covariates respectively, and with 200 observations each. Each explanatory variable vector is generated from the Binomial distribution with a probability percentage of 0.1. The effect size for the regressors for all associated covariates is 1. All covariates are assumed to be associated with the response. These are rather restrictive assumptions and these simulations do not allow one to conclude in generality. Since in any application, explanatory variables will not have the same distribution and will not have the same effect size. Under each situation, the explanatory covariates are spilt into 2 and 5 sub data sets, in turn. under each setting, 10 data sets for checking replication of the results.

For each dataset, and under each setting, we calculated the $E(SSE)$ as discussed in Section 4.2.

4.3.2 Results

Our results showed that when the number of data set splits (k) is increased the values of E(SSE) increase in general (Table 4.1). Moreover, when the number of covariates (p) is increased the values of E(SSE) also increased. This means that the error when fitting the split data sets are higher than

Table 4.1: The upper bound and lower bound of E(SSE) and the mean when $\beta = 1$ (20, 50 and 100 covariates) under 10 replications

No. of Splits	20 covariates	50 covariates	100 covariates	500 covariates
2	151.44(140.46,172.98)	186.23(181.61,198.39)	199.15(188.95,215.28)	222.11(198.21,238.87)
5	212.72(192.67,229.83)	214.49(187.96,225.44)	217.96(189.71,230.97)	236.85(208.76,250.33)
10	230.85(211.37,259.61)	235.35(208.33,250.37)	236.38(208.46,255.31)	253.18(227.64,269.31)
whole	77.98(74.98,84.43)	79.80(73.54,85.16)	82.51(73.48,86.25)	96.13(90.56,102.49)

when using whole dataset. Moreover, when the number of sub-datasets (k) is increased, it yields higher E(SSE) (Figure 4.1). The ratio of E(SSE) between the split sub-dataset and the whole data set are greater than 1 in all situations (Figure 4.2)

4.4 Computational times

Since the benefit of data splitting is claimed to be computational efficiency. The CPU times for running in each splitting and the whole dataset are presented in Table 4.2.

Table 4.2: The CPU times for running in each splitting and the whole dataset (seconds)

No. of Splits	20 covariates	50 covariates	100 covariates	500 covariates
2	1250.23	1450.48	1640.08	1950.67
5	1346.11	1538.62	1751.63	2097.43
10	1436.21	1629.27	1846.94	2136.41
whole	10316.34	11246.65	12893.54	14379.14

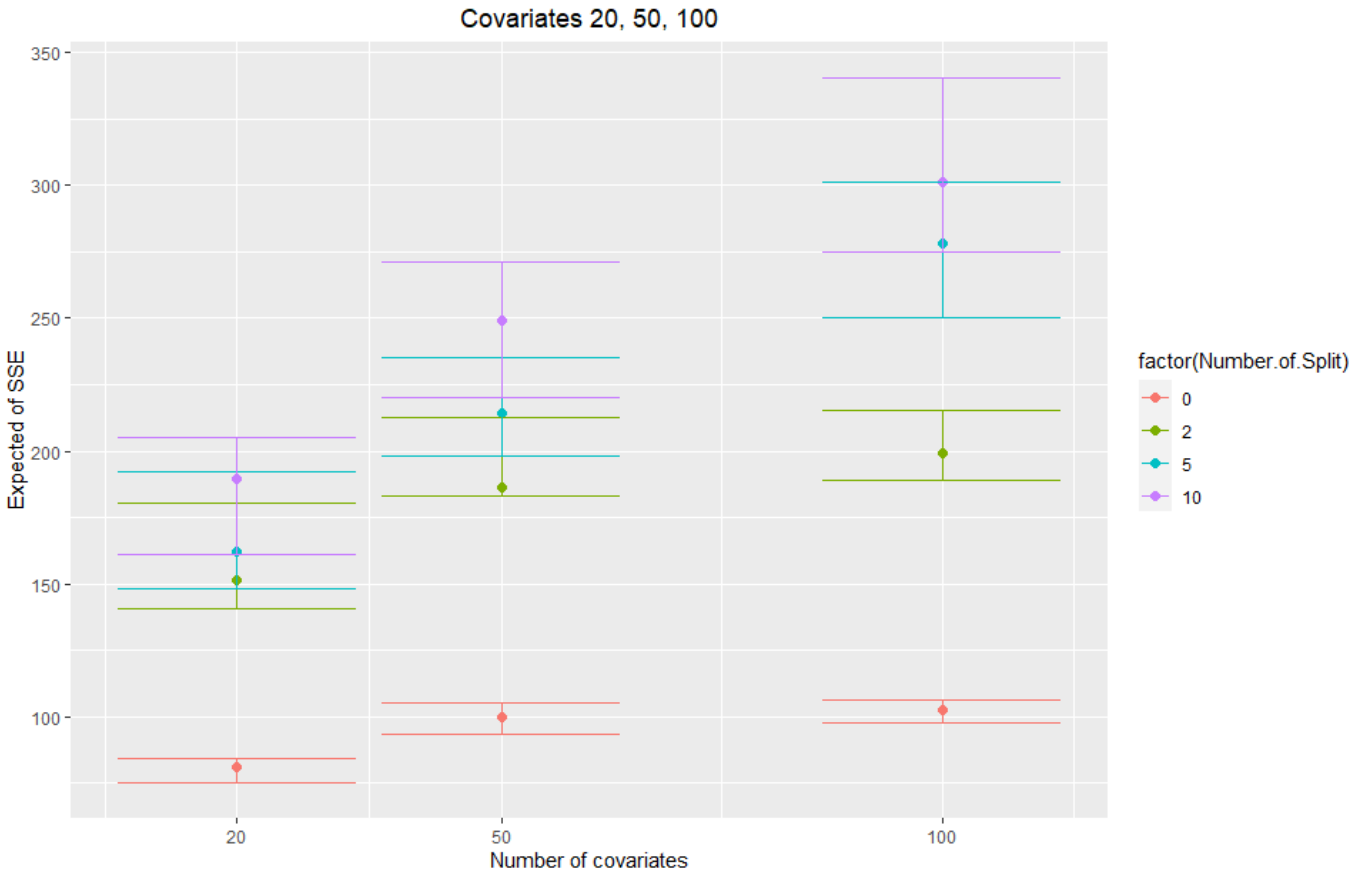


Figure 4.1: The plot of the Expected SSE under 20, 50 and 100 covariates

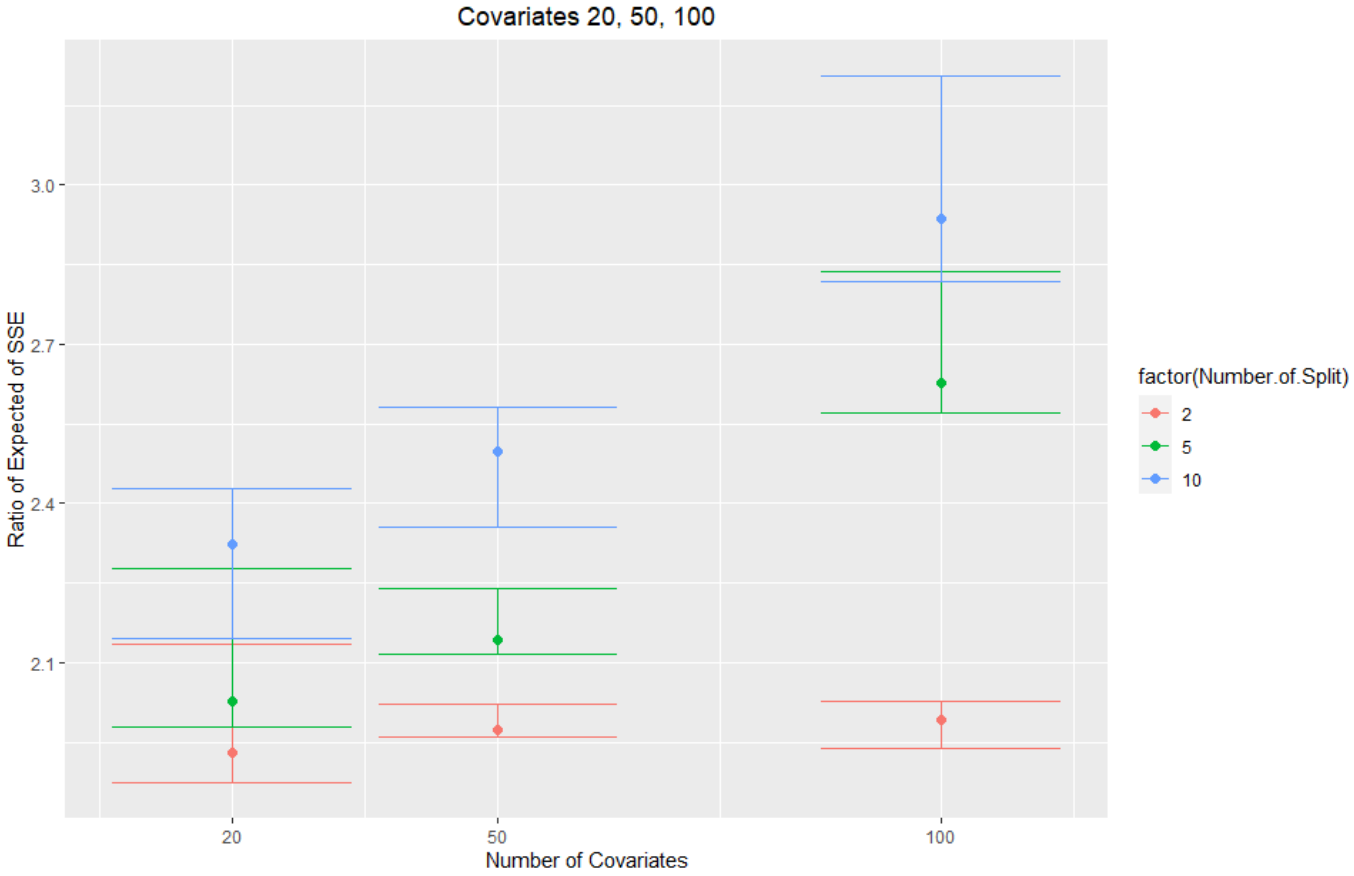


Figure 4.2: The plot of the ratio of Expected of SSE between split dataset and the whole dataset with 20, 50 and 100 covariates

4.5 Splitting datasets in the Bayesian framework

Expected SSE under the Bayesian framework is calculated via using

$$E(SSE) = [n - \frac{(g^2+2g)}{(g+1)^2}p]\sigma^2$$

that relied on the whole data set, where g is a parameter in Zellner' s g-prior on the regression coefficients (adapted from Liang et al., 2007). Clearly, as $g \to \infty$, $E(SSE)$ converges to the same value as in the classical framework (Section 4.1).

The expected value of SSE for the splitting data set can be shown using a similar procedure as before, to be

$$E(SSE) = (nk - \frac{(g^2+2g)}{(g+1)^2}\sum_{j=1}^{k} p_j)\sigma^2 + \sum_{j=1}^{k}[(\boldsymbol{\beta}'X')(I - \frac{(g^2+2g)}{(g+1)^2}P_j)(X\boldsymbol{\beta})].$$

The derivation is given below:

$$\begin{aligned}
E(SSE) &= \sum_{j=1}^{k} E[\mathbf{y}'(I - \frac{(g^2+2g)}{(g+1)^2}\mathbf{y}] \\
&= \sum_{j=1}^{k}\{\text{trace}[(I - \frac{(g^2+2g)}{(g+1)^2}P_j)Var(\mathbf{y})] + E(\mathbf{y}')(I - \frac{(g^2+2g)}{(g+1)^2}P_j)E(\mathbf{y})\} \\
&= \sum_{j=1}^{k}\{\text{trace}[(I - \frac{(g^2+2g)}{(g+1)^2}P_j)\sigma^2 I] + (\boldsymbol{\beta}'X')(I - \frac{(g^2+2g)}{(g+1)^2}P_j)(X\boldsymbol{\beta})\} \\
&= \sum_{j=1}^{k}\{\text{trace}[(I - \frac{(g^2+2g)}{(g+1)^2}\tilde{X}_j(\tilde{X}_j'\tilde{X}_j)^{-1}\tilde{X}_j')]\sigma^2 I + (\boldsymbol{\beta}'X')(I - \frac{(g^2+2g)}{(g+1)^2}P_j)(X\boldsymbol{\beta})\} \\
&= \sum_{j=1}^{k}\{[\text{trace}(I) - \frac{(g^2+2g)}{(g+1)^2}\text{trace}(\tilde{X}_j(\tilde{X}_j'\tilde{X}_j)^{-1}\tilde{X}_j')]\sigma^2 I + (\boldsymbol{\beta}'X')(I - \frac{(g^2+2g)}{(g+1)^2}P_j)(X\boldsymbol{\beta})\} \\
&= \sum_{j=1}^{k}\{[n - \frac{(g^2+2g)}{(g+1)^2}\text{trace}((\tilde{X}_j'\tilde{X}_j)^{-1}\tilde{X}_j'\tilde{X}_j]\sigma^2 I + (\boldsymbol{\beta}'X')(I - \frac{(g^2+2g)}{(g+1)^2}P_j)(X\boldsymbol{\beta})\} \\
&= \sum_{j=1}^{k}\{[n - \frac{(g^2+2g)}{(g+1)^2}\text{trace}(I_j^*)]\sigma^2 I + (\boldsymbol{\beta}'X')(I - \frac{(g^2+2g)}{(g+1)^2}P_j)(X\boldsymbol{\beta})\} \\
&= \sum_{j=1}^{k}\{[n - \frac{(g^2+2g)}{(g+1)^2}P_j]\sigma^2 I + (\boldsymbol{\beta}'X')(I - \frac{(g^2+2g)}{(g+1)^2}P_j)(X\boldsymbol{\beta})\} \\
&= (nk - \frac{(g^2+2g)}{(g+1)^2}\sum_{j=1}^{k} P_j)\sigma^2 + \sum_{j=1}^{k}\{(\boldsymbol{\beta}'X')(I - \frac{(g^2+2g)}{(g+1)^2}P_j)(X\boldsymbol{\beta})\},
\end{aligned}$$

where I_j^* is the $P_j \times P_j$ identity matrix. Again the E(SSE) under the Bayesian framework converges to the classical when $g \to \infty$, equivalent to better emulate the scenario of real GWAS datasets.

We expand our simulation to the correlated case, beginning with a small number of covariates first, there are 20 covariates: X_1 is correlated with X_2 and X_3 , while X_{11} is correlated with X_{12} and X_{13} .

Results from Tables 4.3 and 4.4 show that when there are higher levels of correlation, the

expected SSE is increased due to the effect on the regression coefficients estimation through multicollinearity.

Table 4.3: The upper bound and lower bound of E(SSE) and the mean when $\beta = 1$ (20 covariates) under 10 replications when $\rho = 0.1$

No. of Splits	g=1	g=10	g=100
2	204.356(189.112,215.634)	188.335(177.854,197.364)	170.234(165.148,182.287)
5	262.369(237.452,283.117)	249.865(226.186,261.337)	233.862(211.339,260.382)
10	290.114(276.249,298.964)	275.198(259.632,290.112)	265.841(255.126,279.854)
whole	100.352(92.126,109.345)	96.147(84.238,104.632)	91.287(79.968,98.963)

Table 4.4: The upper bound and lower bound of E(SSE) and the mean when $\beta = 1$ (20 covariates) under 10 replications when $\rho = 0.5$

No. of Splits	g=1	g=10	g=100
2	256.432(225.116,272.553)	221.334(189.398,246.337)	203.257(174.554,211.897)
5	286.331(262.417,302.745)	255.389(236.247,279.845)	240.552(228.968,269.997)
10	311.551(290.235,328.874)	289.211(270.114,304.489)	272.115(260.511,284.116)
whole	125.167(109.887,132.552)	102.338(90.115,111.287)	95.889(86.115,105.432)

Results from Table 4.5 and Figure 4.3 indicate that when the number of splits are increased the expected SSE increases. However, also the values of g are increased , the expected SSE decreases as the prior become less informative. Results under g = 1000 are close to the classical framework (Table 4.1).

Table 4.5: The upper bound and lower bound of E(SSE) and the mean when $\beta = 1$ (20 covariates) under 10 replications

No. of Splits	g=1	g=10	g=100	g=1000
2	198.34(185.61,206.96)	179.90(172.40,188.19)	164.65(159.82,175.24)	152.36(138.53,162.44)
5	255.11(229.35,272.12)	242.11(219.53,254.00)	227.51(207.44,252.44)	213.41(189.51,226.93)
10	286.72(271.11,293.19)	269.61(253.71,282.70)	259.79(249.90,272.64)	235.44(207.36,252.33)
whole	98.43(88.62,105.44)	92.34(81.35,100.63)	87.63(76.96,95.33)	79.86(71.35,85.11)

The results in Figures 4.3 to 4.5 indicate that with a less informative g-prior the expected SSE is close to the classical framework of OLS. Moreover, the ratio of expected SSE between the Bayesian and the classical framework are close to 1 under the less informative prior. The results are reported in Figures 4.6 to 4.8. Moreover, we use the splitting only the Bayesian framework for comparion to the whole data set based on the OLS framework.

In term of the number of splittings, the $E(SSE)$ increased when the number of splits increased. They show that more bias is likely to occur when using a high number of splittings.
Based on our findings, we can recommend the following guidelines for splitting. First, using a Bayesian model, as may be necessary in high-dimensional settings, should not be a problem, and will perform almost as well as an OLS approach as long as the priors taken are not highly informative. Second, since increasing the number of splits leads to more bias, it is best to take as small a number of splits as possible while still having a computationally efficient method to fit the model to the data.

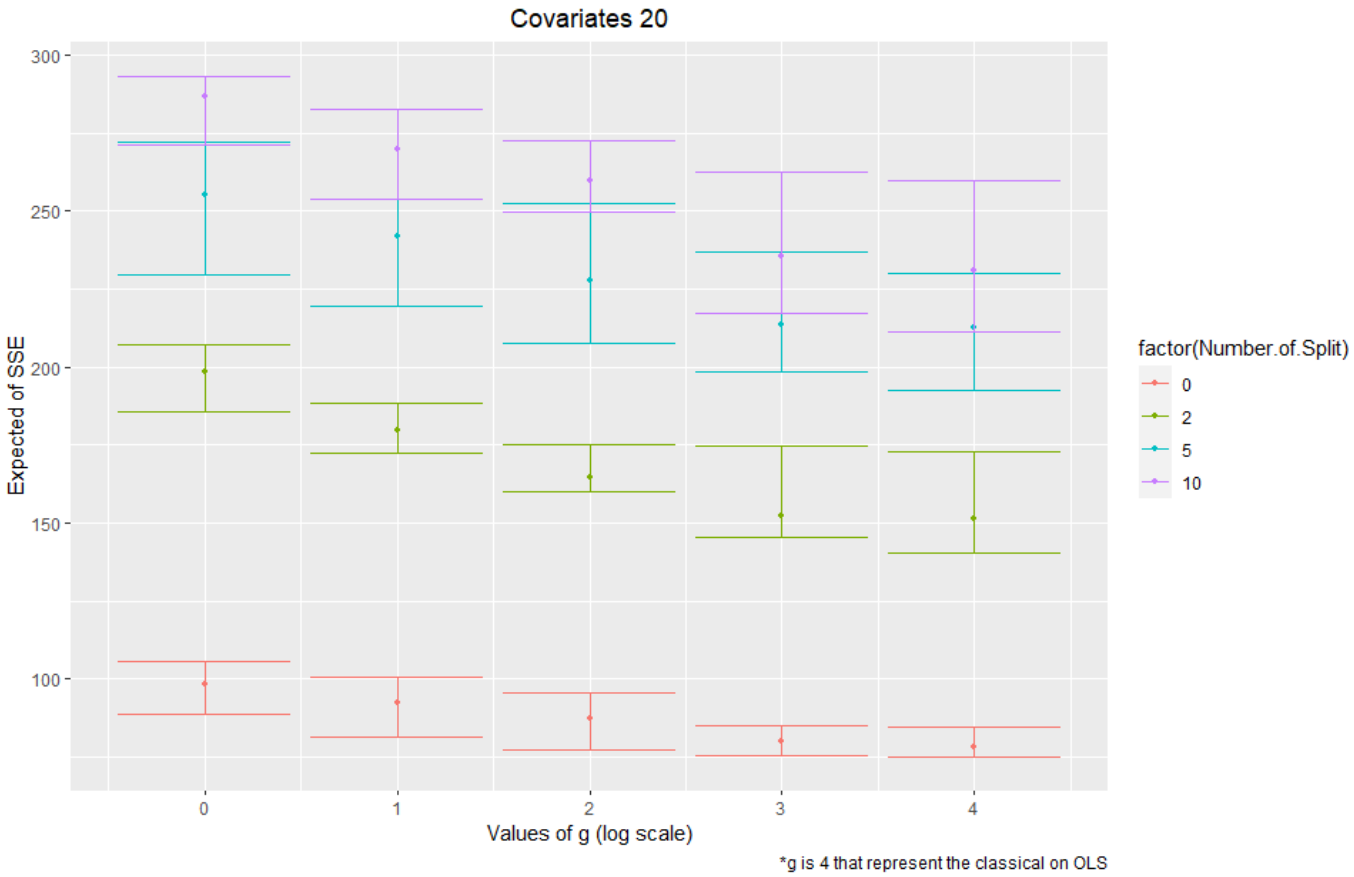


Figure 4.3: The plot of the Expected SSE with 20 covariates

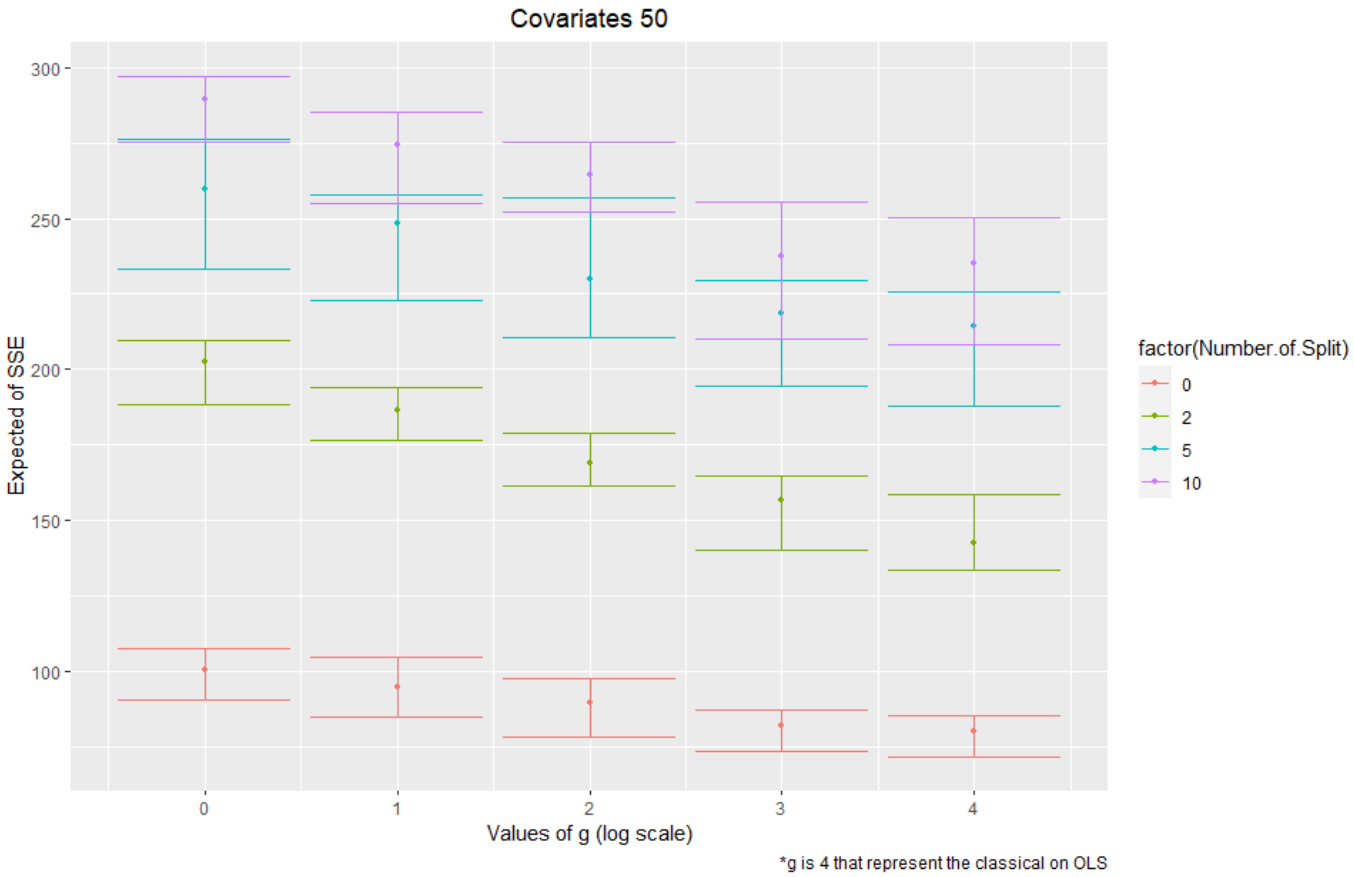


Figure 4.4: The plot of the Expected SSE with 50 covariates

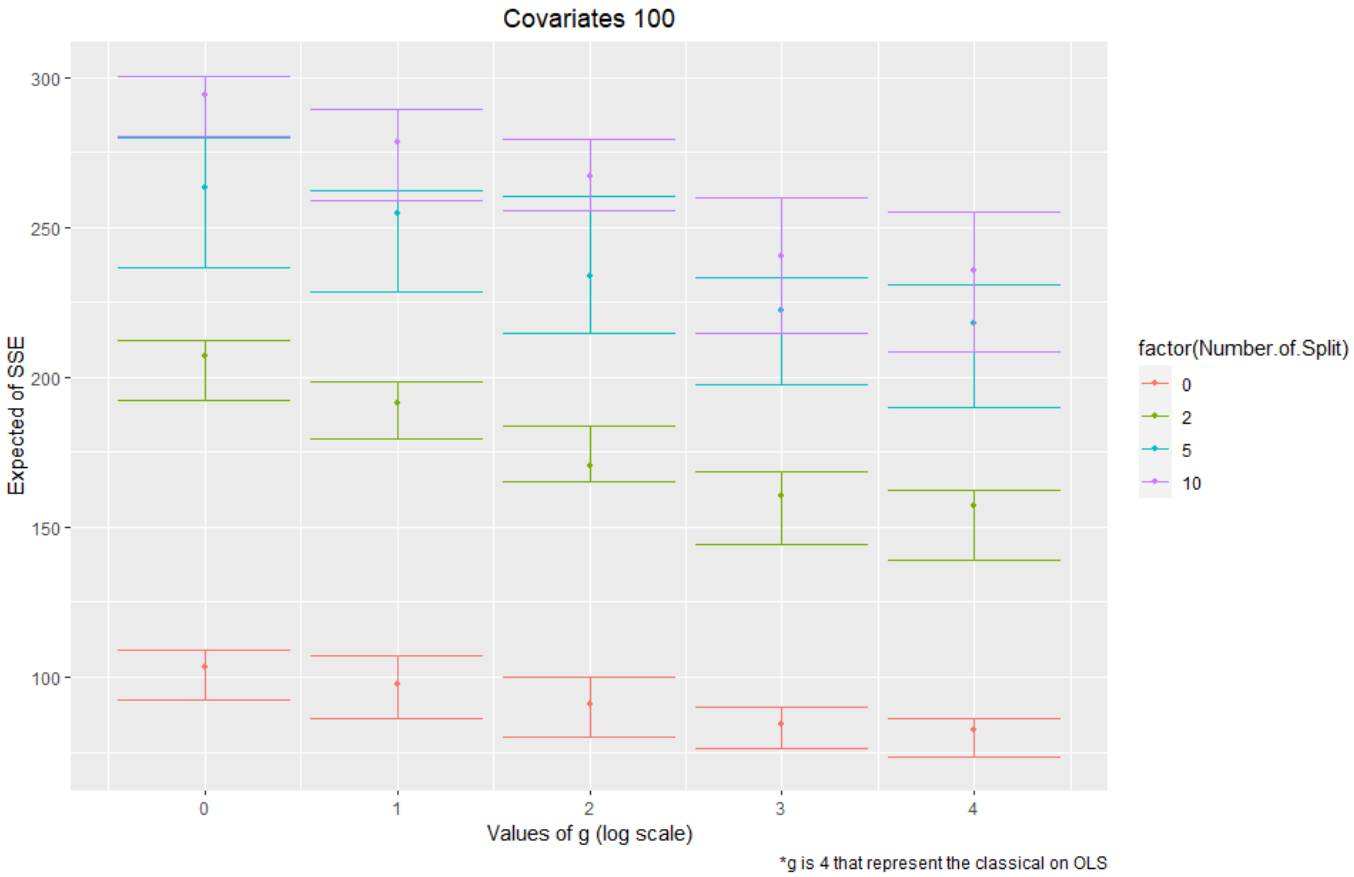


Figure 4.5: The plot of the Expected SSE with 100 covariates

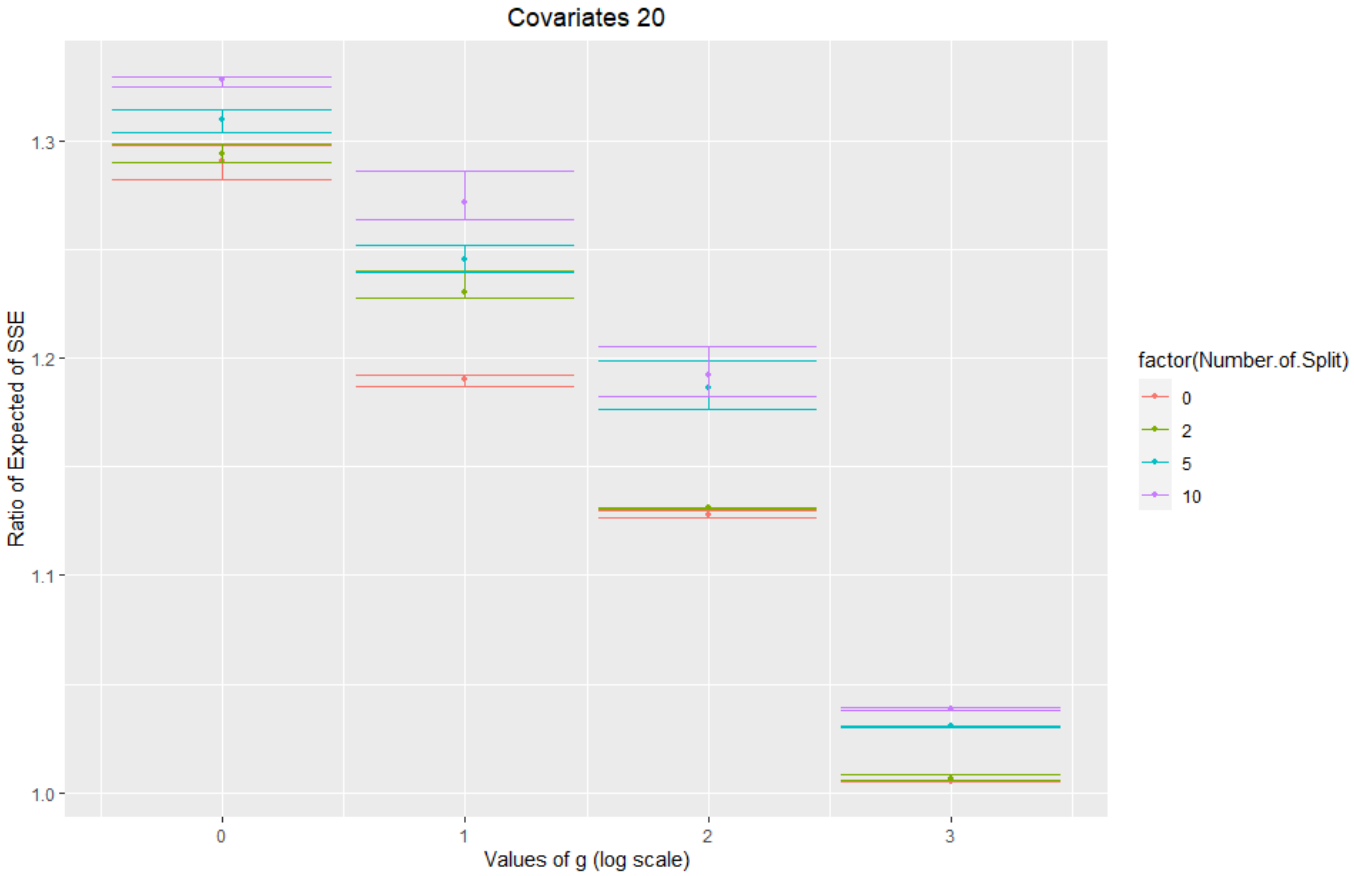


Figure 4.6: The plot of the ratio of Expected SSE under the Bayesian and classical frameworks with 20 covariates

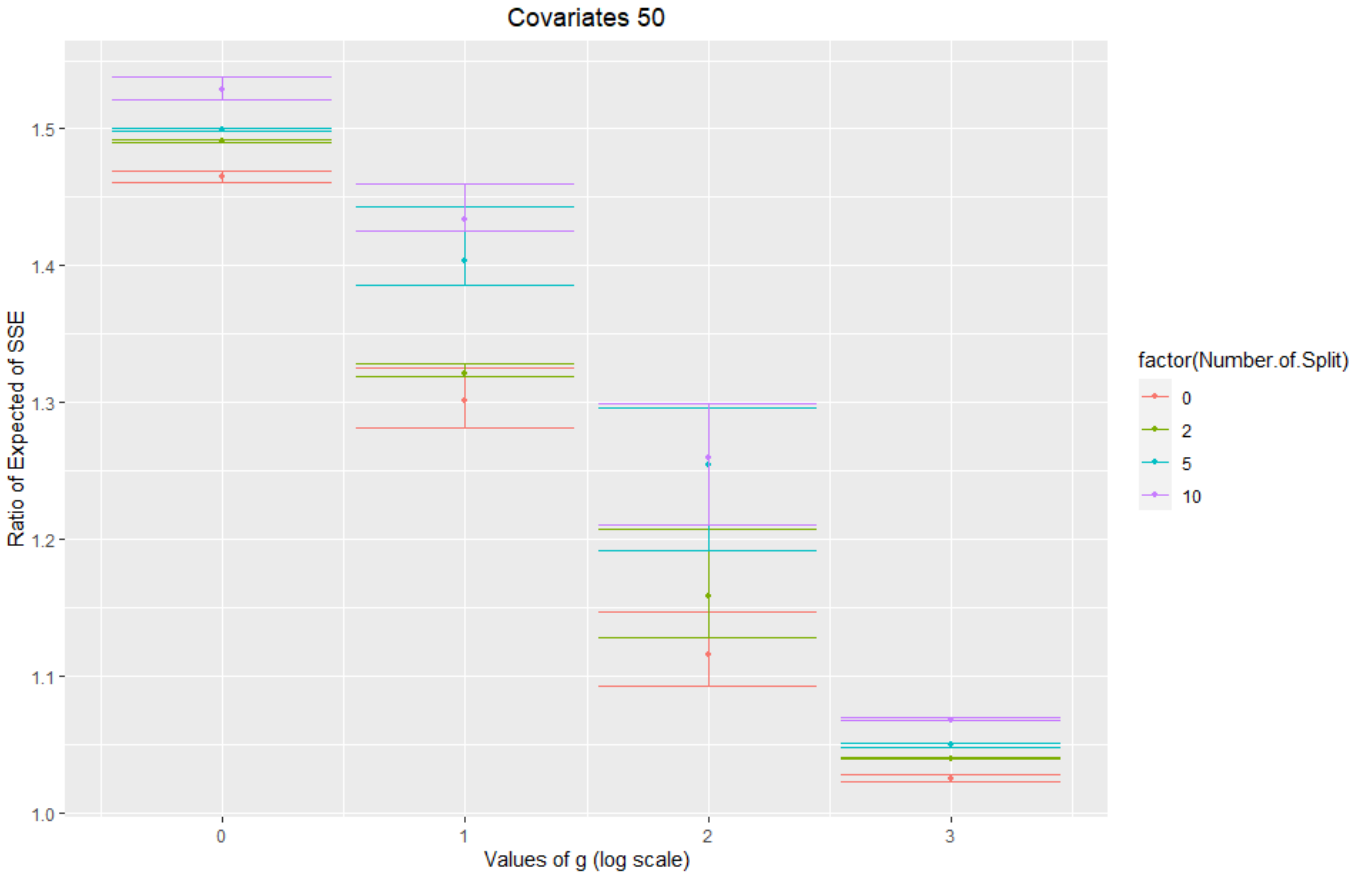


Figure 4.7: The plot of the ratio of Expected SSE under the Bayesian and classical framework with 50 covariates

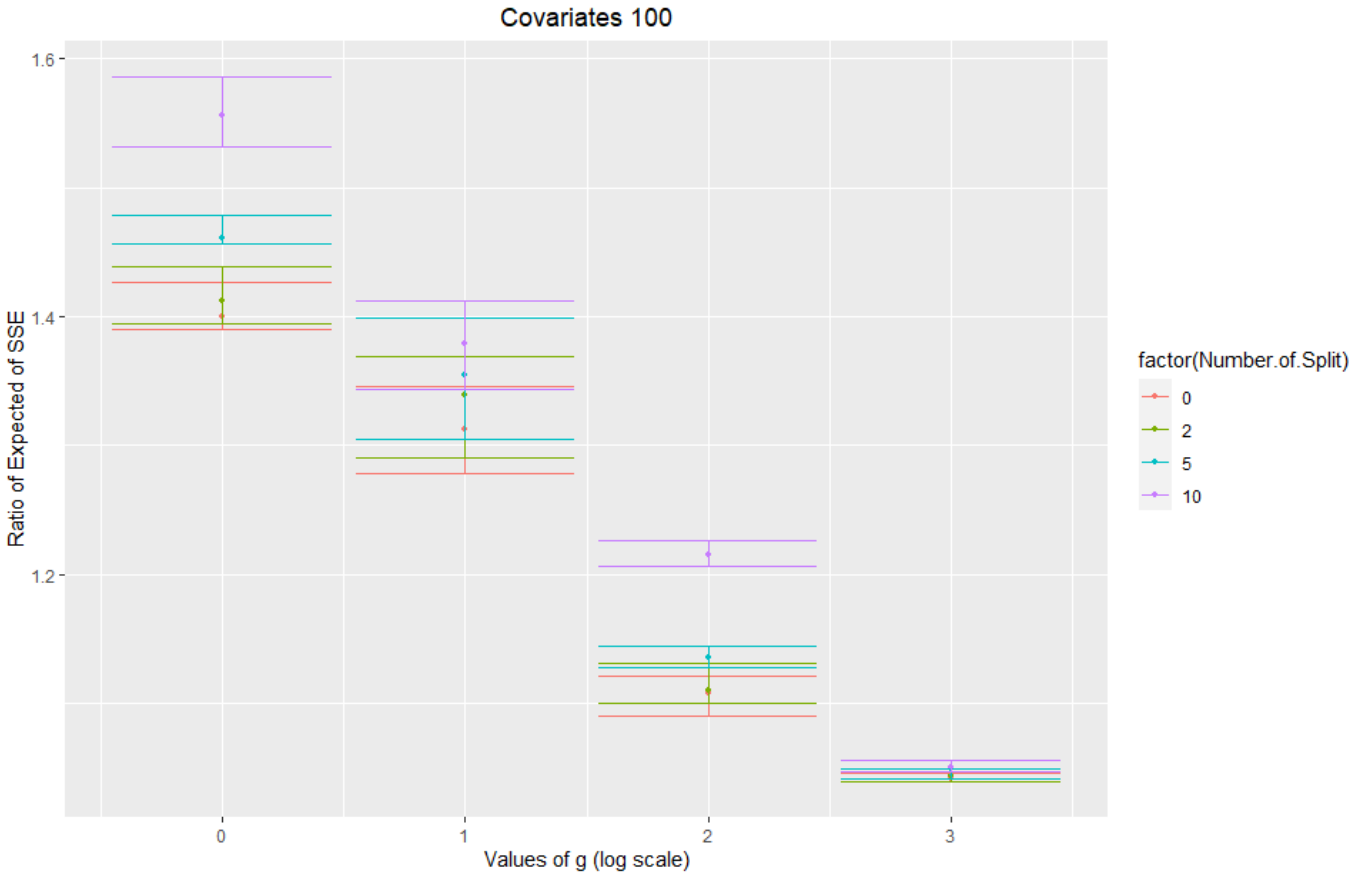


Figure 4.8: The plot of the ratio of Expected SSE under the Bayesian and classical framework with 100 covariates

Chapter 5

Simulation studies

In the following section two simulation studies are presented. They include linear regression models and logistic regression models. The goal of the simulation studies presented here is to study Bayesian variable selection in data sets with realistic features emulating GWAS data.

The criteria for assessment of model performance used in this study are posterior inclusion probabilities.Inclusion probabilities measure the importance of a covariate based on all models in which the covariate is included.

Since SNPs in GWAS are often highly correlated where close to each other on the genome, generating dependent covariates is also considered. The predictors are generated from correlated Binomial distributions using a procedure described in the section below. The genotype at a given SNP contains three possible values, represented in this chapter as 0, 1 or 2.

5.1 Simulation of correlated binomial data

The procedure to generate dependent Binomial random variables we use is based on Bernoulli distribution (Wuber, 2012) :

1. Define the bivariate joint probability function as $P(0,0) = a$, $P(1,0) = 1 - q - a$, $P(0,1) = 1 - p - a$ and $P(1,1) = a + p + q - 1$ where p and q are the probabilities to specify for each Bernoulli random variable, $a = \rho\sqrt{pq(1-p)(1-q)} + pq$ and ρ is the desired correlation between the variables.

2. Given the bivariate joint probabilities, simulate realizations of the random variables and consider the first component as a single Bernoulli random variable and also the second. Then, the resultant two Bernoulli random variables are correlated with each other with a value of ρ.

3. Sum n realizations of each random variable generated in step 2), and finally we obtain 2 Bivariate Binomial random variables correlated with ρ.
   The next step is to generate more than two random variables when the new random variable

is conditional on two random variables that are generated before.

4. Define a new bivariate joint probability function as

$$P(0,0) = \frac{a}{1-p},$$

$$P(1,0) = \frac{1-p-a}{1-p},$$

$$P(0,1)\frac{1-q-a}{p}$$

and

$$P(1,1) = \frac{p+q-1+a}{p}.$$

5. Perform step 2) and 3) with the new probabilities. Finally, the generated variables are correlated Binomial random variables that correlate with the previous random variables.

The idea behind the simulation of random variables is to consider the joint probability of the two random variables under the Bernoulli distribution. Let $X \sim Ber(p)$ and $Y \sim Ber(q)$. Hence, $p(X=0) = 1-p$, $p(X=1) = p$, $p(Y=0) = 1-q$ and $p(Y=1) = q$.

Then the joint probability for four possibilities is given by

$$\begin{aligned} p(X=0, Y=0) &= a, \\ p(X=0, Y=1) &= 1-p-a, \\ p(X=1, Y=0) &= 1-q-a, \\ p(X=1, Y=1) &= p+q-1+a. \end{aligned}$$

using the axiom that the probability of all possible outcomes is 1. Moreover, the sum of the joint probabilities is equal to the marginal probability. For example, let

$$p(X=0) = 0.3, p(X=1) = 0.7, p(Y=0) = 0.4, p(Y=1) = 0.6.$$

Then

$$\begin{aligned} p(X=0, Y=0) &= 0.25, \\ p(X=0, Y=1) &= 0.05, \\ p(X=1, Y=0) &= 0.15, \\ p(X=1, Y=1) &= 0.55. \end{aligned}$$

The crucial part is to specify the level of correlation, the correlation $(\rho_{x,y})$ is defined as

$$\rho_{x,y} = \frac{E(XY) - E(X)E(Y)}{\sqrt{Var(X)Var(Y)}}$$

Since X and Y are Bernoulli, the expectation and variance of X and Y are deduced after the formula from above are put together:

$$E(XY) = \rho_{x,y}\sqrt{pq(1-p)(1-q)} + pq$$

.

Moreover, X and Y are the only terms contributing to $E(XY)$ if $X = 1$ and $Y = 1$. The formula is rearranged in terms of the joint probability concluding correlation (ρ). Thus,

$$a = \rho_{x,y}\sqrt{pq(1-p)(1-q)} + pq - p - q + 1.$$

For simplicity, the above formula is rewritten as

$$a = \rho_{x,y}\sqrt{pq(1-p)(1-q)} + (1-p)(1-q).$$

To simulate more than 2 variables, an additional procedure is required to generate the third variable Z conditional on the value of Y. (and possibly X)

Considering Y as a Bernoulli distribution, there will be 2 possible values: 0 and 1. If Y is 0, there are 2 ways to generate Z. First is to generate 0 with probability

$$\frac{a}{1-p}.$$

The other is to generate 1 with probability

$$\frac{1-p-a}{1-p}.$$

These probabilities are based on the idea of conditional probability. For example,

$$p(Z = 0|Y = 0) = \frac{p(Y = 0, Z = 0)}{p(Y = 0)} = \frac{a}{1-p},$$

the joint probability $p(Y = 0, Z = 0)$ is the same in previous simulation, the marginal distribution $p(Y = 0)$ is based on summing up $p(Y = 0, Z = 0)$ and $p(Y = 0, Z = 1)$.

If Y is 1, there are 2 ways to generate Z. First is to generate 0 with probability

$$\frac{1-q-a}{p},$$

and the other is to generate 1 with probability

$$\frac{p+q-1+a}{p}.$$

These probabilities are based on the same pattern when Y is 0.

The last step will be the same for both procedures, adding the two Bernoulli random variables which will produce a new Binomial random variable that correlates with the previous variable when used in the simulation.

Comparing the relative frequencies of each variable in all simulations, the values were seen to converge to the true proportions of the Binomial distribution. The Chi-square Goodness of fit test was used to test the hypothesis on the equality of proportions in each covariate. The p-value of the test is 0.8136. We can conclude that the observed proportions are not significantly different from the expected proportions.

The empirical correlation can be verified with two cases. The first is through simulation of X and Y. The second is through simulation of Z conditional on Y. The results indicated that the empirical correlation was similar to the true correlation.

5.2 Simulation part 1: linear regression model

The simulation studies will first focus on variable selection in linear regression , based in the Bayesian framework. In the simulation study, the covariates are separated in two groups - the associated covariates that affect the response variable, and the non-associated covariates that do not affect the dependent variable. One characteristic of GWAS data is that SNPs are correlated. Hence, correlated covariates will also be considered in our simulations. For the independent case, covariates are drawn independently. For the correlated cases, where variables are correlated to each other, there will be 3 different patterns of correlation considered. The first is the associated covariate being correlated with other covariates. The second is more than one associated covariate being correlated with other non-associated covariates. The last is where there is correlation among the associated covariates.

5.2.1 Simulation setup

The simulation consists of 500, 1000 and 2000 covariates and there are 500 observations. The Binomial probability parameter p is varied over 0.01, 0.05, 0.1, 0.2 and 0.4. We set the effect size of regressors as 1, 1.5 and 2. x_2, x_8 and x_{12} are assumed to the associated with the response, the others are not.
The simulations involve 3 scenarios as follows:

1. The first case where all covariates are independent.

2. The second case where there are correlations among some covariates. The 3 associated covariates are: x_2 (correlated with x_4 and x_5) x_8 (correlated with x_6 and x_7) and x_{12} (correlated with x_3 and x_9). We set for two levels of different correlation: low (when $\rho = 0.1$) and high (when $\rho = 0.8$).

3. Third, where there are correlations among all the associated covariates x_2, x_8 and x_{12}. The levels of correlation are set the same as the second case (two levels).

For each case, we simulate 10 data sets under each setting, to account for variability across replications.

We use the package *BayesVarSel* in R statistical software in our study. The function *GibbsBvs* in the *BayesVarSel* package is used for Bayesian variable selection in linear regression models.

5.2.2 Model and MCMC diagnostics

The standard normal linear model describes the relationship between the set of all predictors $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_p$ namely $f(\boldsymbol{Y}|\boldsymbol{\beta}, \sigma^2, \boldsymbol{X}) = MN(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 I)$ where $\boldsymbol{Y}$ is a $n \times 1$ vector, $\boldsymbol{X}$ is a $n \times p$ matrix, $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown regression coefficients and σ^2 is an unknown positive scalar. Moreover, there are indicator variables assumed to be given by $\gamma = (\gamma_1, \ldots, \gamma_p)'$ and there are two possible values of each γ_i i.e. 0 and 1 ($i = 1, ..., p$). $\gamma_i = 0$ corresponds to X_i being excluded from the model, and $\gamma_i = 1$ represents that X_i is included in the model .

Before any inference can be made from the MCMC output, convergence diagnostics must be carried out to assess convergence to the stationary distribution. The first diagnostic used is a trace plot. If the MCMC has converged to the posterior distribution, the traceplot will show randomness . The next diagnostic is the autocorrelation plot. Ideally the autocorrelation plot should dramatically decrease after a few lags. The final diagnostic used is the Gelman-Rubin statistic to examine the convergence with multiple chains. If the value is close to 1, one may assume that there is little evidence of non-convergence. A number of convergence diagnostic tools are used as a single tool may not give reliable evidence that no serious violations of MCMC convergence are present. The Figures 5.18 to 5.13 are the example on the autocorrelation plots that decrease after a few lag. Moreover, the Figures 5.14 to 5.16 are the example on the traceplots that show mix random.

The results of the simulation studies are presented in the following sections.

5.2.3 Results of linear regression model

The first set of results are presented as a plot of the marginal inclusion probability in each covariate under 500 covariates (Figures 5.1 to 5.3). The mean Bayes Factor of the true model to the best model and the percentage of best models that correspond to the true model under both the independent and correlated cases is given in Table 5.1 and 5.2. In term of the 80% credible interval, since there are 10 data sets in each setting, we select the 10th percentile as the lower bound and the 90th percentile as the upper bound.

The inclusion probabilities of associated covariates are presented in Figures 5.1 to 5.3. Each row of the plot gives the effect sizes corresponding to the regressors, and the vertical axis in each plot represents the inclusion probability. The horizontal axis in each plot represents the covariates.

The different colour in each bar represents the difference values in the probability parameter of generating in each covariate from a Binomial distribution.

Computing the marginal inclusion probability of each variable helps determine whether the variable should be included in the model.

From the first column of Figure 5.1, we see that when p increases, so does the inclusion probability. The study also suggests higher inclusion probabilities with higher effect size as shown through the increasing trend in the first column of Figure 5.1. When the effect size is low ($\boldsymbol{\beta}$ = 0.1), the inclusion probabilities are also low for x_2, x_8 and x_{12} (illustrated in the first case in the first plot of Figures 5.2 and 5.3).

Under the scenario with correlation, the inclusion probabilities under the low correlation (ρ = 0.1) are higher than that of the high correlation (ρ = 0.8). Most cases are in line with this hypothesis except in the case of low effect size ($\boldsymbol{\beta}$ = 0.1) shown in the first row of Figures 5.1 to 5.3. A similar result is witnessed when the probability p in the Binomial distribution is increased. Inclusion probabilities increase as p is increased.

In the third case (correlation among all associated covariates) and the last case (two associated covariates correlating with another covariate), the study shows an increasing inclusion probability when p increases (shown in the 1st, 2nd, 3rd and 4th column of Figures 5.2 and 5.3). Again, the inclusion probabilities are higher with high effect size, and lower with low effect size ($\boldsymbol{\beta}$ = 0.1), illustrated in the first row in the 1st, 2nd, 3rd and 4th column of Figures 5.2 and 5.3. These results show a similar trend under other correlated cases.

To confirm these findings, the Bayes Factor is calculated for all cases, where the true model is compared to the best model found (Table 5.1). The true model contains (x_2, x_8, x_{12}) . The mean of the Bayes Factor is then considered with 80% credible interval of BF. There are 10 values of Bayes Factor where the lower bound is the 10th percentile and the upper bound is the 90th percentile. If BF =1, it implies that the best model found is the true model. Under lower effect sizes ($\boldsymbol{\beta} = 0.1, 0.2$) and the probability setting ($p = 0.01, 0.05$), the mean BF is lower than 1. This implies that the best model is not the true model (shown in Table 1.1) . However, with higher effect sizes ($\boldsymbol{\beta} = 0.5, 1, 1.5$) and the probability setting ($p = 0.2, 0.4$), the BF means are close to 1 (shown in Table 1.1). Again, the mean has a similar tendency , it implies that the best model found that is not the true model under the low effect sizes ($\boldsymbol{\beta} = 0.1, 0.2$) and the probability setting($p = 0.01, 0.05$).

Under the independent case, when the effect size is increased, the inclusion probability remains the same. This is expected as the higher effect size reflects a better performance of covariate selection. Moreover, the percentage of cases where the true model is found keeps increasing as the Binomial p increases. This is shown in Table 5.1.

Under the correlated situation (one associated covariate with two other non-associated covariates), the percentage of times the best model was the true model is slightly higher than in the independent case (shown in Table 5.2). The correlated data reduces the performance of the model selection due to the multicollinearity issue. The percentage of times the true model is found in the case with high correlation ($\rho = 0.8$) is lower than that of low correlation ($\rho = 0.1$). With high correlation, the chance of selecting the true model is reduced due to correlations between covariates (shown in Table 5.2). In terms of the Bayes Factor, these values are 0, implying that the best model is the null model. Thus, BF under the true model compared to the best model has lower values (shown in the first row of Table 5.2). However, when effect size and probability setting

value increases, the Bayes Factor for the best model and the true model is slightly different. This is particularly true when the effect size and probability setting value are similarly high. In that case, the percentage of correct variable selection is 100 (shown in the last row of Table 5.2).

The case where all associated covariates are correlated is the most complex as there are correlations among associated covariates. Hence, the Bayes Factor between models is 0 in cases where the effect size and values of p are low (shown in the first row of Table 5.2). Hence the best model is the null model. Generally, we see more null models as the best found for high correlation cases. In the case of low effect size ($\boldsymbol{\beta} = 0.1, 0.2$) , we witnessed null models frequently (shown in Table 5.2).

Figure 5.1: The marginal inclusion probability in each associated covariate (x_2, x_8, x_{12}) when there were 500 covariates in linear regression (under the independent case and first of two correlated cases)

The title in each plot is the specification on the independent case and the correlated cases with $\rho = 0.1$ or $\rho = 0.8$. Each row of the plots is the effect sizes on the regressor($\boldsymbol{\beta} = 0.10, 0.20, 0.50, 1$ and 1.5). The vertical axis in each plot represents the inclusion probability. The horizontal axis in each plot represents the covariates. The difference colour in each bar represents the values in probability ($p = 0.01, 0.05, 0.10, 0.20$ and 0.40) of generating in each covariate from a Binomial distribution. In terms of the title in each plot, 1st is under the independent case, other (2nd, 3rd, 4th, 5th and 6th) are under the correlated cases.

Figure 5.2: The marginal inclusion probability in each associated covariate (x_2, x_8, x_{12}) when there were 500 covariates in linear regression (under the three correlated cases)

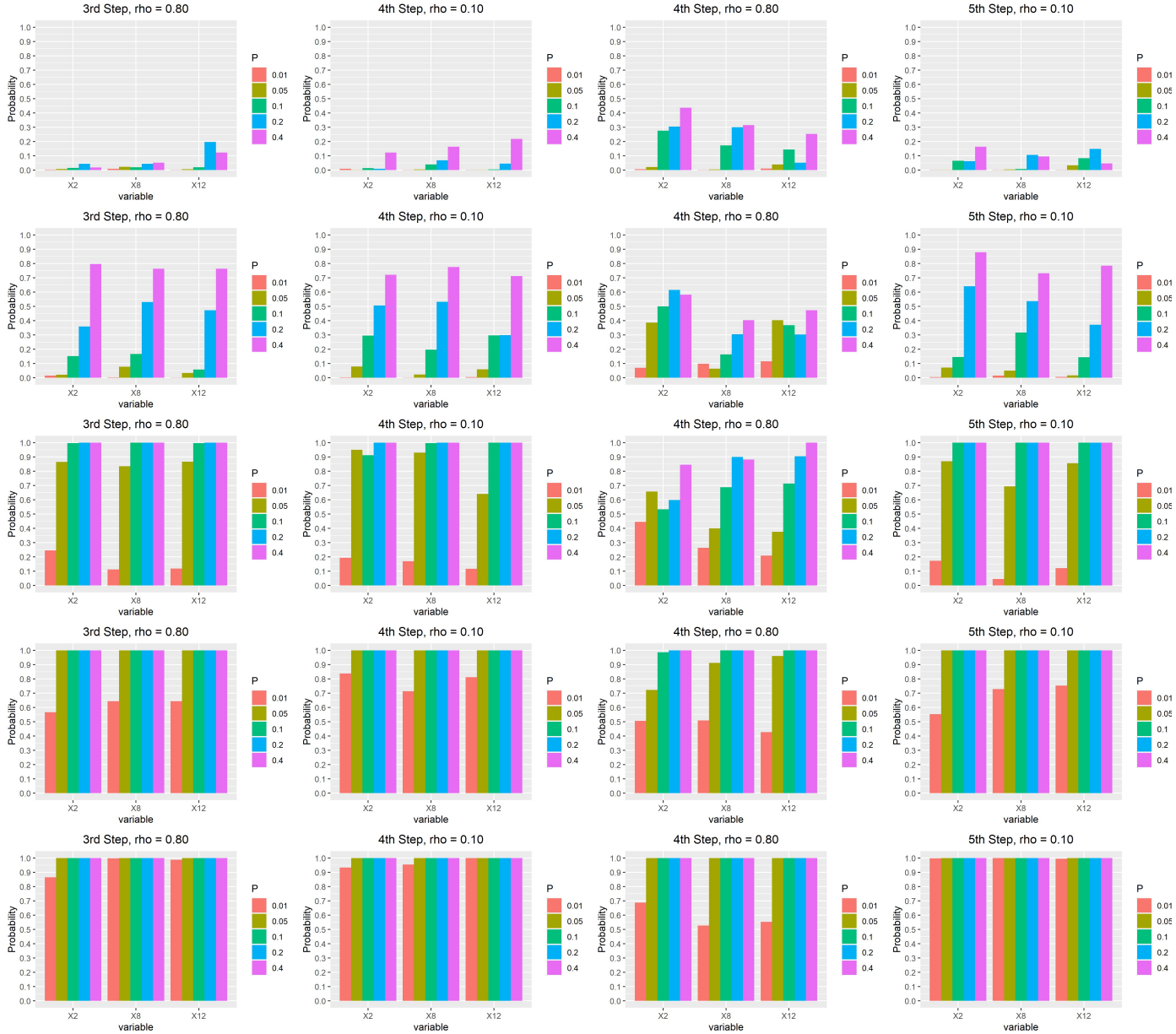
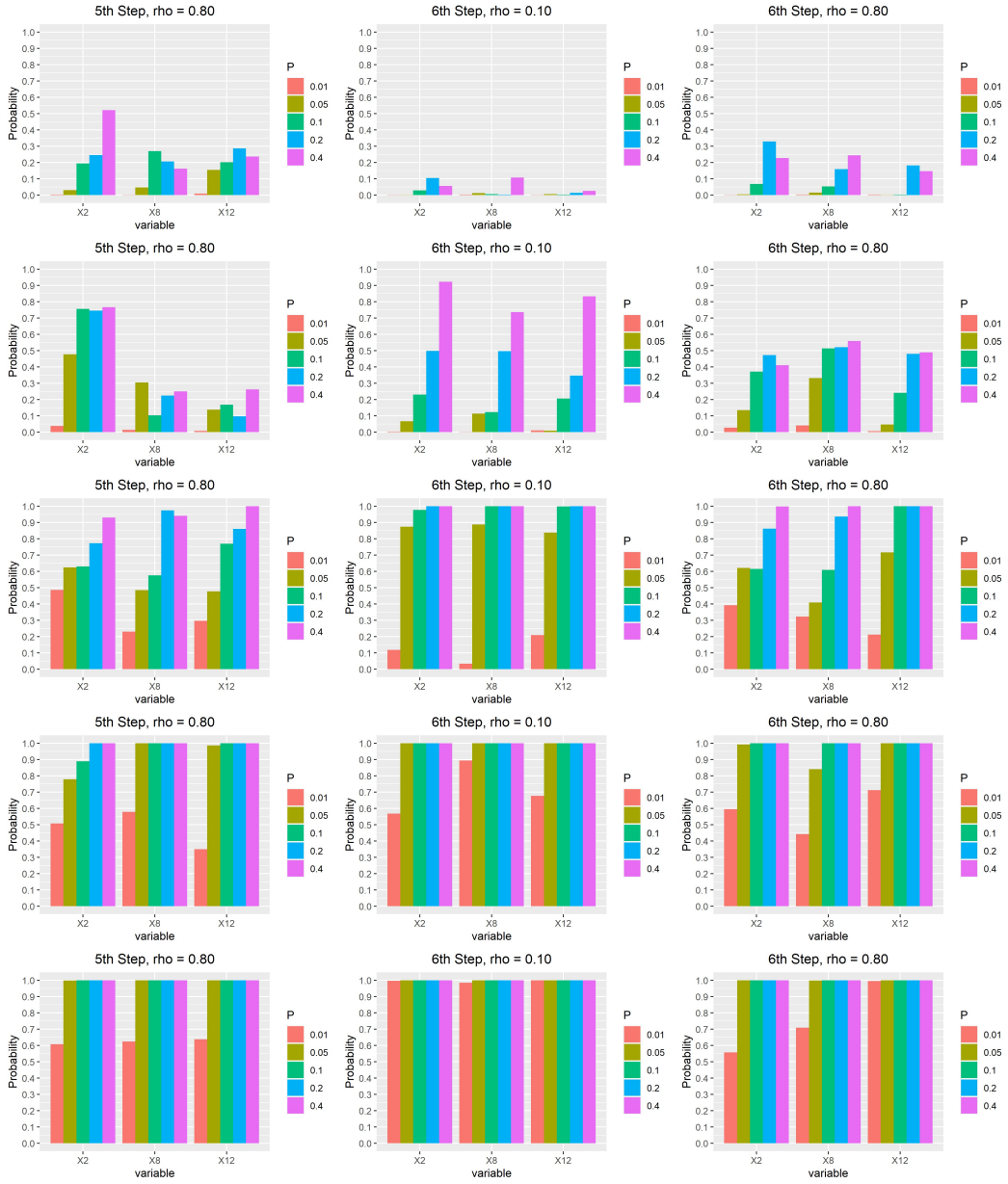


Figure 5.3: The marginal inclusion probability in each associated covariate (x_2, x_8, x_{12}) when there were 500 covariates in linear regression (under the last two correlated cases)



5.2.4 Summary

The simulation studies showed that the inclusion probabilities of associated covariates (i.e. x_2, x_8, x_{12}) are increased when the effect size (β) increses, which is to be expected. These results are shown in the first column on Figure 1 under the independent case and they are reported in other columns on Figure 5.1 for the correlated cases.

The percentage of times the true model is found in the case with high correlation among covariates ($\rho = 0.8$) is lower than that of low correlation ($\rho = 0.1$) (Table 5.1 and 5.2). The high correlation therefore has an impact on reducting the performance of model selection due to multicollinearity .

For the model convergence diagnostic, we present the effective sample size (ESS) which is high for associated covariates (x_2 , x_8 and x_{12}) (Table 5.3 and 5.4).

The effective sample size is the part for estimating the mean, for a time series X of length N

Table 5.1: The mean Bayes Factor and 80% credible interval of BF of the true model compared to the best model The % denotes the percentage of the best model that was the true model (independent covariates)

Beta	p=0.01	p=0.05	p=0.10
0.10	0.0010(0.0009,0.0029)	0.0025(0.0014,0.0037)	0.0158(0.0019,0.0384)
%	0	0	0
0.20	0.09(0.02,0.13)	0.16(0.06,0.33)	0.46(0.35,0.98)
%	0	0	0
0.50	0.52(0.11,0.93)	0.60(0.35,0.94)	0.95(0.82,0.96)
%	0	70	90
1.00	0.98(0.91,1)	0.99(0.97,1)	1(1,1)
%	80	90	100
1.50	0.99(0.95,1)	1(1,1)	1(1,1)
%	80	100	100

Beta	p=0.20	p=0.40
0.10	0.0214(0.0037,0.0663)	0.0310(0.0053,0.0917)
%	0	0
0.20	0.88(0.84,1)	0.95(0.92,1)
%	0	30
0.50	1(1,1)	1(1,1)
%	100	100
1.00	1(1,1)	1(1,1)
%	100	100
1.50	1(1,1)	1(1,1)
%	100	100

the standard error for the mean is $\sqrt{Var(X)/n}$ where n is the effective sample size (Plummer et al., 2022).

Table 5.2: The mean Bayes Factor and 80% credible interval of the BF of the true model compared to the best model. The % denotes the percentage of the best model that was the true model (correlated case and $\rho = 0.8$)

Beta	p=0.01	p=0.05	p=0.10
0.10	0.0001(0.0001,0.0004)	0.0006(0.0003,0.0009)	0.0065(0.0016,0.0072)
%	0	0	0
0.20	0.03(0.01,0.07)	0.04(-0.55,0.09)	0.20(0.06,0.33)
%	0	0	0
0.50	0.14(-0.16,0.44)	0.43(0.17,0.64)	0.52(0.19,0.77)
%	0	0	0
1.00	0.79(0.21,0.80)	0.96(0.91,1)	1(1,1)
%	0	80	100
1.50	0.81(0.44,0.85)	1(1,1)	1(1,1)
%	10	100	100

Beta	p=0.20	p=0.40
0.10	0.0246(0.0031,0.0362)	0.0257(0.0036,0.0277)
%	0	0
0.20	0.35(0.21,0.53)	0.47(0.39,0.73)
%	0	0
0.50	0.79(0.56,0.89)	1(1,1)
%	60	100
1.00	1(1,1)	1(1,1)
%	100	100
1.50	1(1,1)	1(1,1)
%	100	100

Table 5.3: The median, the 10th percentile and the 90th percentile of Effective Sample Size (ESS) (independent covariates)

Beta	p=0.01	p=0.05	p=0.10
0.1	9296.70(9013.15,9494.81)	9547.39(8978.77,10065.78)	9221.95(8611.50,10215.54)
0.2	9217.53(9029.57,10076.78)	9431.68(8599.21,10684.17)	9271.63(9211.58,10506.26)
0.5	9155.20(8960.54,10055.48)	10000(9345.01,10576.58)	9422.67(9200.04,10000)
1	9493.82(8525.86,10052.19)	9294.44(8904.14,10000)	9252.60(8884.34,9413.74)
1.5	9773.90(9474.33,10379)	9334.40(8939.03,10034.88)	9464.85(8987.66,9602.01)

Beta	p=0.20	p=0.40
0.1	9867.17(8988.35,10317.22)	10000(9587.25,10497.06)
0.2	9286.86(9256.42,10008.96)	10000(9232.64,10047.65)
0.5	9315.74(8886.96,9984.22)	9190.39(8812.41,9685.50)
1	9388.81(9112.81,9737.32)	9193.34(9098.84,9674.83)
1.5	9316.54(8876.12,10167.15)	9114.42(8643.73,9583.83)

Table 5.4: The median, the 10th percentile and the 90th percentile of Effective Sample Size (ESS) (correlated case)

Rho	Beta	p=0.01	p=0.05	p=0.10
0.1	0.1	9471.82(9049.45,10403.66)	9530.59(9092.17,10081.86)	9320.86(9060.55,10065.52)
0.8	0.1	9283.80(8876.42,10034.65)	9183.89(8491.47,10632.70)	9664.99(9016.79,10715.68)
0.1	0.2	9162.98(8873.15,9844.85)	9392.31(9186.75,10601)	9936.34(9385.77,10332.92)
0.8	0.2	9470.55(9118.23,10614.19)	9939.39(9648.56,10097.22)	9018.90(9032.25,9840.24)
0.1	0.5	10000(8806.14,10790.87)	10000(9362.43,10030.15)	9294.44(8783.09,9566.99)
0.8	0.5	9094.39(9029.01,9913.41)	8921.62(8515.11,9059.89)	8800.25(8200.51,9408.75)
0.1	1	9966.29(8448.93,10939.99)	8975.98(8827.72,9426.42)	9116.67(8899.83,9574.33)
0.8	1	9300.23(8405.82,9124.24)	9837.32(9506.39,10032.86)	9430.48(9089.53,10054.53)
0.1	1.5	9445.91(8788.68,10003.66)	9422.95(8819.98,10000)	9162.63(8987.32,9349.62)
0.8	1.5	9061.84(8518.82,10546.27)	9241.18(8900.21,10000)	9335.11(8901.33,10000)

Rho	Beta	p=0.20	p=0.40
0.1	0.1	9693.34(9246.26,10489.90)	9961.79(9701.27,10400.56)
0.8	0.1	9735.37(9130.24,9900.19)	9086.58(9029.34,9410.19)
0.1	0.2	9989(9700.06,10461.98)	9980.99(9883.96,10252.73)
0.8	0.2	9434.07(8691.38,9740.07)	9517.57(9074.70,10031.98)
0.1	0.5	9375.87(9078.89,10000)	9178.71(8709.79,9528)
0.8	0.5	10000(9639.44,10534.42)	9849.45(9204.73,10489.61)
0.1	1	9138.62(8670.79,10160.10)	9083.16(8871.53,9524.95)
0.8	1	9279.65(9007.49,10305,98)	8992.91(8421.63,9173.96)
0.1	1.5	9154.72(8659.00,10000)	9300.57(8973.06,10039.20)
0.8	1.5	9185.15(8789.14,9373.72)	9098.22(8703.02,10015.27)

5.3 Simulation part 2: logistic regression model

The section will describe investigations into the performance of Bayesian variable selection in logistic regression models where the response is binary (categorical) and the covariate settings are similar to those discussed in Section 5.2.

5.3.1 Simulation setup

The simulation consists of p =500, 1000 and 2000 covariates and there are n =500 observations. The Binomial probability (for covariates) is set at 0.1. We set the effect size of regressors as 1, 1.5 and 2. At each setting, we generated 10 data sets. In each case, we found that the inclusion probability was very low. Hence, we set higher values in each case. The simulation involves 3 scenarios, similar to the linear regression model. x_1, x_2 and x_3 are assumed to the associated with the response, the others are not.

For the variable selection we use the $logisticVS$ function in the $bvsflex$ package, setting the number of iterations as 200000 based on pilot runs indicating a longer time is needed for MCMC convergence. There are several parameters in the package. The prior mean for $\boldsymbol{\beta}$ is set to 0. We conduct sensitivity studies by changing values of many hyper-parameters, where g for the prior covariance matrix for $\boldsymbol{\beta}$ is set at 0.1, 1, 10 and 100, the prior precision for $\boldsymbol{\beta}$ is set at 100, and the prior mean of Beta prior distribution is 0.06. Under the package, the additional hyper-prior distribution for π is given by $\pi \sim Beta(a, b)$. In addition, there are 3 classes of prior distribution for the hyper-parameter g which are inverse-gamma, hyper-g and none. We choose none which implies that g is assumed to be fixed at the value specified. The $bvsflex$ package on R-forge is used for the variable selection in logistic regression models. -(http://bvsflex.r-forge.r-project.org) (Zucknick, 2013).

We consider a simulation scenario where we have $p = 500$, 1000 and 2000 covariates and there are $n = 500$ observations. Each explanatory variable is generated from a Binomial distribution with parameters $(2, p)$ where p takes the value 0.01. The effect size for the regressor is 2. The dependent variable is generated from the Binomial distribution with n=1 (for the binary values) . Prior to generating response variables, the input variables (x_i) are centered to mean-0 as this model used does not contain an intercept.

With many possible combinations of settings, a few cases are chosen for presentation below.

5.3.2 Results of logistic regression model

These results are shown in terms of inclusion probability of each covariate for the logistic regression model with $g = 1$, both under the independent and correlated cases (Table 5.5). The plot of confidence interval and credible intervals of the regression coefficients under the independent case are shown in Figure 5.4.

The inclusion probability is high when the effect size is high for associated covariate (i.e. x_1, x_2, x_3) with 500 covariates under the independent case. Moreover, the inclusion probability for other covariates are less than those for associated covariates in all situations, since the true model contains all 3 covariates as depicted in Figure 5.5, 5.6 and 5.7. Moreover, when the effect size ($\boldsymbol{\beta}$) increases the inclusion probability increases. However, in some cases with the lower effect size,

i.e., β=1, the inclusion probability of associated covariates is quite low. These values are reported in Figure 3 when $\beta = 1$. Moreover, the inclusion probability is consistently high when the effect size is higher ($\beta = 1.5, 2$) as shown in those figures when $\beta = 1.5, 2$ (Figure 5.6 and 5.7).

In terms of the prior, the posterior inclusion probability appeared higher under the less informative prior, i.e. $g = 50, 70, 80$. However, the inclusion probability was lower under the more informative prior, i.e. $g = 1, 10$. This is due to the fact that the more informative prior has an impact on the posterior distribution.

Moreover, Figure 5.14, 5.15 and 5.16 show the traceplot of all iterations under some settings. Each traceplot confirms that inclusion probabilities are low at low effect size ($\beta = 1$). However, the inclusion probabilities are high with higher effect sizes ($\beta = 2$), as seen in the the third panel for each traceplot.

We further estimate the posterior distribution through the credible intervals of the regression coefficients which are presented in plots of the credible interval separately in each simulation setting when considering just associated covariates (x_1, x_2, x_3) (Figure 5.17, 5.18 and 5.19). These figures indicate that most credible interval covered the estimated values by the glm fitting.

Moreover, autocorrelation plots were considered to diagnose possible lack of MCMC convergence. (Figures 5.8 to 5.13)

In term of the autocorrelation plot, most autocorrelations decrease dramatically after lag 1 . Hence, there does not seem to be evidence of non-convergence to the stationary distribution. However, some plots with $\beta = 2$ do not decrease after lag 1 as shown in Figures 5.8 and 5.12.

The plots of the credible intervals of regression coefficients are shown separately in each simulation setting for only the associated covariates (x_1, x_2, x_3). The results show that all credible intervals cover the true value when the glm is fit. The red lines represent the lower bound and the upper bound for confidence intervals from a logistic regression fitted with only x_1, x_2, x_3. The blue box plots represent the credible intervals from the MCMC output, and the green line illustrates the true values of the simulation setting. These results are reported in Figure 5.4.

Figure 5.4: The plot of the confidence interval under the associated covariates (x_1, x_2, x_3) situation and under the independent case in logistic regression (The red lines represent the lower bound and the upper bound for glm fitting. The blue box plots represent the confidence interval of MCMC output where the green line illustrates the true values of the simulation setting.)

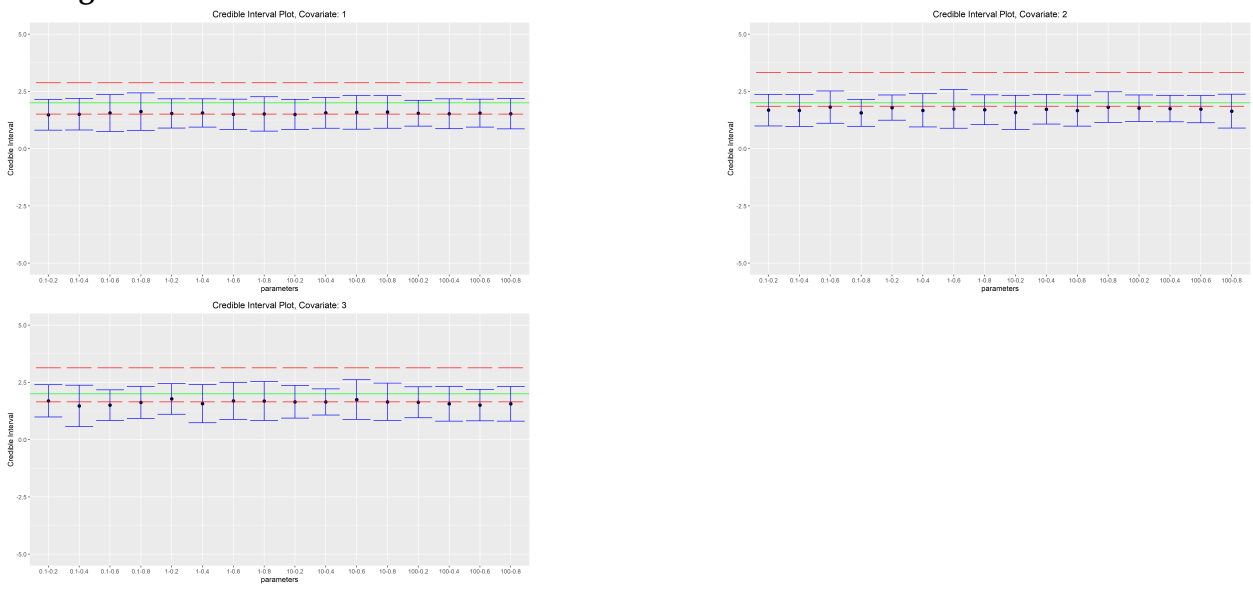


Figure 5.5: The boxplots of inclusion probability for β_1, β_2, β_3 and $g = 10$, $\mu = 0.06$ under the independent case (ten data sets) when $\beta = 1$ from 500, 1000 and 2000 covariates

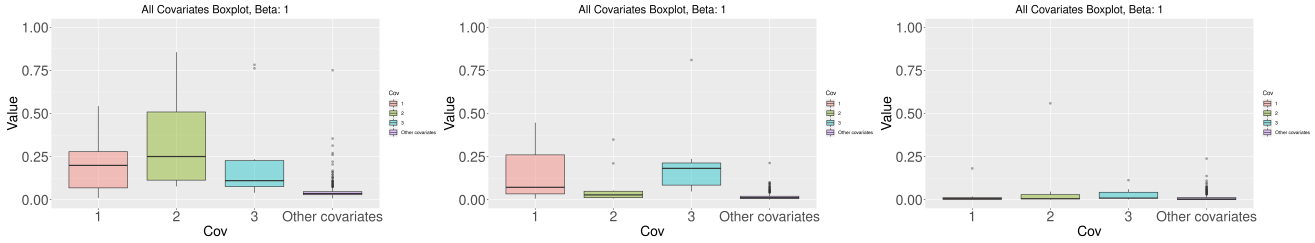


Figure 5.6 : The boxplots of inclusion probability for β_1, β_2, β_3 and $g = 10$, $\mu = 0.06$ under the independent case (ten data sets) when $\beta = 1.5$ from 500, 1000 and 2000 covariates

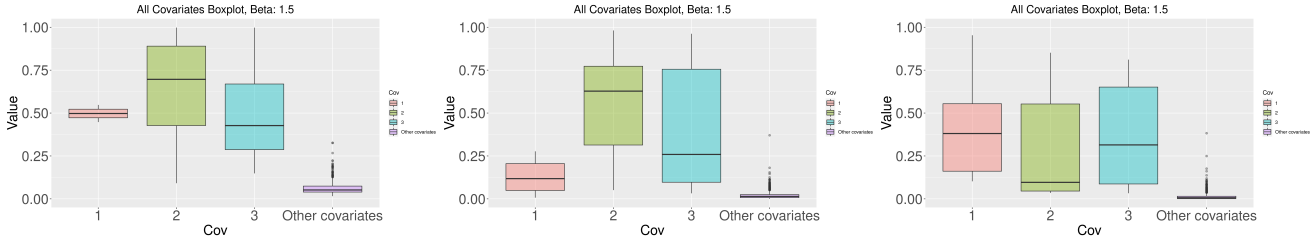


Figure 5.7 : The boxplots of inclusion probability for β_1, β_2, β_3 and $g = 10$, $\mu = 0.06$ under the independent case (ten data sets) when $\beta = 2$ from 500, 1000 and 2000 covariates

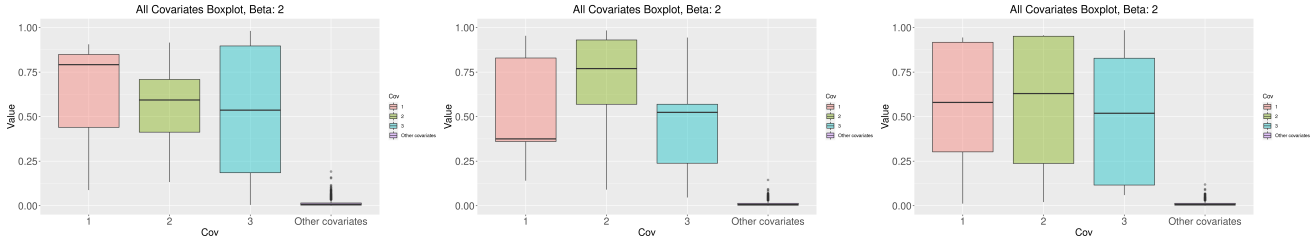
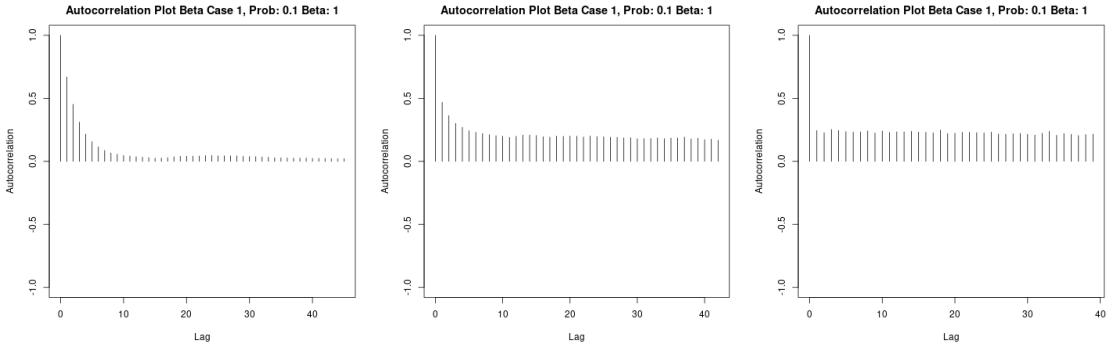
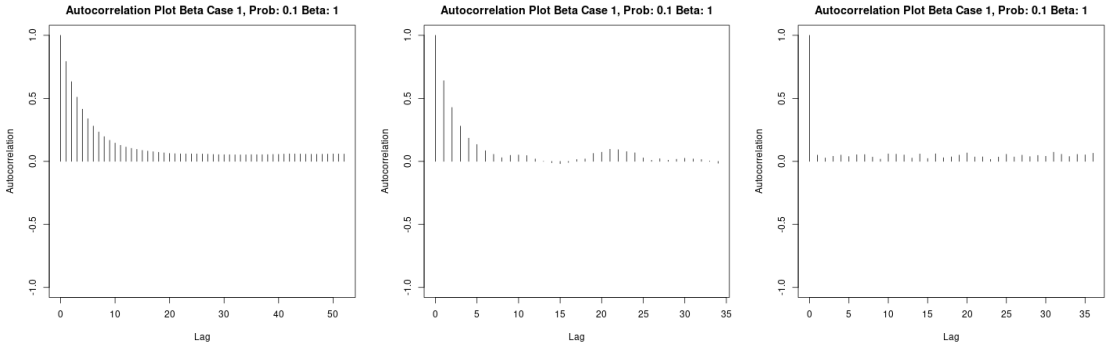


Figure 5.8: The acfplot of β and $g = 10$, $\mu = 0.06$ when $\beta = 1$ (single data set) from 500 covari-



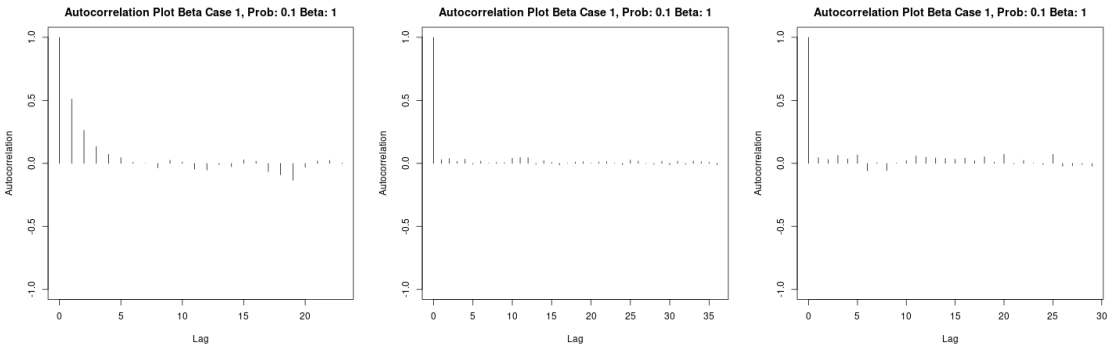
ates

Figure 5.9: The acfplot of β and $g = 10$, $\mu = 0.06$ when $\beta = 1$ (single data set) from 1000 covari-



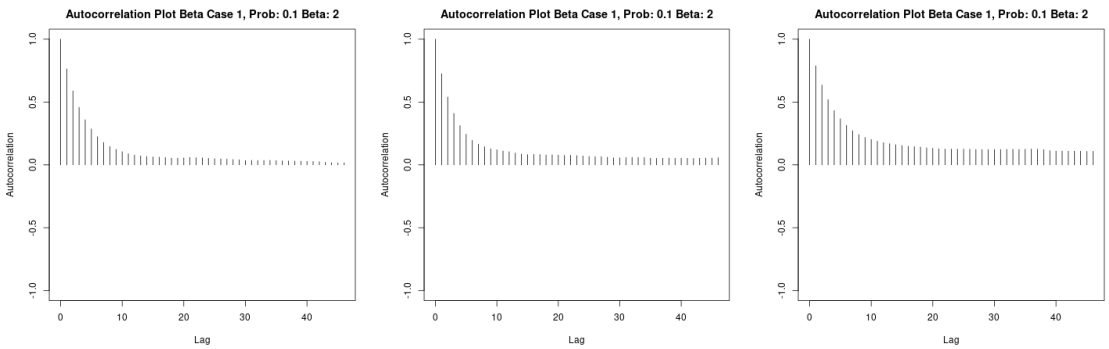
ates

Figure 5.10: The acfplot of β and $g = 10$, $\mu = 0.06$ when $\beta = 1$ (single data set) from 2000 co-



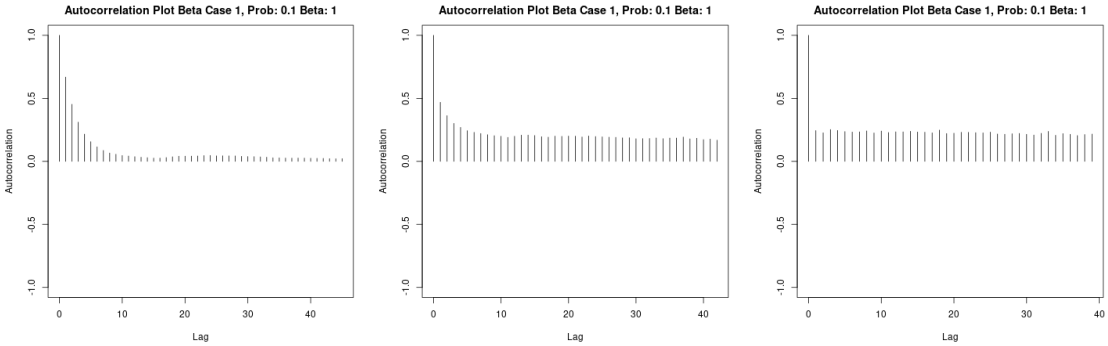
variates

Figure 5.11: The acfplot of β and $g = 10$, $\mu = 0.06$ when $\beta = 2$ (single data set) from 500 covari-



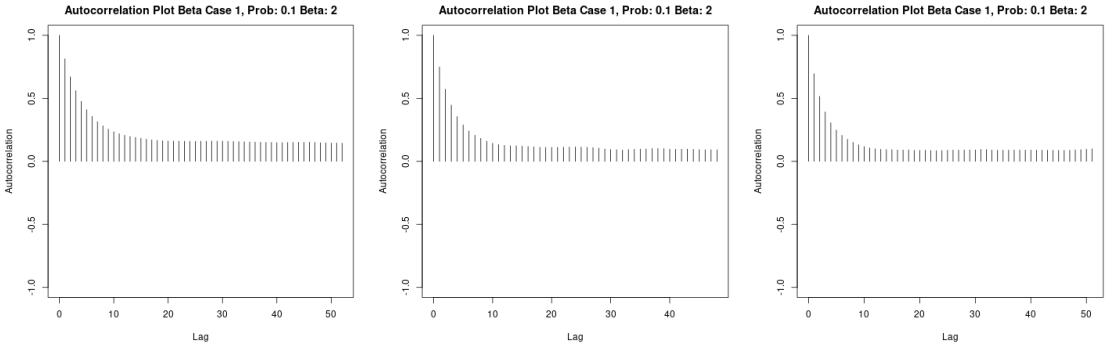
ates

Figure 5.12: The acfplot of β and $g = 10$, $\mu = 0.06$ when $\beta = 2$ (single data set) from 1000 co-



variates

Figure 5.13: The acfplot of β and $g = 10$, $\mu = 0.06$ when $\beta = 2$ (single data set) from 2000 co-



variates

Figure 5.14: The traceplot of $\beta = 2$ and $g = 10$, $\mu = 0.06$ (single data set) from 500 covariates

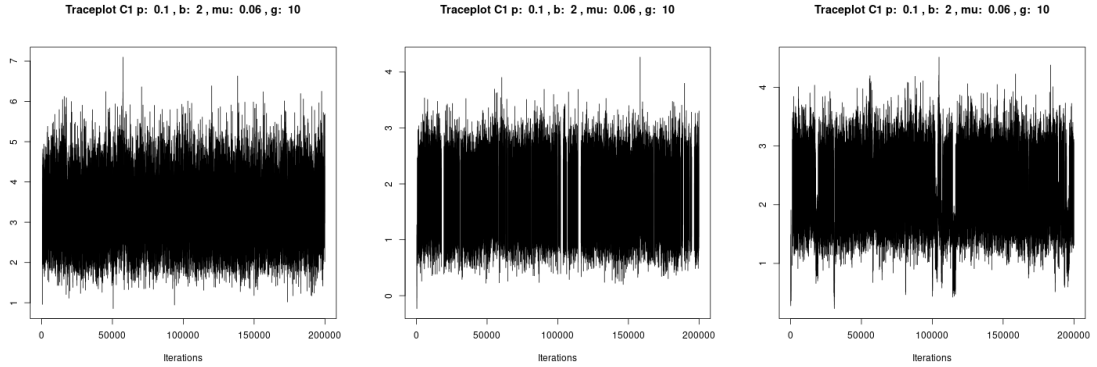


Figure 5.15: The traceplot of $\beta = 2$ and $g = 10$, $\mu = 0.06$ (single data set) from 1000 covariates

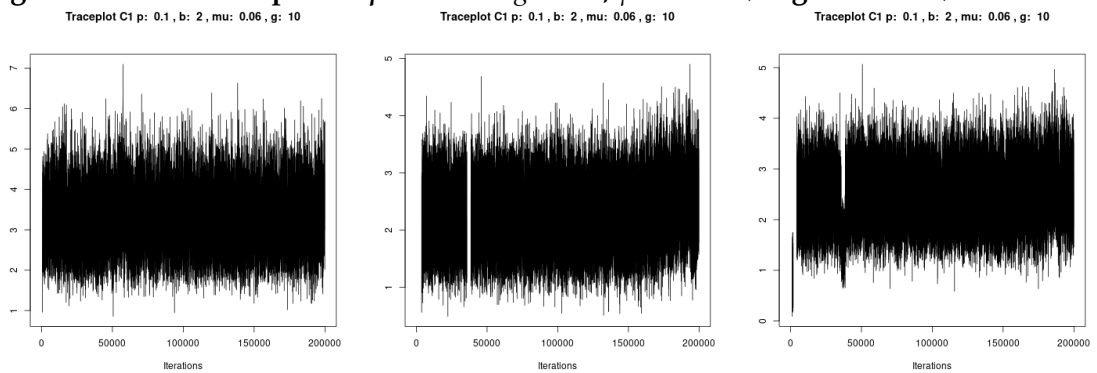
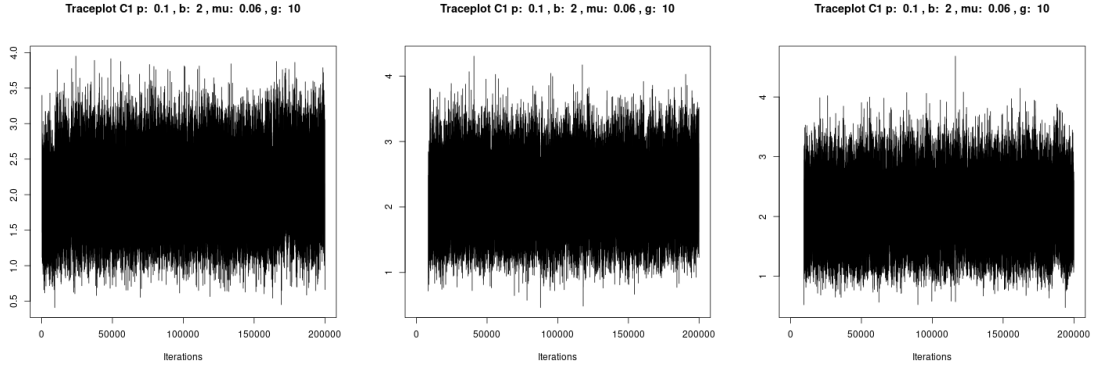


Figure 5.16 : The traceplot of $\beta = 2$ and $g = 10$, $\mu = 0.06$ (single data set) from 2000 covariates



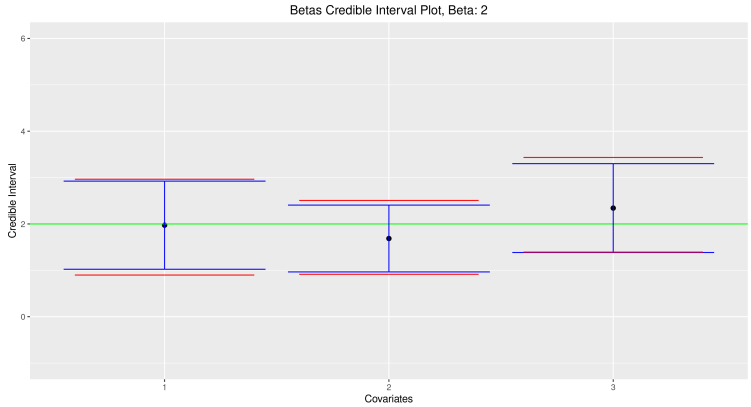


Figure 5.17 : The plot of the credible interval of the associated covariates when $\beta = 2$ under the independent case from 500 covariates

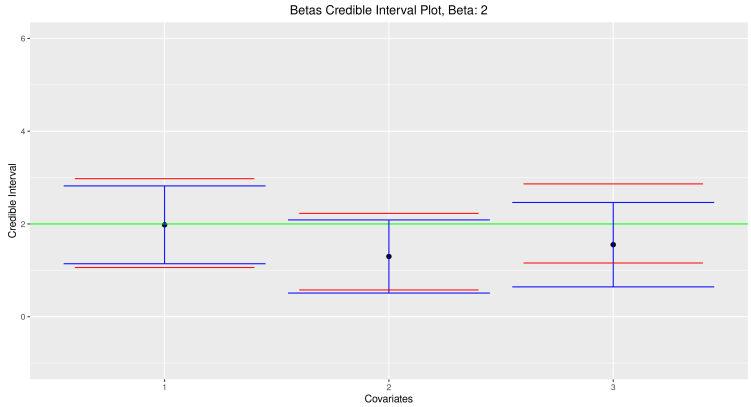


Figure 5.18: The plot of the credible interval of the associated covariates when $\beta = 2$ under the independent case from 1000 covariates

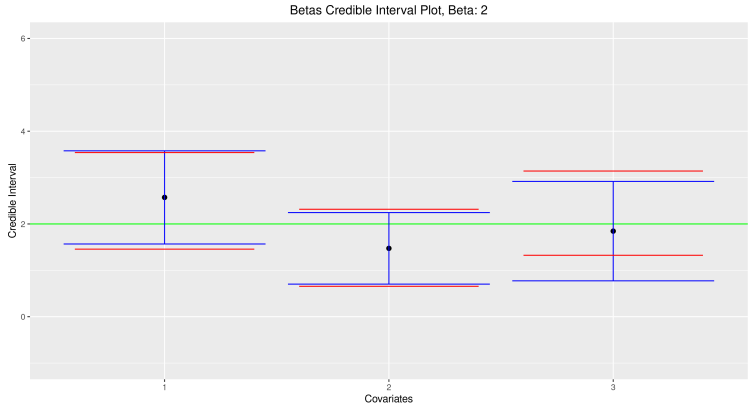


Figure 5.19 : The plot of the credible interval of the associated covariates when $\beta = 2$ under the independent case from 2000 covariates

Table 5.5: The inclusion probability for each covariate for the logistic regression model when $g = 10$ under the correlated case

Prior mean μ	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}
0.2	0.28	0.29	0.30	0.23	0.21	0.19	0.16	0.08	0.06	0.08	0.08	0.07
0.4	0.27	0.29	0.28	0.21	0.29	0.11	0.19	0.13	0.14	0.06	0.08	0.15
0.6	0.25	0.21	0.29	0.21	0.17	0.26	0.18	0.16	0.13	0.12	0.17	0.16
0.8	0.25	0.23	0.27	0.20	0.26	0.19	0.22	0.18	0.13	0.17	0.18	0.17

5.3.3 Summary

Our simulation studies showed that, as expected, the posterior inclusion probabilities of associated covariates (x_1, x_2, x_3) are increased when the effect size (β) is increased. However, the other covariates (non-associated covariates) are very low on the posterior inclusion probabilities. These results are shown in Figures 5.5 to 5.7 (under the independent case) and reported in Figures 5.20 to 5.22 (under the correlated cases.) These simulation studies can be applied on GWAS datasets when the number of associated covariates is small, with moderately large effect sizes, even though the total number of covariates may be large, say in the thousands, with a lower sample size. However, the simulation studies should be extended to higher numbers of covariates and sample size. Also, as we have shown in Chapter 4, with larger datasets, one can use the split-merge approach for greater efficiency without too much added bias.

The credible intervals (CI's) for the estimation for β from the MCMC method were compared with confidence intervals from the glm fitting. We found that CI's from higher effect sizes ($\beta = 2$) credible intervals under the MCMC method and under the glm fitting are similar. They were more accurate when the effect size was higher. The results are on the lower effect sizes ($\beta = 1, 1.5$) which are reported in Appendix. However, they can be applied on the lower effect size

When the overall number of covariates increase, the inclusion probabilities decrease for each covariate. These are reported in Figures 5.5 to 5.7. This is understandable as the variable selection search space increases massively with an increase in p.

Model diagnostics indicate that the posterior samples are likely to be taken from the stationary distribution.

In the effective sample size (ESS) is high for some covariates especially x_1 but it is low for other associated covariates (x_2 and x_3). (Figures 5.25 to 5.28).

The results of simulation studies in this Chapter indicated that Bayesian Variable Selection (BVS) can deal with the multicollinearity problem when the number of covariates is higher than number of observations ($p >> n$), as in the situation with GWAS.

These simulation studies, contain about a few thousand covariates and 500 observations which is somewhat smaller than typical sizes of real datasets. However, with the split-and-merge approach, one can reduce the dimension of each split datasets to make the problem comparable. We next apply our methods to the real dataset in Chapter 6.

Figure 5.20: The boxplots of inclusion probability forβ_1, β_2, β_3 and $g = 10$, $mu = 0.06$ under the second correlated case (x_1 with x_4, x_5, x_2 with x_6, x_7, x_3 with x_8, x_9) when $\beta = 1$, $\rho = 0.1$ from 500, 1000 and 2000 covariates

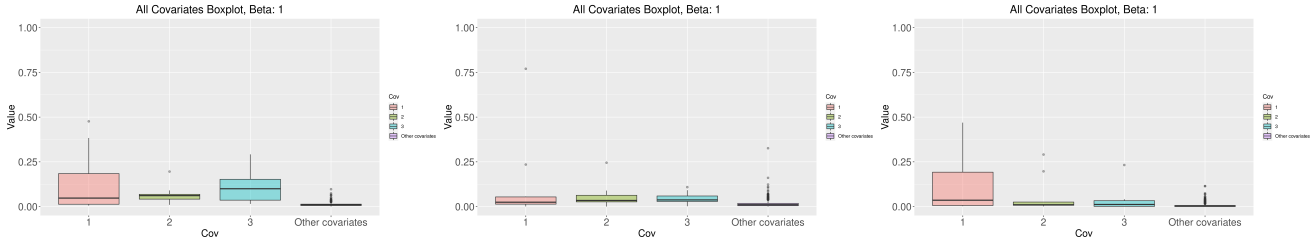


Figure 5.21 : The boxplots of inclusion probability forβ_1, β_2, β_3 and $g = 10$, $mu = 0.06$ under the second correlated case (x_1 with x_4, x_5, x_2 with x_6, x_7, x_3 with x_8, x_9) when $\beta = 1.5$, $\rho = 0.1$ from 500, 1000 and 2000 covariates

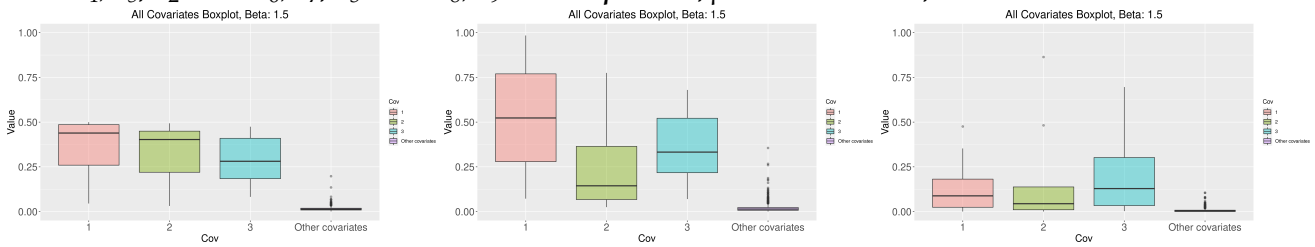


Table 5.6: The inclusion probability for each covariate for the logistic regression model when $g = 1$ and under the independent case

Prior mean μ	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}
0.2	0.56	0.31	0.53	0.11	0.12	0.13	0.15	0.10	0.14	0.13	0.12	0.12
0.4	0.49	0.47	0.50	0.23	0.13	0.14	0.17	0.13	0.19	0.16	0.15	0.13
0.6	0.40	0.42	0.46	0.27	0.19	0.17	0.16	0.24	0.24	0.16	0.19	0.21
0.8	0.39	0.40	0.45	0.29	0.19	0.24	0.18	0.24	0.25	0.24	0.17	0.19

Figure 5.22 : The boxplots of inclusion probability for β_1, β_2, β_3 and $g = 10$, $mu = 0.06$ under the second correlated case (x_1 with x_4, x_5, x_2 with x_6, x_7, x_3 with x_8, x_9) when $\beta = 2$, $\rho = 0.1$ from 500, 1000 and 2000 covariates

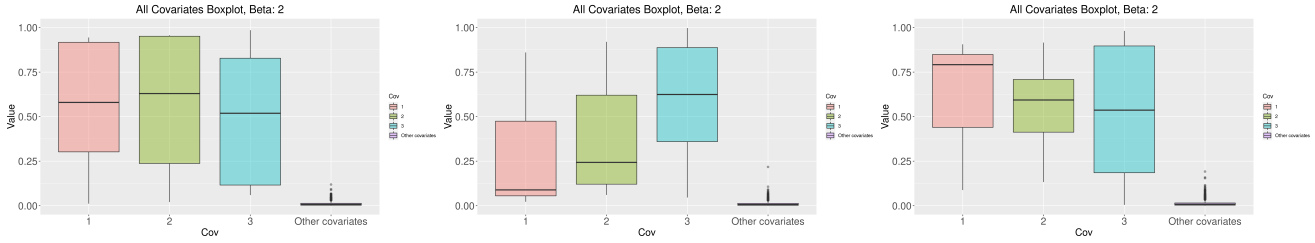


Figure 5.23 : The plot of the credible interval under the associated covariates when $\beta = 2$ under the independent case from 500,1000 and 2000 covariates

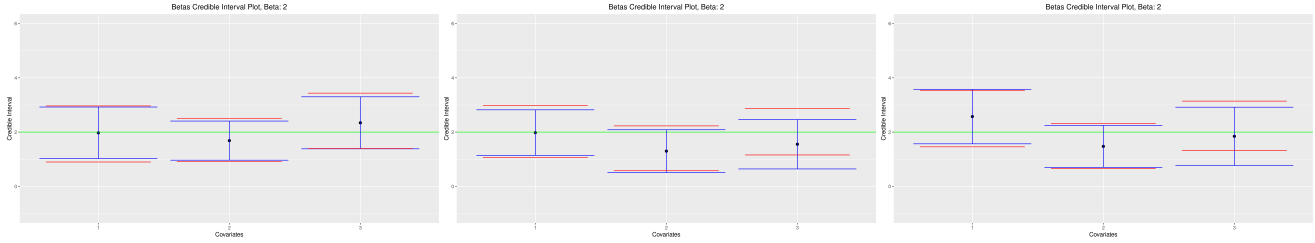


Figure 5.24 : The plot of the credible interval under the associated covariates when $\beta = 2$ under the second correlated case (x_1 with x_4, x_5, x_2 with x_6, x_7, x_3 with x_8, x_9) when $\rho = 0.1$ from 500, 1000 and 2000 covariates

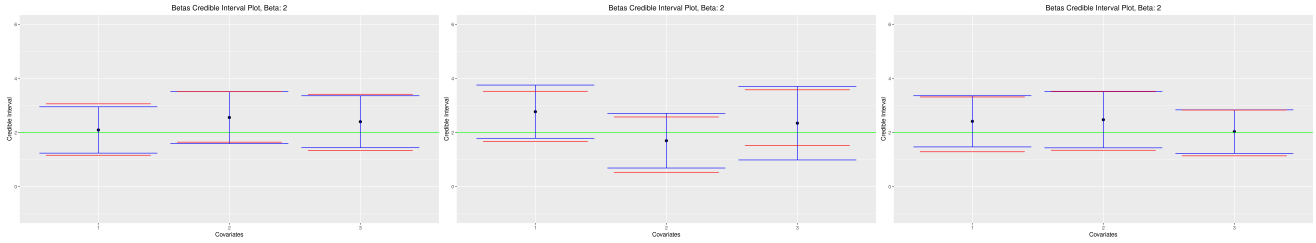


Figure 5.25 : The boxplots of ESS/the total of iterations after burn-in for β_1, β_2, β_3 and $g = 10$, $mu = 0.06$ under the independent case (ten data sets) from 500, 1000 and 2000 covariates

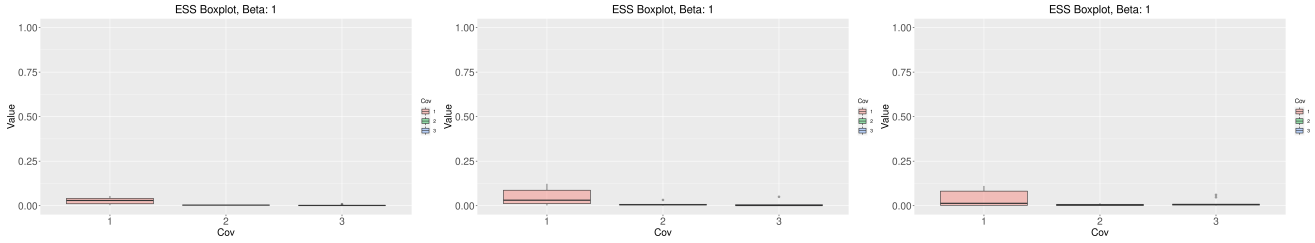


Figure 5.26 : The boxplots of ESS/the number of itreations that contain each β for β_1, β_2, β_3 and $g = 10$, $mu = 0.06$ under the independent case (ten data sets) from 500, 1000 and 2000 covariates

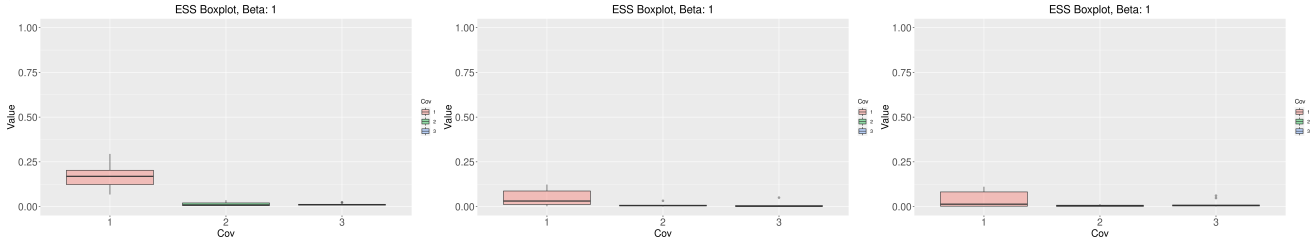


Figure 5.27 : The boxplots of ESS/the total of iterations after burn-in for β_1, β_2, β_3 and $g = 10$, $mu = 0.06$ under the independent case (ten data sets) from 500, 1000 and 2000 covariates

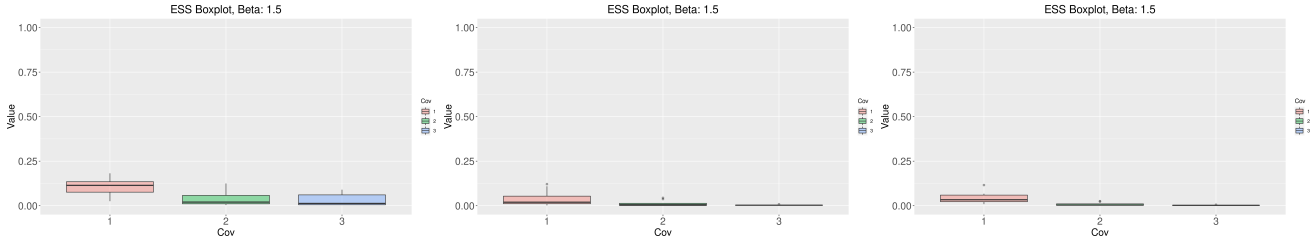
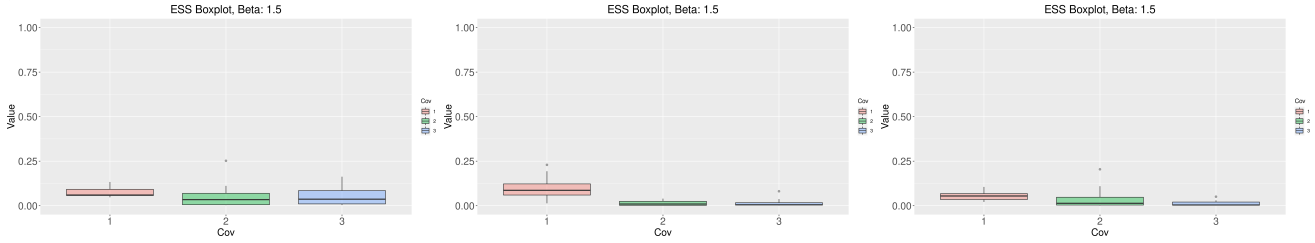


Figure 5.28 : The boxplots of ESS/the number of itreations that contain each β for β_1, β_2, β_3 and $g = 10$, $mu = 0.06$ under the independent case (ten data sets) from 500, 1000 and 2000 covariates



Chapter 6

Analysis of hypertension GWAS data

6.1 Introduction

Implementing simulation studies allowed the investigation and comparison of various methods' performance under specified scenarios, which in turn provide guidelines for their use in the assessment and interpretation of results obtained in the analysis of real data. The previous chapters of this thesis were dedicated to the development of methods and their assessment. In this chapter, we compare methods in a more realistic setting, with application to actual GWAS datasets. We first introduce and describe the dataset that will be analysed later in this chapter.

6.2 Data Deccription and Exploratory Analysis

This dataset is about GWAS of heart disease from Prof.Sandosh Padmanabhan's lab at Cardiovascular Sciences at Glasgow. A partial analysis of the data is covered in Padmanabhan and Joe (2017). There are 5312 observations, however after excluding rows with missing values, 5158 observations remain. Moreover, there are only 3731 that match the genotype data by patient ID. The covariates measured include ID, age, sex, Body Mass Index (BMI), smoking behavior and the existence of previous cardiovascular disease. There are two response variables. The first is severe stage 2 hypertension - where a person has a Systolic Blood Pressure (SBP) of greater than or equal to 140 mmHg or a Diastolic Blood Pressure (DBP) of greater than or equal to 90 mmHg . The second outcome variable is "hypertensive crisis", when a person has a blood pressure higher than 180/120 mmHg, requiring urgent medical care. There are 15221 associated covariates of SNP genotype information for each individual.

A summary of the categorical variables is reported in Table 6.1.

Table 6.1: Summary of categorical variables in heart GWAS data

Variable	Frequency of 0	Frequency of 1	Interpretation
sex	1865	1866	1= women
newsmoke	2905	826	1 = smokers
prevcd	3573	158	1 = having previous cardiovascular disease
hypstage2	45	3686	1 = stage 2 hypertension
hypcrisis	2389	1342	1 = have hypertensive crisis

The continuous variables are summarized in Table 6.2.

Table 6.2: Summary of some continuous variables

Variable	Minimum	Maximum	Mean
age	49	75	60.09
BMI	15.81	55.40	28.12

Next, as preliminary data analysis, we individually tested the association of each covariate with each of the response variables.

We also constructed Chi-square tests of association between each outcome and the categorical covariates. At a 5% level of significance, age and sex appeared significantly associated with hypertensive crisis.

For computational efficiency, considering the high dimension of the SNP genotype variables, we wanted to test the applicability of variable selection under a data splitting scenario (Chapter 5). We separate 3 scenarios. First, we split the total covariates in 5 data sets, with each data set containing about 3000 covariates. Second, the data set was split into 3 data sets, each data set containing about 5000 covariates. Last, the whole data set was considered together, with about 15000 covariates.

Moreover, age variable was centered by subtracting the mean and dividing by the standard deviation. Centering was also done for the SNP covariates before any method was applied.

In testing for the quantitative covariates, we used 2 sample Student t test. The results showed that higher the age of an individual, the more likely to have higher chance of hypertensive crisis. However, the remaining covariates do not show any association. Boxplots of the continuous covariates, grouped by outcome category, also showed similar results (Figures 6.1 and 6.2)

Table 6.3: Summary of test for association between covariates

Variables	test statistic	p-value	Interpretation
prevcd and hyperstage2	$\chi^2 = 5.238e-28$	1	hyperstage2 not associated for previous cardiovascular disease
newsmoke and hyperstage2	$\chi^2 = 6.416e-28$	1	hyperstage2 in smoker and non-smoker not different
prevcd and hypercrisis	$\chi^2 = 0.20443$	0.6512	hypercrisis not associated for previous cardiovascular disease
newsmoke and hypercrisis	$\chi^2 = 1.2117$	0.271	hypercrisis in smoker and non-smoker not different
sex and hyperstage2	$\chi^2 = 2.2548$	0.1332	hyperstage2 and sex no significant association
sex and hypercrisis	$\chi^2 = 39.674$	3e-10	hypercrisis in each sex associated difference

Table 6.4: Summary of test for the quantitative covariates

Variables	test statistic	p-value
age and hyperstage2	$t = -0.3022$	0.7654
age and hypercrisis	$t = -11.364$	2.2e-16
BMI and hyperstage2	$t = 0.54687$	0.5871
BMI and hypercrisis	$t = -1.5368$	0.1245

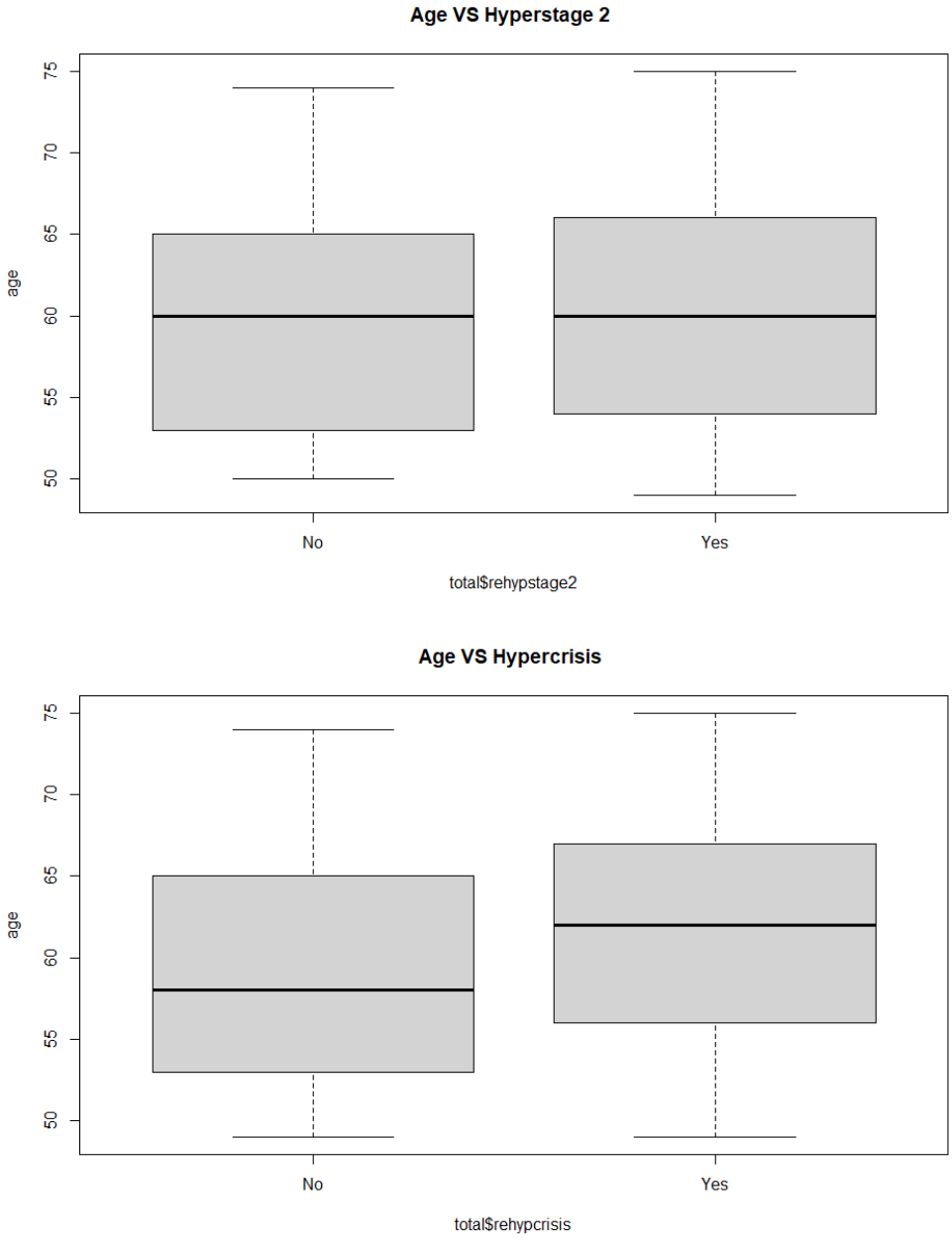


Figure 6.1: The boxplots of age and two response variables (stage 2 of hypertension and hypotensive crisis)

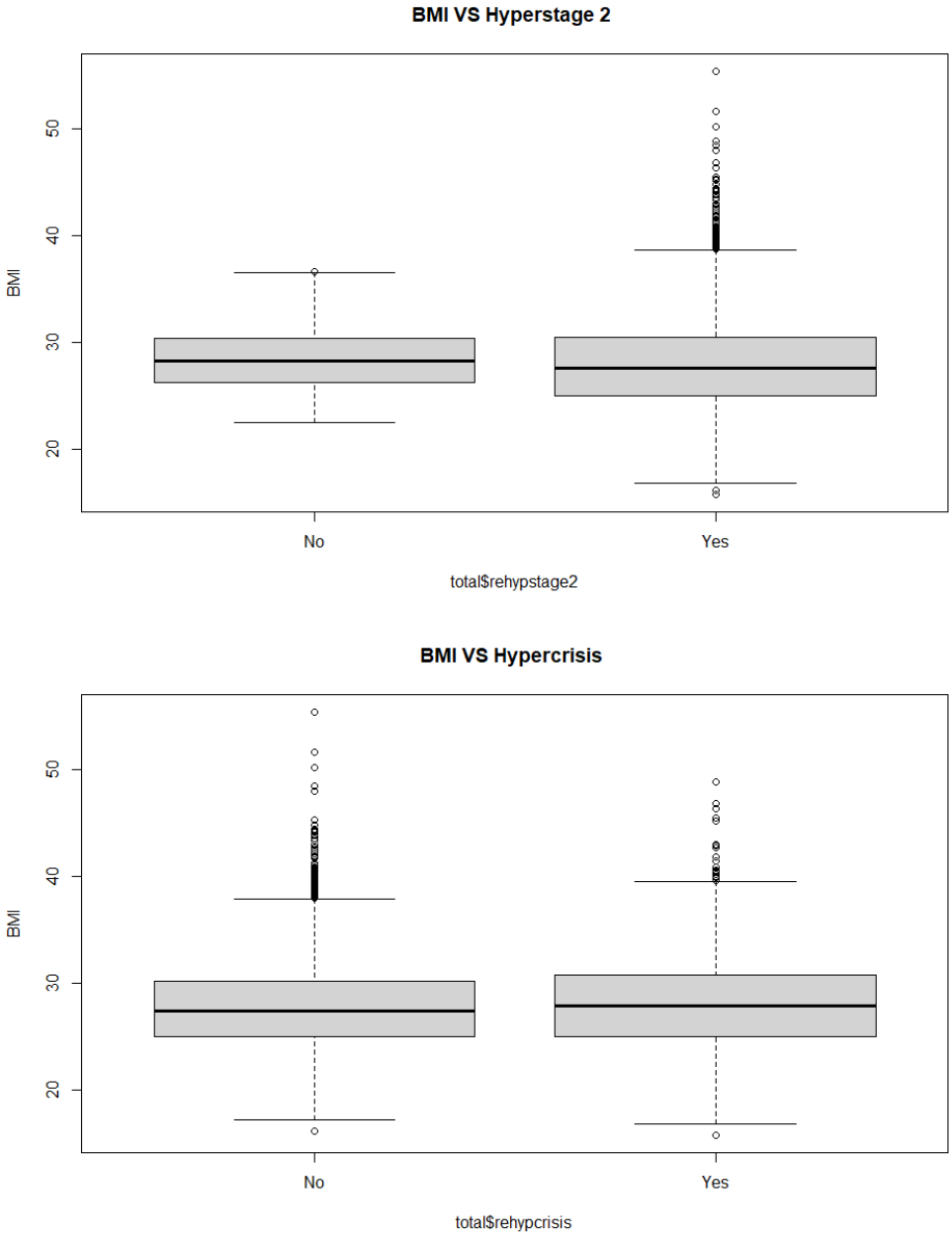


Figure 6.2: The boxplots of BMI and two response variables (stage 2 of hypertension and hypotensive crisis)

6.3 Results

Since outcome variable is categorical, we used logistic regression models and performed Bayesian Variable Selection (BVS) as dicussed in Chapter 3. Moreover, we compared BVS with LASSO penalised regression via using the *glmnet* package. There are two criteria, $\lambda.min$ and $\lambda.1se$ that were selecting the final LASSO model:$\lambda.min$ is the value of λ that gives minimum mean cross-validated error, while $\lambda.1se$ is the value of λ that gives the most regularised model such that the cross-validated error is within one standard error of the minimum (Friedman et al., 2010). We used the $bvsflex$ package used for on R-forge for variable selection in logistic regression models (Zucknick, 2013).

There are about 15000 SNPs covariates in total, that were split into 3, 5 and 15 datasets in turn in this analysis . The results are reported in Tables 6.5 to 6.7 from the first stage analysis of the datasets.

The cut off point of the inclusion probability of each covariate to select the first stage of for *Bayesian* is 0.10.

Under each data set where there are 15 splits, there are about 1000 SNP covariates. These results implied that only a few covariates are selected at the first stage. Moreover, these posterior mean estimate of the regression coefficients for SNPs from Bayesian are higher slightly than *glmnet* under each of the criteria.

Table 6.5: The 95% posterior credible interval of regression coefficient for each SNP selected under BVS and the upper bound and lower bound of the point estimates for each SNP selected when using LASSO by spiltting dataset into 15 subdata sets

Data set	SNP	$\lambda.min$	$\lambda.1se$	*Bayesian*
1	*rs*17137365	0.1234(0.0138,0.1516)	0.0239(0.0071,0.0917)	0.1465(0.0311,0.2034)
1	*rs*8035965	0.0526(0.0147,0.1053)	-	-
1	*rs*2030484	0.0307(0.0197,0.1011)	0.0475(0.0187,0.1125)	-
1	*rs*1453556	-	0.0178(0.0095,0.0579)	-
1	*rs*12909900	-	0.0161(0.0074,0.0697)	-
1	*rs*1548566	-	-	0.0475(0.0104,0.0963)
1	*rs*11632360	-	-	0.0397(0.0162,0.1014)
1	*rs*981347	-	-	0.0306(0.0187,0.0699)
1	*rs*4613037	0.0714(0.0211,0.1137)	0.0804(0.0187,0.1305)	-
1	*rs*10519577	-	0.0169(0.0084,0.0579)	-
1	*rs*1871017	-	0.0792(0.0272,0.1207)	-
1	*rs*10519442	0.0519(0.0174,0.1042)	-	-
1	*rs*8035695	0.1056(0.0294,0.1502)	0.0894(0.0163,0.1289)	-
2	*rs*4238497	0.0179(0.0088,0.0382)	0.0107(0.0055,0.0351)	0.0678(0.0125,0.1486)
2	*rs*968476	0.0481(0.0173,0.1087)	-	0.0415(0.0144,0.1129)
2	*rs*2337124	0.0205(0.0109,0.0895)	0.0184(0.0056,0.0412)	-
2	*rs*1719343	-	0.0457(0.0165,0.1041)	-
2	*rs*11638086	-	0.0760(0.0219,0.1248)	-

Continued on next page

Table 6.5 – *Continued from previous page*

Data set	*SNP*	$\lambda.min$	$\lambda.1se$	*Bayesian*
2	*rs*8034856	-	0.1067(0.0452,0.1562)	-
2	*rs*2241493	-	-	0.0263(0.0114,0.0679)
2	*rs*4779527	-	-	0.0645(0.0174,0.0996)
2	*rs*71166243	-	0.0219(0.0045,0.0874)	-
2	*rs*2037844	-	-	0.0145(0.0028,0.0614)
2	*rs*2337124	0.0261(0.0247,0.1184)	0.0356(0.0148,0.1219)	-
2	*rs*2292547	-	0.0998(0.0257,0.1387)	-
2	*rs*10519816	-	0.0232(0.0119,0.0745)	-
3	*rs*674155	0.0983(0.0245,0.1327)	-	0.1127(0.0396,0.2218)
3	*rs*4779630	0.0357(0.0163,0.1014)	0.0148(0.0071,0.0987)	-
3	*rs*17817518	0.0416(0.0136,0.1225)	-	0.0886(0.0279,0.1432)
3	*rs*12908501	-	0.0286(0.0097,0.0874)	-
3	*rs*8026003	-	0.0377(0.0112,0.0895)	-
3	*rs*442873	0.0469(0.0219,0.0923)	-	-
3	*rs*7163190	-	-	0.0478(0.0178,0.1092)
3	*rs*11854649	-	-	0.0689(0.0211,0.1356)
4	*rs*12438737	0.1158(0.0236,0.1784)	-	0.1971(0.0458,0.2355)
4	*rs*16967222	0.0282(0.0059,0.0974)	0.0136(0.0032,0.0691)	-
4	*rs*2033544	0.0179(0.0063,0.0845)	0.0287(0.0107,0.0988)	-
4	*rs*1847663	-	0.0373(0.0149,0.1026)	-
4	*rs*4924402	0.0322(0.0164,0.1035)	-	-
4	*rs*8026641	-	-	0.0619(0.0591,0.1247)
4	*rs*23050312	-	-	0.0156(0.0029,0.0612)
4	*rs*11070349	0.0479(0.0162,0.1085)	0.0389(0.0109,0.0996)	-
4	*rs*1530837	-	0.0483(0.0149,0.1108)	-
4	*rs*7163310	0.0573(0.0164,0.1264)	-	-
4	*rs*16972615	-	-	0.0122(0.0049,0.0647)
4	*rs*1668575	0.0590(0.0164,0.1149)	-	-
4	*rs*12914570	-	-	0.0663(0.0259,0.1168)
5	*rs*12439639	0.0319(0.0135,0.0798)	-	0.1245(0.0287,0.1935)
5	*rs*8026843	0.0134(0.0086,0.0446)	-	-
5	*rs*2033735	0.0755(0.0197,0.1254)	-	0.1479(0.0175,0.2081)
5	*rs*10519062	0.0234(0.0115,0.0579)	-	-
5	*rs*12437829	0.0419(0.0114,0.0768)	-	-
5	*rs*1160623	-	0.0311(0.0118,0.0694)	-
5	*rs*8040530	-	0.0232(0.0096,0.0617)	-
5	*rs*281299	0.0591(0.0187,0.1102)	-	-
5	*rs*11070583	-	0.0822(0.0115,0.1397)	-
5	*rs*44435197	-	-	0.0834(0.0194,0.1748)
5	*rs*24104010	0.0217(0.0097,0.0638)	-	0.0474(0.0175,0.0894)

Continued on next page

Table 6.5 – *Continued from previous page*

Data set	*SNP*	$\lambda.min$	$\lambda.1se$	*Bayesian*
5	*rs*9944192	0.0816(0.0238,0.1537)	-	-
5	*rs*17522269	-	0.0947(0.0115,0.1602)	-
5	*rs*3784308	-	-	0.0304(0.0178,0.0745)
6	*rs*4774594	0.0297(0.0183,0.0729)	-	0.0496(0.0147,0.0859)
6	*rs*2278295	0.0314(0.0145,0.0678)	0.0185(0.0071,0.0368)	-
6	*rs*562804	0.0466(0.0193,0.0968)	-	0.0876(0.0279,0.1365)
6	*rs*1813171	-	0.0419(0.0174,0.0962)	-
6	*rs*2926881	-	0.0398(0.0182,0.0785)	-
6	*rs*2115825	0.0361(0.0174,0.0712)	-	-
6	*rs*11857095	-	-	0.0427(0.0163,0.0941)
6	*rs*2414371	-	-	0.0639(0.0234,0.1304)
6	*rs*600753	-	-	0.0734(0.0211,0.1431)
7	*rs*7114742	0.0552(0.0102,0.0975)	-	0.1205(0.0238,0.2147)
7	*rs*4494480	0.0169(0.0086,0.0461)	-	-
7	*rs*1632868	0.0316(0.0108,0.0637)	-	0.0974(0.0175,0.1879)
7	*rs*890271	0.0652(0.0254,0.1175)	-	-
7	*rs*1469280	0.0244(0.0074,0.0632)	-	-
7	*rs*4775017	-	0.0574(0.0218,0.1174)	-
7	*rs*4775031	-	0.0417(0.0189,0.1005)	-
7	*rs*10152146	0.0369(0.0187,0.0672)	-	-
7	*rs*4775149	-	0.0298(0.0105,0.0497)	-
7	*rs*4775162	-	-	0.0334(0.0134,0.0744)
7	*rs*12912193	0.0301(0.0107,0.0637)	-	0.0974(0.0175,0.1868)
7	*rs*2059553	0.0625(0.0254,0.1169)	-	-
7	*rs*8033609	-	0.0275(0.0109,0.0479)	-
7	*rs*10519074	-	-	0.0394(0.0115,0.0775)
8	*rs*17237486	0.0175(0.0092,0.0397)	-	0.0279(0.0105,0.0531)
8	*rs*782947	0.0196(0.0059,0.0418)	0.0296(0.0103,0.0617)	-
8	*rs*782913	0.0344(0.0165,0.0745)	0.0126(0.0042,0.0384)	-
8	*rs*2899667	-	0.0517(0.0154,0.1064)	-
8	*rs*697333	0.0912(0.0234,0.1512)	-	-
8	*rs*1727179	-	-	0.0969(0.0354,0.1647)
8	*rs*1320190	-	-	0.0267(0.0129,0.0678)
8	*rs*12593770	0.0293(0.0062,0.0628)	0.0387(0.0106,0.0815)	-
8	*rs*16953154	-	0.0969(0.0179,0.1614)	-
8	*rs*1107183	0.0217(0.0144,0.0782)	-	-
8	*rs*11071841	-	-	0.0424(0.0154,0.0746)
8	*rs*12592821	0.0294(0.0084,0.0612)	-	-
9	*rs*412705	0.0496(0.0192,0.1017)	-	0.0791(0.0235,0.1312)
9	*rs*718878	0.0206(0.0134,0.0654)	0.0639(0.0151,0.1246)	-

Continued on next page

Table 6.5 – *Continued from previous page*

Data set	*SNP*	$\lambda.min$	$\lambda.1se$	*Bayesian*
9	*rs*893473	0.0129(0.0065,0.0459)	0.0328(0.0107,0.0657)	-
9	*rs*1115528	-	0.0145(0.0047,0.0388)	-
9	*rs*9806190	0.0176(0.0064,0.0347)	-	-
9	*rs*11633040	-	-	0.0258(0.0049,0.0647)
9	*rs*4617810	-	-	0.0652(0.0148,0.1216)
9	*rs*477722	0.0265(0.0067,0.0596)	0.0388(0.0107,0.0875)	-
9	*rs*694846	-	0.0541(0.0185,0.1079)	-
9	*rs*904972	0.0762(0.0207,0.1311)	-	-
9	*rs*108331	-	-	0.0819(0.0259,0.1632)
9	*rs*598472	0.0328(0.0164,0.0785)	-	-
10	*rs*5950667	0.0789(0.0186,0.1472)	-	0.0915(0.0396,0.1589)
10	*rs*2912271	0.0365(0.0142,0.0856)	0.0185(0.0071,0.0567)	-
10	*rs*16958237	0.0416(0.0193,0.1047)	-	0.0866(0.0257,0.1497)
10	*rs*4887066	-	0.0311(0.0197,0.0659)	-
10	*rs*8042694	-	0.0512(0.0218,0.1174)	-
10	*rs*744336	0.0367(0.0219,0.0712)	-	-
10	*rs*12904553	-	-	0.0477(0.0157,0.0963)
10	*rs*12439144	-	-	0.0631(0.0223,0.1357)
11	*rs*7183358	0.0594(0.0125,0.1194)	-	0.0469(0.0245,0.0975)
11	*rs*769765	0.0112(0.0065,0.0384)	0.0185(0.0071,0.0441)	-
11	*rs*1685146	0.0376(0.0184,0.0785)	-	0.0887(0.0279,0.1532)
11	*rs*372447	-	0.0714(0.0197,0.1455)	-
11	*rs*2759307	-	0.0122(0.0082,0.0397)	-
11	*rs*2654212	0.0443(0.0198,0.0975)	-	-
11	*rs*4479064	-	-	0.0347(0.0185,0.0815)
11	*rs*7165448	-	-	0.0880(0.0225,0.1532)
11	*rs*11853372	-	-	0.0292(0.0122,0.0548)
12	*rs*4887255	0.0442(0.0208,0.0789)	-	0.0981(0.0315,0.1965)
12	*rs*2028731	0.0264(0.0112,0.0598)	0.0632(0.0214,0.1307)	-
12	*rs*16977968	0.0696(0.0142,0.1158)	0.0389(0.0195,0.0715)	-
12	*rs*7174605	-	0.0564(0.0154,0.1122)	-
12	*rs*1349381	0.0855(0.0167,0.1387)	-	-
12	*rs*17655115	-	-	0.0575(0.0214,0.1147)
12	*rs*8041239	-	-	0.0144(0.0029,0.0316)
12	*rs*7164471	0.0647(0.0133,0.1154)	0.0368(0.0125,0.0755)	-
12	*rs*1372600	-	0.0169(0.0041,0.0412)	-
12	*rs*11073795	0.0371(0.0145,0.0735)	-	-
12	*rs*1126823	-	-	0.0124(0.0039,0.0462)
13	*rs*2601181	0.0416(0.0168,0.0927)	-	0.0961(0.0348,0.1588)
13	*rs*1476078	0.0318(0.0165,0.0811)	0.0186(0.0071,0.0398)	-

Continued on next page

Table 6.5 – *Continued from previous page*

Data set	*SNP*	$\lambda.min$	$\lambda.1se$	*Bayesian*
13	*rs*1053909	0.0451(0.0193,0.0882)	-	0.0876(0.0279,0.1463)
13	*rs*7162155	-	0.0441(0.0197,0.0865)	-
13	*rs*8024027	-	0.0369(0.0182,0.0749)	-
13	*rs*7167984	0.0279(0.0108,0.0588)	-	-
13	*rs*16948067	-	-	0.0417(0.0174,0.0941)
13	*rs*7182315	-	-	0.0389(0.0112,0.0854)
14	*rs*201641	0.0497(0.0192,0.0957)	-	0.0642(0.0163,0.1169)
14	*rs*7180576	0.0347(0.0165,0.0788)	0.0185(0.0071,0.0347)	-
14	*rs*8025603	0.0316(0.0119,0.0754)	-	0.0874(0.0279,0.1532)
14	*rs*2199732	-	0.0847(0.0197,0.1510)	-
14	*rs*10520802	-	0.0517(0.0218,0.1204)	-
14	*rs*13380379	0.0662(0.0219,0.1224)	-	-
14	*rs*12915943	-	-	0.0815(0.0173,0.1539)
14	*rs*8039169	-	-	0.0763(0.0211,0.1466)
15	*rs*1442815	0.0561(0.0195,0.1157)	-	0.0679(0.0296,0.1267)
15	*rs*12164914	0.0310(0.0165,0.0755)	0.0186(0.0071,0.0412)	-
15	*rs*12911516	0.0297(0.0093,0.0614)	-	0.0678(0.0279,0.1351)
15	*rs*2289558	-	0.0344(0.0107,0.0741)	-
15	*rs*12903750	-	0.0568(0.0211,0.1167)	-
15	*rs*8039762	0.0413(0.0197,0.0952)	-	-
15	*rs*8042302	-	-	0.0507(0.0174,0.1165)
15	*rs*12915781	-	-	0.0681(0.0211,0.1388)
15	*rs*11855154	-	-	0.0599(0.0207,0.1205)

Table 6.6: The 95% posterior central credible intervals of regression coefficient each selected *SNP*under *BVS* and the upper bound and lower bound of the point estimates for each *SNP* selected when using *glmnet* package when data split into 5 subdata sets

Data set	*SNP*	$\lambda.min$	$\lambda.1se$	*Bayesian*
1	*rs*1256841	0.03737(0.0145,0.1278)	0.02115(0.0087,0.1156)	0.092421(0.0488,0.2197)
1	*rs*17595461	0.01653(0.0017,0.1159)	-	-
1	*rs*1390786	0.107402(0.0197,0.1496)	0.08745(0.0159,0.1298)	-
1	*rs*17137192	-	0.0478(0.0215,0.1579)	-
1	*rs*754185	-	0.05469(0.0272,0.1662)	-
1	*rs*12439582	-	-	0.123849(0.0113,0.2602)
1	*rs*2119010	-	-	0.01181(0.0034,0.1311)
1	*rs*11634759	-	-	0.03247(0.0156,0.1654)
2	*rs*3887013	0.02066(0.0133,0.1072)	0.01975(0.0055,0.1007)	0.10142(0.0113,0.1096)
2	*rs*17541104	0.0273(0.0175,0.1598)	-	0.10544(0.0172,0.1310)

Continued on next page

Table 6.6 – *Continued from previous page*

Data set	*SNP*	$\lambda.min$	$\lambda.1se$	*Bayesian*
2	*rs*4777858	0.08968(0.0247,0.1085)	0.03588(0.0194,0.1024)	-
2	*rs*776726	-	0.03697(0.0174,0.1241)	-
2	*rs*1546424	-	0.0413(0.0219,0.1417)	-
2	*rs*2220176	-	0.02874(0.0045,0.1278)	-
2	*rs*1365593	-	-	0.0141(0.0028,0.1619)
2	*rs*1863464	-	-	0.02362(0.0186,0.1992)
3	*rs*8024303	0.10792(0.0196,0.1927)	-	0.041967(0.0396,0.2579)
3	*rs*8042590	0.02447(0.0165,0.1234)	0.01845(0.0071,0.1096)	-
3	*rs*1020987	0.0221(0.0193,0.1228)	-	0.08766(0.0279,0.1435)
3	*rs*7181675	-	0.03847(0.0197,0.1315)	-
3	*rs*290334	-	0.04578(0.0218,0.1478)	-
3	*rs*11858141	0.02669(0.0219,0.1217)	-	-
3	*rs*755599	-	-	0.05515(0.0173,0.1491)
3	*rs*622442	-	-	0.083941(0.0211,0.1458)
4	*rs*8039509	0.06404(0.0219,0.1217)	-	0.1971(0.0433,0.2132)
4	*rs*234508	0.03691(0.0159,0.1212)	0.12936(0.0315,0.1996)	-
4	*rs*9806494	0.04565(0.0162,0.1092)	0.03887(0.0204,0.1008)	-
4	*rs*8030638	-	0.02597(0.0149,0.1211)	-
4	*rs*963960	0.03871(0.0164,0.1227)	-	-
4	*rs*2083061	-	-	0.012335(0.0059,0.2647)
4	*rs*713469	-	-	0.04615(0.0292,0.1612)
5	*rs*11247163	0.022559(0.0102,0.1795)	-	0.16373(0.0238,0.1997)
5	*rs*11630747	0.018285(0.0086,0.1416)	-	-
5	*rs*4616271	0.03061(0.0197,0.1637)	-	0.149974(0.0175,0.1895)
5	*rs*2047222	0.06522(0.0254,0.1509)	-	-
5	*rs*1037117	0.02439(0.0114,0.1266)	-	-
5	*rs*11634329	-	0.05978(0.0218,0.1894)	-
5	*rs*1834207	-	0.04369(0.0247,0.1772)	-
5	*rs*2305668	0.03687(0.0187,0.1612)	-	-
5	*rs*4076999	-	0.02975(0.0115,0.1479)	-
5	*rs*4886999	-	-	0.033939(0.0195,0.1774)

The frequency of selected SNPs (with a credible interval for the coefficient not including zero) is similar to the 15 dataset case.

Table 6.7: The 95% posterior central credible intervals of regression coefficient each selected SNPunder BVS and the upper bound and lower bound of the point estimates for each SNP selected when using *glmnet* package when data split into 3 subdata sets

Data set	SNP	$\lambda.min$	$\lambda.1se$	*Bayesian*
1	*rs*12905013	0.078555(0.0084,0.1105)	0.05843(0.0071,0.1027)	0.125748(0.0166,0.1635)
1	*rs*2928719	-	0.04891(0.0054,0.0986)	-
1	*rs*1863464	-	-	0.12338(0.0147,0.1598)
2	*rs*999787	0.08123(0.0109,0.1216)	0.06394(0.0085,0.1153)	0.10944(0.0174,0.1364)
2	*rs*4775077	-	0.04875(0.0081,0.1043)	-
2	*rs*178189639	-	0.03697(0.0075,0.1006)	0.07889(0.0151,0.1077)
2	*rs*8039254	-	-	0.075093(0.0094,0.1028)
3	*rs*8027171	0.08412(0.0118,0.1269)	0.05946(0.0072,0.1011)	0.11204(0.0134,0.1282)
3	*rs*1392161	-	0.04947(0.0061,0.0942)	-
3	*rs*12591031	0.03884(0.0078,0.0915)	-	0.09477(0.0114,0.1164)
3	*rs*2172188	-	-	0.10513(0.0144,0.1239)

A crucial point observed is under the various scenarios of splitting datasets (3,5 and 15) the same SNP covariates are not selected . Hence, they may be varied the number of splits.

Due to the large number of SNP covariates, we tried to filter variables in the split datasets at the first stage. And then, we use these SNP covariates that were selected in the first stage for variable selection in the second stage by combining 3,5 and 15 datasets respectively. These results are reported in Tables 6.8 to 6.10.

Table 6.8: The 95% posterior credible interval of 3 splits first, then 5, then 15 for each SNP under the *Bayesian* method and the upper bound and lower bound of the point estimates for each SNP selected when using *glmnet* package via $\lambda.min$ and $\lambda.1se$. The second stage was run via using BVS and *glmnet* with only the covariates that are selected in the first stage (15 datasets)

SNP	$\lambda.min$	$\lambda.1se$	*Bayesian*	*Index*
*rs*17137365(91)	0.115(0.0247,0.1895)	0.059(0.0176,0.1044)	0.144(0.0358,0.2638)	91
*rs*8035965(163)	0.055(0.0205,0.1127)	-	-	163
*rs*4613037(819)	0.114(0.0314,0.1925)	0.091(0.0265,0.1674)	-	819
*rs*4238497(1113)	0.028(0.0182,0.0524)	0.012(0.0068,0.0387)	0.069(0.0158,0.1499)	96
*rs*968476(1189)	0.045(0.0173,0.1112)	-	0.043(0.0144,0.1084)	172
*rs*2037844(1717)	-	-	0.024(0.0028,0.0752)	700
*rs*2292547(1901)	-	0.102(0.0274,0.1865)	-	884
*rs*674155(2039)	0.099(0.0256,0.1745)	-	0.119(0.0384,0.2345)	7
*rs*4779630(2105)	0.039(0.0186,0.0807)	0.016(0.0095,0.0783)	-	73
*rs*17817518(2119)	0.046(0.0136,0.0988)	-	0.089(0.0266,0.1512)	87

Continued on next page

Table 6.8 – *Continued from previous page*

SNP	$\lambda.min$	$\lambda.1se$	*Bayesian*	*Index*
*rs*12438737(3136)	0.105(0.0355,0.1985)	-	0.199(0.0583,0.2878)	89
*rs*16967222(3189)	0.029(0.0105,0.0955)	0.017(0.0074,0.0714)	-	142
*rs*11070349(3704)	0.054(0.0162,0.1095)	0.035(0.0117,0.0984)	-	657
*rs*12914570(3869)	-	-	0.069(0.0277,0.1237)	822
*rs*12439639(4070)	0.039(0.0155,0.0847)	-	0.105(0.0356,0.2068)	8
*rs*2033735(4231)	0.079(0.0238,0.1349)	-	0.142(0.0325,0.2369)	169
*rs*24104010(4758)	0.029(0.0097,0.0785)	-	0.049(0.0231,0.0925)	689
*rs*3784308(5022)	-	-	0.033(0.0186,0.0796)	953
*rs*4774594(5083)	0.036(0.0194,0.0790)	-	0.051(0.0184,0.0966)	6
*rs*2278295(5164)	0.033(0.0141,0.0714)	0.019(0.0079,0.0396)	-	87
*rs*562804(5341)	0.049(0.0195,0.0996)	-	0.089(0.0356,0.1511)	264
*rs*7114742(6107)	0.059(0.0134,0.1012)	-	0.121(0.0269,0.2202)	15
*rs*4494480(6121)	0.019(0.0088,0.0455)	-	-	29
*rs*1632868(6233)	0.039(0.0158,0.0674)	-	0.099(0.0197,0.1896)	141
*rs*12912193(6789)	0.031(0.0116,0.0696)	-	0.099(0.0183,0.1890)	697
*rs*10519074(7062)	-	-	0.041(0.0115,0.0854)	970
*rs*17237486(7119)	0.017(0.0095,0.0410)	-	0.029(0.0115,0.0582)	12
*rs*12593770(7895)	0.035(0.0165,0.0692)	0.039(0.0106,0.0824)	-	788
*rs*412705(8283)	0.047(0.0204,0.1087)	-	0.076(0.0215,0.1389)	161
*rs*718878(8311)	0.024(0.0147,0.0685)	0.069(0.0185,0.1269)	-	189
*rs*893473(8329)	0.019(0.0095,0.0512)	0.035(0.0117,0.0725)	-	207
*rs*477722(8829)	0.029(0.0095,0.0612)	0.039(0.0118,0.0882)	-	707
*rs*5950667(9160)	0.081(0.0285,0.1532)	-	0.095(0.0412,0.1615)	23
*rs*16958237(9433)	0.046(0.0245,0.0998)	-	0.089(0.0276,0.1574)	296
*rs*12439144(9983)	-	-	0.067(0.0235,0.1373)	846
*rs*7183358(10167)	0.059(0.0165,0.1178)	-	0.049(0.0259,0.0990)	15
*rs*1685146(10244)	0.039(0.0196,0.0811)	-	0.095(0.0295,0.1612)	92
*rs*11853372(11113)	-	-	0.035(0.0185,0.0652)	961
*rs*4887255(11372)	0.046(0.0217,0.0853)	-	0.099(0.0352,0.1996)	205
*rs*2028731(11382)	0.028(0.0153,0.0607)	0.069(0.0254,0.1355)	-	215
*rs*7164471(11795)	0.066(0.0155,0.1241)	0.039(0.0186,0.0792)	-	628
*rs*1126823(11998)	-	-	0.015(0.0047,0.0498)	831
*rs*2601181(12253)	0.046(0.0183,0.0987)	-	0.098(0.0395,0.1635)	71
*rs*1053909(12676)	0.049(0.0195,0.0996)	-	0.089(0.0285,0.1536)	494
*rs*201641(13222)	0.048(0.0195,0.0987)	-	0.066(0.0178,0.1256)	25
*rs*7180576(13292)	0.037(0.0169,0.0795)	0.018(0.0071,0.0368)	-	95
*rs*8025603(13383)	0.039(0.0119,0.0755)	-	0.089(0.0295,0.1624)	186
*rs*1442815(14253)	0.058(0.0214,0.1188)	-	0.069(0.0298,0.1284)	41
*rs*12164914(14344)	0.036(0.0166,0.0774)	0.019(0.0082,0.0455)	-	132

Continued on next page

Table 6.8 – *Continued from previous page*

SNP	$\lambda.min$	$\lambda.1se$	*Bayesian*	*Index*
*rs*12911516(14502)	0.028(0.0087,0.0651)	-	0.072(0.0283,0.1416)	290

Table 6.9: The 95% posterior credible interval of 3 splits first, then 5, then 15 for each SNP under the *Bayesian* method and the upper bound and lower bound of the point estimates for each SNP selected when using *glmnet* package via $\lambda.min$ and $\lambda.1se$. The second stage was run via using BVS and *glmnet* with only the covariates that are selected in the first stage (5 datasets)

SNP	$\lambda.min$	$\lambda.1se$	*Bayesian*	*Index*
*rs*12594495(60)	0.0786(0.0321,0.1698)	0.0583(0.0199,0.1359)	0.1012(0.0494,0.2351)	60
*rs*17555920(388)	0.0356(0.0174,0.1341)	-	-	497
*rs*1719332(1256)	-	0.0681(0.0294,0.1632)	-	1341
*rs*2118157(2047)	-	-	0.0598(0.0210,0.1459)	2343
*rs*8025254(3091)	0.0459(0.0197,0.1248)	0.0296(0.0134,0.1151)	0.1234(0.0337,0.1697)	44
*rs*1757463(3673)	0.0394(0.0207,0.1637)	-	0.1154(0.0341,0.1469)	626
*rs*4774497(4429)	0.0964(0.0317,0.1298)	0.0689(0.0235,0.1364)	-	1382
*rs*1863427(6052)	-	-	0.0418(0.0251,0.2074)	3005
*rs*8039952(6181)	0.1154(0.0271,0.2051)	-	0.0671(0.0479,0.2394)	89
*rs*8041221(6671)	0.0512(0.0279,0.1338)	0.0372(0.0152,0.1142)	-	579
*rs*6494361(7607)	-	-	0.0952(0.0277,0.1519)	1515
*rs*8038734(9279)	0.0711(0.0284,0.1296)	-	0.2018(0.0519,0.2334)	142
*rs*2343675(11125)	0.0418(0.0209,0.1331)	0.1314(0.0416,0.2081)	-	1988
*rs*7180683(11201)	-	-	0.0596(0.0312,0.1704)	2064
*rs*1125280(12310)	0.0347(0.0207,0.1834)	-	0.1714(0.0287,0.2071)	128
*rs*4614672(13630)	0.0462(0.0273,0.1718)	-	0.1571(0.0352,0.1996)	1448
*rs*4076597(14074)	-	0.0375(0.0195,0.1516)	-	1892
*rs*8029926(14104)	-	-	0.0496(0.0251,0.1833)	1922

Table 6.10: The 95% posterior credible interval of 3 splits first, then 5, then 15 for each SNP under the *Bayesian* method and the upper bound and lower bound of the point estimates for each SNP selected when using *glmnet* package via $\lambda.min$ and $\lambda.1se$. The second stage was run via using BVS and *glmnet* with only the covariates that are selected in the first stage (3 datasets)

SNP	$\lambda.min$	$\lambda.1se$	*Bayesian*	*Index*
*rs*12905013(790)	0.086(0.0217,0.1312)	0.064(0.0194,0.1164)	0.141(0.0317,0.1867)	790
*rs*1863464(3915)	-	-	0.139(0.0254,0.1615)	3915
*rs*999787(5810)	0.096(0.0237,0.1375)	0.081(0.0195,0.1237)	0.117(0.0321,0.1512)	733
*rs*17818939(6523)	-	0.057(0.0184,0.1169)	0.091(0.0234,0.1164)	1446
*rs*8027171(10866)	0.101(0.0251,0.1365)	0.074(0.0155,0.1154)	0.129(0.0352,0.1421)	744
*rs*1392161(11635)	0.046(0.0152,0.1074)	-	0.111(0.0234,0.1369)	1513

When we used the whole dataset without splitting, there were not any covariates selected in any of the methods. However, when we spilt the dataset there were some covariates selected-giving us a chance to potentially find associated covariates that may have been missed due to the

signal being too weak in the full dataset.

To check consistency of the selected SNPs, we used 2nd and 3rd run independently for the 3 - split dataset. These results are shown in Table 6.11 (under the first stage) and Table 6.12 (under the second stage). It is promising to see that about ten of the same SNPs are selected in each run, irrespective of split although some are different.

Table 6.11: The credible interval of posterior mean in each *SNP* under the *Bayesian* method by spiltting into 3 data sets when run on 3 different splits in the data (under the first stage). The value in the bracket represents the index of sub-dataset in that run.

SNP	**1st run CI**	**2nd run CI**	**3rd run CI**
*rs*12905013	0.12479(0.0149,0.1675)[1]	0.12461(0.0117,0.1689)[1]	0.12575(0.0166,0.1635)[2]
*rs*1863464	0.12457(0.0132,0.1584)[1]	0.12427(0.0114,0.1572)[1]	0.12338(0.0147,0.1598)[3]
*rs*2290352	0.11327(0.0140,0.1612)[3]	-	-
*rs*1619030	0.11875(0.0138,0.1631)[2]	0.11779(0.0152,0.1604)[1]	-
*rs*10519226	0.11766(0.0147,0.1598)[1]	-	0.11537(0.0149,0.1603)[3]
*rs*1350090	-	0.11572(0.0134,0.1586) [1]	-
*rs*2442464	-	0.11596(0.0157,0.1518)[2]	-
*rs*1055879	-	-	0.11577(0.0184,0.1567)[1]
*rs*745636	-	-	0.11642(0.0149,0.1578)[3]
*rs*999787	0.10936(0.0135,0.1359)[2]	0.10957(0.0181,0.1396)[3]	0.10944(0.0174,0.1364)[2]
*rs*17818939	0.07871(0.0131,0.1082)[2]	0.07832(0.0141,0.1069)[1]	0.07889(0.0151,0.1077)[1]
*rs*8039254	0.075082(0.0052,0.1033)[3]	0.075171(0.0071,0.1050)[3]	0.07509(0.0094,0.1028)[1]
*rs*2165488	0.08931(0.0087,0.1098)[2]	0.086312(0.0089,0.1074)[3]	-
*rs*1865923	0.08531(0.0077,0.1054)[2]	-	0.08476(0.0086,0.1076)[1]
*rs*7173314	0.08212(0.0071,0.1046)[2]	-	-
*rs*1712429	-	0.08451(0.0069,0.1059)[2]	-
*rs*7173622	-	0.08341(0.0067,0.1053)[3]	-
*rs*166357	-	-	0.08671(0.0073,0.1087)[1]
*rs*789776	-	-	0.08507(0.0068,0.1082)[1]
*rs*8027171	0.11213(0.0147,0.1277)[1]	0.11251(0.0141,0.1297)[2]	0.11204(0.0134,0.1282)[2]
*rs*12591031	0.09477(0.0118,0.1167)[1]	0.09451(0.0112,0.1159)[3]	0.09478(0.0114,0.1164)[2]
*rs*1392161	0.10535(0.0166,0.1297)[2]	0.10548(0.0154,0.1281)[1]	0.10513(0.0144,0.1239)[1]
*rs*1719336	0.10611(0.0174,0.1288)[3]	-	0.10511(0.0135,0.1216)[1]
*rs*12902710	0.10498(0.0159,0.1275)[3]	0.10539(0.0144,0.1247)[2]	-
*rs*293376	0.10247(0.0134,0.1195)[2]	0.10432(0.0157,0.1212)[3]	-
*rs*12902470	-	0.10386(0.0129,0.1192)[3]	-
*rs*260543	-	-	0.10116(0.0118,0.1137)[3]
*rs*1874835	-	-	0.10227(0.0108,0.1129)[2]

Under each run, some covariates are selected in all run. However, there is a slightly diffr-

erent on the upper bound and lower bound of credible interval.

Table 6.12: The credible interval of posterior mean in each SNP under the *Bayesian* method by spiltting into 3 data sets when run on 3 different splits in the data (under the second stage).

SNP	**1st run CI**	**2nd run CI**	**3rd run CI**
$rs12905013$	0.129(0.0175,0.2779)	0.156(0.0119,0.2133)	0.153(0.0199,0.2331)
$rs1863464$	0.061(0.0079,0.2001)	0.071(0.0072,0.1727)	0.074(0.0076,0.1733)
$rs1619030$	0.072(0.0088,0.1983)	0.075(0.0092,0.1916)	-
$rs10519226$	0.076(0.0095,0.1873)	-	0.079(0.0098,0.1851)
$rs2442464$	-	-	-
$rs1055879$	-	-	-
$rs745636$	-	-	-
$rs999787$	0.075(0.0099,0.1765)	0.073(0.0071,0.1737)	0.064(0.0066,0.1636)
$rs8039254$	0.080(0.0135,0.1597)	0.095(0.0167,0.1631)	0.086(0.0157,0.1619)
$rs2165488$	0.084(0.0125,0.1613)	0.089(0.0146,0.1627)	-
$rs1865923$	0.087(0.0139,0.1607)	-	0.091(0.0157,0.1649)
$rs8027171$	0.134(0.0185,0.2146)	0.157(0.0216,0.2201)	0.142(0.0205,0.2197)
$rs1392161$	0.114(0.0175,0.1976)	0.134(0.0165,0.2017)	0.126(0.0157,0.2041)
$rs1719336$	0.109(0.0159,0.1954)	-	0.115(0.0162,0.1963)
$rs12902710$	0.105(0.0147,0.1949)	0.111(0.0169,0.1952)	-
$rs293376$	0.107(0.0162,0.1950)	0.110(0.0159,0.1975)	-

The objective of this Chapter is the application of our methods to the real dataset. We used the splitting for the Bayesian variable selection since the MCMC algorithm for the full Bayesian variable selection is fail and cnsume more times when there are ultrahigh on the number of covariates. Then we used to combine those covariates are selected in the first stage to the second stage for Bayesian variable selection again. Moreover, we used the same algorithm to LASSO with two citeria.

The credible intervals of the posterior mean in each SNP under the Bayesian method are wider than both of criteria in LASSO. They indicated that the outperformance on the estimation of regression coefficient ($\boldsymbol{\beta}$). Moreover, for confirmation in the consistency of the results in Bayesian frame work, we ran on 3 runs on the second stage. The results pointed that on the consistency in each run for the Bayesian approach. They are reported in Table 6.12.

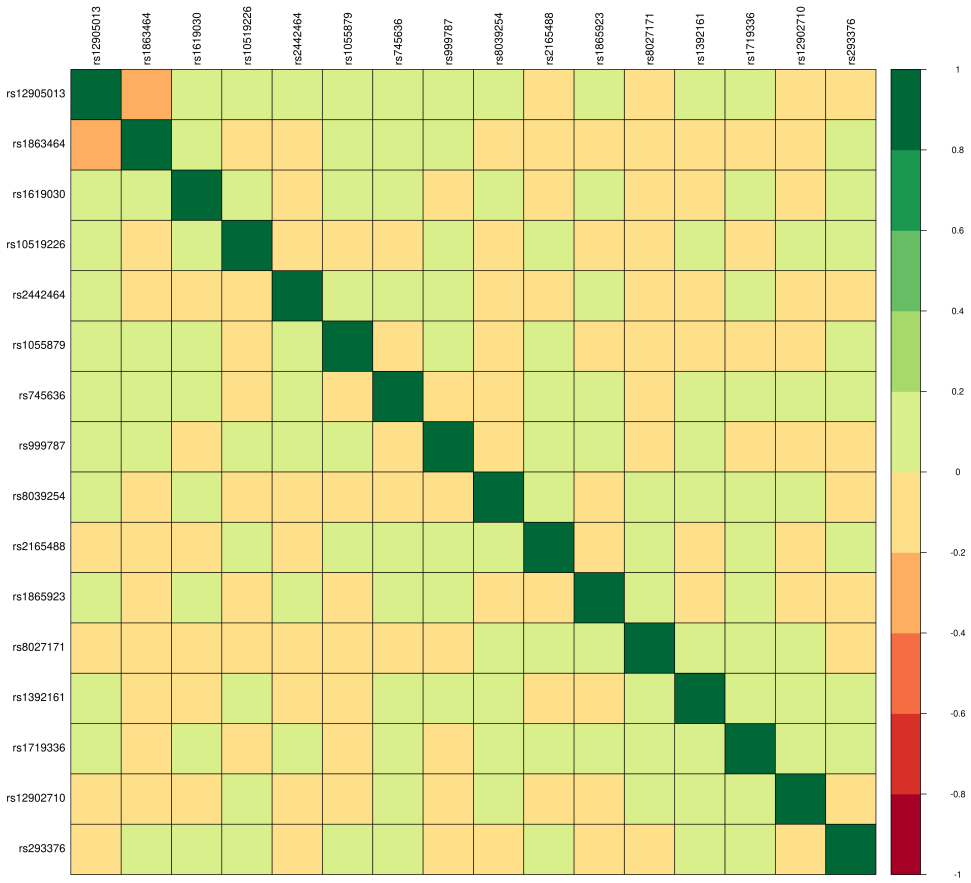


Figure 6.3: The correlation plot of SNPs that ate selected in the second stage via using Bayesian framework

There were 16 SNPs covariates that are selected on the second stage. We wished check the correlation between the selected covariate. The correlation plot (Figure 6.3) indicated that the correlation was less than 0.2 between the selected SNPs. There is not evidence for the multicollinearity problem since those values of correlation are low.

Table 6.13: The MAF of SNP selected under BVS in the second stage.

SNP	$MAF(data)$	$MAF(database)$	$chromosome$	$location$
$rs12905013$	0.359479	0.334907	15	26703938
$rs1863464$	0.180563	0.183554	15	26693341
$rs999787$	0.131716	0.133645	15	55000253
$rs8039254$	0.471455	0.461428	15	55016650
$rs8027171$	0.402279	0.399968	15	83392095
$rs1392161$	0.294418	0.282969	15	87208615

We also compared the Minor Allele Frequency (MAF) of selected SNPs in the data set and the database of SNPs which were close to each other (Table 6.13). These results pointed that MAF of selected SNPs in the data set are close to the values of MAF in database SNPs.

Moreover, we also did a 2nd run and 3rd run for confirmation of the results of variable selection when using the *glmnet* package. The first stage results are reported in Tables 6.14 and 6.15, and the second stage results are shown in Tables 6.16 and 6.17.

Table 6.14: The upper bound and lower bound of the point estimates for each SNP when using LASSO via $\lambda.min$ in 3 sub-data sets running 3 times under different splits of data (under the first stage) The value in the brackets represents the index of sub-dataset in which the SNP is present.

SNP	**1st run**	**2nd run**	**3rd run**
*rs*12905013	0.0892(0.0264,0.1742)[1]	0.0835(0.0236,0.1714)[1]	0.0809(0.0198,0.1697)[2]
*rs*1863464	0.0875(0.0287,0.1715)[1]	0.0896(0.0211,0.1796)[1]	0.0841(0.0209,0.1724)[3]
*rs*17116056	0.0814(0.0259,0.1699)[1]	-	-
*rs*1619030	0.0853(0.0274,0.1689)[2]	0.0811(0.0207,0.1702)[1]	-
*rs*2344848	0.0861(0.0217,0.1651)[2]	-	-
*rs*2165488	0.0817(0.0235,0.1634)[2]	0.0759(0.0211,0.1611)[3]	-
*rs*999787	0.0892(0.0274,0.1678)[2]	0.0853(0.0217,0.1641)[3]	0.0819(0.0194,0.1679)[2]
*rs*1718939	0.0868(0.0259,0.1642)[2]	0.0785(0.0196,0.1544)[1]	0.0803(0.0184,0.1604)[1]
*rs*8039254	0.0815(0.0214,0.1648)[3]	0.0749(0.0188,0.1529)[3]	0.0796(0.0175,0.1601)[1]
*rs*12101585	0.0842(0.0221,0.1626)[2]	-	-
*rs*1719336	0.0873(0.0215,0.1631)[3]	-	0.0783(0.0187,0.1549)[1]
*rs*12902710	0.0816(0.0204,0.1659)[1]	0.0774(0.0192,0.1543)[2]	-
*rs*919961	0.0861(0.0211,0.1643)[2]	-	-
*rs*8027171	0.0817(0.0215,0.1531)[1]	0.0762(0.0189,0.1455)[2]	0.0846(0.0196,0.1503)[2]
*rs*12591031	0.0826(0.0211,0.1577)[1]	0.0744(0.0181,0.1461)[3]	0.0795(0.0156,0.1475)[2]
*rs*1392161	0.0894(0.0254,0.1623)[2]	0.0716(0.0185,0.1479)[1]	0.0765(0.0147,0.1445)[1]
*rs*293376	0.0822(0.0241,0.1631)[2]	0.0755(0.0178,0.1511)[3]	-
*rs*2932201	0.0824(0.0222,0.1599)[2]	0.0763(0.0184,0.1494)[3]	-
*rs*1027549	0.0819(0.0208,0.1624)[1]	-	-
*rs*8027231	0.0825(0.0225,0.1559)[2]	-	-

Table 6.15: Corresponding bounds of point estimates (as Table 6.14) from LASSO using $\lambda.1se$ criterion.

SNP	**1st run**	**2nd run**	**3rd run**
*rs*12905013	0.0789(0.0184,0.1644)[1]	0.0736(0.0112,0.1619)[1]	0.0719(0.0154,0.1653)[2]
*rs*1863464	0.0741(0.0186,0.1635)[1]	0.0769(0.0155,0.1667)[1]	0.0788(0.0192,0.1694)[3]
*rs*1619030	0.0712(0.0153,0.1637)[2]	0.0729(0.0174,0.1609)[1]	-
*rs*2165488	0.0745(0.0157,0.1624)[2]	0.0768(0.0184,0.1633)[3]	-
*rs*2311748	-	0.0675(0.0167,0.1584)[1]	-
*rs*2871864	-	0.0617(0.0159,0.1512)[2]	-
*rs*999787	0.0711(0.0188,0.1658)[2]	0.0743(0.0159,0.1669)[3]	0.0763(0.0204,0.1698)[2]

Continued on next page

Table 6.15 – *Continued from previous page*

SNP	**1st run**	**2nd run**	**3rd run**
*rs*1718939	0.0768(0.0195,0.1634)[2]	0.0785(0.0196,0.1684)[1]	0.0739(0.0188,0.1633)[1]
*rs*8039254	0.0745(0.0147,0.1617)[3]	0.0749(0.0188,0.1629)[3]	0.0752(0.0195,0.1651)[1]
*rs*1719336	0.0773(0.0201,0.1622)[3]	-	0.0783(0.0187,0.1649)[1]
*rs*2280194	0.0716(0.0187,0.1598)[1]	0.0743(0.0192,0.1636)[2]	-
*rs*2297381	-	0.0678(0.0121,0.1521)[3]	-
*rs*1439618	-	0.0664(0.0118,0.1528)[3]	-
*rs*8027171	0.0785(0.0115,0.1622)[1]	0.0762(0.0139,0.1655)[2]	0.0745(0.0151,0.1624)[2]
*rs*12591031	0.0726(0.0129,0.1637)[1]	0.0744(0.0181,0.1661)[3]	0.0792(0.0204,0.1695)[2]
*rs*1392161	0.0745(0.0155,0.1653)[2]	0.0716(0.0184,0.1598)[1]	0.0721(0.0164,0.1659)[1]
*rs*12902710	0.0742(0.0186,0.1631)[3]	0.0755(0.0178,0.1611)[2]	-
*rs*293376	0.0724(0.0162,0.1599)[2]	0.0763(0.0184,0.1594)[3]	-
*rs*1050255	-	0.0683(0.0116,0.1516)[3]	-
*rs*170781	-	0.0632(0.0125,0.1524)[2]	-

Under each data set split, some of the same covariates are selected with BVS (those covariates are selected in all runs under BVS).

Table 6.16: The upper bound and lower bound of the point estimates for each SNP when using LASSO via $\lambda.min$ in 3 sub-data sets running 3 times under different splits of data (under the second stage).

SNP	$1^{st}run$	$2^{nd}run$	$3^{rd}run$
*rs*12905013	0.0956(0.0289,0.1797)	0.0986(0.0295,0.1803)	0.0942(0.0255,0.1764)
*rs*1863464	0.0975(0.0295,0.1786)	0.0965(0.0283,0.1774)	0.0988(0.0286,0.1781)
*rs*1719336	0.0935(0.0268,0.1699)	-	0.0914(0.0204,0.1683)
*rs*999787	0.0978(0.0269,0.1725)	0.0989(0.0271,0.1752)	0.0951(0.0201,0.1721)
*rs*1718939	0.0897(0.0278,0.1689)	0.0862(0.0205,0.1673)	0.0867(0.0207,0.1669)
*rs*8039254	0.0896(0.0223,0.1695)	0.0821(0.0196,0.1614)	0.0842(0.0194,0.1639)
*rs*12902710	0.0853(0.0234,0.1657)	0.0862(0.0155,0.1649)	-
*rs*8027171	0.0867(0.0253,0.1698)	0.0823(0.0214,0.1662)	0.0847(0.0206,0.1676)
*rs*12591031	0.0886(0.0221,0.1696)	0.0796(0.0197,0.1598)	0.0876(0.0165,0.1624)
*rs*1392161	0.0804(0.0189,0.1629)	0.0803(0.0197,0.1607)	0.0841(0.0199,0.1619)
*rs*2932201	0.0866(0.0232,0.1626)	0.0827(0.0197,0.1619)	-

Table 6.17: Corresponding bounds of point estimates (as Table 6.16) from LASSO using $\lambda.1se$ criterion.

SNP	$1^{st}run$	$2^{nd}run$	$3^{rd}run$
rs12905013	0.0897(0.0214,0.1689)	0.0835(0.0195,0.1669)	0.0867(0.0175,0.1643)
rs1863464	0.0859(0.0225,0.1654)	0.0843(0.0213,0.1637)	0.0821(0.0207,0.1629)
rs1719336	0.0867(0.0218,0.1639)	-	0.0816(0.0199,0.1651)
rs999787	0.0878(0.0269,0.1705)	0.0889(0.0251,0.1696)	0.0833(0.0235,0.1681)
rs1718939	0.0897(0.0278,0.1689)	0.0812(0.0205,0.1653)	0.0851(0.0201,0.1669)
rs8039254	0.0896(0.0223,0.1695)	0.0898(0.0196,0.1657)	0.0831(0.0202,0.1633)
rs12902710	0.0853(0.0234,0.1657)	-	0.0857(0.0201,0.1649)
rs8027171	0.0897(0.0253,0.1598)	0.0798(0.0196,0.1612)	0.0851(0.0199,0.1669)
rs12591031	0.0846(0.0211,0.1601)	0.0805(0.0187,0.1596)	0.0817(0.0198,0.1612)
rs1392161	0.0804(0.0178,0.1609)	0.0795(0.0197,0.1567)	0.0821(0.0204,0.1608)
rs2932201	0.0814(0.0159,0.1621)	0.0811(0.0145,0.1654)	-

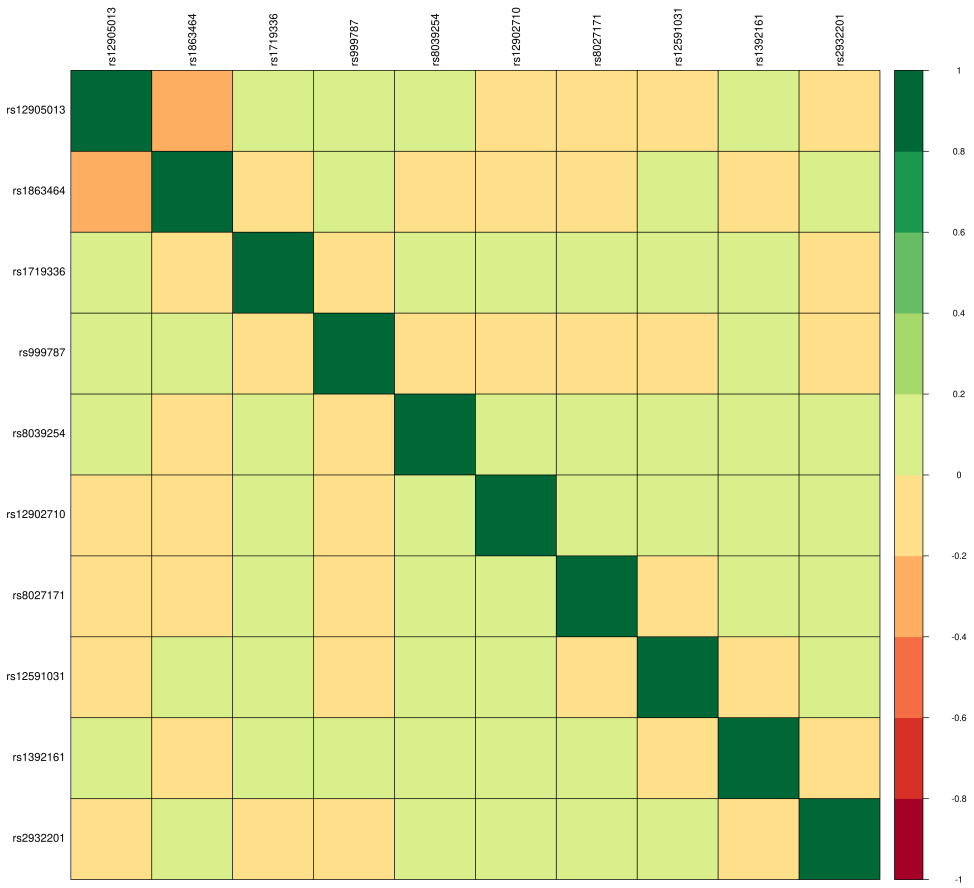


Figure 6.4: The correlation plot of SNPs that ate selected in the second stage via using glmnet

Again, we checked the correlation between each pair of covariates that are selected from the second stage when we used LASSO variable selection. Figure 6.4 indicates that the correlation coefficients in each pair of covariate is less than 0.2. There is lacking evidence on the multi-collinearity. They are expected results.

The final comparison of variable selection of SNPs between BVS and LASSO is reported in Table 6.18. From this, it can be seen that there are 7 SNPs that are selected in all methods. 9 SNPs are chosen using the Bayesian framework. Moreover, there are 3 that are selected by the glmnet under both criteria. From the final results show that there are 7 SNPs for the consistency selected. They should be the associate SNPs with hypertension diseases.

Table 6.18: The summary of *SNP* when using Bayesian, *glmnet* package via $\lambda.min$ and $\lambda.1se$ (under the second stage)

SNP	*Bayesian*	$\lambda.min$	$\lambda.1se$
*rs*12905013	✓	✓	✓
*rs*1863464	✓	✓	✓

Continued on next page

Table 6.18 – *Continued from previous page*

SNP	*Bayesian*	$\lambda.min$	$\lambda.1se$
*rs*1619030	✓	-	-
*rs*10519226	✓	-	-
*rs*2442464	✓	-	-
*rs*1055879	✓	-	-
*rs*745636	✓	-	-
*rs*1719336	-	✓	✓
*rs*999787	✓	✓	✓
*rs*1718939	-	✓	✓
*rs*8039254	✓	✓	✓
*rs*2165488	✓	-	-
*rs*1865923	✓	-	-
*rs*12902710	✓	✓	✓
*rs*8027171	✓	✓	✓
*rs*12591031	-	✓	✓
*rs*1392161	✓	✓	✓
*rs*1719336	✓	-	-
*rs*293376	✓	-	-

6.4 Computational times

Since the benefit of data splitting is claimed to be computational efficiency. The CPU times for running in each splitting and the whole dataset are presented in Table 6.19. Those results pointed that the computational times are under splitting dataset dramatically lower than under the whole dataset. Moreover, the computation times are used for running Bayesian framework that are higher than the *glmnet* in both criteria. Since the variable selection in Bayesian need to use MCMC algorithm that consumes many time in each running. From the performance and computational cost, we suggest 3 to 5 splitting for the splitting in the first stage since they are a moderate number of splittings. Since when we will use the higher number of splittings, they give more on the compuational times and the error in high $E(SSE)$.

Table 6.19: The CPU times for running in each splitting and the whole dataset (seconds)

No. of Splits	$\lambda.min$	$\lambda.1se$	Bayesian
3	583.67	594.12	20163.37
5	611.36	615.84	21583.46
15	641.66	656.19	23129.14
whole	1059.43	1168.16	198574.11

Chapter 7

Discussion and Conclusion

7.1 Introduction

The main aim of this thesis is to contribute to the current developments in the area of statistical genetics in the specific problem of determining SNPs associated with clinical outcomes in high-dimensional GWAS. We investigated the performance of several existing methods for this and introduce a new computationally efficient approach of splitting the dataset for the analysis of genome-wide association study data. We illustrate these methods and evaluate them through simulation studies and also using real GWAS data from a study of hypertension. In this chapter we provide a summary of the results obtained in this thesis, discuss their implications on genetic data analysis and provide some insight into interesting further research topics based on the findings in this thesis.

7.2 Summary of results from simulation studies

Computing the marginal inclusion probability of each variable helps determine whether the variable should be included in the model. Under the case of independent covariates, the inclusion probabilities of associated covariates appear to increase when the effect size increases. For example the inclusion probabilities of associated covariates are about 0.5 when the effect size is 1, whereas the inclusion probabilities of associated covariates are nearly 1 when the effect size is 1.5. When the effect size is low ($\boldsymbol{\beta}$ = 0.5), the inclusion probabilities are also low for the associated covariates. The estimation of regression coefficients for the associated covariates and the non-associated covariate values are accurate (similar to the true effect size). Moreover, we see that when the Binomial proportion for the generative model (p) increases to be closer to 0.5, the inclusion probability increases. The study also suggests higher inclusion probabilities with higher effect sizes ($\boldsymbol{\beta}$ = 1.5, 2). For the correlated cases, the inclusion probabilities under the low correlation are higher than that of the high correlation. For example, inclusion probabilities of about 0.6 when $\rho = 0.1$ drop to inclusion probabilities of about 0.2 when $\rho = 0.8$. The inclusion probabilities of associated covariates are increased when the effect size increases. A similar result is witnessed when the probability p of the generative Binomial distribution is increased. Inclusion probabilities increase as p is increased.

The credible intervals under Bayesian framework are wider when compared to the LASSO

estimates in both criteria ($\lambda.min$ and $\lambda.1se$) with the lower effect size ($\boldsymbol{\beta} = 0.1, 0.3$). However, when the effect sizes increase ($\boldsymbol{\beta} = 0.5, 1$) the credible intervals from the Bayesian method are slightly different when compared to the LASSO using both criteria.

7.3 Summary for real data analyses

Some covariates are selected in both methods - BVS and LASSO, a fraction of these overlap and could potentially indicate the stronger signals in the data. The credible intervals from BVS as well as the posterior estimates are slightly different to the LASSO-based estimates.

The splitting of data sets appears necessary to detect any signals in this dataset due to the size of the data. Different splits of the data yield largely consistent results, indicating that this is a potentially promising approach for similary sized GWAS dataset applications.

However, more extensive simulations and analyses would need to be conducted to assess the power of the method in a variety of situations.

7.4 Limitations of the Study

In our study, the simulation studies used to evaluate the variable selection in Chapter 5 were restricted to 500 SNP covariates due to limitations in time and the available computation power. The real dataset used for analysis in Chapter 6 has about a total of 15000 covariates. We may expect to see big differences between results based on 500 SNP covariates and hundreds of thousands of SNPs, so we would want to further the investigation of methods with much larger datasets and extend our simulation studies to a genome-wide scale.

Moreover, the number of splits for the real dataset were considered to be 3,5 and 15 but this could be varied. Finding the optimum number of splitting datasets could be a useful further direction of this work.

7.5 Further Research and future directions

The simulation studies used to evaluate the splitting technique in Chapter 4 were restricted by the number of covariates due to limitations in time and the available computational power. However, for a dataset on a genome-wide scale for large real-life datasets, we based our expectation on how the methods would perform on results that we obtained in much smaller simulation studies, resembling fine mapping more than genome-wide studies. We would want to further the investigation of methods on much larger datasets and extend our results to a genome-wide scale. A major challenge would be to incorporate realistic models of population structure into the model framework, as well as realistic interaction effects.

Appendix A

Figures

Figure Appedix 1 : The plot of the credible interval under the associated covariates when $\beta = 1$ the second correlated case (x_1 with x_4, x_5, x_2 with x_6, x_7, x_3 with x_8, x_9) when $\rho = 0.1$ from 500,1000 and 2000 covariates

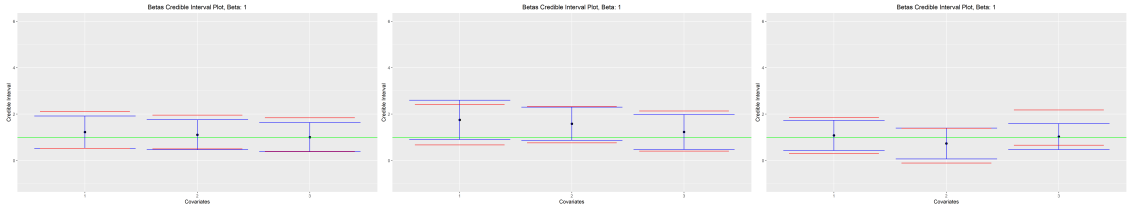
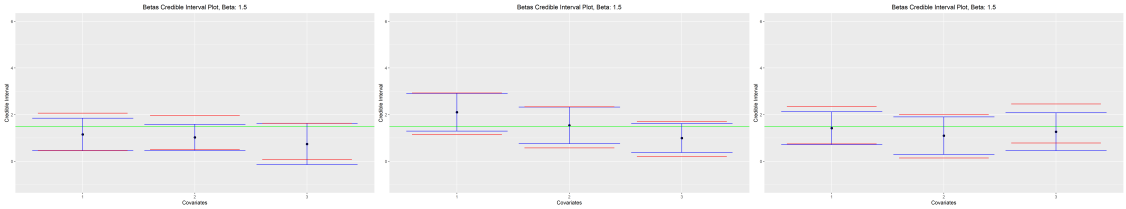


Figure Appendix 2 : The plot of the credible interval under the associated covariates when $\beta = 1.5$ under the second correlated case (x_1 with x_4, x_5, x_2 with x_6, x_7, x_3 with x_8, x_9) when $\rho = 0.1$ from 500, 1000 and 2000 covariates



Bibliography

[1] Abdellaoui, A., Hugh-Jones, D., Yengo, L., Kemper, K.E., Nivard, M.G., Veul, L. and Holtz, Y. (2019). Genetic correlates of social stratification in Great Britain. *Nat Hum Behav* ,3: 1332-1342.

[2] Altshuler, D. and Donnelly, P. (2005). A haplotype map of the human genome. *Nature*, 437: 1299-1320.

[3] Andrews, S.J., Fulton-Howard, B., Zietsch, B.P., Fraying, T.M., Wray, N.R., Yang, J., Verweij, K.J., Visscher, P.M. and Goate, A. (2020). Interpretation of risk loci from genome-wide association studies of Alzheimer's disease. *Lancet Neurol*, 19(4): 326-335.

[4] Asimit, J. and Zeggini, E. (2010). Rare variant association analysis methods for complex traits. *Annu Rev Genet*, 44: 293-308.

[5] Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A., and Abecasis, G.R. (2015). A global reference for human genetic variation. *Nature*, 526(7571): 68-74.

[6] Avery, O.T., Macleod, C.M. and McCarty, M. (1944). Studies on the chemical nature of the substance inducing transformation of pneumococcal types : Induction of Transformation by a Desoxyribonucleic Acid fraction isolated from pneumococcus type III. *J Exp Med*, 79(2): 137-158.

[7] Baselmans, B.M., Jansen, R., Ip, H.F. and Dongen, J.V. (2019). Multivariate genome-wide analyses of the well-being spectrum. *Nat Genet*, 51(3): 445-451.

[8] Bayrhuber, H. and Kull, U. (1989). *Linder Biologie*. Lebrbuch fur die Obserstufe. Stuttgart.

[9] Benjamin, D.J., Heffetz, O., Kimball, M.S. and Rees-Jones, A. (2012). What Do You Think Would Make You Happier? What Do You Think You Would Choose?. *Am Econ Rev*, 102(5): 2083-2110.

[10] Bingham, N.H. and Fry, J.M. (2010). *Regression: Linear Models in Statistics*. Springer. London.

[11] Bush, W.S and Moore, J.H. (2012). Chapter 11: Genome-Wide Association Studies. *PLoS Comput Biol*, 8(12): 1-11.

[12] Chaitankar, V., Karakulah, G., Ratnapriya, R., Giuste, I.O., Brooks, M. and Swarooop, A. (2016). Next generation sequencing technology and genomewide data analysis: Perspectives for retinal research. *Prog Retin Eye Res*, 55: 1-31.

[13] Cui, G., Wong, M.L., and Zhang, G. (2010). Bayesian variable selection for binary response models and direct marketing forecasting. *Expert Systems with Applications*, 37(12): 7656-7662.

[14] Friedman, J.H., Hastie, T. and Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, 33(1): 1-22.

[15] Fry, A., Littlejohns, T.J., Sudlow, C., Doherty, N., Adamska, L., Sprosen, T., Collins, R. and Allen, N.E. (2017). Comparison of Sociodemographic and Health-Related Characteristics of UK Biobank Participants With Those of the General Population. *Am J Epidemiol* , 186(9): 1026-1034.

[16] Gareth, J., Witten, D., Hastie,T. and Tibshirani,R. (2014). *An Introduction to Statistical Learning: With Applications in R.* Springer Publishing Company. New York.

[17] Guo, Z., Yang, Q., Huang, F., Zheng, H., Sang, Z., Xu, Y., Zhang, C., Wu, K., Tao, J., Prasanna, B.M., Olsen, M.S, Wang, Y., Zhang, J. and Xu, Y. (2021). Development of high-resolution multiple-SNP arrays for genetic analyses and molecular breeding through genotyping by target sequencing and liquid chip. *Plant Comm*, 2(6): 1-15.

[18] Hardy, G.H. (1908). Mendelian proportions in a mixed population. *Science*, 28: 49-50.

[19] Henderson, P., Wilson, D.C., Satsangi, J. and Stevens, C. (2012). A role for vimentin in Crohn disease. *Autophagy*, 8(11): 1695-1696.

[20] Hindorff, L.A., Sethupathy, P., Junkins, H.A., Ramos, E.M., Mehta, J.P., Collins, F.S., and Manolio, T.A. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *PNAS*, 106(23): 9362-9367.

[21] Hirschhorn, J.N. and Daly, M.J. (2005). Genome-wide association studies. *Nat Rev Genet*, 6(2): 95-108.

[22] Kashani, A. and Schwartzl, D.A. (2019). The Expanding Role of Anti-IL-12 and/or Anti-IL-23 Antibodies in the Treatment of Inflammatory Bowel Disease. *Gastroenterol Hepatol*, 15(5): 255-265.

[23] Keats, B.J.B. and Sherman, S.L. (2013). Chapter 13 - Population Genetics, in *Emery and Rimoin's Principles and Practice of Medical Genetics*. Sixth edition. Academic Press. London.

[24] Khera, A.V., Chaffin, M., Aragam, K.G., Haas, M.E., Roselli, C., Choi, S.H., Natarajan, P., Lander, E.S., Lubitz, S.A., Ellinor, P.T. and Kathiresan, S. (2018). Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat genet*, 50: 1219-1224.

[25] Klein, R.J., Zeiss, C., Chew, E.Y., Tsai, J.Y., Sackler, R.S., Haynes, C., Henning, A.K., Sangiovanni, J.P., Mane, S.M., Mayne, S.T., Bracken, M.B., Ferris, F.L., Ott, J., Barstable, C. and Hoh, J. (2005). Complement factor H polymorphism in age-related macular degeneration. *Science*, 308(5720): 385-389.

[26] Korte, A. and Farlow, A. (2013). The advantages and limitations of trait analysis with GWAS: a review. *Plant Methods*, 9(29): 1-9.

[27] Lam, M., Awasthi,S., Watson, H.J., Goldstein, J., Panagiotaropoulou. G., Trubetskoy, V., Karlsson, R., Frei, O., Fan, C.C., Witte, W.D., Mota, N.R., Mullins, N., Brugger, K., Lee, S.H., Wray, N.R., Skarabis, N., Huang, H., Neale, B., Daly, M.J., Mattheisen, M., Walters, R. and Ripke, S. (2020). RICOPILI: Rapid Imputation for COnsortias PIpeLIne. *Bioinformatics*, 36(3): 930-933.

[28] Lan, N., Lu, Y., Zhang, Y., Pu, S., Xi, H., Nie, X., Liu, J. and Yuan, W. (2020). FTO - A Common Genetic Basis for Obesity and Cancer. *Front Genet*, 16(11): 1-12.

[29] Liang, F., Paulo,R., Molina,G., Clyde,M.A. and Berger, J.O. (2007). Mixtures of g-priors for Bayesian Variable Selection. *Journal of the American Statistical Association*, 103: 410-423.

[30] Lawson, K.A., Sousa, C.M., Zhang, X., Kim, E., Akthar, R., Caumanns, J.J., Yao, Y., Mikolajewicz, N.M., Ross, C., Brown, K.R., Zid, A.A., Fan, Z.P., Hui, S. Krall, J.A., Simons, D.M., Slater, C.S., Jesus, V.D., Tang, L., Singh, R., Goldford, J.E., Martin, S., Huang, Q., Francis, E.A., Hasbid, A. and Moffat, J. (2020). Functional genomic landscape of cancer-intrinsic evasion of killing by T cells. *Nature*, 586: 120-126.

[31] Lee, Y.N., Frugoni, F., Dobbs, K., Walter, J.E., Giliani, S., Gennery, A.R., Herz,W.A., Haddad, E., LeDeist, F., Blessing, J.H., Henders, L.A., Pai, S.Y., Nelson, R.P., Ghoneimy, D.H., El-Feky, R.A., Reda, S.M., Hossny, E., Soler-Palacin, P., Fuleihan, R.L., Patal, N.C., Massaad, M.J., Geha, R.S., Pack, J.M., Palma, P., Cancrini, C., Chen, K., Vihinen, M., Alt, F.W. and Notarangelo, L.D. (2014). A systematic analysis of recombination activity and genotype-phenotype correlation in human recombination-activating gene 1 deficiency. *J Allergy Clin Immunol*, 133(4): 1099-1108.

[32] Li, Y., Willer, C., Sanna, S. and Abecasis, G. (2009). Genotype imputation. *Annu Rev Genomics Hum Genet*, 10:387-406.

[33] Li, Y., Zhao, H., Wilkins, K., Hughes, C. and Damon, E.K. (2010). Real-time PCR assays for the specific detection of monkeypox virus West African and Congo Basin strain DNA. *J Virol Methods*, 169(1): 223-227.

[34] Liu, H.M., Zheng, J.P., Yang, D., Liu, Z.F., Li, Z., and Hu, Z.Z. (2021).) Recessive/dominant model:Alternative choice in case-control-based genomewide association studies. *PLoS ONE*, 16(7): e0254947.

[35] MacArthur, J., Bowler, E., Cerezo, M., Gil, L., Hall, P., Hastings, E., Junkins, H., McMahor, A., Milano, A., Morales, J., Pendington, Z.M., Welter, D., Burdett, T., Hindorff, L., Flicek, P., Cuningham, F. and Parkinson, H. (2017). The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res*, 45 (Database issue): D 896-D 901.

[36] Manolio, T.A., Collins, F.S., Cox, N.J., Goldstein, D.B., Hindorff, L.A., Hunter, D.J., McCarthy, M.I., Ramos, E.M., Cardon, C.R., Chakravarti, A., Cho, J.H., Guttmacher, A.E., Kong, A., Kruglyak, L., Mardis, E., Rotimi, C.N., Slatkin, M., Valle, D., Whittermore, A.S., Boehuke, M., Clark, A.G., Eichler, E.E., Gibson, G., Haines, J.L., Mackay, T.F., McCarroll, S.A. and Visscher, P.M. (2009). Finding the missing heritability of complex diseases. *Nature*, 461: 747-753.

[37] Novembre, J., Johnson, T., Byrc, K., Kutalik, Z., Boyko, A.R., Auton, A., Indap, A., King, K.S., Bergmann, S., Nelson, M.R., Stephens, M. and Bustamante, C.D. (2008). Genes mirror geography within Europe. *Nature*, 456: 98-101.

[38] Olsen, E.A., Whittaker, S., Kim, Y.H., Duvic, M., Prince, H.M., Lessin, S.R., Wood, G.S., Willemze, R., Demierre, M.F., Pimpinelli, N., Bernengo, M.G., Romero, P.L., Bagot, M., Estrach, T., Guitart, J., Khobler, R., Sanches, J.A., Iwatsuki, K., Sugaya, M., Dummer, R.M., Pittelkow, M., Hoppe, R., Parker, S., Geskin, L., Brown, L.P., Girardi, M., Burg, G., Ranki, A., Vermeer, M., Horwitz, S., Heald, P., Rosen, S., Cerroni, L., Dreno, B. and Vonderheid, E.C. (2011). Clinical end points and response criteria in mycosis fungoides and Sézary syndrome: a consensus statement of the International Society for Cutaneous Lymphomas, the United States Cutaneous Lymphoma Consortium, and the Cutaneous Lymphoma Task Force of the European Organisation for Research and Treatment of Cancer. *J Clin Oncol*, 29(18): 2598-2607.

[39] Palmer, L. and Cardon, L. (2005). Shaking the tree: mapping complex disease genes with linkage disequilibrium. *Lancet*, 366(9492): 1223-1234.

[40] Pirinen, E., Canto, C., Houtkooper, R.H., Youn, D.Y., Yamamoto, H., Andreux, P.A., Rose, P.C., Gademann, K., Rinsch, C., Schoonjans, K., Sauve, A.A. and Auwerx, J. (2012). The NAD(+) precursor nicotinamide riboside enhances oxidative metabolism and protects against high-fat diet-induced obesity. *Cell Metab*, 15(6): 838-847.

[41] Plummer, M., Best, N., Cowles, K., Vines, K., Sarkar,D., Bates, D., Almond, R. and Magnusson,A. (2022). Package 'coda'.

[42] Press, W.H., Teukolsky, S.A., Vetterling, W.T. and Flannery, B.P. (1992). *Numerical Recipes in C.* Cambridge University Press. Cambridge.

[43] Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A. and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet*, 38(8): 904-909.

[44] Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., Bakker, P.W., Daly, M.J. and Sham, D.C. (2007). PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *Am J Hum Genet*, 81(3): 559-575.

[45] Robinson, M.A. (1998). *Encyclopedia of Immunology*.(2nd ed.).

[46] Ruth, K.S., Day, F.R., Tyrrell, J., Thompson, D.J., Wood, A.R., Mahajan, A., Beaumont, R.N., Wittemans,L., Martin, S., Busch, A.S., Erzurumluoglu, A.M., Hollis, B., Mara, T.A., McCarthy, M.I., Largenberg, C., Easton, D.F., Wareham, N.J., Burgess, S., Murray, A., Ong, K.K., Frayling, T.M. and Perry, J.R. (2020). Using human genetics to understand the disease impacts of testosterone in men and women. *Nature Medicine*, 26: 252-258.

[47] Siminovitch, K.A. (2004). PTPN22 and autoimmune disease. Nat genet, 36(12): 1248-1249.

[48] Smith, M. and Kohn, R. (1996). Nonparametric regression using Bayesian variable selection. *J. Econometrics*, 75: 317-343.

[49] Sulem, P., Gudbjartsson, D.F., Stacey, S.N., Helgason, A., Rafnar, T., Magnusson, K.P., Manolescu, A., Karason, A., Palsson, A., Thorleifsson, G., Jakobsdottir, M., Steinberg, S., Palsson, S., Jonasson, F., Sigurgeirsson, B., Thorisdottir, M., Ragnarsson, R., Benediktsdottir, K.R.,

Aben, K.K., Kiemeney, L.A., Olaffson, J.H., Gulcher, J., Kong, A., Thorsteinsdottir, U. and Stefansson, K. (2007). Genetic determinants of hair, eye and skin pigmentation in Europeans. *Nat genet*, 39(12): 1443-1452.

[50] Teare, M.D. and Barrett, J.H. (2005). Genetic linkage studies. *Lancet*, 366(9490): 1036-1044.

[51] Timpson, N.J., Greenwood, C.M., Soranzo, N., Lawson, D.L. and Richards, J.B. (2018). Genetic architecture: the shape of the genetic contribution to human traits and disease. *Nat Rev Genet*, 19(2): 110-124.

[52] Uffelmann, E., Huang, Q.Q., Munung, S.M., Vries, J.D., Okada, Y., Martin, A.R., Martin, H.C., Lappalainen, T. and Posthuma, D. (2021). Chapter 11: Genome-wide association studies. *Nat Rev Methods Primers*, 1 59: 1-21.

[53] Watanabe, K., Stringer, S., Frei, O. and Markov, M.U., Leeuw, C.D., Tinca, J.C., Sluis, S.V., Andreassen, O.A., Neale, B.M. and Posthuma, D. (2019). A global overview of pleiotropy and genetic architecture in complex traits. *Nat Genet*, 51: 1339-1348.

[54] Watson, J.D. and Crick, F.H. (1953). Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature*, 171: 737-738.

[55] Weinberg, W. (1908). Über den Nachweis der Vererbung beim Menschen. *Jahresh Wuertt Ver*, 64: 369-382.

[56] Willer, C.J., Li, Y. and Abecasis, G.R. (2010). METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics*, 26(17): 2190-2191.

[57] Zhou, W., Nielsen, J.B., Fritsche, L.G., Dey, R., Gabrielssen, M.E., Wolford, B.N., LeFaive, J., VandeHaar, P., Gagliano, S.A., Gifford, A., Bastarache, L.A., Wei, W.Q., Denny, J.C., Lin, M., Hveem, K., Kang, H.M., Abecasis, G.R., Willer, C.J. and Lee, S. (2018). Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nat Genet*, 50: 1335-1341.

[58] Zucknick, M. (2013). https://r-forge.r-project.org/projects/bvsflex/. Accessed: 30-04-2023.